

Analyse des essais contrôlés randomisés (ECR) multi-niveaux d'Education Endowment Foundation au Royaume-Uni.

Analysis of multi-level randomized controlled trials (RCTs) from the Education Endowment Foundation in the United Kingdom.

Nathalie Roques

## Résumé

Les méthodes statistiques utilisées pour calculer les résultats de 11 essais contrôlés randomisés conduits (ECR) par Education Endowment Foundation (EEF) au Royaume-Uni sont analysées. Les modèles mathématiques et les données publiées ont dans un premier temps été identifiés. Devant l'absence d'informations pour plusieurs de ces études, des hypothèses de calculs ont été émises qui ont permis de recalculer les tailles d'effet. Il a alors été montré que certaines préconisations publiées par les documents cadre d'EEF n'ont pas été suivies. Des pistes sont proposées pour améliorer la qualité des comptes-rendus de ce type d'études.

Mots-clés : essai contrôlé randomisé, sciences de l'éducation, intervention, taille d'effet, régression linéaire multiple

## Abstract

The statistical methods used to calculate the results of 11 randomized controlled trials (RCTs) conducted by the Education Endowment Foundation (EEF) in the United Kingdom are analyzed. The mathematical models and published data were first identified. Given the absence of information for several of these studies, calculation hypotheses were made which made it possible to recalculate the effect sizes. It was then shown that certain recommendations published by the EEF framework documents were not followed. Ways are proposed to improve the quality of the reports of this type of study.

Keywords: randomized controlled trial, educational sciences, intervention, effect size, multiple linear regression

Permettre aux jeunes d'acquérir les compétences nécessaires à leur développement tant personnel que professionnel est un objectif commun à de nombreux pays dans le monde. Identifier les caractéristiques d'une organisation de l'enseignement et de la scolarité qui favorise les apprentissages est dans cette optique un élément essentiel de la politique des États. Une façon de répondre à cette demande consiste à prendre en compte les résultats d'études quantitatives qui évaluent l'impact d'une intervention ou traitement sur les élèves, et notamment des études causales que sont les essais contrôlés randomisés (ECR). L'importance donnée aux résultats des ECR dans le domaine des Sciences de l'éducation a fortement augmenté depuis les années 2000, tout particulièrement dans les pays anglo-saxons, berceaux de l'*evidence based education* et de l'*evidence based policy*. Le début de ce millénaire a non seulement vu le nombre d'ECR augmenter fortement (comme de celui de leurs synthèses), mais aussi la naissance d'organismes financés par des États dans le but de réaliser, de synthétiser et de disséminer les résultats de grandes études comparatives (White, 2019). On citera ici le What Works Clearinghouse (WWC) aux USA et Education Endowment Foundation (EEF) au Royaume-Uni. L'intérêt récent témoigné par le Conseil Scientifique National de l'Education National (CSEN) pour les publications d'EEF (CSEN, 2023 ; Dehaene et Pasquinelli, n.d., p.17) devrait inciter aujourd'hui les acteurs de la communauté éducative française à s'intéresser davantage à ce type de recherche qui semble prendre de l'ampleur.

C'est dans ce contexte international qu'EEF joue depuis sa naissance en 2011 un rôle tout à fait inédit. En un peu plus de 10 ans, cet organisme a encadré et financé la réalisation de 218 ECR à hauteur de 130 millions de livres (soit environ 150 millions d'euros). Ces études ont concerné 1,8 millions d'élèves scolarisés dans 23 000 établissements scolaires sur le territoire britannique (EEF, n.d.) et représentent 19% des ECR réalisés dans le monde cette dernière décennie (Evovald, 2021). Chacune de ces études a fait l'objet d'une publication sur le site internet d'EEF ([www.educationendowmentfoundation.org](http://www.educationendowmentfoundation.org)) où l'on trouve également de nombreux documents annexes qui explicitent les procédures mises en œuvre. La validité interne de ces études repose entièrement sur la qualité de leur design et sur la rigueur des analyses statistiques qui produisent les résultats dont les lecteurs vont se saisir. Si les résultats des ECR d'EEF ont déjà fait l'objet d'analyses en ce qui concerne leur ampleur et l'interprétation de leurs résultats (Lortie-Forgue, 2019 ; Demack et al., 2021 ; Ashraf et al., 2021 ; Roques, à paraître), les analyses statistiques telles qu'elles ont été conduites puis décrites dans les rapports publiés, n'ont à ce jour pas été commentées.

## 1. Objectifs et méthodes

L'objectif premier de cet article était d'explicitier les méthodes analytiques mises en œuvre dans les ECR d'EEF afin de permettre aux lecteurs d'en comprendre les principaux résultats et d'évaluer la portée qu'il convient d'accorder à leurs conclusions. Devant la difficulté rencontrée pour identifier et présenter clairement les méthodes analytiques des ECR d'EEF sélectionnés, caractériser et interroger les zones d'ombre mises à jour s'est alors imposé comme objectif complémentaire.

EEF conduit plusieurs types d'ECR qui concernent la plupart du temps soit les mathématiques, soit la littéracie. Cet article a choisi de sélectionner les ECR

- de grande ampleur (les *effectiveness trials*, voir plus loin) dont les validités interne et externe sont *a priori* supérieures à celles des autres études,

- traitant de l'enseignement des mathématiques, et produisant des résultats plus facilement transférables à un autre pays que ceux qui concernent la littéracie.

La description des ECR sélectionnés sera suivie d'une présentation des méthodes statistiques telles que prévues par EEF dans leurs documents cadres. Les modèles statistiques utilisés par les chercheurs pour calculer les résultats des ECR sélectionnés sont ensuite analysés : les modèles statistiques et les données pertinentes publiés sont répertoriés ; devant l'absence de certaines de ces informations des hypothèses de calculs sont émises ; enfin les résultats ont alors été recalculés et comparés à ceux publiés par EEF.

## 2. Les études d'efficacité en conditions réelles d'EEF

Certains des ECR conduits par EEF répondent à un cahier des charges contraignant et ont été baptisés études d'efficacité en condition réelles (*effectiveness trials*). Ces études sont menées dans des conditions ordinaires de mise en œuvre sur un large territoire : les développeurs de l'intervention restent à distance de l'étude 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 de concert et le coût de l'intervention est calculé. Ces études font souvent suite à une étude d'efficacité (*efficacy trial*) où les développeurs de l'intervention jouent un rôle important et qui se déroule dans des conditions idéales sur un territoire plus restreint. Pour chaque étude, un rapport d'évaluation (*evaluation report*) est publié. Il comprend entre autres une évaluation de l'impact (*impact evaluation*) qui va retenir notre attention ici. Pour certaines études, un protocole et un plan d'analyses statistiques (PAS) sont également rédigés (ils sont publiés avant que les résultats ne soient connus). En janvier 2023, dix études d'efficacité en conditions réelles ayant comme sujet les mathématiques ont été repérées sur le site internet d'EEF (tableau 1 ; les dates mentionnées dans ce tableau sont les dates de publication des rapports). 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<sup>1</sup>. L'intervention *Ark Mathematics Mastery* a fait l'objet de deux études (et donc de deux rapports) : l'une concerne les élèves du primaire et la seconde les élèves du secondaire. L'intervention *Maths Champion* a été l'objet d'une nouvelle étude dont le rapport a été publié en juillet 2023 (et qui n'a pas été analysée ici). Dans la suite de cet article et pour alléger le texte, il sera fait référence aux rapports des interventions en utilisant la lettre permettant de les identifier (première colonne du tableau 1). Les références de ces rapports sont données en annexe 1.

Tableau 1

Rapports, protocoles et plan d'analyse statistique des 10 études d'efficacité en conditions réelles

Rapports	Intervention	Date	PAS	Protocole
A	1stClass@Number	2018	Oui (2017)	Oui (n.d.)
B	Affordable Maths Tuition	2016	Oui* (2015)	Oui (n.d.)
C1	Ark Mathematics Mastery primary	2015	Non	Oui (n.d.)
C2	Ark Mathematics Mastery secondary	2015	Non	Oui (n.d.)
D	Catch up numeracy	2019	Oui (2017)	Oui (n.d.)
E	Chess in primary school	2016	Non	Oui (2016)

<sup>1</sup> Il n'était pas question ici de vérifier l'évaluation de ces ECR par EEF, et c'est donc leur classement tel qu'il est proposé et consultable sur leur site internet qui a été retenu.

F	ICCAMS	2021	Oui (2018)	Deux protocoles : original (2015) puis amendé (2018)
G	Mathematical Reasoning	2018	Non	Oui (n.d.)
H	Maths Champions	2018	Oui (2017)	Deux versions disponibles : 2.0 (2016) et 3.0 (2017)**.
I	Shared Maths	2015	Non	Oui (n.d.)
J	Tutor Trust Affordable Tutoring	2018	2 versions (2017)	Un protocole d'évaluation et un protocole sur l'évaluation de la mise en œuvre et du processus.

PAS = plan d'analyse statistiques. n.d. = non daté. \* = publié en annexe du rapport. \*\* = 3 fichiers pdf sont téléchargeables mais deux de ces fichiers sont la même version 3.0

### 3. L'analyse statistiques prévue par EEF

Plusieurs documents sont publiés sur le site d'EEF qui donnent des informations sur la réalisation de leurs ECR et l'analyse statistique des données produites. Les éléments présentés ci-dessous sont issus du *Statistical analysis guidance for EEF evaluations* (EEF, 2022) qui sera dénommé ci-après guide EEF 2022. Les auteurs de ce document insistent sur la nécessité de suivre, pour chacune des évaluations d'une intervention, un cadre commun car « *les résultats issus des évaluations ne sont pas considérés isolés les uns des autres, mais examinés et comparés d'un examen à l'autre. Pour cette raison, il est important que, quand cela est possible, les analyses soient comparables d'une étude à l'autre*<sup>2</sup>. » La version publiée aujourd'hui remplace une version antérieure datée de 2018 (EEF, 2018a) qui elle-même était la mise à jour d'un document daté de 2015. Le site EEF ne conserve aucune archive. La version de 2018 a été téléchargée en 2022 (avant la publication de la nouvelle mise à jour), par contre la version de 2015 n'a pas pu être consultée. Les éléments clés que nous allons décrire ci-dessous sont identiques dans les deux dernières versions, à l'exception de l'interprétation de l'incertitude de la taille d'effet. Les rapports analysés dans cet article n'ont pour la plupart pas pu s'appuyer sur le guide EEF 2018 (et encore moins sur le guide EEF 2022).

Une analyse en intention de traiter doit être conduite : les informations collectées proviennent de tous les sujets présents du début à la fin de l'expérience, quel que soit leur parcours lors de l'expérience (il peut arriver par exemple qu'un élève n'ait pas reçu le traitement alors qu'il était affecté au groupe traitement). Le niveau de compétence et de connaissance des élèves est mesuré en fin d'expérience par un score obtenu après passation d'un test standardisé : ces scores posttests sont les variables à expliquer. Le niveau des élèves avant l'expérience étant fortement corrélé à leur niveau après l'expérience, les études doivent prendre en compte ce lien en mesurant les compétences et connaissances des élèves avant l'expérience et en collectant des scores prétests. Les scores posttests seront ensuite ajustés aux scores prétests en utilisant un modèle de régression linéaire. Enfin, quand l'affectation aléatoire dans les groupes intervention ou contrôle se fait au niveau des établissements et que l'analyse se fait au niveau des élèves, les analyses doivent tenir compte du fait que des élèves regroupés dans un même établissement ont des scores corrélés.

#### 3.1. Calcul de la taille d'effet

Afin de pouvoir comparer les études EEF entre elles, mais aussi d'alimenter la recherche internationale en Sciences de l'éducation, l'effet d'une intervention sera estimé en calculant une taille d'effet (*Effect Size, ES*). La taille d'effet estimée *ES* est la différence standardisée des moyennes des scores posttests

<sup>2</sup> *Results from evaluations are not seen in isolation but reviewed and compared across projects. For this reason, it is important that, whenever possible, analyses should be comparable across studies.*

ajustés aux scores prétests ; elle est présentée comme étant le g de Hedges. C'est l'une des cinq équations publiées dans le guide EEF 2022.

$$ES = \frac{(\bar{Y}_T - \bar{Y}_C)_{ajustée}}{\sqrt{\frac{(n_1 - 1)sd_1^2 + (n_2 - 1)sd_2^2}{n_1 + n_2 - 2}}} \quad (1)$$

- $(\bar{Y}_T - \bar{Y}_C)_{ajustée}$  est la différences des moyennes entre les groupes traitement (intervention) et contrôle, ajustée par une analyse de la covariance (ANCOVA)
- $sd_1^2$  est la variance du groupe 1 (définie de la même façon pour l'autre groupe)
- $n_1$  est le nombre de sujets du groupe 1 (défini de la même façon pour l'autre groupe).

Le dénominateur est présenté comme étant l'écart-type groupé inconditionnel (non ajusté) des deux groupes traitement et contrôle. Les auteurs insistent sur le caractère inconditionnel des écarts-types car utiliser les variances conditionnelles, c'est-à-dire conditionnés ou ajustés aux covariables, et notamment aux scores prétests, conduirait à estimer l'effet de l'intervention pour des élèves de même niveau initial, ce qui n'est pas l'objectif d'EEF. Les études avec affectation aléatoire au niveau des établissements et analyse au niveau des élèves doivent également tenir compte de la multiplicité des sources de variation des scores posttests et une des solutions consiste à utiliser un modèle multi-niveau<sup>3</sup>. Les écarts-types utilisés sont basés sur la variance totale de manière à ce que la taille d'effet tienne compte du regroupement des élèves (la plupart du temps dans des établissements) et des différences potentielles entre ces groupes. Ne pas le faire conduirait à surestimer la taille d'effet.

### 3.2. Calcul du coefficient de corrélation intra-classe (ICC)

Pour chaque niveau de regroupement, un coefficient de corrélation intra-classe (*intra-class correlation, ICC*) doit être calculé en utilisant là aussi les variances inconditionnelles (non ajustées). Aucune équation n'est publiée.

### 3.3. Calcul de l'erreur type et de l'intervalle de confiance

Les tailles d'effet calculées à partir des données mesurées sur des échantillons sont les estimations de paramètres de la population. L'incertitude statistique de ces résultats doit être présentée par le calcul d'un intervalle de confiance à 95% et d'une valeur-p. Une erreur type des tailles d'effet est donc calculée, mais aucune équation n'est publiée. Il est juste précisé que les intervalles de confiance qui en sont alors déduits doivent tenir compte du regroupement des élèves dans les établissements car « *ne pas le faire pourrait conduire à des intervalles de confiance trop faibles [...] (et) cela pourrait conduire à des annonces incorrectes et biaisées sur l'étendue des effets possibles* »<sup>4</sup>. Comme cela a été évoqué précédemment, l'interprétation de ces informations a été modifiée en 2022 : en effet, à partir de cette date les résultats ne doivent plus être décrits comme statistiquement significatifs (si  $p < 0,05$ ) ou statistiquement non significatifs (si  $p \geq 0,05$ ). C'est l'estimation ponctuelle (la taille d'effet) accompagnée d'une évaluation de sa validité interne définie par un niveau de preuve (Roques, à paraître) qui permet le mieux d'estimer l'effet d'une intervention. Deux documents publiés sous l'égide d'EEF (EEF, 2020 ; EEF, 2018b) sont probablement à l'origine de ce changement. Le plus ancien s'interrogeait sur le nombre non négligeable d'études EEF qui, suivant les procédures *ad hoc* d'un essai contrôlé randomisé correctement dimensionné, aboutissaient à des résultats positifs mais

<sup>3</sup> Aussi dénommé modèle hiérarchique linéaire (*hierarchical linear model, HLM*)

<sup>4</sup> *Not doing so could lead to the CI being too narrow ... (and) can lead to incorrect and biased claims about the range of possible effects.*

statistiquement non significatifs (selon la présentation classique de ce type de résultat). Dans le plus récent, la recommandation énoncée par le guide EEF 2022 était déjà clairement formulée, les auteurs conseillant d'abandonner la présentation dichotomique habituellement de mise. Aucun de ces documents n'est cité dans le guide EEF 2022. Tous les rapports d'intervention analysés dans cet article ayant été publiés avant 2022, aucun n'a pu mettre en œuvre cette recommandation.

### 3.4. Analyse des données d'intérêt principal

Une analyse des données d'intérêt principal (*primary outcome*) doit être réalisée, où les seules variables explicatives sont les scores prétests, l'appartenance au groupe intervention ou contrôle et les caractéristiques du design de l'étude (c'est-à-dire le regroupement des élèves dans des établissements). Les études d'efficacité en conditions réelles doivent limiter leur analyse à une seule mesure des compétences académiques (afin de limiter le risque d'erreur de type 1, c'est-à-dire de faux positif), et ce sont les résultats de cette analyse qui sont publiés dans le résumé. D'autres variables dépendantes (*secondary outcome*) peuvent être analysées, comme par exemple les résultats à un autre test, mais elles sont clairement identifiées et les résultats les concernant n'apparaissent pas dans le résumé. Enfin des analyses complémentaires, comme des analyses de sous-groupes, peuvent être conduites pour évaluer l'impact, par exemple, du niveau socioéconomique des élèves (en distinguant les élèves éligibles aux repas gratuits ou encore élèves FSM pour *Free School Meal*) ou de leur genre. L'unique équation linéaire (2) proposée dans le guide EEF 2022 concerne ce type d'analyse; aucune explication sur les notations utilisées n'est donnée (voir l'équation (3) dans le chapitre suivant pour plus d'informations).

$$y_{ij} = \alpha + \beta_1 Treat_j + \beta_2 FSMEVER_{ij} + \beta_3 Treat_j * FSMEVER_{ij} + \beta_4 PreTest_{ij} + \varepsilon_{ij} \quad (2)$$

Nous nous intéresserons par la suite uniquement aux méthodes et résultats concernant les analyses de données d'intérêt principal.

## **4. L'analyse statistique des données des 11 études d'efficacité en conditions réelles d'EEF**

L'objectif de ce chapitre est de définir les modèles statistiques qui ont été utilisés par les chercheurs pour calculer les résultats publiés dans chacun des 11 rapports concernant les études d'efficacité en conditions réelles d'EEF présentées ci-dessus. Ces résultats sont les tailles d'effet et leurs intervalles de confiance publiés dans les résumés, ainsi que les coefficients de corrélation intra classe, soit un total de 33 résultats pour les 11 rapports. Tous ont bien été publiés dans les rapports, à l'exception des coefficients de corrélation intra classe des rapports C1 et C2.

Deux questions se posent et qui concernent pour la première les modèles mathématiques utilisés (et donc les équations employées) et pour la seconde les données<sup>5</sup> utilisées. Pour y répondre, les trois étapes définies ci-dessous ont été suivies.

- Dans un premier temps, la lecture des 11 rapports d'intervention mais aussi de certains plans d'analyses statistiques a permis d'identifier des modèles statistiques ainsi que les données susceptibles d'être utilisées pour calculer les résultats. Des informations complémentaires ont également été demandées aux auteurs des 11 rapports.
- Devant l'absence de certaines informations, des hypothèses concernant essentiellement les équations employées ont été émises dans un second temps.

---

<sup>5</sup> Dans les chapitres précédents, le terme « donnée » faisait référence aux scores des élèves ; dans ce chapitre, ces données sont des valeurs qui n'ont pas pu être recalculées par l'auteur de cet article.

- Dans un troisième temps, les recalculs des tailles d'effet, de leur intervalle de confiance et des coefficients de corrélation intra classe ont enfin permis d'évaluer la pertinence de ces hypothèses.

Le tableau 2 synthétise l'ensemble des informations récoltées après lecture des différents rapports.

#### 4.1. Les informations publiées

Ces informations concernent les modèles statistiques mais aussi les données nécessaires pour effectuer les recalculs.

##### Les modèles linéaires publiés

Les études sont toutes des essais contrôlés randomisés au niveau des établissements (*cluster-randomized controled trials*) : l'affectation aléatoire des élèves dans le groupe intervention ou le groupe contrôle s'est faite au niveau des établissements et l'analyse des scores au niveau des élèves. Pour les 9 rapports A, B, C1, C2, F, G, H, I et J, deux niveaux sont considérés dans le modèle statistique : le niveau 1 qui est le niveau de l'élève, et le niveau 2 qui est le niveau des établissements. Le rapport D considère deux niveaux supplémentaires : le niveau assistant d'éducation (*Teaching Assistant*) et le niveau région. Le rapport E ne prend pas en compte le regroupement des élèves dans les établissements. En ce qui concerne la publication d'une équation linéaire, on note que :

- Les 3 rapports B, H et I ne publient aucune équation linéaire.
- Le rapport D publie une équation prenant en compte 3 niveaux de regroupements.
- Le rapport E publie une équation linéaire ne prenant pas en compte le regroupement des élèves dans les établissements.
- Les 6 rapports A, C1, C2, F, G et J publient une équation du type de l'équation (3) présentée ci-dessous (tableau 2, ligne 2).

Les explications proposées ici ont été récoltées après lecture des différents rapports et s'appuient également sur une littérature plus vaste concernant les analyses multiniveaux (Givord et Guillerm, n.d. ; Roques, 2023). Elles concernent le cas le plus fréquemment rencontré d'une analyse à deux niveaux. On note  $Y$  la variable à expliquer dépendante (le score posttest des élèves) considérée comme une variable aléatoire.  $X$  et  $T$  sont les deux variables explicatives indépendantes score prétest et condition (qui indique si l'élève appartient au groupe traitement ou au groupe contrôle). Ces deux variables ne sont pas aléatoires : le score prétest a un effet fixe et la condition d'un élève est contrôlée par le chercheur ; aucune inférence n'est attendue. Le score prétests est une variable continue (et peut prendre n'importe quelle valeur dans un intervalle donné), la variable condition est discrète (elle ne peut prendre que deux valeurs, 0 ou 1).

Le modèle linéaire publié dans les 6 rapports d'évaluation A, C1, C2, F, G et J est le suivant (avec parfois des notations différentes) :

$$Y_{i,j} = \beta_0 + \beta_1 T_i + \beta_2 X_{i,j} + U_j + \varepsilon_{i,j} \quad (3)$$

- $i$  est le rang de l'élève.
- $j$  est le rang de l'établissement.
- $Y_{i,j}$  est le score posttest observé de l'élève  $i$  dans l'établissement  $j$ .
- $T_i$  est la variable indicatrice « appartenance au groupe traitement » ; elle prend la valeur 1 quand l'élève est dans le groupe traitement et la valeur 0 quand l'élève n'est pas dans ce groupe. On la dénomme aussi variable condition.
- $X_{i,j}$  est le score prétest de l'élève mesuré avant l'intervention.

- $\beta_0$  est l'intercept.
- $\beta_1$  est la différence des moyennes des scores posttests des groupes intervention et traitement ajustée à la covariable score prétest (puisque cette dernière est présente dans le modèle). Ce coefficient de régression sera dénommé par la suite coefficient condition.
- $\beta_2$  est le coefficient de régression de la variable score prétest.
- $U_j$  est l'erreur (variable aléatoire non observée) au niveau des regroupements d'élèves (niveau 2) ; elle traduit l'effet aléatoire des établissements.
- $\varepsilon_{i,j}$  est l'erreur (variable aléatoire non observée) au niveau élève (niveau 1)

Les hypothèses sur lesquelles s'appuie le modèle, notamment pour conduire des tests statistiques sont les suivantes :

- Les erreurs  $U_j$  et  $\varepsilon_{i,j}$  sont indépendantes des variables explicatives.
- Les erreurs  $\varepsilon_{i,j}$  (au niveau élève) sont indépendantes entre elles et normalement distribuées autour de leur moyenne égale à 0 avec une même variance  $\sigma_\varepsilon^2$  ; cette variance élève (ou encore variance intra) ne dépend pas de l'établissement ;  $\varepsilon_{i,j} \sim \mathcal{N}(0, \sigma_\varepsilon)$ .
- Les erreurs  $U_j$  (au niveau établissement) sont normalement distribuées autour de leur moyenne égale à 0 avec comme variance  $\sigma_U^2$  qui est la variance établissement (ou encore variance inter) ;  $U_j \sim \mathcal{N}(0, \sigma_U)$ .
- La variance totale est égale à  $\sigma_\varepsilon^2 + \sigma_U^2$ .

Les effets des variables explicatives traitement et score prétest sont supposés fixes (leurs valeurs sont circonscrites à un ensemble de modalités fini) et les coefficients de régression  $\beta_0, \beta_1$  et  $\beta_2$  ne pas varier en fonction des établissements. Les effets du regroupement des élèves dans les établissements sont considérés comme aléatoires (et non fixes) : ces établissements sont un échantillon représentatif d'un ensemble plus vaste sur lequel on souhaite inférer. On calcule alors des estimations des paramètres  $\beta_0, \beta_1, \beta_2, \sigma_\varepsilon^2$  et  $\sigma_U^2$  ainsi que leurs erreurs types.

Comme nous l'avons noté précédemment, le rapport E présente une équation linéaire sans terme  $U_j$ , et dans le rapport D des termes supplémentaires ont été considérés pour prendre en compte les deux autres niveaux d'agrégation des élèves.

### Le calcul de la taille d'effet et de son erreur type

Dans les 8 rapports A, C1, C2, D, F, G, H, et J la taille d'effet est présentée comme étant le g de Hedges (c'est-à-dire que dans le texte on retrouve indiqué à un moment ou un autre les termes « g de Hedges »). Il reste donc trois rapports qui ne mentionnent pas le g de Hedges : le rapport B, dans lequel il est explicitement écrit que la taille d'effet calculée n'est pas un g de Hedges, les rapports E et I où aucun nom particulier n'est donné à la taille d'effet.

Parmi les 8 rapports affirmant que la taille d'effet calculée est le g de Hedges, on note que (tableau 2, ligne 4) :

- Le rapport G, publie une équation qui s'apparente à l'équation (1) et comporte un facteur correctif  $\gamma$  permettant de tenir compte du regroupement des élèves.
- Les 3 rapports C1, C2 et F, publie une équation qui s'apparente à l'équation (1) mais qui est erronée.
- Les 4 rapports A, D, H et J ne publient pas d'équation sur le modèle de l'équation (1).

Dans les 3 rapports B, D, F une autre équation est publiée qui permet de calculer la taille d'effet. Il s'agit du rapport du coefficient condition (qui est la différence des moyennes des deux groupes, ajustée aux

scores prétests), sur la racine carrée de la somme des composantes de la variance totale qui est égale à l'écart-type total (équation (4) ; tableau 2, ligne 5).

$$ES = \frac{\beta_1}{\sqrt{\sigma_u^2 + \sigma_e^2}} \quad (4)$$

*ES* désigne ici la taille d'effet réelle de la population, et non son estimation comme c'était le cas pour l'équation (1)<sup>6</sup>. Pour calculer l'estimation de la taille d'effet, on remplacera les paramètres  $\beta_1$ ,  $\sigma_u^2$  et  $\sigma_e^2$  par leurs estimations. Cette équation trouve probablement sa source dans un article de Hedges (2007) qui publie une équation du même type. Le guide EEF 2022 et les rapports H et J citent cette référence.

On note que :

- Le rapport B qui annonçait ne pas calculer un *g* de Hedges donne une formule s'apparentant à l'équation (4) mais erronée (sans radical au dénominateur).
- Le rapport F donne également une formule du type de l'équation (1).
- Les rapports D et F annoncent calculer un *g* de Hedges.
- Le dénominateur de l'équation (4) dans le rapport D comporte 4 termes (puisque deux niveaux de regroupement supplémentaires sont pris en compte dans cette étude).
- Le rapport I donne une équation similaire mais très sommaire et qui ne permet aucun recalcul<sup>7</sup>.

Au final, les rapports A, E, H, I et J ne publient aucune équation permettant de recalculer la taille d'effet.

Aucune information n'a été trouvée ni sur le calcul de l'erreur type de la taille d'effet ni sur son intervalle de confiance à 95% (voir plus loin pour les hypothèses de calcul).

#### Le calcul du coefficient de corrélation intra-classe (ICC)

Comme nous l'avons déjà signalé<sup>8</sup>, la variance totale est décomposée en une variance inter-établissement  $\sigma_{ij}^2$  et une variance intra-établissement  $\sigma_e^2$ . Un coefficient de corrélation intra-classe est alors défini et calculé, qui est la proportion de la variance totale due au regroupement des élèves dans les établissements. Ce coefficient est égal à 0 quand toute la variance vient du niveau élève (la variance totale est égale à la variance intra ce qui revient à dire que la variance inter est nulle). Il est égal à 1 quand toute la variance vient du niveau établissement (la variance totale est égale à la variance inter, ce qui revient à dire que chaque élève représente à lui seul son établissement ou que tous les élèves d'un même établissement ont le même score, et que la variance intra est nulle). Aucune équation n'est publiée dans les rapports (voir plus loin pour l'hypothèse de calcul).

#### Les données publiées

Les informations publiées et susceptibles d'être utilisées pour calculer les résultats sont synthétisées dans le tableau 2 (il s'agit des lignes 6 à 10 regroupées sous le titre Données publiés). Pour recalculer les tailles d'effet en utilisant l'équations (1) ou (4) on a besoin dans les deux cas de connaître le coefficient condition (tableau 2, ligne 5) ou la différence des moyennes des scores posttests ajustées aux scores prétests (tableau 2, ligne 6). On doit également connaître

---

<sup>6</sup> En toute rigueur, il faudrait noter cette taille d'effet  $\delta$  et suivre en cela l'usage en vigueur dans la littérature.

<sup>7</sup> la taille d'effet est le rapport de la différence ajustée aux scores prétests des moyennes des scores posttests sur l'écart-type au niveau élève obtenue à partir d'un modèle multiniveau non ajusté, sans référence à la décomposition de la variance totale

<sup>8</sup> Voir aussi le SAP du rapport F.

- pour utiliser l'équation (1), les écarts-types des groupes traitement et contrôle (tableau 2, ligne 8) ;
- pour utiliser l'équation (4), les variances des erreurs de niveau 1 et de niveau 2 (tableau 2, ligne 9).

Tous les rapports publient les coefficients conditions ainsi que leurs intervalles de confiance (et parfois leur erreur type) à l'exception des rapports G et H qui publient la différence des moyennes ajustée aux scores prétests et leurs intervalles de confiance. Dans tous les cas, il est indiqué que ces données ont été ajustées aux scores prétests.

En ce qui concerne la dispersion des scores posttests :

- Les rapport C1, C2 et E ne publient aucune donnée.
- Le rapport G est le seul à publier les écarts-types des groupes intervention et contrôle.
- Les sept rapports A, B, D, F, H, I et J publient les composantes intra et inter de la variance totale, c'est-à-dire une variance élève et une variance établissement.

Quand on cherche à déterminer si ces variances sont conditionnelles ou non conditionnelles, les choses sont moins claires. Comme on l'a déjà souligné ci-dessus, les guides EEF publiés en 2022 et 2018 mentionnent que les variances incluses dans les calculs doivent être non conditionnelles. Seul le rapport I affirme explicitement utiliser les variances non conditionnées aux covariables. Pour les autres rapports, aucune information n'est disponible (tableau 2, ligne 10). On notera que le rapport F et publie des variances calculées à partir d'un modèle conditionnel complet (*full conditional model*) et également des variances calculées à partir d'un modèle vide (*null model (empty)*).

La plupart du temps les variances élève et établissement sont présentées dans un tableau donnant tous les résultats de l'analyse statistique du modèle complet (c'est-à-dire avec la covariable score prétest) ; il est donc vraisemblable que les variances soient ajustées aux covariables. Pour chacun des 11 rapports, des compléments d'information ont été demandés par courriel à la personne identifiée comme responsable de l'étude. Ils concernaient :

- pour les 7 rapports A, B, D, F, H, I et J les variances publiées (conditionnelles ou non conditionnelles)
- pour les 3 rapports C1, C2 et E, qui ont le même superviseur, les données elles-mêmes (les variances ne sont pas publiées)
- pour le rapport G la méthode de calcul de la variance de la taille d'effet.

Seules deux réponses éclairantes ont été apportées : les auteurs des rapports B et D ont tous les deux confirmé que les variances utilisées étaient conditionnées aux scores prétests. Les résultats de cette enquête sont donnés en annexe 2.

Au final, parmi les 7 rapports publiant les variances élèves et établissement, seul le rapport I a utilisé des variances non ajustées<sup>9</sup>.

#### 4.2. Les hypothèses

Pour recalculer les tailles d'effet, leurs intervalles de confiance et les coefficients de corrélation intra classe (ICC), on a besoin de connaître les équations utilisées d'une part et les données indispensables au calcul d'autre part. On vient de voir que pour ces premières, des informations disparates sont

---

<sup>9</sup> Le rapport de la dernière étude de l'impact de l'intervention *Maths Champion* utilise également les variances non ajustées.

données quant aux tailles d'effet, et qu'aucune équation n'est publiée ni pour calculer l'intervalle de confiance de la taille d'effet ni le coefficient de corrélation intra classe. Il a donc fallu formuler des hypothèses à leur sujet.

#### Calcul de la taille d'effet

Comme on l'a vu au chapitre 4.2., les rapports B, D, et F publient une équation du type de l'équation (4) et le rapport G publie une équation du type de l'équation (1). Les données publiées par ces 4 rapports ont permis d'exploiter ces modèles et de recalculer les tailles d'effet. Le rapport F publie également une équation du type de l'équation (1) mais ne donne pas les écarts-types des groupes traitement et contrôle. Pour les rapports C1 et C2 les scores post-tests étant standardisés, les auteurs présentent le coefficient condition comme étant égal à la taille d'effet et la taille d'effet n'a donc pas été recalculée.

Les 5 rapports A, E, H, I et J ne publient aucune équation permettant de calculer la taille d'effet. Aucune donnée n'est publiée dans le rapport E, donc aucune hypothèse n'a pu être faite ni aucun recalcul. Les données publiées par les 4 autres rapports m'ont conduites à penser que c'est l'équation (4) qui a été utilisée, et c'est cette dernière que j'ai employée dans les recalculs de leurs tailles d'effet (tableau 2, ligne 11).

#### Calcul de l'intervalle de confiance

On a déjà souligné qu'aucune information n'a été publiée concernant le calcul de l'erreur type des tailles d'effet comme de leur intervalle de confiance à 95%, et que cela concerne aussi bien les guides EEF que les rapports consultés. Mon hypothèse est que

- dans le rapport G, l'intervalle de confiance est calculé en utilisant des équations permettant de tenir compte du regroupement, telles que celles publiées par le What Works Clearinghouse (WWC, 2022, voir en annexe 3),
- pour les 7 rapports pour lesquels l'équation (4) a été utilisée pour recalculer la taille d'effet, la variance de la taille d'effet (le carré de l'erreur type) a pu être calculée en divisant la variance du coefficient condition par la variance totale (c'est-à-dire la somme des variances des erreurs de niveau 1 et 2). Ce qui revient à proposer l'équation (5) suivante :

$$V(ES) = \frac{V(\beta_1)}{\sigma_u^2 + \sigma_e^2} \quad (5)$$

où  $V(ES)$  est la variance de la taille d'effet  $ES$ ,  $V(\beta_1)$  la variance du coefficient condition (égale au carré de son erreur type),  $\sigma_u^2$  la variance inter (ou établissement) et  $\sigma_e^2$  la variance intra (ou élève)<sup>10</sup>. En supposant les estimations des différences de moyennes et les tailles d'effet normalement distribuées on a alors les intervalles de confiances à 95 %

$$\text{Intervalle de confiance de } \beta_1 = [\beta_1 - 1,96 \sqrt{V(\beta_1)}; \beta_1 + 1,96 \sqrt{V(\beta_1)}] = [\inf \beta_1; \sup \beta_1] \quad (6)$$

$$\text{Intervalle de confiance de } ES = \left[ ES - 1,96 \sqrt{\frac{V(\beta_1)}{\sigma_u^2 + \sigma_e^2}}; ES + 1,96 \sqrt{\frac{V(\beta_1)}{\sigma_u^2 + \sigma_e^2}} \right] \text{ Et donc}$$

$$\text{Intervalle de confiance de } ES = \left[ \frac{\inf \beta_1}{\sqrt{\sigma_u^2 + \sigma_e^2}}; \frac{\sup \beta_1}{\sqrt{\sigma_u^2 + \sigma_e^2}} \right] \quad (7)$$

<sup>10</sup> La remarque faite concernant l'équation (4) et l'utilisation des paramètres et non des estimations dans les équations est également valable ici. En toute rigueur,  $V(\beta_1)$  désigne ici la variance de l'estimateur du paramètre  $\beta_1$ ; dans les calculs, ce sont les estimations des variances inter et intra qui sont utilisées.

Tableau 2

Bilan des informations publiées dans les 11 rapports EEF

1	Rapports	A	B	C1	C2	D	E	F	G	H	I	J
2	Équation linéaire (3)	Oui	Non	Oui	Oui	Non*	Non*	Oui	Oui	Non	Non	Oui
3	g de Hedges cité	Oui	Non	Oui	Oui	Oui	Non	Oui	Oui	Oui	Non	Oui
4	ES Équation (1)	Non	Non	Oui**	Oui **	Non	Non	Oui **	Oui ***	Non	Non	Non
5	ES Équation (4)	Non	Oui (erreur)	Non	Non	Oui	Non	Oui	Non	Non	Non	Non
6	Données publiées Coefficient condition #	Oui	Oui	Oui	Oui	Oui	Non	Oui	Non	Non	Oui	Oui
7	Données publiées Différence des moyennes	Non	Oui	Non	Non	Non	Non	Non	Oui	Oui	Non	Non
8	Données publiées Écarts-types des groupes	Non	Non	Non	Non	Non	Non	Non	Oui	Non	Non	Non
9	Données publiées Variances inter et intra	Oui	Oui	Non	Non	Oui	Non	Oui	Non	Oui	Oui	Oui
10	Données publiées Nature des variances	Inconnue	Ajustée <sup>@</sup>	NC	NC	Ajustée <sup>@</sup>	NC	Ajustée	NC	Inconnue	Non ajustée	Inconnue
11	Recalculs ES et IC	<i>(4) et (7)</i>	<i>(4) et (7)</i>	NC et <i>(6)</i>	NC et <i>(6)</i>	<i>(4) et (7)</i>	Non	<i>(4) et (7)</i>	<i>(1) et WWC</i>	<i>(4) et (7)</i>	<i>(4) et (7)</i>	<i>(4) et (7)</i>
12	Recalculs ICC	<i>(7)</i>	<i>(7)</i>	Non	Non	<i>(7)</i>	Non	<i>(7)</i>	Non	<i>(7)</i>	<i>(7)</i>	<i>(7)</i>

ES = taille d'effet. IC = intervalle de confiance. ICC = coefficient de corrélation intra-classe. WWC = What Works Clearinghouse (équations publiées par le WWC, annexe 3). NC = non concerné. \* : un modèle différent est publié. \*\* : la formule publiée contient une erreur. \*\*\* : un facteur correctif est ajouté. # : accompagné d'un intervalle de confiance (IC) et/ou de son erreur type. @ : réponse par courriel (annexe 2).

La ligne 1 donne les références des rapports (voir tableau 1). La ligne 2 indique si une équation linéaire du type de l'équation (3) a été publiée. En ce qui concerne la taille d'effet estimée *ES*, la ligne 3 indique si le *g* de Hedges est mentionné dans le texte du rapport et les lignes 4 et 5 précisent le cas échéant l'équation publiée. Les lignes 6 à 10 concernent les données publiées. Enfin les deux dernières lignes indiquent s'il a été possible de recalculer la taille d'effet et son intervalle de confiance mais également le coefficient de corrélation intra-classe ICC en précisant le numéro des équations utilisées.

Les numéros d'équations présentés en *italique* dans les lignes 11 et 12 correspondent à des hypothèses faites par l'auteur de cet article.

L'équation (7) est publiée dans un document disponible sur le site d'EEF (Ashraf, 2021, p.13) et est également mentionnée dans le rapport de la dernière étude de l'intervention *Maths Champion* publié en 2023. Pour l'utiliser, on a besoin de connaître l'intervalle de confiance du coefficient condition (ou son erreur type) ainsi que les composantes de la variance totale des scores. C'est le cas, comme nous l'avons déjà dit, pour les rapports A, B, D, F, H, I et J (tableau 2, lignes 9 et 11).

Pour les rapports C1 et C2 qui publient des coefficients conditions standardisés et leur erreur type, c'est l'équation (6) qui a été employée.

#### Calcul du coefficient de corrélation intra classe (ICC)

En ce qui concerne l'ICC, aucune équation n'étant publiée, ni dans les guides EEF, ni dans les rapports, les recalculs se sont basés sur l'équation (8) pour les sept rapports A, B, D, F, H, I et J qui fournissaient les variances inter et intra (tableau 2, ligne 12)

$$ICC = \frac{\sigma_U^2}{\sigma_\varepsilon^2 + \sigma_U^2} \quad (8)$$

#### 4.3. Les recalculs

Un fichier Excel est disponible sur le site de l'auteure qui présente l'ensemble des recalculs pour chacun des rapports.

#### Recalculs des tailles d'effet et de leurs intervalles de confiance

Les tailles d'effet des rapports C1, C2 et E n'ont pas été recalculées : pour les deux premiers, les scores étant standardisés, le coefficient condition est donc égal à la taille d'effet et les recalculs n'ont concernés que les intervalles de confiance (en utilisant tout simplement l'erreur type du coefficient condition) ; pour le dernier, c'est l'absence de données qui est en cause. Pour les 8 autres rapports, les tailles d'effets recalculées (voir tableau 3) sont toutes égales au centième près à celles publiées dans les résumés des rapports, à l'exception du rapport J (taille d'effet recalculée égale à 0,20 ; taille d'effet publiée égale à 0,19) et du rapport B. Pour ce dernier, la taille d'effet recalculée est bien égale à une taille d'effet publiée dans le rapport, mais ce n'est pas celle qui est présentée dans le résumé. Aucune trace de la taille d'effet publiée dans le résumé n'est retrouvée ailleurs dans ce rapport.

On notera que pour le rapport A, c'est la variance intra (*within variance*) qui a été utilisée dans le calcul de la taille d'effet publiée dans le résumé ; une taille d'effet est également publiée (*total Hedges'g*), qui elle a utilisé la somme des variances de niveau 1 et de niveau 2 (et qui de ce fait est inférieure à la taille d'effet du résumé). Dans le rapport I, ce sont les variances non ajustées qui permettent de retrouver la taille d'effet publiée dans le résumé (des variances ajustées sont également publiées). En ce qui concerne les intervalles de confiance (tableau 3), il y a égalité parfaite pour les rapports A, C2, D et F ; les différences sont de l'ordre du centième pour les rapports B, C1, G, H, I et J.

#### Coefficient de corrélation intra classe (ICC)

Les ICC ont été recalculés quand cela était possible (tableau 4), c'est-à-dire pour les 7 rapports A, B, D, F, H, I et J qui tous publiaient les variances élèves et les variances établissements. Dans les 3 rapports B, D et F, les variances sont ajustées aux scores prétests, dans 1 rapport (le rapport I), elles sont non ajustées et dans les 3 autres rapports aucune information n'est communiquée. Tous les coefficients recalculés sont égaux aux coefficients publiés selon la précision choisie par les auteurs (au centième ou au millième) à l'exception du rapport A (le coefficient recalculé est égal à 0,21 pour un résultat publié de 0,22) et du rapport D. Ce dernier analyse plus de 2 niveaux et indique utiliser les variances obtenues à partir d'un modèle vide de covariables pour calculer les ICC (p.17). Les variances publiées en annexe

du rapport semblent issues d'un modèle avec covariables et les notations sont confuses ; aucun recalcul ne correspond aux ICC publiés (voir annexe 4).

Tableau 3  
Recalcul des tailles d'effet et de leur IC

		ES recalculée			ES publiée		
		ES	Inf IC	Sup IC	ES	Inf IC	Sup IC
A	1stClass@Number (avec variance établissement)	<i>0,16</i>	<i>-0,07</i>	<i>0,38</i>	<i>0,16</i>	<i>-0,07</i>	<i>0,38</i>
	1stClass@Number (sans variance établissement)	0,18	-0,08	0,43	0,18	-0,08	0,43
B	Affordable Maths Tuition	<i>0,00</i>	<i>-0,31</i>	<i>0,31</i>	<i>0,00</i>	<i>-0,31</i>	<i>0,32</i>
		x	x	x	-0,03	-0,35	0,28
C1	Ark Mathematics Mastery primary	x	-0,01	0,20	0,10	-0,01	0,21
C2	Ark Mathematics Mastery secondary	x	-0,04	0,15	0,06	-0,04	0,15
D	Catch up numeracy	<i>-0,04</i>	<i>-0,22</i>	<i>0,13</i>	<i>-0,04</i>	<i>-0,22</i>	<i>0,13</i>
		x	x	x	-0,04	-0,21	0,13
F	ICCAMS	0,04	-0,07	0,15	0,04	-0,07	0,15
G	Mathematical Reasoning	0,08	-0,02	0,17	0,08	-0,03	0,18
H	Maths Champions	0,10	-0,13	0,32	0,10	-0,13	0,33
I	Shared Maths Year 3	0,01	-0,07	0,08	0,01	-0,07	0,09
	Shared Maths Year 5	0,02	-0,06	0,09	0,02	-0,06	0,10
J	Tutor Trust Affordable Tutoring	0,20	-0,04	0,43	0,19	-0,05	0,44

ES = taille d'effet ; Inf IC = borne inférieure de l'intervalle de confiance ; Sup IC = borne supérieure de l'intervalle de confiance. Les résultats recalculés et publiés sont signalés en *italique* quand ils ne correspondent pas aux tailles d'effet publiées dans le résumé du rapport. Tous les autres recalculs correspondent aux tailles d'effet publiées dans le résumé du rapport.

Notes :

B	En utilisant les données des tableaux 8 et 11 du rapport on obtient la taille d'effet publiée dans le tableau 8 du rapport qui n'est pas la taille d'effet publiée dans le résumé.
D	Les recalculs concernent l'approche classique ; la taille d'effet publiée dans le résumé utilise une approche Bayésienne pour laquelle aucun recalcul n'a été conduit. Le dénominateur de la taille d'effet est égal à la racine carrée de la somme des 4 composantes de la variance totale.
F	La taille d'effet publiée correspond au recalcul avec les variances du modèle complet ; les variances d'un modèle vide (de covariables vraisemblablement) par ailleurs publiées ne semblent pas avoir été utilisées par les auteurs.
G	La taille d'effet et son intervalle de confiance sont recalculés à partir des équations publiées par le What Works Clearinghouse (WWC, 2022) présentées en annexe 3.
H	Les données sont publiées dans le texte du rapport p.42
J	Les données sont publiées dans le texte du rapport p.40

Tableau 4

Recalcul des coefficients intra classe (ICC)

		Variance établissement	Variance élève	ICC recalculé	ICC publié
A	1stClass@Number	3,921	14,695	0,21	0,22
B	Affordable Maths Tuition	3,161284	6,517809	0,327	0,327
D	Catch up numeracy	Voir annexe 4			
F	ICCAMS (modèle complet)	2,884	36,318	0,074	0,074
	ICCAMS (modèle vide)	8,523	103,832	0,076	x
H	Maths Champions	4,2	20,6	0,17	0,17
I	Shared Maths Year 3, variances non ajustées	21,51	268,4	0,074	0,074
	Shared Maths Year 3, variances ajustées	5,34	108,94	0,047	0,047
	Shared Maths Year 5, variances non ajustées	36,4	333,37	0,098	0,098
	Shared Maths Year 5, variances ajustées	7,12	90,88	0,073	0,073
J	Tutor Trust Affordable Tutoring	7,2	17,4	0,29	0,29

La précision des résultats présentés est celle choisie par les auteurs des rapports (au centième près pour certains, au millième près pour d'autres).

#### Bilan des recalculs

Au final, sur les 33 recalculs qu'il était envisageable d'effectuer, il a été possible d'en faire dans 25 cas (tableau 5). Pour les rapports B et D, le recalcul ne concerne pas la taille d'effet publiée dans le résumé. Pour le rapport A, deux tailles d'effet ont été recalculées (dont celle publiée dans le résumé). Les tailles d'effet recalculées sont toutes égales à celles qui ont été publiées. Sur les 24 bornes inférieures ou supérieures des intervalles de confiance, 15 sont strictement égales à celles qui sont publiées, les 9 autres s'écartant de 0,01 de cette valeur. En ce qui concerne les 10 coefficients de corrélation recalculés, là aussi les résultats sont identiques aux valeurs publiées (à l'exception du rapport D comme signalé plus haut).

Tableau 5 : bilan des recalculs

Rapports		A	B	C1	C2	D	E	F	G	H	I	J
Recalculs de/du	la taille d'effet	oui	oui	NC	NC	oui	<b>non</b>	oui	oui	oui	oui	oui
	son intervalle de confiance	oui	oui	oui	oui	oui	<b>non</b>	oui	oui	oui	oui	oui
	coefficient ICC	oui	oui	<b>non*</b>	<b>non*</b>	oui	<b>non</b>	oui	<b>non</b>	oui	oui	oui

\* : ces coefficients n'ont pas été publiés.

Finalement les résultats des recalculs sont très proches des résultats publiés dans pratiquement tous les cas, ce qui valide les hypothèses émises qui concernaient essentiellement les équations utilisées dans les rapports. Pour ces dernières, on peut dresser le bilan suivant :

- Les trois équations (4), (7) et (8) (tableau 6) ont permis de recalculer les tailles d'effet, leur intervalle de confiance et les coefficients ICC pour les 7 rapports A, B, D, F, H, I et J. Parmi eux, les 5 rapports A, B, H, I et J ne publient pas d'équations du type (4) (le rapport B donne une version erronée), et aucun n'explique le calcul de l'intervalle de confiance.
- Les auteurs du rapport G sont les seuls à avoir utilisé l'équation (1) publiée par les guides EEF.

- Pour les 2 rapports C1, C2 les recalculs n'ont concernés que les intervalles de confiance (équation (6)).
- Aucun recalcul n'a été conduit pour le rapport E faute d'information.

Pour les 8 rapports A, C1, C2, D, F, G, H et J la taille d'effet calculée est présentée comme étant le g de Hedges. Cette dénomination semble recouvrir ici des réalités différentes. Ainsi le rapport G utilise une équation différente des autres rapports pour lesquels des recalculs ont été possibles (les rapports A, D, F, H, et J) ; et le rapport B qui dit explicitement ne pas calculer le g de Hedges semble avoir utilisé la même équation que les rapports A, D, F, H et J.

Tableau 6

Les 3 équations utilisées dans les rapports A, B, D, F, H, I et J

Taille d'effet ( $ES$ )	Intervalle de confiance (IC) de la taille d'effet	Coefficient de corrélation intra classe ( $ICC$ )
$ES = \frac{\beta_1}{\sqrt{\sigma_u^2 + \sigma_e^2}} \quad (4)$	$IC \text{ de } ES = \left[ \frac{\inf \beta_1}{\sqrt{\sigma_u^2 + \sigma_e^2}} ; \frac{\sup \beta_1}{\sqrt{\sigma_u^2 + \sigma_e^2}} \right] \quad (7)$	$ICC = \frac{\sigma_U^2}{\sigma_\varepsilon^2 + \sigma_U^2} \quad (8)$

## 5. Conclusions

En un peu plus de 10 ans, les ECR réalisés par EEF ont probablement changé le paysage de nos connaissances basées sur les recherches quantitatives dans le domaine des Sciences de l'éducation. Le fait que tous les résultats soient systématiquement publiés, quelque soient leur ampleur et quelle que soit l'interprétation qui puisse en être tirée, est sans conteste une qualité majeure qui se doit d'être soulignée. Ce travail ne passe pas inaperçu. Des voix se lèvent qui soulignent le peu d'information dont nous disposons après lecture des résultats publiés par EEF, car au final les tailles d'effet sont souvent modestes et statistiquement non significatives (Demack et al., 2021 ; Ashraf et al., 2021 ; Lortie-Forgues et Inglis, 2019 ; Roques, à paraître). Les explications proposées font souvent la part belle aux qualités élevées des procédures mises en œuvre, ces qualités qualifiant aussi bien l'implémentation des interventions que l'analyse statistiques des données qui en sont le fruit. On sait que les ECR conduits sur des échantillons de grande taille et qui mesurent les effets d'une intervention en utilisant des tests standardisés, et non des tests « maison », produisent des résultats plus modestes quand on les compare aux résultats d'études qui ne s'embarrassent pas de telles exigences (Kraft, 2023 ; Cheung et Slavin, 2016). Les études EEF entrent clairement dans cette catégorie, 70% de leurs ECR incluant plus de 1000 participants (Edovald, 2021). Mais si les méthodes et procédures qui charpentent leurs réalisations ont déjà été analysées et discutées (Connoly, 2018 ; Dawson, 2018 ; Edovald, 2021), aucune réplique n'a été à ce jour entreprise dans le but d'explicitier leur analyse statistique, analyse sur laquelle repose en grande partie leur validité interne.

Finalement, sur les 11 rapports analysés, aucun n'a publié toutes les informations suffisantes pour permettre le recalcul des tailles d'effet, de leur intervalle de confiance et du coefficient intra classe (les deux premiers résultats étant affichés en première ligne dans les résumés). Et un seul a suivi les recommandations d'EEF sur l'utilisation des variances non conditionnelles. Il est vrai que cette dernière recommandation n'était peut-être pas énoncée dans les documents cadres sur lesquels ces études se sont appuyées (et qui ne sont plus consultables aujourd'hui). S'il a été si compliqué de présenter d'une façon claire et organisée les méthodes statistiques employées par les auteurs de ces études, c'est que la clarté et la rigueur attendues a parfois fait défaut, notamment en ce qui concerne la description des analyses. La discussion qui suit va porter dans un premier temps sur les documents cadres qui

soutiennent (ou devraient le faire) la réalisation des ECR comme la publication de leurs résultats, pour ensuite revenir sur certaines des méthodes et procédures suivies par les auteurs de ces études primaires et terminer par quelques suggestions.

### 5.1. Les guides EEF

Il est impossible ici de ne pas évoquer les WWC Procedures and Standards Handbooks publiés par le WWC depuis sa création en 2001. Ces documents explicitent en détails la réalisation des méta-analyses du WWC, et cet organisme américain fait actuellement office d'étalon or dans ce domaine notamment pour la qualité de ces textes régulièrement mis à jour<sup>11</sup>. L'exposé détaillé de leurs analyses statistiques débute par le calcul de l'estimation de la taille d'effet de chacune des études sélectionnées dans une méta-analyse, et de sa précision. De nombreux cas de figures sont abordés qui permettent de tenir compte entre autres du design des études. Dans la version publiée en décembre 2022 (WWC, 2022), toutes les équations<sup>12</sup> sont rassemblées dans l'annexe E et sont accompagnées d'exemples numériques. Aucune d'entre elles n'utilise les composantes de la variance totale des scores posttests calculées à partir de modèles multiniveaux. Le WWC m'a cependant confirmé qu'il était possible de calculer un écart-type groupé (utilisé au dénominateur de l'équation (1)) à partir de ces composantes, à condition qu'elles soient uniquement ajustées à la variable condition, ce qui implique que le modèle ne contienne pas alors d'autres covariables (par exemple la variable prétest doit être exclue du modèle linéaire à cette étape)<sup>13</sup>. On a vu que cette recommandation, également reprise dans les guides EEF, n'a pas été suivie par la majorité des auteurs des rapports EEF analysés ici.

Il a déjà été souligné qu'aucune information sur la détermination de la précision des tailles d'effet n'a été publiée dans les guides EEF. Quand j'ai interrogé EEF à ce sujet<sup>14</sup>, il m'a été répondu qu'effectivement cette absence pourrait être corrigée à l'avenir (soit en fournissant une équation, soit en donnant une référence explicite). Les guides EEF ne sont clairement pas du même niveau que les Handbooks du WWC. Un autre point important concerne leur archivage. Contrairement au WWC qui met en ligne les anciennes versions de ses Handbook, il n'est pas possible de connaître les procédures suivies par les ECR qui ont été réalisés par EEF avant la dernière mise à jour du guide. On pourra rétorquer ici que cet archivage ne sera pas de grande utilité puisque les rapports ne font presque jamais référence au guide EEF qui a été (ou aurait dû être) consulté et appliqué. Parmi les 11 rapports analysés dans cet article, seul le rapport de l'intervention *Maths Champion* fait référence au guide sur lequel il s'est appuyé (et il s'agit de la version de 2015 aujourd'hui introuvable). Aucune référence à ces guides n'a été identifiée dans les autres rapports. De leur côté, tous les rapports d'intervention publiés par le WWC font référence à la version du Handbook appliquée.

### 5.2. Les modèles et les données

Pour la majorité des 11 rapports analysés ici, la disponibilité limitée des informations sur les modèles analytiques employés ne permet pas aux lecteurs de comprendre comment les résultats publiés ont été calculés. Seuls 4 rapports publient suffisamment d'informations pour permettre le recalcul de la taille d'effet (et l'une d'elle contient une erreur). Une certaine confusion semble également régner quant à la dénomination de la taille d'effet calculée. Pour 8 rapports il s'agit du  $g$  de Hedges. Si cette dénomination a subi dans le temps des fluctuations quant à sa définition, aujourd'hui de nombreux auteurs (dont les experts d'EEF et du WWC mais aussi les auteurs d'Introduction to Meta-Analysis

---

<sup>11</sup> Le Handbook actuellement en ligne est la 7<sup>ème</sup> version publiée.

<sup>12</sup> Au nombre d'une vingtaine pour les ECR et les études quasi-expérimentales.

<sup>13</sup> Réponse du Help Desk Team du WWC reçue le 9 mars 2023.

<sup>14</sup> Réponse du 6 avril 2023 de Kim Bohling, directrice de l'équipe Evaluation à EEF.

publié en 2021 et qui fait office de référence en la matière<sup>15</sup>) considèrent que le  $g$  de Hedges est le rapport de la différence des moyennes (ajustées ou non aux scores prétests) sur l'écart-type groupé calculé à partir des écarts-types des scores posttests non ajustés des élèves (voir annexe 3). Ce choix est dicté par l'objectif principal poursuivi par les auteurs de synthèses quantitatives : synthétiser et comparer des études qui ont été réalisées dans des contextes multiples. La taille d'effet calculée est alors une estimation du nombre d'écarts-types qu'il faut ajouter à la moyenne des scores des élèves de la population contrôle pour obtenir la moyenne des scores de la population traitée. On a constaté que sur les 8 rapports pour lesquels des recalculs ont été possibles, un seul a clairement utilisé les variances élèves et établissements non ajustées. La même remarque peut être faite en ce qui concerne le calcul du coefficient de corrélation intra classe (ICC). On peut comprendre bien entendu que des études qui ont été réalisées il y a plus de 10 ans pour certaines, n'emploient pas des méthodes d'analyse statistiques préconisées ces dernières années. Enfin l'absence complète d'information sur le calcul de l'erreur type de la taille d'effet dans les guides EEF tout comme dans les rapports, est également un point négatif qui mériterait d'être corrigé. On terminera cette liste en remarquant que certains des résultats (variances, tailles d'effet) ne sont mentionnés que dans le texte et non dans des tableaux facilement identifiables.

Il est probable que le manque d'information que nous venons de souligner soit au moins en partie dû au fait que les calculs sont faits par des logiciels d'analyse statistiques et que les rapports ne sont pas rédigés par des personnes en charge de l'analyse statistiques des données. Mais là aussi, le WWC montre le chemin, en imposant à ses examinateurs un modèle de rédaction des rapports explicite.

### 5.3. Implications

Les éléments présentés dans la liste ci-dessous permettraient d'améliorer la qualité des rapports EEF ce qui renforcerait la confiance que les acteurs de la communauté éducative devraient leur témoigner.

#### Pour les documents cadres

- Archiver les documents qui ne sont plus à jour afin de permettre aux lecteurs de comprendre les procédures qui ont été mises en œuvre dans tous les ECR et d'alimenter la réflexion des chercheurs.
- Expliciter les modèles analytiques utilisés en donnant les équations tenant compte du design des études ainsi que des données disponibles, y compris pour l'erreur type de la taille d'effet.

#### Pour les essais contrôlés randomisés (ECR)

- Uniformiser la présentation des données de façon à faciliter la recherche d'informations notamment quand plusieurs études sont consultées.
- Faire référence au guide EEF sur lequel s'appuie l'étude.
- Expliciter le modèle analytique utilisé, notamment en ce qui concerne le calcul de la taille d'effet.
- Publier les équations permettant de calculer les tailles d'effet et leurs erreurs types si ces équations ne sont pas données dans le guide EEF, par exemple dans le plan d'analyse statistique.

---

<sup>15</sup> On notera que Larry V. Hedges (qui a donc donné son nom à cette méthode de calcul) est l'un des auteurs de ce livre et fait également partie des experts consultés par le WWC.

- Former les chercheurs de façon à ce que les analyses soient conduites de façon similaire (en prenant exemple là encore sur le WWC qui exige de ses examinateurs d'être formés et certifiés).

#### 5.4. Limites

Cet article a été écrit suite à la lecture des différents rapports mentionnés ci-dessus par une enseignante en mathématiques curieuse de comprendre les résultats publiés. Aucune question de recherche n'a donc été pré spécifiée, et cette analyse *post hoc* ne constitue en aucune manière une évaluation définitive des 11 rapports étudiés. Ces documents qui évaluent et détaillent les procédures statistiques mais aussi la mise en œuvre d'une intervention sont denses et complexes, et il n'a été question ici que d'un petit segment de la masse d'informations qu'ils contiennent. Il serait intéressant de poursuivre les investigations de manière à nourrir la réflexion des acteurs de la communauté éducative française.

## Références

- Ashraf, B., A. Singh, G. Uwimpuhwe, T. Coolen-Maturi, J. Einbeck, S. Higgins, & A. Kasim, (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>
- Borenstein, M., L. Hedges, J. Higgins, & H. Rothstein, (2021). *Introduction to Meta-Analysis*. Wiley.
- Cheung, A. C., & R. E. Slavin, (2016). How methodological features affect effect sizes in education. *Educational Researcher*, 45(5), 283-292. <https://doi.org/10.3102/0013189X16656615>
- Connolly, P., C. Keenan, & K. Urbanska, (2018). The trials of evidence-based practice in education: A systematic review of randomised controlled trials in education research 1980–2016. *Educational Research*, 60(3), 276-291. <https://doi.org/10.1080/00131881.2018.1493353>
- CSEN. (2023). [La boîte à idées du CSEN-Quelques pistes pédagogiques fondées sur la recherche](https://www.reseau-canope.fr/conseil-scientifique-de-leducation-nationale-site-officiel/outils-pedagogiques/boite-a-idees-du-csen.html).  
<https://www.reseau-canope.fr/conseil-scientifique-de-leducation-nationale-site-officiel/outils-pedagogiques/boite-a-idees-du-csen.html>
- Dawson, A., E. Yeomans, & E. R. Brown, (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>
- Demack, S., B. Maxwell, M. Coldwell, A. Stevens, C. Wolstenholme, S. Reaney-Wood, B. Stiel, & H. Lortie-Forgues (2021). Review of EEF projects.  
<https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-evaluation-reports-and-research-papers/syntheses-of-eef-evaluations/review-of-eef-projects>
- Dehaene, S., & E. Pasquinelli, (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](https://www.reseau-canope.fr/fileadmin/user_upload/Projets/conseil_scientifique_education_nationale/Ressources_pedagogiques/La_recherche_translationnelle_en_education.pdf)
- Lortie-Forgues, H., & M. Inglis, (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned?. *Educational Researcher*, 48(3), 158-166.  
<https://doi.org/10.3102/0013189X19832850>
- EEF. (2018). Statistical uncertainty in Randomised Controlled Trials.  
<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>

EEF, n.d., EEF annual report septembre 2021- august 2022, <https://d2tic4wvo1iusb.cloudfront.net/production/documents/annual-reports/EEF-Annual-Report-2022.pdf?v=1688892940>

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

Edoald, T., & C. Nevill, (2021). Working out what works: The case of the Education Endowment Foundation in England. *ECNU Review of Education*, 4(1), 46-64. <https://doi.org/10.1177/2096531120913039>

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics* 32, 4: 341 - 370 <https://doi.org/10.3102/1076998606298043>

Lortie-Forgues, H., & M. Inglis, (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned?. *Educational Researcher*, 48(3), 158-166. <https://doi.org/10.3102/0013189X19832850>

Roques, N. (2023). La régression linéaire multiple dans les essais contrôlés randomisés et les méta-analyses en sciences de l'éducation

Roques, N. (à paraître). Les essais contrôlés randomisés au Royaume-Uni : évaluer les interventions efficaces pour favoriser l'apprentissage des mathématiques. *Éducation & Didactique*.

White, H. (2019). The twenty-first century experimenting society: the four waves of the evidence revolution. *Palgrave Communications*, 5(1), 1-7. <https://doi.org/10.1057/s41599-019-0253-6>

WWC. (2022). WWC Procedures and Standards Handbook 5.0. [https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final\\_WWC-HandbookVer5\\_0-0-508.pdf](https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5_0-0-508.pdf)

## **Annexe 1. Les 11 rapports EEF**

**Rapport A.** 1stClass@Number. Evaluation report and executive summary.

<https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/1stclassnumber>

**Rapport B.** Affordable Online MathsTuition. Evaluation report and executive summary.

<https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/affordable-maths-tuition>

**Rapport C1.** Mathematics Mastery Primary. Evaluation Report.

<https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/mathematics-mastery-primary>

**Rapport C2.** Mathematics Mastery Secondary. Evaluation Report.

<https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/mathematics-mastery-secondary>

**Rapport D.** Catch Up® Numeracy. Evaluation report and executive summary.

<https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/catch-up-numeracy-2015>

**Rapport E.** Chess in Schools. Evaluation report and executive summary.

<https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/chess-in-primary-schools>

**Rapport F.** Evaluation Report. Increasing Competence and Confidence in Algebra and Multiplicative Structures (ICAMS). <https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/increasing-competence-and-confidence-in-algebra-and-multiplicative-structur>

**Rapport G.** Mathematical Reasoning. Evaluation report and executive summary.

<https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/mathematical-reasoning>

**Rapport H.** Maths Champions. Evaluation report and executive summary.

<https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/maths-champions>

**Rapport I.** Durham Shared Maths Project. Evaluation report and Executive summary.

<https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/shared-maths>

**Rapport J.** Tutor Trust: Affordable Primary Tuition. Evaluation report and executive summary.

<https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/tutor-trust-effectiveness-trial>

Le rapport publié en 2023 sur la **dernière étude de l'intervention *Maths Champion***

Evaluation Report. Independent evaluation of Maths Champions in nursery to develop children's early numeracy : A two-armed cluster randomised controlled trial

<https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/maths-champions-effectiveness>

## Annexe 2. Demande d'informations par mail

Des informations complémentaires ont été demandées par mail aux 9 responsables des 11 interventions (les rapports C1, C2 et E ont été supervisées par la même personne). Les échanges se sont étalés de décembre 2022 à avril 2023. Pour les interventions C1, C2 et E il s'agissait d'obtenir les variances (élèves et établissement) non publiées ; pour les rapports B, D, H, I et J la demande concernait la nature des variances publiées (afin de savoir si elles étaient conditionnelles ou non), pour les rapports A, F et G les questions portaient sur les méthodes de calculs retenues. Cinq demandes (F, G, H, I et J) sont restées sans réponse (absence de réponse ou adresse inconnue), 1 réponse consistait pour moi à relire les rapports (rapports C1, C2, E), 1 réponse n'a pas aboutit (rapport A), 2 réponses confirment que les variances utilisées sont bien conditionnelles (et non inconditionnelles) pour les rapports B et D.

	Intervention	Questions	Adresse Courrie	Réponses
A	1stClass@Number	Quelles variances utilisées dans calculs ?	terezinha.nunes@education.ox.ac.uk	Echanges, finalement pas de contact avec statisticien possible
B	Affordable Maths Tuiton	Nature des variances ?	<a href="mailto:david.torgerson@york.ac.uk">david.torgerson@york.ac.uk</a> puis <a href="mailto:catherine.hewitt@york.ac.uk">catherine.hewitt@york.ac.uk</a>	Variances calculées à partir du modèle avec covariables
C1	Ark Mathematics Mastery primary	Variance ?	<a href="mailto:j.jerrim@ioe.ac.uk">j.jerrim@ioe.ac.uk</a>	Les réponses sont dans les rapports
C2	Ark Mathematics Mastery secondary	Variance ?	<a href="mailto:j.jerrim@ioe.ac.uk">j.jerrim@ioe.ac.uk</a>	Les réponses sont dans les rapports
D	Catch up numeracy	Nature des variances ?	<a href="mailto:jeremy.hodgen@ucl.ac.uk">jeremy.hodgen@ucl.ac.uk</a>	Variances calculées à partir du modèle complet (les préconisations EEF ont changé après).
E	Chess in primary school	Variance ?	<a href="mailto:j.jerrim@ucl.ac.uk">j.jerrim@ucl.ac.uk</a>	Même interlocuteur que C
F	ICCAMS	Méthode de calcul retenue	<a href="mailto:maria.pampaka@manchester.ac.uk">maria.pampaka@manchester.ac.uk</a>	Pas de réponse
G	Mathematical Reasoning	Calcul de l'erreur type ?	<a href="mailto:l.stokes@niesr.ac.uk">l.stokes@niesr.ac.uk</a>	Pas de réponse
H	Maths Champions	Nature des variances ?	<a href="mailto:hannah.ainsworth@york.ac.uk">hannah.ainsworth@york.ac.uk</a>	Erreur adresse
I	Shared Maths	Nature des variances ?	<a href="mailto:emily.tanner@natcen.ac.uk">emily.tanner@natcen.ac.uk</a>	Erreur adresse
J	Tutor Trust Affordable Tutoring	Nature des variances ?	<a href="mailto:carole.torgerson@durham.ac.uk">carole.torgerson@durham.ac.uk</a>	Pas de réponse

### Annexe 3. Equations du What Works Clearinghouse

Les équations publiées par le What Works Clearinghouse (WWC, 2022) présentées dans cette annexe ont été utilisées pour les recalculs du rapport G (voir leur utilisation dans le fichier Excel sur le site de l'auteure). Elles permettent de tenir compte des scores prétests et du regroupement des élèves dans leurs établissements. De manière à faciliter leur repérage dans le document original du WWC, les notations et la numérotation telle qu'utilisées dans leur ouvrage d'origine a été conservée.

$$SD_p = \sqrt{\frac{(n_i - 1)SD_i^2 + (n_c - 1)SD_c^2}{n_i + n_c - 2}} \quad (2)$$

$$df = \frac{\left[ (N - 2) - 2 \left( \frac{N}{M} - 1 \right) \rho_{ICC} \right]^2}{(N - 2)(1 - \rho_{ICC})^2 + \frac{N}{M} \left( N - 2 \frac{N}{M} \right) \rho_{ICC}^2 + 2 \left( N - 2 \frac{N}{M} \right) \rho_{ICC} (1 - \rho_{ICC})} \quad (21)$$

$\omega = 1 - \frac{3}{4df - 1} \quad (3)$	$\eta = 1 + \left( \frac{N}{M} - 1 \right) \rho_{ICC} \quad (22)$
$\gamma = 1 - \frac{2 \left( \frac{N}{M} - 1 \right) \rho_{ICC}}{N - 2} \quad (20)$	$SE[g] = \omega \sqrt{\frac{n_i + n_c}{n_i n_c} \eta (1 - R^2) + \frac{g^2}{2df}} \quad (26)$
$g = \frac{\omega b}{SD_p} \sqrt{\gamma} \quad (20 \text{ bis})$	

Les notations utilisées sont les suivantes :

$\rho_{ICC}$	Coefficient de corrélation intra-classe
$b$	Différence des moyennes non standardisée entre les groupes comparaison et contrôle
$df$	Degré de liberté
$g$	Taille d'effet standardisée, $g$ de Hedges
$M$	Nombre total de grappes dans une étude avec affectation au niveau des grappes
$N$	Nombre total de sujets dans l'échantillon analytique
$n_c$	Nombre de sujets dans l'échantillon analytique du groupe comparaison
$n_i$	Nombre de sujets dans l'échantillon analytique du groupe contrôle
$R^2$	Corrélation multiple entre les covariables et le score post-test
$SD_c$	Écart-type au niveau individuel du groupe comparaison
$SD_i$	Écart-type au niveau individuel du groupe intervention
$SD_p$	Écart-type groupé des scores au niveau individuel
$SE[g]$	Erreur type de la taille d'effet (le $g$ de Hedges)

#### Annexe 4. Calcul des coefficients intra classes pour l'intervention Catch Up Numeracy (rapport D)

Les intitulés et les nombres en **gras** sont ceux publiés dans le rapport D.

		Annexe E, tableau 5 (approche Baysienne)	Annexe E, tableau 8 et 9 de (approche classique)	ICC publié (tableau 6)
<b>Variance <i>group</i>*</b>		<b>6</b>	<b>0</b>	
<b>Variance <i>school</i></b>		<b>6,6</b>	<b>6,6</b>	
<b>Variance <i>TA</i></b>		<b>10,7</b>	<b>10,3</b>	
<b>Variance <i>student</i></b>		<b>38,4</b>	<b>38,3</b>	
Variance totale		61,7	55,2	
Recalculs des coefficients ICC	<i>Group</i>	0,10	0,00	<b>0,06**</b>
	<i>School</i>	0,11	0,12	<b>0,25</b>
	<i>TA</i>	0,17	0,19	<b>0,09</b>

TA : Teacher Assistant ; \* = ce terme n'est pas explicité dans le rapport ; \*\* = coefficient de corrélation intra classe au niveau Région.

Aucun des coefficients recalculés ici ne correspond à ceux publiés dans le tableau 6 p.26 du rapport. De nombreux résultats sont également présentés en annexe du rapport D, avec des dénominations différentes qui compliquent la compréhension de l'ensemble. Il est possible que la modification du plan d'analyse statistique soit en partie responsable de cette présentation confuse : l'équation linéaire publiée par le plan d'analyse statistique prend en compte les 4 régions britanniques d'où sont issus les établissements en incluant trois variables muettes dans le modèle alors que l'équation linéaire publiée dans le rapport D utilise en lieu et place un intercept « *randomisation group* ».