Mesure et évaluation en éducation

Introduction au calcul de la taille d'effet globale d'une intervention dans les méta-analyses en sciences de l'éducation
Introduction to calculating the overall effect size of an intervention in meta-analyses in educational sciences
Introdução ao cálculo do tamanho do efeito global de uma intervenção em meta-análises nas ciências da educação

Nathalie Roques

Volume 45, Number 3, 2022

URI: https://id.erudit.org/iderudit/1106852ar
DOI: https://doi.org/10.7202/1106852ar

See table of contents

Publisher(s)

ADMEE-Canada

ISSN

0823-3993 (print)
2368-2000 (digital)

Explore this journal

Cite this article

Roques, N. (2022). Introduction au calcul de la taille d'effet globale d'une intervention dans les méta-analyses en sciences de l'éducation. *Mesure et évaluation en éducation*, *45*(3), 1–32. https://doi.org/10.7202/1106852ar

Article abstract

When several comparison group studies relate to the same intervention in the school environment, the synthesis of their results makes it possible to evaluate the effect of this intervention and to meet the expectations of practitioners, but also to guide the future research. It is for this purpose that quantitative summaries or meta-analyses are carried out. For each of the selected studies, an effect size is calculated; when the variable to be explained is a posttest score, we use Hedges' g (standardized difference of the means of the two groups). The overall effect size is then estimated by following in most cases the random effects model. Finally, to determine the interest of the intervention, the results of the meta-analysis are interpreted according to several grids. A worked example is provided to help the understanding of the analytical models implemented in a meta-analysis in simple cases.

Introduction au calcul de la taille d'effet globale d'une intervention dans les méta-analyses en sciences de l'éducation

Introduction to calculating the overall effect size of an intervention in meta-analyses in educational sciences

Introdução ao cálculo do tamanho do efeito global de uma intervenção em meta-análises nas ciências da educação

Nathalie Roques
ID ORCID : 00000-0001-9739-2490
Collège Lacassagne, Lyon, France

Mots clés : effet fixe, effets aléatoires, évaluation, intervention, méta-analyse, taille d'effet

Quand plusieurs études par comparaison de groupes (intervention et témoin) portent sur une même intervention en milieu scolaire, la synthèse de leurs résultats permet d'évaluer l'effet de cette intervention et de répondre aux attentes des praticiens, mais aussi d'orienter les recherches futures. C'est dans ce but que sont réalisées des synthèses quantitatives ou encore des méta-analyses. Pour chacune des études sélectionnées, une taille d'effet est calculée, qui est le g de Hedges (différence standardisée des moyennes des deux groupes) quand la variable à expliquer est un score post-test. La taille d'effet globale est alors estimée en suivant dans la plupart des cas le modèle des effets aléatoires. Enfin, pour conclure quant à l'intérêt de l'intervention, les tailles d'effet sont traduites en nombre de mois de progrès. Un exemple numérique est proposé pour faciliter la compréhension des modèles analytiques mis en œuvre dans une méta-analyse pour les cas les plus simples.

Key words: effect size, evaluation, fixed effect, intervention, meta-analysis, random effects

When several comparison group studies relate to the same intervention in the school environment, the synthesis of their results makes it possible to evaluate the effect of this intervention and to meet the expectations of practitioners, but also to guide the future research. It is for this purpose that quantitative summaries or meta-analyses are carried out. For each of the selected studies, an effect size is calculated; when the variable to be explained is a posttest score, we use Hedges' g (standardized difference of the means of the two groups). The overall effect size is then estimated by following in most cases the random effects model. Finally, to determine the interest of the intervention, the results of the meta-analysis are interpreted according to several grids. A worked example is provided to help the understanding of the analytical models implemented in a meta-analysis in simple cases.

Palavras-chave: avaliação, efeito fixo, efeitos aleatórios, intervenção, meta-análise, tamanho do efeito

Quando vários estudos por comparação de grupos (intervenção e controlo) se referem à mesma intervenção em ambiente escolar, a síntese dos seus resultados permite avaliar o efeito desta intervenção e ir ao encontro das expectativas dos práticos, mas também orientar futuras investigações. É para esse fim que são realizados resumos quantitativos ou meta-análises. Para cada um dos estudos selecionados é calculado um tamanho de efeito, que é o g de Hedges (diferença padronizada entre as médias dos dois grupos) quando a variável a ser explicada é uma pontuação pós-teste. O tamanho do efeito global é então estimado seguindo, na maioria dos casos, o modelo de efeitos aleatórios. Finalmente, para concluir sobre o interesse da intervenção, os tamanhos dos efeitos são traduzidos no número de meses de progresso. Um exemplo numérico é proposto para facilitar a compreensão dos modelos analíticos implementados numa meta-análise para os casos mais simples.

Note de l'auteure : La correspondance liée à cet article peut être adressée à nathalie.roques@yahoo.fr

Introduction

De nombreuses études par comparaison de groupes ont été menées en sciences de l'éducation ces 60 dernières années dans le but de mesurer les influences d'une pratique liée à l'enseignement (comme une méthode pédagogique ou l'utilisation d'un manuel scolaire ou d'une application déployée sur ordinateur) sur ce que les élèves apprennent (Hedges & Schauer, 2018). Quand plusieurs études de ce type analysent les effets d'une même intervention, synthétiser leurs résultats permet alors d'évaluer l'efficacité globale de cette dernière. Cette information intéresse bien évidemment les acteurs de la communauté éducative, mais également les scientifiques qui orientent leurs recherches en se basant sur les résultats de travaux antérieurs. Les synthèses qui font appel à des procédures statistiques sont des méta-analyses (terme utilisé pour la première fois en 1976 par Gene Glass qui le définit comme une « analyse d'analyses »). Leurs réalisations s'appuient sur un cadre conceptuel et méthodologique qui a fait l'objet de nombreuses recherches menées en grande partie aux États-Unis (Chalmers, 2015 ; Pigott & Polanin, 2020). Des organismes comme le What Works Clearinghouse (WWC) aux USA et la Education Endowment Foundation (EEF) au Royaume-Uni ont également contribué au développement de ce pan de la recherche en publiant de nombreux documents[1] qui encadrent la réalisation de leurs méta-analyses. Les résultats de ces dernières sont également des éléments d'information importants que les praticiens comme les pouvoirs publics ne peuvent pas négliger (voir par exemple la traduction d'un Guide des Pratiques du WWC (Roques, 2022a)).

L'objectif de cet article est de faciliter la compréhension des procédures mises en œuvre dans les méta-analyses de manière à ce qu'un large public soit capable d'en interpréter correctement les résultats et, au-delà, d'améliorer les pratiques d'analyse pour une science de l'éducation de

1. Librement accessibles sur leurs sites internet https://ies.ed.gov/ncee/wwc/ et https://educationendowmentfoundation.org.uk/

qualité accrue. Si ces procédures sont bien documentées dans de nombreux textes anglo-saxons, bien peu d'informations sont disponibles en français. Ce texte devrait contribuer à combler cette lacune.

Une présentation générale des méta-analyses

Il faut distinguer deux niveaux de recherche. Le premier concerne les études qui comparent en milieu scolaire ordinaire deux groupes d'élèves, un groupe pour lequel une intervention est mise en place (le groupe intervention) et un groupe qui n'est pas soumis à cette intervention (le groupe témoin) : ce sont des études par comparaison de groupes. Cette catégorie regroupe des essais contrôlés randomisés (ECR), où la répartition des élèves dans les deux groupes est aléatoire, et des études quasi-expérimentales quand cette répartition n'est pas aléatoire. Certaines méta-analyses sélectionnent également des études par régression de la discontinuité ou des études de cas, mais elles sont moins fréquemment rencontrées et nous les laisserons de côté ici. D'autres types d'études sont par contre systématiquement exclus des méta-analyses, comme les études qui n'évaluent l'intervention que sur un seul groupe d'élèves (sans groupe témoin).

Quand plusieurs études indépendantes ont été menées sur une même question de recherche (ce sont les études primaires), on passe ensuite au deuxième niveau de recherche en réalisant une méta-analyse (qui est une étude secondaire). Ces deux niveaux de recherche utilisent les mêmes concepts théoriques pour conduire une analyse statistique des données et ont un même objectif : caractériser l'effet d'une intervention sur les élèves.

Les méta-analyses sont réalisées en suivant trois étapes majeures : 1) l'identification, la sélection et l'évaluation des études primaires, 2) le calcul de résultats permettant d'évaluer l'effet global d'une intervention, 3) la publication de ces résultats. La description qui suit reprend les grandes lignes directrices encadrant la réalisation des méta-analyses conduites par le WWC (2022).

La première étape : l'identification, la sélection et l'évaluation des études analysant une même intervention

Des critères sont clairement définis et énoncés *a priori* (en amont de l'examen des études) de manière à ce que l'identification, la sélection et l'évaluation des études soient transparentes, systématiques et exhaustives. Les motifs d'exclusion sont explicités et la liste des études exclues est publiée. Par exemple, le WWC classe les résultats des études identifiées

en trois catégories : les résultats conformes sans réserve aux normes WWC, les résultats conformes avec réserve aux normes WWC ou les résultats non conformes aux normes WWC. Les résultats des études non conformes seront exclus de la méta-analyse. Nous verrons par la suite que l'évaluation des études retenues (les études conformes aux normes WWC) aura un impact sur la caractérisation de l'effet de l'intervention.

La deuxième étape : la synthèse et l'interprétation des données

L'analyse statistique des données (qui sont les résultats publiés par les études primaires sélectionnées) peut alors débuter. Dans un premier temps et pour chacune des études sélectionnées, une taille d'effet est calculée; elle quantifie l'ampleur de l'effet de l'intervention sur les élèves mesurée par l'étude. Les tailles d'effet de chacune des études sont ensuite agrégées et une taille d'effet globale est calculée. Toutes ces tailles d'effet (calculées au niveau de chacune des études mais aussi au niveau de leur ensemble) sont ensuite transformées en indicateurs plus simples à comprendre dans le but d'en faciliter l'interprétation. Cet article détaillera plus particulièrement cette deuxième étape.

La troisième étape : la publication des résultats

Les résultats statistiques sont enfin publiés sur des sites internet conçus pour s'adresser au grand public. Il est également possible de télécharger des rapports qui détaillent les procédures et les méthodes suivies, pour ceux qui souhaitent en savoir plus. Ce troisième point distingue nettement les méta-analyses des analyses primaires (d'audience le plus souvent restreinte à un cercle réduit de chercheurs) dont elles sont le fruit. Les sites internet comme Find What Works du WWC et Education Endowment Foundation sont de bons exemples de cet effort de communication.

L'ensemble des concepts et des méthodes mis en œuvre pour réaliser des méta-analyses dans le domaine des sciences de l'éducation sont explicités de manière complète et détaillée dans *The Handbook of Research Synthesis and Meta-Analysis* (Cooper et al., 2019). Pour aller à l'essentiel, la lecture de *Introduction to Meta-Analysis* (Borenstein et al., 2009) permet de saisir le cadre méthodologique qui soutient la réalisation de ces synthèses quantitatives et plus particulièrement de comprendre les procédures statistiques déployées, mais aussi le sens à donner aux résultats. Les nombreux exemples résolus (pour lesquels des fichiers csv sont téléchargeables sur le site internet www.meta-analysis.com) font de cet ouvrage un outil pédagogique particulièrement utile aux néophytes. Enfin, la dernière

version du *WWC Procedures and Standards Handbook 5.0* (WWC, 2022) présente de façon claire et synthétique l'ensemble des formules applicables aux résultats des études sélectionnées dans ses annexes F et G.

Cet article n'a d'autre vocation que de proposer une initiation à certaines procédures d'analyse statistique en limitant volontairement leur champ d'application à des situations très simples. Les équations présentées ci-dessous qui permettent de calculer les estimations ponctuelles des tailles d'effet et d'estimer leur précision, dans un premier temps pour une étude primaire, puis pour en ensemble de ces études, ne permettent pas de traiter l'ensemble des questions auxquelles un méta-analyste doit faire face. Pour relever ce défi, une expertise professionnelle basée sur des connaissances approfondies en analyse statistique mais aussi sur une solide expérience du terrain est indispensable.

La taille d'effet d'une étude primaire

Dans ce qui suit, les données brutes exploitées par les études primaires sélectionnées sont des scores d'élèves obtenus après passation d'un test à la fin de l'expérience, ou scores post-tests (ce sont des données continues). Il s'agit d'un cas simple où les élèves sont affectés dans l'un des deux groupes au niveau individuel et où les scores ne sont pas ajustés à des covariables. Dans la plupart des études de grande ampleur, des classes ou des établissements sont affectés dans ces deux groupes et les scores post-tests sont ajustés aux scores prétests (scores des élèves obtenus avant l'intervention). Cette complexité du terrain est évoquée en fin d'article où des pistes qui permettent d'en tenir compte sont proposées. Afin de comparer puis de synthétiser tous ces résultats, il est indispensable de produire dans un premier temps un indicateur commun pour chacune des études : la taille d'effet. Considérons deux échantillons. L'échantillon **a** subit l'intervention dont nous cherchons à évaluer les bénéfices (c'est le groupe intervention) et l'échantillon **b** ne subit pas l'intervention (c'est le groupe témoin). Nous sommes dans la situation où les tailles d'échantillon n_a et n_b, les moyennes m_a et m_b ainsi que les écart-types s_a et s_b des scores post-tests des élèves des deux groupes, sont connus. Ces échantillons sont représentatifs de deux populations, la population **a** (la population traitée, qui est une population fictive) et la population **b** (la population non traitée qui est la population réelle). Au niveau de ces populations, μ_a et μ_b sont les moyennes des scores

post-tests et σ leur écart-type[2]. Ces nombres sont des paramètres de ces populations et ne sont pas connus des chercheurs. Par définition, la taille d'effet de l'intervention au niveau de la population (aussi nommée taille d'effet réelle) est la différence des moyennes standardisée, c'est-à-dire la différence entre les deux moyennes, divisée par l'écart-type des populations. Cela revient à déterminer la différence des moyennes comme un nombre d'écarts-types. Ce paramètre est noté δ

$$\delta = \frac{\mu_a - \mu_b}{\sigma} \quad (1)$$

Comme toujours en statistiques inférentielles, nous cherchons à estimer ce paramètre à partir des observations faites sur les deux échantillons et à évaluer la qualité de cette estimation. Pour répondre à la première demande, nous calculons une estimation ponctuelle de la taille d'effet réelle ; pour répondre à la seconde demande, nous calculons l'erreur type de cette estimation et l'intervalle de confiance à 95 % qui pourra lui être associé.

Les estimations ponctuelles d'une taille d'effet

Les moyennes des populations seront estimées par les moyennes des échantillons m_a et m_b. Notons que c'est la différence des moyennes qui intéresse le chercheur ici et, dans les études utilisant un modèle de régression linéaire multiple, cette différence est égale au coefficient de corrélation de la variable indicatrice « intervention ». En ce qui concerne l'estimation de l'écart-type de la population σ, deux méthodes de calcul sont utilisées et ont donné naissance à deux familles d'estimations de δ : le *d* de Cohen (Cohen, 1988) et le *g* de Hedges (Hedges, 1981), d'une part, le Δ de Glass (Glass & Smith, 1977), d'autre part. Les dénominations fluctuent d'un article à l'autre : dans le cas présent, les définitions utilisées par Borenstein et al. (2009) ont été employées. Pour les deux premières, c'est l'utilisation (ou non) d'un facteur correctif pour des échantillons de petite taille qui fera la différence.

Quand il est raisonnable de penser que les écarts-types des groupes intervention et témoin sont des estimations de l'écart-type de la population, ce dernier est estimé par *s* l'écart-type groupé, qui est la racine carrée de la moyenne des variances pondérée par leurs degrés de liberté. Nous calculons alors *d*, une estimation de la taille d'effet δ :

2. Les modèles statistiques que nous allons décrire supposent que les scores des deux populations ont le même écart-type, qu'ils sont indépendants et normalement distribués.

$$d = \frac{m_a - m_b}{s} \quad (2) \quad \text{avec} \quad s = \sqrt{\frac{(n_a - 1)s_a^2 + (n_a - 1)s_b^2}{n_a + n_b - 2}} \quad (3)$$

En 1981, Hedges montre que le *d* de Cohen est biaisé (son espérance n'est pas égale à δ la taille d'effet de la population) et que ce biais est substantiel pour les échantillons de petites tailles[3]. L'estimation non biaisée est obtenue en multipliant *d* par un facteur multiplicatif correctif ω. Une nouvelle estimation de δ, le *g* de Hedges, ont alors obtenue :

$$g = \omega d = \omega \frac{m_a - m_b}{s} \quad (4) \quad \text{avec} \quad \omega = 1 - \frac{3}{4(n_a + n_b) - 9} \quad (5)$$

C'est l'estimation utilisée par le WWC et EEF pour les études par comparaison de groupes utilisant des données continues avec affectation au niveau individuel. Pour des échantillons de grande taille, le facteur correctif ω est très proche de 1 et est souvent négligé.

Quand l'écart-type du groupe témoin (le groupe **b**) est une bonne estimation de l'écart-type de la population ou quand les écarts-types des groupes témoin et intervention sont très différents (et que l'écart-type groupé ne semble pas estimer correctement l'écart-type de la population), alors le Δ de Glass est utilisé pour estimer la taille de l'effet δ

$$\Delta = \frac{m_a - m_b}{s_b} \quad (6)$$

Ce choix se justifie aussi en soulignant que, dans les études par comparaison de groupes, l'intervention influe sur la moyenne des scores, bien sûr, mais également sur leur écart-type, et qu'il est dans ce cas préférable d'utiliser l'écart-type du groupe témoin comme estimation de l'écart-type de la population. C'est par exemple l'option retenue par Slavin et al. (2009) dans leur méta-analyse sur les méthodes pédagogiques efficaces pour l'enseignement des mathématiques.

Exemple numérique (1/4)

Cet exemple est inspiré de Borenstein et al. (2009, p. 88). Il a été partagé en quatre parties pour suivre au mieux le texte. Le tableau 1 présente les données fictives de six études repérées par les lettres A à F. Il est

3. C'est-à-dire pour des tailles d'échantillons inférieures à 30.

possible de télécharger un fichier Excel sur www.mathadoc.fr (à consulter notamment pour les questions d'arrondis)[4]. Les résultats sont donnés au millième près, ou sont des valeurs exactes.

Tableau 1
Données de six études fictives

Étude	n_a	n_b	m_a	m_b	s_a	s_b
A	65	65	98	92	21	22
B	200	200	94	82	19	17
C	60	60	94	92	22	20
D	40	40	98	88	28	26
E	50	45	98	88	21	22
F	85	85	96	92	21	22

Les calculs du* d *de Cohen, du* g *de Hedges, du Δ de Glass pour l'étude A

Le calcul des tailles d'effet s'effectue selon les trois méthodes vues ci-dessus. Il faut obtenir l'écart-type groupé s pour calculer le d de Cohen et le g de Hedges :

$$s^2 = \frac{(65-1)\times 21^2 + (65-1)\times 22^2}{65+65-1} = 462{,}5 \quad ; \quad d = \frac{98-92}{\sqrt{462{,}5}} = 0{,}279$$

ω le terme correctif qui permet de calculer le g de Hedges, s'obtient de la façon suivante :

$$\omega = 1 - \frac{3}{4(65+65)-9} = 0{,}994 \quad ; \quad g = 0{,}994 \times 0{,}279 = 0{,}277$$

Pour calculer le Δ de Glass, il faut diviser la différence des moyennes par l'écart-type du groupe témoin :

$$\Delta = \frac{98-92}{22} = 0{,}273$$

4. https://www.mathadoc.fr/wp-content/uploads/2023/08/Exemple-numerique_ROQUES_2023.xlsx

Finalement, les résultats sont rassemblés dans le tableau 2 (les taille d'effet sont également présentées avec un arrondi au centième, précision communément adoptée dans les articles publiant les résultats de méta-analyses).

Tableau 2
Tailles d'effet de six études fictives

Étude	É-T groupé	Facteur correctif ω	*d* de Cohen		*g* de Hedges		Δ de Glass	
			au millième	au centième	au millième	au centième	au millième	au centième
A	21,506	0,994	0,279	0,28	0,277	0,28	0,273	0,27
B	18,028	0,998	0,666	0,67	0,664	0,66	0,706	0,71
C	21,024	0,994	0,095	0,10	0,095	0,10	0,100	0,10
D	27,019	0,990	0,370	0,37	0,367	0,37	0,385	0,39
E	21,479	0,992	0,466	0,47	0,462	0,46	0,455	0,46
F	21,506	0,996	0,186	0,19	0,185	0,19	0,182	0,18

Ici, les trois méthodes de calcul donnent pour une même étude des tailles d'effet différentes de 0,05 au plus (c'est le cas de l'étude B). Avec un arrondissement au dixième près, plus aucune différence n'est décelable.

Le calcul des erreurs types

L'erreur type de l'estimation d'un paramètre est l'écart-type de sa distribution d'échantillonnage (par exemple imaginons qu'un grand nombre d'expériences sont faites, toutes de la même manière, et qu'un grand nombre d'estimations de la taille d'effet sont calculées à partir des observations).

Nous calculons une variance de la distribution d'échantillonnage de chacune des trois estimations *d*, *g*, et Δ,.

La variance du *d* de Cohen (notée V_d) est

$$V_d = \frac{n_a + n_b}{n_a n_b} + \frac{d^2}{2(n_a + n_b)} \quad (7)$$

Le premier terme reflète l'incertitude dans l'estimation de la différence des moyennes, le second reflète l'incertitude dans l'estimation de l'écart-type σ. Et comme $g = \omega d$, nous calculons alors la variance de g (notée V_g)[5]

$$V_g = \omega^2 V_d = \omega^2 \left[\frac{n_a + n_b}{n_a n_b} + \frac{d^2}{2(n_a + n_b)} \right] = \omega^2 \frac{n_a + n_b}{n_a n_b} + \frac{g^2}{2(n_a + n_b)} \qquad (8)$$

Par définition $\omega < 1$ donc $\omega^2 < 1$ et $V_g < V_d$. Cela signifie que la précision du g de Hedges est supérieure à la précision du d de Cohen.

Et enfin, pour le Δ de Glass, la variance de Δ notée V_Δ se calcule comme suit :

$$V_\Delta = \frac{n_a + n_b}{n_a n_b} + \frac{\Delta^2}{2(n_b - 1)} \qquad (9)$$

Il faut remarquer que $2(n_b - 1) < 2(n_a + n_b)$ et que $V_\Delta > V_d > V_g$.

Notons enfin que pour toutes ces variances, plus les échantillons sont de tailles importantes, plus les variances sont faibles et la précision des estimations des tailles d'effet augmente.

Les racines carrées de ces variances permettent alors de calculer les écarts-types de ces estimateurs, qui sont des erreurs types s_d, s_g et s_Δ. Le tableau 3 récapitule les différentes formules de ce chapitre.

Tableau 3
Tailles d'effet et erreurs types pour les trois méthodes de calcul

	Tailles d'effet	Erreurs types
d de Cohen	$\dfrac{m_a - m_b}{\sqrt{\dfrac{(n_a - 1)s_a^2 + (n_a - 1)s_b^2}{n_a + n_b - 2}}}$	$\sqrt{\dfrac{n_a + n_b}{n_a n_b} + \dfrac{d^2}{2(n_a + n_b)}}$
g de Hedges	$\left[1 - \dfrac{3}{4(n_a + n_b) - 9}\right] \dfrac{m_a - m_b}{\sqrt{\dfrac{(n_a - 1)s_a^2 + (n_a - 1)s_b^2}{n_a + n_b - 2}}}$	$\sqrt{\left[1 - \dfrac{3}{4(n_a + n_b) - 9}\right] \dfrac{n_a + n_b}{n_a n_b} + \dfrac{g^2}{2(n_a + n_b)}}$
Δ de Glass	$\dfrac{m_a - m_b}{s_b}$	$\sqrt{\dfrac{n_a + n_b}{n_a n_b} + \dfrac{\Delta^2}{2(n_b - 1)}}$

5. Le WWC utilise une formule légèrement différente, avec comme numérateur du second terme $\omega^2 g^2$ à la place de g^2.

Comme nous venons de le voir, le *g* de Hedges est non-biaisé et sa variance est la plus petite. C'est ce qui explique qu'il soit choisi pour estimer les tailles d'effet dans presque toutes les études par comparaison de groupes récentes. Dans la suite de cet article, seules les méthodes et les formules utilisant le *g* de Hedges seront présentées.

L'estimation par intervalle de confiance et test d'hypothèse

Nous verrons plus loin que la précision des tailles d'effet a toute son importance quand il s'agit de caractériser l'effet d'une intervention.

La distribution du *g* de Hedges est une loi de Student non centrée asymétrique qu'il est possible d'approcher par une loi normale pour des degrés de liberté suffisamment grands (Hedges, 1981). Ainsi, δ est estimé en calculant un intervalle de confiance à 95%

$$\left[g - 1{,}96 s_g \; ; \; g + 1{,}96 s_g\right]$$

Si cet intervalle ne contient pas la valeur zéro, nous pourrons conclure à un effet statistiquement significatif au niveau de confiance 0,95. Dans le cas contraire, nous dirons que le résultat est statistiquement non significatif. Cette présentation dichotomique des résultats utilisée par le WWC peut prêter à discussion. Par exemple, la EEF a abandonné cette classification en 2022 (EEF, 2022 ; Roques, à paraître) et se contente de donner l'estimation ponctuelle accompagnée de son intervalle de confiance (dénommé intervalle de compatibilité) sans autre commentaire. Dans la suite de cet article qui s'appuie largement sur les procédures statistiques déployées et publiées par le WWC, les résultats seront présentés comme statistiquement significatifs (ou non).

Il est également possible de suivre la procédure des tests d'hypothèse pour estimer la signification statistique à associer au *g* de Hedges. Ici, un test *Z* sera conduit pour décider s'il est possible d'écarter l'hypothèse nulle H_0 qui est « la taille d'effet au niveau de la population est égale à zéro », ou encore « $\delta = 0$ ». Pour des échantillons de tailles suffisamment grandes, la variable centrée réduite suit une loi normale centrée-réduite sous H_0. La valeur observée Z_{obs} est comparée à la valeur critique au niveau de confiance choisi (p. ex., pour un risque de première espèce $\alpha = 0{,}05$, z_{obs}, est comparée à 1,96 et à –1,96). La valeur *–p* correspondante est également calculée et comparée à α.[6]

6. Dans la lignée de ce qui a été écrit plus haut, depuis 2022, la EEF se contente de publier les valeurs – *p* sans les commenter.

Enfin, les résultats sont représentés par un diagramme en forêt. L'axe des abscisses est gradué en nombre de tailles d'effet. Pour chaque étude, l'estimation de la taille d'effet (carré noir) et son intervalle de confiance à 95 %, qui est représenté par un segment, sont présentés sur une ligne (voir la figure 1 de l'exemple numérique).

Exemple numérique (2/4)

Le calcul des variances et des intervalles de confiance

Nous nous intéresserons d'abord à l'étude A pour calculer les bornes de l'intervalle de confiance et la valeur *–p* associée au test *Z* avec $\alpha = 0{,}05$. La variance de *g* est calculée :

$$V_g = \frac{65 + 65}{65 \times 65} \times 0{,}994^2 + \frac{0{,}277^2}{2(65+65)} = 0{,}031 \text{ et donc } \frac{g}{s_g} = \frac{0{,}277}{\sqrt{0{,}0307}} = 1{,}583$$

Nous calculons une valeur *–p* égale à 0,113. La taille d'effet calculée est statistiquement non significativement différente de 0. L'intervalle de confiance au seuil de 0,95 est [– 0,066 ; 0,621] qui inclut la valeur 0.

Pour l'étude B, le *g* de Hedges et sa variance sont calculés de la même façon :

$$V_g = 0{,}011 \text{ et } \frac{g}{s_g} = \frac{0{,}6644}{\sqrt{0{,}0105}} = 6{,}480$$

Ici, la taille d'effet est statistiquement et significativement différente de 0 (valeur $-p < 0{,}001$) et l'intervalle de confiance au seuil de 0,95 qui est [0,463 ; 0,865] n'inclut pas la valeur 0.

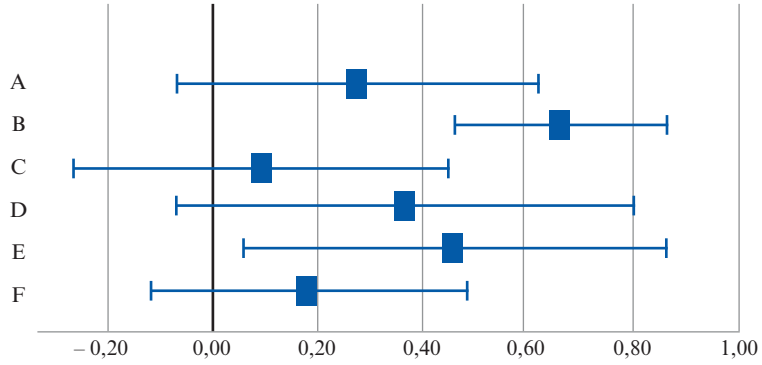
Les résultats des six études fictives sont rassemblés dans le tableau 4. Pour la moitié des études, l'intervalle de confiance inclut la valeur 0 et les tailles d'effet sont statistiquement non significatives (ce qui correspond à une valeur-p supérieure à 0,05).

La figure 1 représente ces résultats sous la forme d'un diagramme en forêt. Les résultats non significatifs sont ceux pour lesquels l'intervalle de confiance est coupé par la droite verticale passant par 0.

Tableau 4
Intervalles de confiance des tailles d'effet des six études fictives

					Intervalle de confiance	
Étude	V_g	s_g	gls_g	**p**	**Borne inférieure**	**Borne supérieure**
A	0,031	0,175	1,583	0,113	-0,066	0,621
B	0,011	0,103	6,480	0,000	0,463	0,865
C	0,033	0,182	0,521	0,603	-0,261	0,450
D	0,050	0,223	1,641	0,101	-0,071	0,804
E	0,043	0,207	2,236	0,025	0,057	0,867
F	0,023	0,153	1,210	0,226	-0,115	0,485

Figure 1
g de Hedges et intervalles de confiance pour les six études fictives



Les méta-analyses

Nous supposons ici que des études primaires indépendantes ont été sélectionnées et que chacune d'entre elles a donné lieu au calcul d'une taille d'effet comme nous venons de le voir. C'est le cas le plus simple. Une étude peut parfois donner lieu au calcul de plusieurs tailles d'effet et, dans ce cas, ces tailles d'effet ne sont pas indépendantes. Cette situation requiert un traitement particulier qui est en dehors des objectifs de cet article.

Le but poursuivi par le méta-analyste est de même nature que celui que poursuit l'auteur d'une étude primaire : calculer un indicateur de position centrale qui est l'estimation d'un paramètre de la population (p. ex.,

la moyenne des scores pour les études primaires) et évaluer la dispersion des données autour de cet indicateur (comme l'écart-type des scores pour le premier type d'études). Les statisticiens conduisant les méta-analyses publiées par des organismes tels que la EEF, le Center for Research and Reform in Education (bestevidence.org/) ou l'organisation Campbell (www.campbellcollaboration.org/) utilisent tous le modèle des effets aléatoires. Ce modèle statistique peut être considéré comme une amélioration du modèle de l'effet fixe qui a l'avantage d'être plus simple à comprendre. C'est pour cette raison que ce chapitre débute par une présentation du modèle de l'effet fixe, qui n'est pas celui qui nous intéresse réellement. Si la liste des organismes cités ci-dessus n'inclut pas le WWC, c'est tout simplement que leurs méta-analyses ne comptent pas suffisamment d'études pour appliquer le modèle des effets aléatoires (voir plus loin).

Dans ce qui suit, k études partageant des caractéristiques communes, comme l'intervention étudiée et les compétences testées, ont été publiées. Pour chaque étude i (compris entre $i = 1$ et k), une taille d'effet a été calculée et est ici le g de Hedges. Pour l'étude i, la taille d'effet est notée g_i, sa variance V_{gi} et son erreur type s_{gi} ; n_{ai} et n_{bi} désignent les tailles d'échantillons des groupes intervention et contrôle et N_i la taille totale de l'échantillon de l'étude i ; donc $N_i = n_{ai} + n_{bi}$.

Le modèle de l'effet fixe

Ce modèle ne peut être appliqué que si les études sélectionnées ont toutes été menées dans des conditions similaires, sur des populations identiques ayant subi une même intervention. La plupart du temps, les études sélectionnées par une méta-analyse en sciences de l'éducation ne satisfont pas à ces critères, mais comprendre ce modèle permettra de comprendre le modèle des effets aléatoires présenté plus loin. Dans cette situation qui peut être qualifiée de théorique, les études sont des répétitions d'un même protocole qui permet d'estimer une seule et unique taille d'effet réelle δ c'est-à-dire la taille d'effet de l'intervention sur la population entière dont sont issus les échantillons. Les différences observées entre les tailles d'effet expérimentales ou quasi-expérimentales sont donc le fruit du hasard qui prévaut dans tout échantillonnage. La taille d'effet globale notée M, estimation de la taille d'effet réelle unique δ, est calculée. De la même façon que pour une étude primaire, la variance de cette taille d'effet globale (notée V_M) est également calculée pour définir un intervalle de confiance ou pour procéder à un test d'hypothèse. Ces derniers éléments auront toute leur importance quand il s'agira de caractériser l'effet de l'intervention.

La taille d'effet globale M calculée en utilisant le modèle de l'effet fixe est égale à la moyenne des tailles d'effet calculées pour chacune des études, pondérée par les inverses de leur variance qui sont donc les poids affectés à chaque étude (noté P_i)[7].

$$M = \frac{\sum_i P_i g_i}{\sum_i P_i} \quad (11) \; ; \; P_i = \frac{1}{V_{gi}} = \frac{1}{s_i^2} \quad (12) \; ; \; V_{gi} = \frac{n_{ai} + n_{bi}}{n_{ai} n_{bi}} + \frac{g_i^2}{2(n_{ai} + n_{bi})} \quad (8)$$

Plus la variance de la taille d'effet de l'étude i est faible, plus son poids dans le calcul de la taille d'effet globale est important. Il est possible de montrer que le poids de la taille d'effet de l'étude augmente si

- N_i la taille totale de l'échantillon augmente,
- pour une taille totale N_i fixée, les tailles des deux échantillons se rapprochent.

La variance de M et son écart-type (noté s_M) sont alors calculés.

$$V_M = \frac{1}{\sum_i P_i} \quad (13) \; ; \; s_M = \sqrt{V_M} = \frac{1}{\sqrt{\sum_i P_i}} \quad (14)$$

Admettons que la statistique M est normalement distribuée. Il sera possible de mener un test Z et de déterminer si M est statistiquement et significativement différente de zéro en ayant fixé un risque α. L'hypothèse nulle est ici que « la taille d'effet réelle est nulle » ou encore « $\delta = 0$ ». La valeur observée z_{obs} est calculée ainsi que la valeur –p associée.

$$Z_{obs} = \frac{\mathrm{M}}{s_M} = \frac{\sum_i P_i g_i}{\sqrt{\sum_i P_i}} \quad (15)$$

Nous pouvons également définir un intervalle de confiance au niveau de confiance 0,95.

$$\left[\mathrm{M} - 1{,}96 s_M ; \; M + 1{,}96 s_M\right]$$

Si cet intervalle comprend la valeur zéro, nous conclurons à une absence de signification statistique.

7. Dans tout ce qui suit, on a $\sum_i = \sum_{i=1}^{k}$

Le modèle des effets aléatoires

Ce modèle est choisi quand les études primaires sélectionnées ont été menées sur des populations différentes (p. ex., d'âges ou de pays différents) en appliquant des protocoles divers (p. ex., les durées des interventions ou les designs des études sont différents). La plupart du temps, les études primaires rassemblées lors de méta-analyses réalisées dans le domaine des sciences de l'éducation sont hétérogènes et correspondent bien à cette description. C'est donc ce modèle qui est le plus souvent choisi. Dans ce cas, les différences entre les résultats expérimentaux (ou quasi-expérimentaux) obtenus d'une étude à l'autre sont attribuables comme toujours à l'échantillonnage, mais aussi aux caractéristiques diverses évoquées ci-dessus. Nous considérerons que la taille d'effet calculée pour chaque étude *i* estime une taille d'effet réelle unique (notée δ_i) associée aux caractéristiques particulières de l'étude *i* et que ces tailles d'effet réelles δ_i sont elles-mêmes normalement distribuées autour d'une taille d'effet globale réelle δ, avec un écart-type noté τ. Cette taille d'effet globale réelle δ est le paramètre que nous cherchons à estimer.

Nous devons considérer deux distributions :

- la distribution normale des tailles d'effet calculées pour une étude donnée autour de la taille d'effet réelle δ_i de cette étude avec une erreur type (qui est aussi la racine carrée de la variance intra-étude) et que nous savons estimer (voir ci-dessus);
- la distribution normale des tailles d'effet réelles δ_i de l'ensemble des études autour de la taille d'effet globale δ avec un écart-type τ_i (qui est aussi la racine carrée de la variance inter-études τ^2), taille d'effet globale et écart-type que nous cherchons à estimer. Nous noterons ici M^* et T ces estimations.

Il est déjà possible de dresser le bilan suivant concernant les notations et le vocabulaire utilisés :

Pour chaque étude *i*	
Taille d'effet réelle	δ_i
Estimation de la taille d'effet réelle	g_i
Estimation de la variance intra-étude	V_{gi}

Pour l'ensemble des *k* études de la méta-analyse	
Taille d'effet globale réelle	δ
Estimation de la taille d'effet globale réelle	M^*
Variance inter-étude	τ^2
Estimation de la variance inter-étude	T^2

Le modèle des effets aléatoires prenant en compte les différences entre les populations analysées par les études primaires, il est donc possible d'inférer au-delà de ces populations considérées ici comme des échantillons d'un ensemble plus large, ce qui n'était pas possible avec le modèle de l'effet fixe précédemment étudié.

Au niveau des calculs, les principes sont les mêmes que ceux mis en œuvre dans le modèle de l'effet fixe et l'estimation de la taille d'effet réelle moyenne est toujours égale à la moyenne des tailles d'effet expérimentales pondérée par les inverses de leur variance. Il faudra ici rajouter aux variances intra-étude l'estimation de la variance inter-études. La variance de la taille d'effet calculée pour l'étude *i* dans ce modèle est notée V_{gi}^* et la variance et l'erreur type de M^* sont notées V_{M^*} et s_{M^*}.

$$V_{gi}^* = V_{gi} + T^2 \quad (16) \qquad P_i^* = \frac{1}{V_{gi^*}} \quad (17) \qquad M^* = \frac{\sum_i P_i^* g_i}{\sum_i P_i^*} \quad (18)$$

$$V_{M^*} = \frac{1}{\sum_i P_i^*} \quad (19) \qquad s_{M^*} = \frac{1}{\sqrt{\sum_i P_i^*}} \quad (20)$$

Comme pour le modèle précédent, en considérant que M^* est normalement distribuée, nous procédons à un test Z avec comme hypothèse nulle que « la taille d'effet moyenne de la population est égale à zéro » ou encore « $\delta = 0$ » et en associant à ce test une valeur $-p$ avec

$$Z_{obs} = \frac{M^*}{s_{M^*}} \quad (21)$$

Un intervalle de confiance au niveau 0,95 est aussi calculé.

$$\left[M^* - 1{,}96 s_{M^*};\ M^* + 1{,}96 s_{M^*}\right]$$

Soulignons que, par définition, $V_{gi}^* \geq V_{gi}$, donc $s_{M^*} \geq s_M$, ce qui revient à dire que le modèle des effets aléatoires, incluant de fait la variabilité des études dans la variabilité totale, est moins précis que le modèle des effets fixes.

Il va falloir calculer T^2, l'estimation de τ^2. Il est ici admis que :

$$T^2 = \frac{Q - (k-1)}{C} \quad (22) \quad avec\ Q = \sum_i \left[\frac{g_i - M}{s_{g_i}} \right] \quad (23) \quad et\ C = \sum_i P_i - \frac{\sum_i P_i^2}{\sum_i P_i} \quad (24)$$

Si $Q - (k-1) < 0$, la valeur nulle sera attribuée à T^2 (τ^2 ne peut pas être négative).

Il faut un nombre suffisant d'études pour pouvoir estimer T^2 avec assez de précision[8]. Si le nombre d'études sélectionnées est trop faible, c'est le modèle de l'effet fixe qui devra être utilisé, mais il ne permettra pas d'inférer au-delà des populations analysées dans les études primaires.

Les méta-analyses réalisées par la EEF pour identifier les méthodes pédagogiques efficaces (voir les pages du Teaching and Learning Toolkit sur leur site internet) rassemblent un grand nombre d'études et utilisent le modèle des effets aléatoires pour calculer les tailles d'effet globales. Les méta-analyses du WWC, quant à elles, ne comptent le plus souvent qu'une poignée d'études primaires, car elles ne concernent qu'une seule intervention clairement identifiée et doivent respecter un cahier des charges exigeant pour être sélectionnées. La taille d'effet globale est donc calculée en suivant le modèle de l'effet fixe. Et même, si les hypothèses de départ restent celles du modèle des effets aléatoires (les tailles d'effet réelles des études primaires sont reconnues être différentes les unes des autres), les conclusions des méta-analyses du WWC ne pourront pas être étendues au-delà des populations analysées par les études sélectionnées. Ce modèle mixte a été baptisé par le WWC modèle des effets fixes (WWC, 2022).

Exemple numérique (3/4)

Les données des tableaux 2 et 4 sont utilisées pour appliquer successivement le modèle de l'effet fixe puis le modèle des effets aléatoires aux six études fictives.

8. Aucune précision quant au nombre d'études requis n'a été trouvée dans la littérature consultée.

Le modèle de l'effet fixe

Le poids est calculé pour chacune des études. Pour l'étude A, $P = 1/0{,}031=32{,}568$. Les poids des cinq autres études sont obtenus de la même façon (tableau 5).

Tableau 5
Poids des six études fictives (modèle de l'effet fixe)

Étude	$P = 1/V_g$
A	32,568
B	95,129
C	30,352
D	20,055
E	23,449
F	42,700

$$M = \frac{32{,}568 \times 0{,}277 + 95{,}129 \times 0{,}664 + \ldots + 42{,}700 \times 0{,}185}{32{,}572 + 95{,}129 + \ldots + 42{,}700} = 0{,}414$$

$$V_M = \frac{1}{32{,}572 + 95{,}129 + \ldots + 42{,}700} = 0{,}004$$

$$\frac{M}{s_m} = \frac{0{,}414}{\sqrt{0{,}004}} = 6{,}475$$

Une valeur – *p* inférieure à 0,001 est alors calculée. La taille d'effet globale calculée est statistiquement et significativement différente de 0 (test *Z* avec $\alpha = 0{,}05$). L'intervalle de confiance au seuil de 0,95 est [0,289 ; 0,540] qui n'inclut pas la valeur 0.

Le modèle des effets aléatoires

Il faut calculer la variance intra-étude T^2. $k = 6$, donc $k - 1 = 5$

$$Q = \left[\frac{0{,}277 - 0{,}414}{0{,}175}\right]^2 + \left[\frac{0{,}664 - 0{,}414}{0{,}103}\right]^2 + \ldots + \left[\frac{0{,}185 - 0{,}414}{0{,}153}\right]^2 = 12{,}005$$

$$C = 244{,}256 - \frac{32{,}572^2 + \ldots + 42{,}700^2}{244{,}256} = 187{,}729 \quad \textit{et donc } T^2 = \frac{12{,}005 - 5}{187{,}729} = 0{,}037$$

Les variances des tailles d'effet sont calculées en rajoutant à la variance intra-étude la variance inter-étude T^2. Par exemple, pour l'étude A,

$$V_g^* = 0{,}031 + 0{,}037 = 0{,}068 \quad \text{et} \quad P^* = 1/0{,}068 = 14{,}703$$

Les poids des cinq autres études s'obtiennent de la même façon (tableau 6).

Tableau 6
Poids des six études fictives (modèle des effets aléatoires)

Étude	Vg^*	$P^* = 1/V_g^*$
A	0,068	14,702
B	0,048	20,910
C	0,070	14,233
D	0,087	11,471
E	0,080	12,506
F	0,061	16,466

Nous procédons de la même façon que pour le modèle de l'effet fixe pour calculer ensuite les autres résultats.

$$M^* = \frac{14{,}703 \times 0{,}277 + 20{,}910 \times 0{,}664 + \ldots + 16{,}466 \times 0{,}185}{14{,}703 + 20{,}910 + \ldots + 16{,}466} = 0{,}358$$

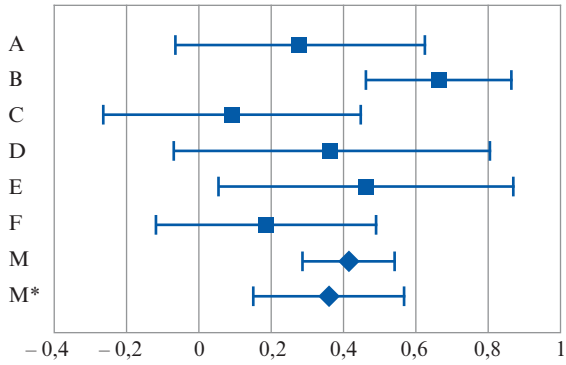
$$V_M^* = \frac{1}{14{,}703 + 20{,}910 + \ldots + 16{,}466} = 0{,}011 \quad \text{et} \quad \frac{M^*}{s_M^*} = \frac{0{,}358}{\sqrt{0{,}011}} = 3{,}404$$

Une valeur *–p* inférieure à 0,001 est alors calculée. La taille d'effet globale calculée est statistiquement et significativement différente de 0 (test Z avec $\alpha = 0{,}05$). L'intervalle de confiance au seuil de 0,95 est [0,152 ; 0,565] qui n'inclut pas la valeur 0.

Un diagramme en forêt (figure 2) présente les résultats de ces deux méta-analyses. Les losanges noirs représentent les estimations ponctuelles des effets globaux. Il est à noter que l'intervalle de confiance de la taille

d'effet est plus grand quand le modèle des effets aléatoires est choisi. Quel que soit le modèle utilisé, les intervalles de confiance des tailles d'effet des études sont calculés à partir de la variance intra-étude V_g (et non V_g^*).

Figure 2
Intervalles de confiance des tailles d'effet des études et des tailles d'effet globales pour les deux modèles



Le bilan

Le tableau 7 présente les éléments essentiels de ces deux modèles statistiques pour les quatre points clés suivants : les hypothèses initiales qui influent sur le choix du modèle, la modélisation des tailles d'effet réelles des études, les poids intervenant dans le calcul de la taille d'effet globale et les limites concernant les inférences.

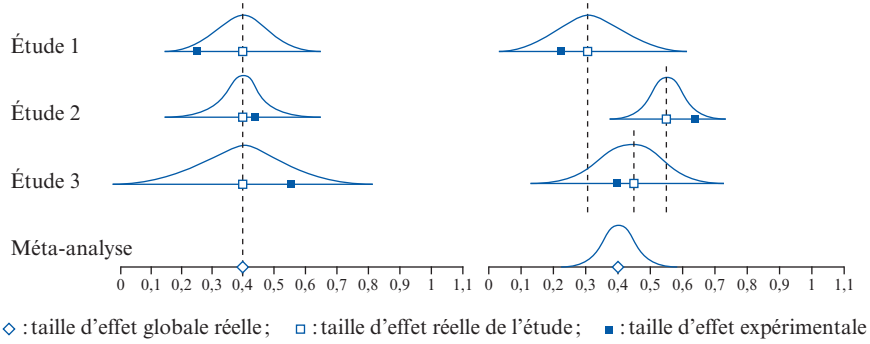
La figure 3 propose une représentation graphique des distributions normales des estimations des tailles d'effet de trois études fictives selon les deux modèles. À gauche c'est le modèle de l'effet fixe qui a été mis en œuvre, à droite, le modèle des effets aléatoires[9]. Dans le modèle de l'effet fixe, la taille d'effet réelle est la même pour toutes les études (les carrés blancs sont sur une même ligne verticale) et c'est également la taille d'effet réelle globale (losange blanc). Dans le modèle des effets aléatoires, les tailles d'effet réelles sont différentes d'une étude à l'autre et normalement distribuées autour de la taille d'effet réelle globale.

9. Cette figure est inspirée des figures 12.4 et 11.3 de Borenstein et al. (2009).

Tableau 7
Comparaison entre modèle de l'effet fixe et modèle des effets aléatoires

Points clés	**Modèle de l'effet fixe**	**Modèle des effets aléatoires**
Hypothèses initiales	Les études sont semblables (répétition d'un même protocole sur une même population).	Les études sont différentes : les interventions, les procédures ou les populations sont différentes.
Tailles d'effet réelles des études	Il y a une seule et unique taille d'effet réelle commune à toutes les études : $\delta_1=\delta_2\ldots=\delta$.	Les tailles d'effet réelles δ_i associées à chaque étude sont différentes et normalement distribuées autour de δ.
Poids et taille d'effet globale	Les poids utilisés pour calculer la taille d'effet globale M sont les inverses des variances intra-étude V_{gi} ; on a $P_i = \frac{1}{V_{gi}}$	Les poids utilisés pour calculer la taille d'effet globale M^* sont les inverses des sommes des variances intra-étude et de la variance inter-études ; on a $V_{gi}^* = V_{gi} + T^2$ et $P_i^* = \frac{1}{V_{gi}^*}$
Inférences	Il est impossible d'inférer au-delà des populations des études analysées.	Il est possible d'inférer au-delà des populations étudiées considérées comme un échantillon d'une population plus vaste.

Figure 3
Exemple de trois études fictives
(À gauche le modèle de l'effet fixe, à droite le modèle des effets aléatoires)



Les interprétations des résultats

Comme cela a déjà été signalé au début de l'article, la finalité des synthèses des études par comparaison de groupes est avant tout de répondre aux attentes concrètes des praticiens comme des pouvoirs publics en leur permettant de trouver les interventions efficaces. La troisième étape de cette analyse statistique consiste à bâtir un pont entre les résultats statistiques qui en sont le fruit et les salles de classe. Pour faciliter cette interprétation des résultats, les tailles d'effet calculées (aussi bien pour une étude primaire que pour une méta-analyse) sont traduites en indicateurs plus intuitifs. De plus, l'effet d'une intervention est parfois classé en fonction de plusieurs paramètres comme les résultats de l'étude primaire (ou de la méta-analyse) mais aussi la qualité de l'étude primaire (ou la quantité et la qualité des études primaires sélectionnées pour une méta-analyse).

Interpréter une taille d'effet

Nous pouvons traduire la taille d'effet comme étant la différence des moyennes des deux groupes (intervention et témoin) en nombre d'écarts-types. Les personnes habituées à considérer des courbes normales peuvent déjà y voir un élément éclairant. Mais pour des personnes peu familières avec les statistiques, il existe d'autres interprétations plus compréhensibles. Certains utilisent encore la grille d'interprétation de Cohen (Cohen, 1988) qui indique qu'une taille d'effet supérieure à 0,8 est importante, qu'elle est moyenne entre 0,2 et 0,8 et faible sous ce seuil. Cette grille de lecture basée notamment sur une comparaison des tailles de jeunes filles réparties en groupes d'âge ne semble pas adaptée aux sciences de l'éducation où d'autres indicateurs sont actuellement utilisés. Par exemple, la EEF traduit la taille d'effet en un nombre de mois de progrès. Il s'agit plus précisément du nombre de mois dont un élève moyen du groupe intervention a progressé par rapport à un élève moyen du groupe témoin. La EEF considère que, pour la plupart des scores mesurés au niveau national, un élève britannique progresse de 1 écart-type en une année scolaire; donc, un mois d'études équivaut à 1/12 d'écart-type soit 0,09 écart-type. Le tableau 8, qui associe chaque nombre de mois de progrès à un intervalle des tailles d'effet, concerne les études primaires (EEF, n. d.). Un tableau légèrement différent (ces différences concernent les deux premières colonnes du tableau 8) est publié pour les tailles d'effet globales calculées pour les méta-analyses du *Toolkit* (EEF, 2023).

Tableau 8
Nombre de mois de progrès pour une étude primaire

	Nombre de mois de progrès								
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Taille d'effet minimale	-0,04	0,05	0,10	0,19	0,27	0,36	0,45	0,53	0,62
Taille d'effet maximale	0,04	0,09	0,18	0,26	0,35	0,44	0,52	0,61	0,69

Le WWC propose quant à lui de traduire la taille d'effet par un indice d'amélioration (IA) qui est le changement attendu du rang centile d'un élève moyen du groupe témoin qui serait ensuite soumis à l'intervention. Il peut aussi s'agir de la différence entre le rang centile d'un élève du groupe témoin qui obtient le score d'un élève moyen du groupe intervention et le rang centile d'un élève moyen du groupe témoin. Le calcul de l'indice d'amélioration se fait en deux étapes : il faut d'abord calculer l'indice U3 de Cohen, qui est la fraction des élèves du groupe témoin surpassés par l'élève moyen du groupe intervention (et qui se calcule grâce aux propriétés des courbes normales). En l'absence d'intervention, cet indice est égal à 50 %. Puis, il faut soustraire 50 % à l'indice U3.

Exemple numérique (4/4)

À partir des *g* de Hedges des six études de l'exemple numérique (tableau 2), les indices d'amélioration sont calculés et les nombres de mois de progrès correspondants déterminés (tableau 9).

Tableau 9
Indices d'amélioration (IA) et nombre de mois de progrès des six études fictives (les tailles d'effet sont arrondies au centième comme cela est d'usage)

Étude	***g***	**U3 (%)**	**IA (%)**	**Nombre de mois de progrès**
A	0,28	61	11	4
B	0,66	75	25	8
C	0,09	54	4	1
D	0,37	64	14	5
E	0,46	68	18	6
F	0,19	57	7	3

Par exemple, pour l'étude A, nous pourrons dire que la moyenne du groupe intervention est supérieure de 0,28 écart-type à la moyenne du groupe témoin, ou que l'élève moyen du groupe intervention a progressé de 11 centiles dans la cohorte, ou qu'il a fait autant de progrès grâce à l'intervention qu'il en aurait fait sans intervention en quatre mois.

Le tableau 10 présente les résultats pour les tailles d'effet globales M et M*.

Tableau 10
Indices d'amélioration (IA) et nombre de mois de progrès pour les tailles d'effet globales

Taille d'effet globale	U3 (%)	IA (%)	Nombre de mois de progrès
M = 0,41	66	16	5
M* = 0,36	64	14	5

Caractériser l'effet d'une intervention

L'objectif ici est d'associer un niveau de preuve à l'effet tel qu'il a été calculé, aussi bien pour une étude primaire que pour une méta-analyse. La description qui suit reprend les lignes directrices encadrant l'interprétation des résultats des analyses du WWC (WWC, 2022). Nous avons déjà vu que le WWC classe les études en évaluant leur design et donc leur validité interne. Dans le cas présent, il s'agit de caractériser l'effet d'une intervention en tenant compte à la fois du design de l'étude (ou des études sélectionnées dans le cas d'une méta-analyse) mais aussi de l'ampleur de l'effet (c'est-à-dire de la taille d'effet) et de la précision avec laquelle cet effet a été évalué, par exemple en calculant un intervalle de confiance. Pour une étude primaire comme pour une méta-analyse, le WWC classe l'effet de l'intervention dans l'une des cinq catégories suivantes : 1) preuves fortes (*tier 1*), 2) preuves modérées (*tier 2*), 3) preuves prometteuses (*tier 3*), 4) effets incertains et 5) effets négatifs. Par exemple, pour obtenir le niveau le plus élevé, la taille d'effet (la taille d'effet globale pour une méta-analyse) doit être positive et statistiquement significative, l'échantillon total doit avoir un effectif supérieur à 350 et inclure plusieurs sites. De plus, dans le cas d'une étude primaire, celle-ci doit être conforme sans réserve aux normes WWC ; dans le cas d'une méta-analyse, les résultats d'études conformes sans réserve aux normes WWC doivent représenter plus de 50 % du poids des tailles d'effets calculées et aucun effet négatif ne doit

avoir été publié. Quand une taille d'effet positive est calculée mais qu'elle n'est pas statistiquement significative, pour une étude primaire comme pour une méta-analyse, l'effet de l'intervention est classé par le WWC comme présentant des effets incertains.

5. La complexité du terrain

Les méthodes statistiques présentées dans cet article concernent la situation la plus simple que puisse rencontrer un chercheur :

- les élèves sont affectés au niveau individuel à un groupe (intervention ou témoin),
- leur niveau initial n'a pas besoin d'être pris en compte,
- les moyennes et les écarts-types des scores des groupes intervention et témoin sont publiés, et
- les études ne fournissent qu'un seul résultat.

Comme chacune de ces conditions est le plus souvent contredite en pratique, il faut tenir compte de la réalité du terrain. Les informations données dans le tableau 11 ci-dessous n'ont d'autre vocation que de fournir quelques pistes.

Prenons comme exemple une étude avec affectation au niveau individuel qui a mené une régression linéaire multiple permettant de tenir compte des scores prétests des élèves. Si cette étude ne publie pas les écarts-types non ajustés des groupes intervention et témoin mais le résultat d'un test t ajusté aux scores prétests ainsi que le coefficient de corrélation du modèle linéaire R^2, le g de Hedges pourra être calculé de la façon suivante (WWC, 2022) :

$$g = \omega t \sqrt{\frac{n_i + n_c}{n_i n_c}(1 - R^2)} \quad (26)$$

Et la variance de la taille d'effet est alors égale à

$$V_g = \omega^2 \frac{n_a + n_b}{n_a n_b}(1 - R^2) + \frac{g^2}{2(n_a + n_b)} \quad (27)$$

Tableau 11
La complexité du terrain

Le cas idéal	La réalité	Que faire ?
Les élèves sont affectés au niveau individuel à un groupe	Ce sont des classes ou des établissements qui sont affectés à un groupe et l'analyse se fait au niveau de l'élève.	À l'intérieur d'un même groupe (classe ou établissement) les scores ne sont pas indépendants. Des formules permettent de corriger la taille d'effet mais surtout sa variance, plus particulièrement affectée (WWC, 2022).
Les niveaux initiaux des élèves n'ont pas besoin d'être pris en compte (p. ex., dans un ECR).	Les niveaux initiaux des élèves doivent être pris en compte car ils peuvent influer sur les résultats mesurés après l'intervention (p. ex., dans les études quasi-expérimentales).	Le niveau initial des élèves est mesuré par leurs scores prétest et les moyennes des scores post-tests seront ajustées aux scores prétests (p. ex., en conduisant une ANCOVA ou une régression linéaire multiple).
Les moyennes et les écarts-types des scores des groupes intervention et témoin sont publiés.	Il manque une de ces données et l'étude publie par exemple un test t ou F, ou l'écart-type de tout l'échantillon.	Des formules permettent dans certains cas de calculer la taille de l'effet et sa variance (WWC, 2022).
Chaque étude ne fournit qu'un seul résultat.	Plusieurs résultats sont publiés, concernant plusieurs sous-domaines ou ont été mesurés à différents moments après l'intervention ou sur des sous-échantillons disjoints d'élèves.	L'analyse de l'étude permettra d'identifier le résultat principal ou une taille d'effet est calculée pour plusieurs sous-échantillons ou pour plusieurs résultats principaux (WWC, 2022). D'autres modèles permettent d'agréger des tailles d'effet dépendantes (Hedges et al., 2010).

Conclusion

Cet article avait comme objectif de poser les règles de calcul permettant d'estimer la taille d'effet globale d'une intervention dans le cas simple d'une méta-analyse n'ayant sélectionné que des études expérimentales ou quasi-expérimentales avec affectation des élèves au niveau individuel, sans tenir compte d'éventuelles covariables. Les estimations ponctuelles et leurs variances calculées en suivant le modèle des effets aléatoires ne sont que les premiers éléments de la description quantitative d'un ensemble d'études traitant d'un sujet commun. De nos jours, les regards se tournent plus volontiers vers une analyse de l'hétérogénéité de ces indicateurs (Pigott, 2020; Roques, 2022b). En s'inspirant de modèles statistiques mis en œuvre pour des études primaires, des analyses de sous-groupes (qui sont en fait des ANOVA) ou des méta-régressions (qui sont des régressions linéaires multiples) sont alors conduites (Tipton et al., 2018) et permettent d'explorer les influences que certaines variables modératrices peuvent avoir sur les tailles d'effet. Ces méthodes ont été mises en œuvre par exemple dans la méta-analyse conduite par la collaboration Campbell en 2021 (Dietrichson et al., 2021) et qui porte sur l'enseignement des mathématiques et de la lecture en primaire. Nous pouvons également citer les méta-analyses menées par la EEF pour son Toolkit ou encore la méta-analyse de Slavin et al. (2009). L'objectif n'est plus alors d'analyser l'effet d'une intervention, mais d'identifier des caractéristiques (un élément spécifique commun à plusieurs interventions, par exemple, ou un domaine précis des apprentissages) associées à l'ampleur de l'effet calculé.

Les derniers mots concerneront la complexité parfois peu visible qui caractérise les procédures, les analyses statistiques et les concepts structurant les méta-analyses. La synthèse d'études quantitatives peut en effet séduire un public large de non-initiés, car elle s'apparente dans ses grandes lignes à un simple calcul de moyenne. Cette simplicité de façade représente finalement un défi pour les méta-analystes qui doivent souvent expliciter leurs procédures pour justifier de la qualité de leur synthèse (Berlin & Golub, 2014). Le WWC l'a bien compris. En effet, il exige de ses examinateurs qu'ils soient certifiés après avoir suivi une formation interne obligatoire et s'efforce de développer, de publier et de mettre à jour des documents cadres qui constituent une référence incontournable dans ce domaine.

Révision linguistique : Marie-Claire Legaré
Mise en page : Emmanuel Gagnon
Résumé en portugais : Eusébio André Machado

Réception : 08 décembre 2022
Version finale : 31 mai 2023
Acceptation : 17 juillet 2023

LISTE DE RÉFÉRENCES

Berlin, J., & Golub, R. (2014). Meta-analysis as Evidence. Building a Better Pyramid. *JAMA*, *312*(6), 603-605. https://doi.org/10.1001/jama.2014.8167

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to Meta-Analysis*. Wiley.

Chalmers, I. (2015). Commentary for history special issue of Research Synthesis Method. *Research Synthesis Method*, *6*(3), 268-71. https://doi.org/10.1002/jrsm.1144

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Academic Press.

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

Dietrichson, J., Filges, T., Seerup, J. K., Klokker, R. H., Viinholt, B. C. A., Bøg, M., & Eiberg, M. (2012). Targeted school-based interventions for improving reading and mathematics for students with or at risk of academic difficulties in Grades K - 6: A systematic review. *Campbell Systematic Reviews,* e1152. https://doi.org/10.1002/cl2.1152

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

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

EEF. (2023). Teaching and learning. Early years toolkit guide https://d2tic4wvo1iusb.cloudfront.net/documents/toolkit/Toolkit_guide_v1.2_-_2023.pdf?v=1677829895

Glass, G. (1976). Primary, Secondary, and Meta-Analysis of Research. Educational Researcher, 5(10), 3-8. https://doi.org/10.3102/0013189X005010003

Glass, G., & Smith, M. L. (1977). Meta-Analysis of Psychotherapy Outcome Studies. *American Psychologist*, *32*(9), 752–760. https://psycnet.apa.org/doi/10.1037/0003-066X.32.9.752

Hedges, L. (1981). Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics*, *6*(2), 107-128. https://doi.org/10.2307/1164588

Hedges, L., & Schauer, J. (2018). Randomised trials in education in the USA. *Educational Research*, *60*(3), 265-275. https://doi.org/10.1080/00131881.2018.1493350

Hedges, L., Tipton, E., & Johnson, M. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Method*, *1*(1), 39-65. https://doi.org/10.1002/jrsm.5

Pigott, T., & Polanin, J. (2020). Methodological Guidance Paper: High-Quality Meta-Analysis in a Systematic Review. *Review of Educational Research*, *90*(1), 24 –46. https://doi.org/10.3102/0034654319877153

Roques, N. (2022a). Aider les élèves en mathématiques dès l'école primaire. Guide des pratiques du What Works Clearinghouse. L'Harmattan.

Roques, N. (2022b). La régression linéaire multiple dans les études par comparaison de groupes en milieu scolaire. https://www.mathadoc.fr/wp-content/uploads/2023/08/Regression-lineaire-multiple-et-sciences-de-leducation.pdf

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

Slavin, R., Lake, C., & Groff C. (2009). Effective programs in Middle and High School Mathematics: a best-evidence Synthesis. *Review of Educational Research*, *79*(2), 839-911. https://doi.org/10.3102/0034654308330968

Tipton, E., Pustejovsky, J., & Ahmadi, H. (2018). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods*, *10*(2), 161-179. https://doi.org/10.1002/jrsm.1338

WWC. (2022). WWC *Procedures and Standards Handbook 5.0.* https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5_0-0-508.pdf