

Les formules statistiques utilisées par le What Works Clearinghouse



What Works Clearinghouse
Procedures and Standards
Handbook, Version 5.0

A Publication of the National Center for Education Evaluation at IES

Nathalie ROQUES

Traduction partielle des annexes E, F et G du *What Works Clearinghouse procedures and standards handbook, version 5.0* (2022)

NOTES CONCERNANT LA TRADUCTION	5
PARTIE 1. FORMULES STATISTIQUES POUR CHAQUE RESULTAT D'UNE ETUDE.....	7
1. Études avec affectation au niveau individuel.....	7
Formules générales de la taille d'effet	8
Différences des moyennes non ajustées	9
Tests statistiques pour les comparaisons de moyennes non ajustées	11
Différences des moyennes ajustées aux covariables.....	11
Tests statistiques pour les comparaisons de moyennes ajustées aux covariables.....	13
Autres méthodes de calcul de l'écart-type groupé.....	14
2. Études avec affectation au niveau des grappes	16
Définition de la taille d'effet pour les études avec affectation au niveau des grappes.....	16
La corrélation intraclasse : application à un exemple	17
Erreurs standards pour les analyses non ajustées.....	18
Erreurs standards pour les analyses ajustées aux covariables	19
Analyses des données au niveau des grappes	21
Autres formules pour des études avec affectation au niveau des grappes.....	24
3. Calcul des valeurs-p et significativité statistique	25
4. Calcul de l'indice d'amélioration.....	26
PARTIE2. FORMULES STATISTIQUES POUR DES RESULTATS AGREGES	27
1. Synthèse de sous-échantillons au sein d'une étude.....	27
2. Synthèse des résultats principaux d'une étude dans un domaine donné.....	28
3. Synthèse des résultats de plusieurs études pour d'un domaine donné	29
GLOSSAIRE.....	34

Notes concernant la traduction

Le What Works Clearinghouse réalise des méta-analyses dans le domaine des sciences de l'Éducation depuis sa création en 2002. Les méthodes mises en œuvre par cet organisme pour sélectionner, analyser puis synthétiser les résultats d'études quantitatives sont explicitées dans leurs manuels des procédures et des normes régulièrement mis à jour. Les annexes D, E et F du manuel des procédures et des normes du WWC¹ (version 5.0, publié en 2022) rassemblent les formules utilisées par les statisticiens du WWC et détaillent leurs conditions d'utilisation. Elles ont été partiellement traduites ici : les informations retenues concernent les résultats continus (tels que des scores d'élèves) analysés par des études par comparaison de groupes, car ce sont les types de résultats et d'études les plus fréquemment rencontrés. Ainsi, les éléments concernant par exemple les résultats dichotomiques ou les études par régression de la discontinuité n'ont pas été traduits et les références éventuelles à ces éléments ont été supprimés du texte. Pour faciliter la lecture de cette traduction, les deux annexes ont été fusionnées ; la numérotation des équations, des figures et des tableaux ainsi que les notations utilisées sont celles de la version originale. Pour comprendre les différentes formules proposées, des connaissances en statistiques inférentielles sont indispensables (l'utilisation de tests statistiques et les résultats d'analyses par régression linéaire multiple sont fréquemment invoqués). Les personnes qui ne sont pas familières de ces méthodes de calculs peuvent consulter *Mesurer l'effet d'un traitement. Les méta-analyses en sciences de l'éducation* (ROQUES, avril 2021) et *La régression linéaire multiple dans les études par comparaison de groupes en milieu scolaire* (ROQUES, septembre 202) téléchargeable sur mathadoc.fr

Une traduction du chapitre I de ce même *Manuels des procédures et des normes WWC v.5.0* est disponible sur le site www.mathadoc.fr. Ce chapitre présente un aperçu complet du WWC, de son fonctionnement et des principes qui sous-tendent les normes du WWC.

Cette traduction été rédigée par Nathalie ROQUES en janvier 2023 ne saurait en aucun cas engager la responsabilité du What Works Clearinghouse.

¹ What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE), à télécharger sur <https://ies.ed.gov/ncee/wwc/Handbooks>. Cette version a été révisée en décembre 2022.

Partie 1. Formules statistiques pour chaque résultat d'une étude

Les auteurs des études peuvent rendre compte de leurs analyses de plusieurs façons, en publiant des résultats plus ou moins comparables et plus ou moins utiles dans le cadre d'une synthèse. Le What Works Clearinghouse (WWC) vise à rendre compte des résultats d'une étude de manière cohérente, en utilisant une métrique commune et en tenant compte des différences entre les analyses, différences qui peuvent affecter leurs résultats. Cette première partie explique comment le WWC calcule les tailles d'effet qui capturent l'ampleur des effets d'une intervention, et les erreurs standards qui capturent l'incertitude dans l'estimation de ces effets.

Cette partie comprend les chapitres suivants : (1) études avec [affectation au niveau individuel](#), (2) études avec [affectation au niveau des grappes](#), (3) calcul des [valeurs-p](#) et détermination de la significativité statistique et (4) calcul de l'[indice d'amélioration](#).

1. Études avec affectation au niveau individuel

Le scénario le plus simple pour calculer les tailles d'effet WWC et les erreurs standards concerne les études avec affectation au niveau individuel. Par exemple, une étude pourrait comparer les moyennes d'un groupe intervention et d'un groupe comparaison pour une mesure des compétences en lecture des élèves. L'un des défis que l'on rencontre lorsqu'on souhaite caractériser l'effet d'une intervention concerne la comparabilité entre les différentes mesures. Les différences des moyennes brutes peuvent ne pas être comparables les unes aux autres car utiliser des échelles ou des procédures de notation différentes.

Le WWC utilise comme mesure de la taille d'effet le g de Hedges, qui est une mesure standardisée de l'effet de l'intervention. Cette mesure consiste à présenter la différence des moyennes des groupes intervention et comparaison en un nombre d'unités d'écart-types. Autrement dit, la différence des moyennes brute est divisée par la variabilité au sein des groupes, plaçant les tailles d'effet sur une échelle commune. Cette approche permet au WWC de synthétiser les résultats dans une même étude présentant plusieurs résultats mais aussi pour un ensemble d'études.

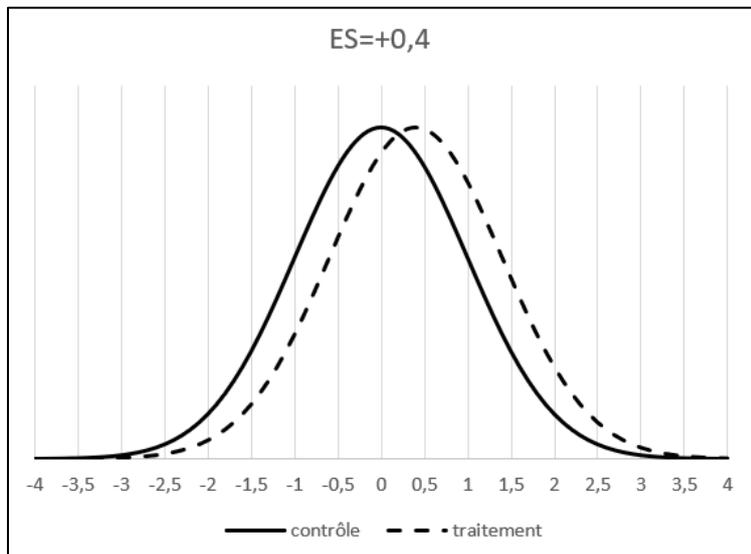
Informations nécessaires pour calculer le g de Hedges

- Une estimation de la différence des moyennes non standardisée (de préférence ajustées aux covariables)
- Les écarts-types pour chacun des groupes intervention et comparaison (toujours basés sur des statistiques non ajustées)
- Les tailles d'échantillon pour chacun des groupes intervention et comparaison

La figure E.1 représente graphiquement une taille d'effet avec $g = 0,40$, ce qui revient à dire que les résultats des tests de lecture se sont améliorés de 0,4 écart-type. La distribution du groupe intervention (ligne pointillée) est décalée vers la droite par rapport à la distribution du groupe comparaison (ligne continue). Considérons l'ampleur de cet effet pour un élève qui obtient un score médian dans le groupe comparaison. Une amélioration de 0,4 écart-type ferait passer le score de cet

élève du 50e centile au 66e centile. Le WWC définit cette augmentation de +16 du rang centile comme l'indice d'amélioration (voir plus loin).

Illustration E.1. Une amélioration de 0,4 écart-type



Formules générales de la taille d'effet

Le WWC calcule le g de Hedges (la taille de l'effet) en divisant une estimation de la différence des moyennes non standardisée b par l'écart-type groupé au sein du groupe SD_p (comme une moyenne pondérée par la variabilité au sein des groupes intervention et comparaison). Pour la différence des moyennes non standardisée b , le WWC préfère les estimations ajustées aux covariables qui contrôlent la différence entre les niveaux initiaux. Ces estimations sont souvent plus précises et comportent moins de biais que les différences des moyennes non ajustées. Cependant, les écarts-types utilisés pour calculer la taille d'effet standardisée doivent toujours être basés sur des statistiques non ajustées pour assurer la comparabilité entre les études. Les écarts-types de scores en provenance de données collectées au niveau national ou d'un état ou d'un district sont également acceptés par le WWC.

La formule permettant de calculer le g Hedges intègre un facteur de correction pour petit échantillon ω , qui est généralement proche de 1, mais légèrement inférieur à 1 lorsque la taille de l'échantillon est petite (Hedges, 1981). Ce facteur de correction est nécessaire pour produire des estimations non biaisées de la taille d'effet au niveau de la population. La formule générale de calcul du g de Hedges pour les résultats continus dans les études avec affectation au niveau individuel peut s'écrire comme suit :

[E.1]

$$g = \frac{\omega b}{SD_p}$$

Les formules pour l'écart-type groupé SD_p et le facteur de correction pour petit échantillon ω sont données par ce qui suit :

[E.2]

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

[E.3]

$$\omega = 1 - \frac{3}{4df - 1}$$

où n_i et n_c sont les tailles d'échantillon pour les groupes intervention et comparaison, SD_i et SD_c sont les écarts-types des groupes intervention et comparaison, et df est le degré de liberté. La formule suivante donne le degré de liberté pour les études avec affectation au niveau individuel :

[E.4]

$$df = N - 2$$

où $N = n_i + n_c$ est le nombre total d'individus. Le g de Hedges est très similaire au fameux d de Cohen ; la différence provient de l'utilisation pour le calcul du g de Hedges du facteur correctif ω qui permet de corriger le biais pour échantillons de petite taille. Le WWC applique toujours cette correction, y compris pour les échantillons importants. Cependant, ce facteur sera très proche de 1 dans des échantillons de grande taille. Par exemple, ce facteur de correction sera $\omega = 0,99$ pour une étude avec affectation au niveau individuel pour un échantillon total de 100 élèves. Si les auteurs de l'étude ont calculé une taille d'effet standardisée non corrigée pour ce biais (telle que le d de Cohen), le WWC multiplier la taille d'effet publiée par l'auteur par le facteur de correction ω si la taille d'effet ne peut pas être calculée d'une autre manière (par exemple en utilisant les moyennes et les écarts-types qui peuvent ne pas être publiés).

Les chapitres suivants présentent des variantes des formules précédentes pour prendre en compte différents types de résultats, ces derniers pouvant en effet être publiés de multiples façons dans les études avec affectation au niveau individuel. D'autres formules sont ensuite présentées qui concernent les études avec affectation au niveau des grappes.

Différences des moyennes non ajustées

Le WWC autorise l'utilisation des moyennes non ajustées dans le calcul des tailles d'effet dans certains scénarios limités, comme pour les essais contrôlés randomisés (ECR) avec un faible risque de biais dû à un changement de composition des groupes (ou encore avec une faible attrition). Les moyennes non ajustées sont acceptables, bien que non optimales, dans de tels cas, car le design de ces études limite les biais dans l'estimation de l'effet de l'intervention. De tels ECR donnent des estimations non biaisées de la moyenne, même sans ajustement pour des différences entre les niveaux initiaux.

Le WWC utilisera donc des moyennes non ajustées pour calculer les tailles d'effet si (a) l'étude n'a pas rapporté suffisamment d'informations pour calculer les tailles d'effet ajustées aux covariables et (b) un ajustement statistique prenant en compte les différences entre les niveaux initiaux n'est pas nécessaire pour satisfaire aux normes WWC. Quand on remplace dans l'équation E.1 b par la différence des moyennes non ajustée $\bar{y}_i - \bar{y}_c$ on obtient le g de Hedges suivant :

[E.5]

$$g = \frac{\omega(\bar{y}_i - \bar{y}_c)}{\sqrt{\frac{(n_i - 1)SD_i^2 + (n_c - 1)SD_c^2}{n_i + n_c - 2}}}$$

Cette équation utilise la formule de l'écart-type groupé au dénominateur pour montrer explicitement quelles statistiques sont nécessaires au calcul du g de Hedges en utilisant cette approche. Cette formule ne peut pas être utilisée si les auteurs de l'étude ne fournissent pas toutes les statistiques incluses, telles que les écarts-types des groupes intervention et comparaison.

La formule suivante fournit l'erreur standard pour les tailles d'effet calculées à l'aide de moyennes non ajustées (Borenstein & Hedges, 2019)² :

[E.6]

$$SE[g] = \omega \sqrt{\frac{n_i + n_c}{n_i n_c} + \frac{g^2}{2(n_i + n_c)}}$$

L'erreur standard exprime l'incertitude dans l'estimation de la taille de l'effet. Des erreurs standards plus importantes indiquent une plus grande incertitude (et généralement des tailles d'échantillon plus petites). Les chapitres suivants décrivent comment le WWC utilise l'erreur standard pour évaluer la significativité statistique des résultats d'une étude et le poids qu'on leur associe dans les méta-analyses conduites pour synthétiser plusieurs études.

Considérons un ECR avec un faible risque de biais dû à un changement de composition qui donne les statistiques de réussite en lecture suivantes dans le tableau E.1 :

Tableau E.1. Statistiques descriptives pour un essai contrôlé randomisé avec faible attrition

Groupes	<i>M</i>	<i>SD</i>	<i>n</i>
Intervention	59,12	12,32	39
Comparaison	54,34	11,57	41

M est la moyenne, *SD* est l'écart-type et *n* est la taille d'échantillon.

$$\omega = 1 - \frac{3}{4(39 + 41 - 2) - 1} = 0,99$$

$$g = \frac{0,99(59,12 - 54,34)}{\sqrt{\frac{(39 - 1)12,32^2 + (41 - 1)11,57^2}{39 + 41 - 2}}} = \frac{0,99 \times 4,78}{11,94} = 0,40$$

$$SE[g] = 0,99 \sqrt{\frac{39 + 41}{39 \times 41} + \frac{0,40^2}{2(39 + 41)}} = 0,22$$

² Le WWC utilise la taille d'effet corrigée g pour calculer le terme de droite g^2 dans la formule de l'erreur standard. Cette approche contraste légèrement avec l'approche de Borenstein et Hedges (2019), qui utilise la taille de l'effet non corrigée (le d de Cohen) pour calculer ce terme de droite. Les deux approches sont des approximations car la formule de la variance de la population utilise la taille d'effet de la population (Hedges & Olkin, 1985), qui n'est pas connue en pratique. Ces distinctions ont tendance à ne pas avoir beaucoup d'importance car le terme de gauche contribue généralement beaucoup plus à la variance que le terme de droite (et ω est généralement proche de 1).

Tests statistiques pour les comparaisons de moyennes non ajustées

Pour les ECR avec faible attrition, lorsque les moyennes ou les écarts-types ne sont pas publiés, le WWC peut calculer le g de Hedges en utilisant les résultats du test t ou de l'analyse de la variance (ANOVA) F . Les formules suivantes s'appliquent uniquement aux comparaisons entre les moyennes non ajustées des groupes (des formules différentes s'appliquent aux statistiques de test qui tiennent compte de certaines covariables).

La formule suivante permet de calculer la taille d'effet en utilisant le résultat d'un test t :

[E.7]

$$g = \omega t \sqrt{\frac{n_i + n_c}{n_i n_c}}$$

La formule suivante permet de calculer la taille d'effet en utilisant le résultat du test ANOVA F :

[E.8]

$$g = \omega \sqrt{\frac{F(n_i + n_c)}{n_i n_c}}$$

où le signe est déterminé par le signe de la différence principale. La formule d'erreur standard appropriée pour ces deux formules reste l'équation E.6, qui suppose une taille d'effet basée sur des comparaisons non ajustées.

L'étude donnée en exemple dans le tableau E.1 aurait pu ne publier que les statistiques de test $t = 1,79$ ou $F = 3,20$; la taille de l'effet serait toujours $g = 0,40$ avec ces statistiques alternatives.

Enfin, une valeur- p bilatérale exacte peut être convertie en une statistique t puis utilisée dans l'équation E.7. Par exemple, une valeur- p de 0,077 pour un échantillon total de 80 élèves (voir tableau E.1) correspond à $t = 1,79$, en utilisant 78 comme degrés de liberté (voir équation E.4). Les valeurs- p unilatérales peuvent être converties en valeurs- p bilatérales en les multipliant par deux. Le WWC, cependant, n'appliquera pas cette conversion pour des valeurs- p inexactes telles que « $p < 0,05$ », car elles peuvent correspondre à une plage de tailles d'effet.

Différences des moyennes ajustées aux covariables

Les formules précédentes basées sur des comparaisons de moyennes non ajustées ne peuvent pas être utilisées pour les études qui nécessitent un ajustement tenant compte de la différence entre les niveaux initiaux, comme les études au design quasi-expérimentales (QED) et les ECR avec forte attrition. Les tailles d'effet de ces études doivent plutôt utiliser des statistiques ajustées aux covariables qui contrôlent les différences entre les niveaux initiaux. Le WWC préfère également les tailles d'effet ajustées aux covariables pour les ECR présentant un faible risque de biais en raison d'un changement de composition afin d'augmenter la précision de l'estimation de la taille d'effet.

L'utilisation de moyennes ajustées aux covariables pour calculer les tailles d'effet repose sur la même formule générale du g de Hedges présentée dans l'équation E.1. La différence des moyennes non standardisées b sera cette fois la différence des moyennes ajustées aux covariables. Par exemple, l'exemple de l'étude du tableau E.1 aurait pu utiliser une régression linéaire multiple pour contrôler le

niveau de lecture initial, et publier une estimation de l'impact non standardisée (un coefficient de régression) de 3,59 pour la différence des moyennes ajustées en fonction des covariables. Le g de Hedges serait calculé comme suit :

$$\frac{0,99 \times 3,59}{11,94} = 0,30$$

La seule différence avec le calcul antérieur basé sur des moyennes non ajustées est que la différence des moyennes b est maintenant le coefficient de régression par opposition à la différence des moyennes non ajustées. Notez que la correction pour petit échantillon ω et l'écart-type groupé intra-groupe restent les mêmes (et que l'écart-type groupé est calculé en utilisant des statistiques non ajustées).

Alternativement, l'étude aurait pu mener une analyse de covariance (ANCOVA) et rapporter des moyennes ajustées aux covariables de 58,53 et 54,94 pour les groupes intervention et comparaison, respectivement. La différence des moyennes ajustées aux covariables resterait égale à 3,59 dans ce cas et le g de Hedges ajusté aux covariables resterait également inchangé.

Si les auteurs de l'étude ont rapporté l'erreur standard $SE[b]$ pour la différence des moyennes non standardisée b , alors l'erreur standard pour la taille d'effet standardisée peut être calculée comme suit :

[E.9]

$$SE[g] = \omega \sqrt{\left(\frac{SE[b]}{SD_p}\right)^2 + \frac{g^2}{2(n_i + n_c)}}$$

Le terme de gauche dans cette formule normalise l'erreur standard non standardisée en la divisant par l'écart-type regroupé. Le terme de droite est le même que dans l'équation E.6 et reflète l'incertitude introduite par l'utilisation de l'écart-type groupé (une statistique d'échantillon) dans le calcul de la taille d'effet. Considérons que l'exemple de l'étude du tableau E.1 ait rapporté une erreur standard du coefficient de régression égale à 2,07. L'erreur standard de taille d'effet serait alors :

$$SE[g] = 0,99 \sqrt{\left(\frac{2,07}{11,94}\right)^2 + \frac{0,40^2}{2(39 + 41)}} = 0,17$$

Notez que cette valeur de 0,17 est inférieure à l'erreur standard de 0,22 calculée précédemment pour la taille de l'effet calculée à partir des moyennes non ajustées. Cette différence reflète le fait que l'ajustement aux covariables réduit généralement l'incertitude dans l'estimation de l'effet de l'intervention, réduisant l'erreur standard et reflétant une augmentation de la précision.

Si c 'est la statistique t calculée pour un coefficient de régression non standardisé qui est publiée, l'erreur standard pour les différences moyennes non standardisées peut être calculée en utilisant $SE[b] = b/t$. Les études qui mettent en œuvre une ANCOVA peuvent également publier l'erreur standard de la moyenne du groupe intervention $SE[y_i]$ et l'erreur standard de la moyenne du groupe comparaison $SE[y_c]$. Dans ce cas, l'erreur standard de la différence des moyennes peut être calculée comme suit et utilisée dans l'équation E.9 :

[E.10]

$$SE[b] = \sqrt{SE[y_i]^2 + SE[y_c]^2}$$

Une étude peut ne pas publier suffisamment d'informations pour déterminer l'erreur standard de la différence des moyennes ajustées aux covariables, rendant l'équation E.9 inutilisable. Cependant, si les auteurs de l'étude ont rapporté le coefficient de corrélation du modèle de régression linéaire R^2 , l'erreur standard de la taille d'effet ajustée aux covariables peut alors être calculée comme suit :

[E.11]

$$SE[g] = \omega \sqrt{\frac{n_i + n_c}{n_i n_c} (1 - R^2) + \frac{g^2}{2(n_i + n_c)}}$$

Le WWC préfère les valeurs R^2 ajustées au nombre de prédicteurs saisis et qui corrigent ainsi l'excès de confiance dans les prédictions du modèle, aux valeurs de R^2 non ajustées. Le WWC traitera les valeurs R^2 ajustées négatives comme égales à 0 %. Si la valeur ajustée n'est pas publiée, le WWC utilisera des valeurs R^2 non ajustées. Si les auteurs de l'étude de l'exemple avaient indiqué que la valeur de R^2 ajustée était de 40 %, l'erreur standard de la taille d'effet serait alors égale à 0,17 en utilisant l'équation E.11.

Si ni l'erreur standard non standardisée ni la valeur R^2 ne sont disponibles, alors le WWC adoptera une approche prudente (conservative) pour calculer l'erreur standard de la taille d'effet et attribuera une valeur égale à 0 % pour R^2 dans l'équation E.11. Cette approche prudente surestimera l'ampleur de l'erreur standard mais protège contre l'erreur de type I. Le tableau E.2 résume la hiérarchisation des différentes formules en fonction des statistiques publiées par les auteurs de l'étude.

Tableau E.2. Utilisation des formules d'erreur standard ajustées aux covariables

$SE[b]$	R^2	Equation appropriée
Publiée(a)	Publiée	Utiliser l'équation [E.9]
Publiée (a)	Non publiée	Utiliser l'équation [E.9]
Non publiée	Publiée	Utiliser l'équation [E.11]
Non publiée	Non publiée	Utiliser l'équation [E.11] avec $R^2 = 0\%$

a. Inclus les cas où $SE[b]$ peut être calculée en utilisant [E.10].

Tests statistiques pour les comparaisons de moyennes ajustées aux covariables

Une étude pourrait effectuer une analyse ajustée aux covariables, telle qu'une régression linéaire multiple, mais ne pas rapporter suffisamment d'informations pour permettre le calcul d'une taille d'effet à l'aide des formules précédentes. Par exemple, l'article de l'étude de l'exemple pourrait ne pas avoir publié les écarts-types non ajustés. Dans de tels cas, le WWC pourrait calculer la taille de l'effet (le g de Hedges) à l'aide d'un test t ajusté à la covariable ou d'une statistique F comme suit :

[E.12]

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

[E.13]

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

L'utilisation de ces formules n'est possible que si les auteurs de l'étude ont publié le coefficient de corrélation R^2 , car les statistiques de test ajustées aux covariables utilisent la variance intra-groupe ajustée aux covariables, et non la variance intra-groupe non ajustée. L'intégration du terme $1 - R^2$ dans l'équation permet de calculer une taille d'effet comparable à d'autres tailles d'effet calculées à partir d'écart-types non ajustés. Comme indiqué précédemment, le WWC permettra également de convertir des valeurs- p bilatérales exactes en statistiques t pour appliquer ces formules, mais le WWC n'appliquera pas les conversions pour des valeurs- p inexactes telles que $p < 0,05$.

Autres méthodes de calcul de l'écart-type groupé

Les chapitres précédents présentent des méthodes de calculs (basées sur des statistiques de test) de la taille d'effet, c'est-à-dire du g de Hedges, quand l'écart-type groupé ne peut pas être calculé directement à l'aide de l'équation E.2. Ce chapitre développe davantage cette idée pour décrire d'autres manières de calculer l'écart-type groupé SD_p . Ces procédures permettent de calculer des écart-types des résultats collectés sans ajustement aux mesures initiales.

Écart-type total. Une étude peut rapporter l'écart-type total mais pas l'écart-type pour chacun des groupes séparément. L'écart-type total peut être converti en écart-type groupé comme suit :

[E.14]

$$SD_p = \sqrt{\frac{N-1}{N-2}SD^2 - \frac{n_i n_c}{N(N-2)}b^2}$$

où $N = n_i + n_c$ est la taille totale de l'échantillon de l'étude. Cette égalité est vraie si b est la différence des moyennes non ajustées (aux covariables) et presque équivalente si b est la différence des moyennes ajustées aux covariables. Le WWC autorise l'utilisation de cette formule dans les deux cas.

Si l'étude de l'exemple (tableau E.1) avait publié un écart-type total tel que $SD = 12,11$ mais n'avait pas publié les écart-types de chacun des groupes, alors l'écart-type groupé serait calculé comme suit (avec $b = 4,78$ la différence moyenne non ajustée) :

$$SD_p = \sqrt{\frac{80-1}{80-2} \times 12,11^2 - \frac{39 \times 41}{80(80-2)} \times 4,78^2} = 11,94$$

Cet écart-type groupé intra-groupe égal à 11,94 est identique à la valeur calculée à l'aide de l'équation E.2 avec les écart-types de chacun des groupes.

Scores-z standardisés. Une étude peut normaliser l'échelle des réponses en soustrayant la moyenne générale et en divisant par l'écart-type-total (non ajusté) pour générer ce que l'on appelle souvent des scores z . Le WWC remplacera alors SD par 1 dans l'équation E.14 pour traiter ce cas.

Coefficients de régression standardisés. Une étude peut rapporter des coefficients de régression standardisés (parfois appelés pondérations bêta). Ces coefficients standardisés diffèrent des coefficients non standardisés basés sur les scores z . En effet, considérons un résultat avec un écart-type total SD_y et un prédicteur avec un écart-type SD_x , alors le coefficient standardisé β est calculé de la façon suivante :

[E.15]

$$\beta = b \frac{SD_x}{SD_y}$$

Cette formule montre que les coefficients standardisés et les scores-z impliquent tous deux une division par l'écart-type du résultat total, mais que les coefficients standardisés impliquent également une multiplication par l'écart-type du prédicteur. Une formule de taille d'effet pour le g de Hedges qui corrige les deux problèmes est donnée par :

[E.16]

$$g = \frac{\omega\beta}{\sqrt{1-\beta^2}} \sqrt{\frac{N(N-2)}{n_i n_c}}$$

Erreurs standards des moyennes de groupe. Une étude peut rapporter les erreurs standards des moyennes de groupe, mais pas les écarts-types de chacun des groupes. Pour les analyses non ajustées, les erreurs standards de chaque groupe peuvent être converties en écarts-types comme suit :

[E.17]

$$SD_i = SE[y_i]\sqrt{n_i} \text{ et } SD_c = SE[y_c]\sqrt{n_c}$$

Pour les analyses ajustées aux covariables, un facteur $1 - R^2$ supplémentaire est requis :

[E.18]

$$SD_i = SE[y_i]\sqrt{n_i(1-R^2)} \text{ et } SD_c = SE[y_c]\sqrt{n_c(1-R^2)}$$

L'étude doit publier la valeur de R^2 pour effectuer cette conversion des écarts-types dans les analyses ajustées aux covariables. Les écarts-types convertis pour chaque groupe peuvent ensuite être utilisés pour calculer l'écart-type groupé.

Écarts-types de gain de score. Une étude peut ne publier que les moyennes et les écarts-types des gains de score (différence entre le score post-test et le score pré-test). Les écarts-types des gains de scores sont généralement inférieurs aux écarts-types post-test non ajustés, ce qui surestime la taille de l'effet. Si les auteurs de l'étude ont rapporté ρ_{cor} , le coefficient de corrélation des variables scores prétests et scores posttests, le WWC corrigera ce problème comme suit :

[E.19]

$$SD_p = \frac{SD_{p.gain}}{\sqrt{2(1-\rho_{cor})}}$$

où $SD_{p.gain}$ est l'écart-type groupé pour le gain de score. Cette conversion se rapproche de l'écart-type groupé des scores posttests. L'erreur standard de la taille d'effet pour les analyses basées sur les gains de scores est détaillée dans un chapitre ultérieur sur la différence dans les différences.

Le WWC n'appliquera pas cette conversion si l'étude n'a pas rapporté le coefficient de corrélation ρ_{cor} . Le WWC ne rapportera pas non plus les tailles d'effet basées sur les écarts-types des scores de gain non convertis, étant donné leur manque de comparabilité avec d'autres tailles d'effet calculées par le WWC. Par conséquent, les études basées sur les scores de gain doivent publier soit les écarts-types des scores posttests non ajustés (de préférence), soit les écarts-types des gains de scores et le coefficient de corrélation scores prétests / scores posttests (informations requises pour utiliser l'équation E.19).

Écart-types du niveau initial. Le WWC permettra à la direction de l'équipe conduisant l'examen de l'étude d'utiliser des écart-types du niveau initial à la place des écart-types de résultats si (a) les écart-types de résultats sont manquants et (b) le niveau initial et le résultat ont utilisé la même mesure (et utilisent les mêmes procédures de notation). Dans certains cas, la direction de l'équipe peut être particulièrement préoccupée par le fait que les écart-types du niveau initial peuvent différer considérablement des écart-types des résultats. L'une des préoccupations concerne les effets dits « planchers » (c'est-à-dire que les performances au prétest sont proches du minimum, ce qui peut être le cas quand de nouveaux contenus sont introduits). Une autre préoccupation concerne les effets dits « plafonds » (c'est-à-dire que les performances au post-test sont proches du maximum, ce qui peut être le cas quand l'intervention concerne la maîtrise). Bien que ces situations soient rares, la direction de l'équipe a le pouvoir discrétionnaire de ne pas utiliser les écart-types du niveau initial, y compris quand aucun autre écart-type n'est disponible (ce qui rend de ce fait le calcul de la taille de l'effet impossible). Si l'équipe n'est pas dans une telle situation, le WWC autorisera l'utilisation des écart-types du niveau initial pour calculer les tailles d'effet.

Écart-types d'un seul groupe. Dans certains cas rares, une étude peut rapporter l'écart-type d'un seul groupe, par exemple l'écart-type du groupe comparaison mais pas celui du groupe intervention. Bien que le WWC préfère calculer les tailles d'effet à l'aide de l'écart-type groupé, il est important de noter que les approches statistiques classiques invoquent systématiquement l'hypothèse d'homogénéité de la variance, c'est-à-dire supposent que les écart-types au niveau de la population sont les mêmes dans les groupes intervention et comparaison. On peut donc considérer que l'écart-type du groupe comparaison estime le même écart-type au niveau de la population que l'écart-type groupé ; la seule différence étant que les estimations de la taille d'effet calculées sont alors légèrement moins performantes. Le WWC autorisera donc le calcul de la taille d'effet à partir de l'écart-type d'un seul groupe, bien que l'utilisation de l'écart-type groupé soit généralement préférable.

2. Études avec affectation au niveau des grappes

Toutes les formules statistiques précédentes concernent des études avec affectation au niveau individuel. Les considérations statistiques sont plus complexes pour les études avec affectation au niveau des grappes, par exemple quand ce sont des écoles qui sont affectées au groupe intervention ou au groupe comparaison. Par rapport à l'affectation au niveau individuel, l'affectation au niveau des grappes produit généralement plus d'incertitude (erreurs standards plus importantes) pour les estimations de l'effet de l'intervention.

Le chapitre suivant décrit les procédures utilisées par le WWC pour traiter ces questions, et prend également en compte les différentes façons dont les auteurs des études peuvent mener et publier leurs analyses. Par exemple, les auteurs d'une étude peuvent avoir analysé les données au niveau individuel et tenir compte des grappes, analysé les données au niveau individuel et ne pas tenir compte des grappes, ou calculé les moyennes au niveau des grappes.

Définition de la taille d'effet pour les études avec affectation au niveau des grappes

Le WWC définit la taille d'effet pour les études avec affectation au niveau des grappes en prenant en compte la variabilité totale, c'est-à-dire la variabilité au sein des grappes et la variabilité entre les grappes (voir l'équation 3 dans Hedges, 2007). La prise en compte de la variabilité totale donne des tailles d'effet comparables à celles calculées dans les études avec affectation au niveau individuel menées sur plusieurs sites. Ce point est important parce que la plupart des études avec affectation au niveau individuel que le WWC examine sont menées dans plus d'une école. L'utilisation de la seule

variabilité inter-grappe donnerait des tailles d'effet plus importantes que celles généralement calculées dans les études avec affectation au niveau individuel.

Une manière courante de calculer les tailles d'effet à partir d'études avec affectation au niveau des grappes consiste à utiliser une estimation de la différence des moyennes non standardisée b et de l'écart-type groupé au niveau individuel SD_p :

[E.20]

$$g = \frac{\omega b}{SD_p} \sqrt{\gamma} = \frac{\omega b}{SD_p} \sqrt{1 - \frac{2 \left(\frac{N}{M} - 1\right) \rho_{ICC}}{N - 2}}$$

où N est le nombre total d'individus, M est le nombre total de grappes et ρ_{ICC} est le coefficient de corrélation intraclasse qui représente le degré de regroupement. Cette formule peut être utilisée avec une différence des moyennes b non ajustée comme ajustée (aux covariables). La formule est similaire à la formule générale du g de Hedges avec affectation au niveau individuel (voir l'équation E.1) mais ajoute un autre facteur de correction de biais $\sqrt{\gamma}$ (voir l'équation 15 dans Hedges, 2007). Ce facteur peut être considéré comme un ajustement pour les petits nombres de grappes. Bien que généralement faible, cet ajustement est nécessaire pour produire des estimations approximativement non biaisées de la taille d'effet pour la population, en particulier pour les études comportant peu de grappes.

La correction pour petit nombre de grappes $\sqrt{\gamma}$ s'ajoute à l'ajustement pour petite taille d'échantillon ω , donnée précédemment par l'équation E.3. Le calcul du degré de liberté df est cependant différent pour les tailles d'effet basées sur les écarts-types au niveau individuel dans les études avec affectation au niveau des grappes :

[E.21]

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

Cette formule permet le calcul de df qui peut ensuite être utilisé dans l'équation E.3 pour calculer ω .

La corrélation intraclasse : application à un exemple

Le chapitre précédent a introduit un paramètre clé : le coefficient de corrélation intraclasse ρ_{ICC} . Ce paramètre peut varier de 0 à 1 et représente le degré de regroupement statistique des individus au sein des grappes. Plus techniquement, c'est la proportion de la variance totale attribuable à la variance inter-groupes. Le WWC utilisera le coefficient de corrélation intraclasse publié par l'auteur de l'étude lorsqu'il est disponible, mais cette valeur n'est souvent pas publiée. Sur la base de la littérature empirique dans le domaine de l'éducation, le WWC a adopté 0,20 comme valeur par défaut du coefficient de corrélation intraclasse pour les résultats mesurant la réussite et 0,10 pour tous les autres types de résultats (Hedges & Hedberg, 2007 ; Schochet, 2008). Dans certains protocoles d'examen, des valeurs par défaut différentes peuvent être définies et utilisées, avec justification.

Admettons que l'étude donné en exemple (tableau E.1) et portant sur la réussite en lecture était une étude avec affectation au niveau des classes avec $M = 8$ classes affectées au hasard (au groupe intervention ou au groupe contrôle). Supposons enfin que l'étude n'ait pas publié le coefficient de corrélation intraclasse, ce qui signifie que le WWC utiliserait le coefficient de corrélation intraclasse

par défaut de 0,20 pour les résultats mesurant la réussite. Les calculs préliminaires permettant de calculer la taille de l'effet peuvent être écrits comme suit :

$$\sqrt{\gamma} = \sqrt{1 - \frac{2 \left(\frac{80}{8} - 1 \right) 0,20}{80 - 2}} = 0,98$$

$$df = \frac{\left[(80 - 2) - 2 \left(\frac{80}{8} - 1 \right) 0,20 \right]^2}{(80 - 2)(1 - 0,20)^2 + \frac{80}{8} \left(80 - 2 \frac{80}{8} \right) 0,20^2 + 2 \left(80 - 2 \frac{80}{8} \right) 0,20(1 - 0,20)} = 59,44$$

$$\omega = 1 - \frac{3}{4 \times 59,44 - 1} 0,99$$

L'écart-type groupé au niveau individuel reste le même que celui calculé précédemment ($SD_p = 11,94$), et la différence des moyennes non ajustées basée sur les statistiques du tableau E.1 est $b = 4,78$. La taille de l'effet g de Hedges non ajusté pour cet exemple d'étude d'affectation au niveau des grappes serait donc :

$$g = \frac{0,99 \times 4,78}{11,94} 0,98 = 0,39$$

Notez que cette taille d'effet de $g = 0,39$ est presque la même que la valeur $g = 0,40$ calculée précédemment en supposant une affectation au niveau individuel. De même, si la différence des moyennes ajustée aux covariables était utilisé à la place (soit $b = 3,59$) alors la taille de l'effet calculée à partir de la formule E.20 serait $g = 0,29$, ce qui est très proche de la valeur analogue de $g = 0,30$ calculée précédemment pour une affectation au niveau individuel. Cependant, comme le montre le chapitre suivant, les erreurs standards des tailles d'effet sont beaucoup plus importantes avec affectation au niveau des grappes qu'avec affectation au niveau individuel.

Erreurs standards pour les analyses non ajustées

L'affectation de grappes d'individus à des groupes intervention et comparaison réduit la taille effective de l'échantillon d'une étude. Intuitivement, le nombre d'unités randomisées est plus petit pour une affectation au niveau des grappes que pour une affectation au niveau individuel. Les erreurs standards sont plus grandes et les taux de faux positifs pourraient augmenter si les auteurs de l'étude ne tiennent pas compte dans leurs analyses du regroupement des individus en grappes.

L'augmentation de la variance de la taille de l'effet dépend du coefficient de corrélation intraclasse. Si le coefficient de corrélation intraclasse est de 0, alors la variance reste inchangée, ce qui revient à dire que le regroupement des élèves en grappes n'a pas d'influence au niveau statistique. Plus la valeur du coefficient de corrélation intraclasse augmente, et plus la variance de la taille d'effet augmente. L'effet dû au design que l'on note η est un facteur multiplicatif qui représente l'augmentation de la variance que l'on peut attribuer au fait d'avoir regroupé les élèves en grappes :

[E.22]

$$\eta = 1 + \left(\frac{N}{M} - 1 \right) \rho_{ICC}$$

Pour l'étude de l'exemple du chapitre précédent, ce paramètre est donné par :

$$\eta = 1 + \left(\frac{80}{8} - 1\right) \times 0,20 = 2,80$$

Cette valeur signifie que, pour un même nombre d'élèves, la variance de la taille de l'effet sera environ 2,8 fois plus grande pour une étude avec affectation au niveau des grappes que pour une étude avec affectation au niveau individuel. L'erreur standard sera d'environ $\sqrt{2,80} = 1,67$ fois plus grande car l'erreur standard est égale à la racine carrée de la variance.

Pour les analyses non ajustées, l'erreur standard pour la taille de l'effet donnée dans l'équation E.20 est donnée par :

[E.23]

$$SE[g] = \omega \sqrt{\left(\frac{n_i + n_c}{n_i n_c}\right) \eta + \frac{g^2}{2df}}$$

Le degré de liberté df pour cette formule est donné par l'équation E.21. Cette formule est appropriée, que l'analyse des auteurs de l'étude ait ou non tenu compte du regroupement. En d'autres termes, l'erreur standard de la taille d'effet calculée tiendra compte du regroupement, même si l'analyse des auteurs de l'étude ne l'a pas fait. Cette formule s'applique aux analyses sans ajustement pour les niveaux initiaux. Cette formule suppose également que le nombre d'individus est le même pour chaque groupe, mais elle sera toujours approximativement correcte pour les études avec des tailles de groupe inégales³ (Hedges, 2007).

Pour l'exemple de l'étude précédent, l'erreur standard pour une analyse non ajustée serait :

$$SE[g] = 0,99 \sqrt{\frac{39 + 41}{39 \times 41} \times 2,8 + \frac{0,39^2}{2 \times 59,44}} = 0,37$$

Cette erreur standard de 0,37 est beaucoup plus grande que la valeur de 0,22 calculée précédemment avec l'affectation au niveau individuel, reflétant une incertitude accrue quant à l'estimation de l'effet de l'intervention.

Erreurs standards pour les analyses ajustées aux covariables

Le WWC calculera les erreurs standards ajustées aux covariables pour les études d'affectation au niveau des grappes, en étendant les procédures présentées précédemment à partir des erreurs standards des coefficients de régression (équation E.9) et des valeurs R^2 (équation E.11). Une difficulté ici est de savoir si le modèle utilisé par les auteurs de l'étude a correctement pris en compte le regroupement, ce qui peut modifier considérablement le calcul des erreurs standards. Les formules suivantes rendent compte de ces différents choix analytiques.

Les auteurs des études peuvent tenir compte du regroupement de plusieurs façons acceptables. Par exemple, ils peuvent inclure des effets aléatoires au niveau de l'unité d'affectation dans un modèle

³ Souvent, ces études ne rapportent pas suffisamment d'informations pour appliquer des formules plus compliquées qui traitent explicitement des tailles de grappes différentes (Hedges, 2007). Par conséquent, le WWC appliquera l'équation E.23 et d'autres formules de ce chapitre aux études avec des tailles de grappes égales et des tailles de grappes différentes.

linéaire hiérarchique, qui sont parfois aussi appelés modèles à effets mixtes ou modèles multiniveaux (Raudenbush & Bryk, 2002). Les auteurs de l'étude peuvent également utiliser un modèle de régression des moindres carrés ordinaire à un niveau, puis ajuster les erreurs standards au niveau de l'unité d'affectation à l'aide d'erreurs standards robustes aux grappes (McNeish et al., 2017). L'une ou l'autre approche peut produire des erreurs standard des coefficients de régression qui tiennent compte du regroupement.

Si les auteurs de l'étude ont publié une erreur standard tenant compte du regroupement $SE_{CC}[b]$ pour la différence des moyennes non standardisée, le WWC calculera l'erreur standard de taille d'effet comme suit :

[E.24]

$$SE[g] = \omega \sqrt{\left(\frac{SE_{CC}[b]}{SD_p}\right)^2 \gamma + \frac{g^2}{2df}}$$

où γ est le facteur correctif permettant d'ajuster les résultats pour les petits nombres de grappes introduit précédemment dans l'équation E.20.

Les auteurs des études ne tiennent souvent pas compte du regroupement dans les études avec affectation au niveau des grappes, même s'ils le devraient. Par exemple, une étude peut mener un modèle de régression ordinaire des moindres carrés sans ajuster les erreurs standards pour le regroupement dans une étude avec affectation au niveau des classes. Les erreurs standards de ces modèles ne peuvent pas être utilisées dans l'équation E.24 car l'erreur standard serait trop faible. Cependant, le WWC peut calculer l'erreur standard de la taille d'effet comme suit si les auteurs de l'étude ont publié une erreur standard non corrigée $SE_{UC}[b]$ pour la différence des moyennes non standardisée :

[E.25]

$$SE[g] = \omega \sqrt{\left(\frac{SE_{UC}[b]}{SD_p}\right)^2 \eta + \frac{g^2}{2df}}$$

où η est l'effet dû au design calculé par l'équation E.22. Cet effet prend en compte le fait que l'erreur standard publiée par l'auteur est trop faible, et il est souvent beaucoup plus grand que 1.

Une étude peut également publier le coefficient de corrélation multiple R^2 à partir d'un modèle à un seul niveau (comme une ANCOVA ou une régression des moindres carrés ordinaire) et ne pas publier l'erreur standard du coefficient de régression. Dans ce cas, le WWC peut calculer l'erreur standard ajustée à la covariable comme suit :

[E.26]

$$SE[g] = \omega \sqrt{\frac{n_i + n_c}{n_i n_c} (1 - R^2) \eta + \frac{g^2}{2df}}$$

Cette formule d'erreur standard est appropriée, que le modèle des auteurs de l'étude tienne ou non compte du regroupement (car cette formule ne repose pas sur l'erreur standard déclarée par l'auteur). Suivant des directives similaires en cela à l'affectation au niveau individuel, le WWC remplacera R^2 par 0 % dans cette formule si à la fois le coefficient de régression et R^2 ne sont pas publiés, ce qui rend la formule identique à l'erreur standard pour les analyses non ajustées (équation E.23). Le WWC suivra

également la même hiérarchisation relative aux formules d'erreur standard indiquées dans le tableau E.2 (en favorisant l'utilisation de l'erreur standard du coefficient de régression par rapport à l'utilisation de R^2).

Considérons que dans l'étude donnée en exemple présentée précédemment un coefficient de régression $b = 3,59$ ait été publié comme différence des moyennes non standardisées ajustée aux covariables ainsi qu'une erreur standard du coefficient tenant compte du regroupement $SE_{CC}[b] = 3,54$. La taille de l'effet ajustée aux covariables serait $g = 0,29$ selon l'équation E.20. L'erreur standard de la taille d'effet ajustée aux covariables serait calculée comme suit :

$$SE[g] = 0,99 \sqrt{\left(\frac{3,54}{11,94}\right)^2 \left(1 - \frac{2\left(\frac{80}{8} - 1\right) \times 0,20}{80 - 2}\right) + \frac{0,29^2}{2 \times 59,44}} = 0,29$$

Cette erreur standard de la taille d'effet ajustée aux covariables égale à 0,29 est inférieure à l'erreur standard non ajustée de 0,37 calculée dans le cas d'une affectation au niveau des grappes sans ajustements aux covariables, ce qui indique une précision accrue. Cependant, elle est toujours plus grande que l'erreur standard avec affectation au niveau individuel non ajusté comme ajusté aux covariables (0,22 et 0,17, respectivement), ce qui reflète le rôle important de l'affectation au niveau des grappes dans la détermination de la précision de la taille de l'effet.

Alternativement, l'étude aurait pu publier une erreur standard non corrigée $SE_{UC}[b] = 2,07$, ce qui aurait donné la même erreur standard de taille d'effet de 0,29 en utilisant l'équation E.25 (et $\eta = 2,80$ comme calculé précédemment). Publier $R^2 = 40\%$ et utiliser l'équation E.26 aurait donné la même erreur standard de la taille d'effet.

Analyses des données au niveau des grappes

Les chapitres précédents supposaient que les auteurs de l'étude avaient analysé les données au niveau individuel ou bien avaient eu, d'une façon ou d'une autre, accès aux écarts-types au niveau individuel. Cependant, ces données peuvent ne pas être publiées. De nombreuses analyses de données dans le domaine de l'éducation utilisent des moyennes accessibles au public et récoltées au niveau des établissements et peuvent ne pas avoir accès à des données au niveau individuel. Cette limitation signifie que les auteurs de l'étude n'ont accès qu'aux écarts-types au niveau des grappes, qui sont généralement beaucoup plus petits que les écarts-types au niveau individuel.

Les écarts-types au niveau des grappes et au niveau individuel ne sont donc pas interchangeables. Les tailles d'effet calculées avec les écarts-types au niveau des grappes seront beaucoup plus importantes que celles calculées avec les écarts-types au niveau individuel. Le WWC corrigera ce problème en utilisant la corrélation intraclasse pour ajuster les tailles d'effet au niveau des grappes et produire un équivalent de la taille d'effet au niveau individuel (voir les équations 25 et 26 dans Hedges, 2007). Cet équivalent du g de Hedges calculé au niveau individuel peut être calculé comme suit :

[E.27]

$$g = \frac{\omega b}{SD_B} \sqrt{\frac{M\eta}{N}}$$

où SD_B est l'écart-type groupé au niveau des grappes (basé sur l'application de l'équation E.2 aux écarts-types au niveau des grappes et au nombre de grappes dans chaque groupe). Le degré de liberté df pour calculer ω pour cette formule est :

[E.28]

$$df = M - 2$$

Cette formule contraste avec la formule donnée par l'équation E.21. La différence provient du fait que l'écart-type groupé calculé au niveau des grappes est distribué avec un degré de liberté inférieur à celui utilisé quand on calcule l'écart-type groupé au niveau individuel.

Pour les analyses non ajustées, l'erreur standard de la taille d'effet au niveau individuel est donnée par :

[E.29]

$$SE[g] = \omega \sqrt{\left(\frac{n_i + n_c}{n_i n_c}\right) \eta + \frac{g^2}{2df}}$$

Notez que cette formule d'erreur standard est presque identique à celle utilisées pour les tailles d'effet calculée avec les écarts-types au niveau individuel (équation E.23). La seule différence est le degré de liberté utilisé dans le terme de droite qui est ici égal à $M - 2$ et non le résultat donné par l'équation E.21. Cette différence a tendance à ne pas avoir beaucoup d'importance car le terme de droite contribue généralement peu à la variance globale de la taille de l'effet. En d'autres mots, les tailles d'effet initialement calculée au niveau des grappes et converties en équivalents au niveau individuel ont presque la même variabilité que les tailles d'effet calculées avec les écarts-types au niveau individuel.

Pour les analyses avec ajustement aux covariables, l'erreur standard de la taille de l'effet peut être calculée comme suit si l'erreur standard de la différence des moyennes non standardisées $SE[b]$ est publiée :

[E.30]

$$SE[g] = \omega \sqrt{\left(\frac{SE[b]}{SD_B}\right)^2 \frac{M\eta}{N} + \frac{g^2}{2df}}$$

Si l'étude publie plutôt le coefficient de corrélation multiple R^2 , alors l'erreur standard de la taille de l'effet peut être calculée à l'aide de l'équation E.26 (mais en en utilisant l'égalité $df = M - 2$).

La taille d'effet calculée aura une certaine sensibilité à la valeur du coefficient de corrélation intraclasse choisie, mais la détermination de la significativité statistique sera robuste à cette valeur. La taille de l'effet et l'erreur standard dépendent de la corrélation intraclasse par un facteur multiplicatif qui est exactement le même dans les deux cas. Cette dépendance s'annule exactement lors du calcul de la statistique de test t – qui est égal à la taille d'effet divisée par son erreur standard - utilisée pour déterminer la valeur- p calculée par le WWC. Par conséquent, la valeur- p calculée par le WWC pour les tailles d'effet basées sur les écarts-types au niveau des grappes ne dépend pas de la valeur du coefficient de corrélation intraclasse.

Cette conversion ouvre une voie aux études n'ayant pas accès aux données au niveau individuel et leur permet de contribuer aux synthèses du WWC. Sinon, ces études seraient exclues lors du calcul des tailles d'effet dans une méta-analyse. Une considération importante est que la taille d'effet calculée peut être systématiquement trop grande ou trop petite selon que le coefficient de corrélation intraclasse estimé ou imputé est supérieur ou inférieur au véritable coefficient de corrélation intraclasse de la population. Si la taille d'effet calculée est trop petite, la taille d'effet globale de la méta-analyse sera calculée de manière prudente⁴. Si la taille d'effet calculée est trop grande, le biais introduit dans l'estimation de la taille d'effet globale sera partiellement compensé par le fait que les erreurs standard sont également trop grandes, ce qui réduit le poids attribué à l'étude dans la méta-analyse (voir plus loin).

On va considérer que les auteurs de l'étude de l'exemple ont publié les statistiques suivantes pour la réussite en lecture (tableau E.3).

Tableau E.3. Statistiques descriptives pour une analyse au niveau du cluster

Groupes	Ecart-types au niveau des grappes	Nombres de grappes	Nombres d'individus
Intervention	3.42	12	121
Comparaison	3.89	10	99

La taille de l'effet pour une différence des moyennes égale à $b = 2,56$ serait calculée à l'aide de l'équation E.27 comme suit :

$$SD_B = \sqrt{\frac{(12 - 1)3,42^2 + (10 - 1)3,89^2}{12 + 10 - 2}} = 3,64$$

$$df = 12 + 10 - 2 = 20$$

$$\omega = 1 - \frac{3}{4 \times 20 - 1} = 0,96$$

$$\eta = 1 + \left(\frac{220}{22} - 1\right) \times 0,20 = 2,80$$

$$g = \frac{0,96 \times 2,56}{3,64} \sqrt{\frac{22 \times 2,80}{220}} = 0,36$$

Pour les analyses non ajustées, l'erreur standard de la taille de l'effet serait calculée à l'aide de l'équation E.29 :

$$SE[g] = 0,96 \sqrt{\left(\frac{121 + 99}{121 \times 99}\right) 2,80 + \frac{0,36^2}{2 \times 20}} = 0,22$$

Pour les analyses ajustées aux covariables, si on considère que les auteurs de l'étude ont publié une erreur standard non standardisée égale à 1,05, l'erreur standard de la taille d'effet serait calculée à l'aide de l'équation E.30 :

⁴ Note de la traduction : en limitant le risque de type I

$$SE[g] = 0,96 \sqrt{\left(\frac{1,05}{3,64}\right)^2 \frac{22 \times 2,80}{220} + \frac{0,36^2}{2 \times 20}} = 0,16$$

Autres formules pour des études avec affectation au niveau des grappes

D'autres formules permettent de calculer la taille de l'effet ou l'écart-type groupé dans les études avec affectation au niveau des grappes. Par exemple, les formules permettant l'utilisation de l'écart-type total (équation E.14) ou de l'écart-type du gain de score (équation E.19) s'appliquent de la même manière aux études avec affectation au niveau individuel ou au niveau des grappes.

Les formules basées sur les statistiques de test ou les erreurs standard, cependant, doivent être modifiées lorsqu'elles sont appliquées aux études avec affectation au niveau des grappes. Par exemple, considérons la procédure présentée précédemment pour calculer la taille d'effet à l'aide d'une statistique t pour les comparaisons de moyennes non ajustées (équation E.7). Lorsqu'elle est appliquée à des études avec affectation au niveau des grappes, des équations différentes seront utilisées selon que les auteurs de l'étude ont publié une statistique t corrigée et tenant compte du regroupement en grappe t_{CC} ou une statistique t_{UC} non corrigée ne tenant pas compte du regroupement en grappes. La taille de l'effet (le g de Hedges) peut être calculée comme suit dans ces deux scénarios différents :

[E.31]

$$g = \omega t_{CC} \sqrt{\eta \frac{n_i + n_c}{n_i n_c}}$$

[E.32]

$$g = \omega t_{UC} \sqrt{\gamma \frac{n_i + n_c}{n_i n_c}}$$

Des extensions similaires s'appliquent aux statistiques de test ajustées aux covariables (et les équations données ci-dessus devront alors comporter un facteur $\sqrt{(1 - R)^2}$) ou aux statistiques F (et dans ce cas on remplace t par \sqrt{F}). Les statistiques de test basées sur des données au niveau des grappes doivent être traitées comme des statistiques corrigées et tenant compte du regroupement.

La méthode de calcul des écarts-types utilisant les erreurs standards des moyennes de chacun des groupes telles qu'elles sont publiées par l'auteur (équations E.17 et E.18) dépend également de la prise en compte ou non du regroupement par les auteurs de l'étude dans leur analyse. Pour les erreurs standard des moyennes de groupe sans ajustement aux covariables, les écarts-types pour chacun des groupes peuvent être calculés comme suit :

[E.33]

$$SD_i = SE_{CC}[y_i] \sqrt{\frac{\gamma n_i}{\eta}} \text{ et } SD_c = SE_{CC}[y_c] \sqrt{\frac{\gamma n_c}{\eta}}$$

[E.34]

$$SD_i = SE_{UC}[y_i] \sqrt{n_i} \text{ et } SD_c = SE_{UC}[y_c] \sqrt{n_c}$$

Des extensions similaires à celles précédemment proposées s'appliquent aux équations utilisant les erreurs standards des moyennes de groupes ajustées aux covariables (comme ci-dessus mais en ajoutant un facteur $\sqrt{(1 - R)^2}$).

3. Calcul des valeurs-p et significativité statistique

Le WWC publie les résultats de l'étude en se concentrant sur les tailles d'effet moyennes calculées au niveau d'un domaine et qui résument l'ensemble des résultats pour chacun des domaines. Une étude pourrait avoir plusieurs résultats principaux dans un domaine donné, comme avoir deux mesures distinctes de la compréhension en lecture. La partie suivante explique comment le WWC calcule alors ces moyennes au niveau du domaine et détermine leur significativité statistique.

En outre, le WWC présentera la significativité statistique de chacun des résultats calculés pour une étude, aussi bien pour les résultats principaux que pour les résultats complémentaires. La significativité statistique de ces résultats est une information contextuelle supplémentaire, mais elle n'affecte pas les conclusions qualitatives du WWC. En effet, la caractérisation des résultats de l'étude par le WWC dépend de la significativité statistique au niveau du domaine, mais pas au niveau des résultats.

Le WWC accepte généralement les valeurs- p et la significativité statistique des résultats de l'étude publiées par l'auteur. Cependant, le WWC estimera la significativité statistique dans trois cas courants :

- L'étude ne publie pas d'estimation de la significativité statistique.
- Les calculs de l'étude présentent un défaut clairement identifiable, par exemple n'ajustent pas les données pour tenir compte des différences des niveaux initiaux dans les groupes alors que cela s'avère nécessaire.
- L'étude n'a pas pris en compte le regroupement dans une analyse au niveau individuel dans une étude avec affectation au niveau des grappes.

Par exemple, une étude pourrait publier des analyses non ajustées aux différences des niveaux initiaux alors que cet ajustement est nécessaire. Mais la significativité statistique publiée par l'auteur ne pourrait pas être acceptée par le WWC. Dans ces cas, le WWC détermine la significativité statistique de chacun des résultats de l'étude en calculant d'abord t , le rapport entre la taille de l'effet g et son erreur standard $SE[g]$:

[E.43]

$$t = \frac{g}{SE[g]}$$

Le degré de liberté permettant de déterminer la valeur- p pour cette statistique t est calculé en utilisant les formules présentées précédemment quand l'affectation se fait au niveau individuel (équation E.4) ou quand l'affectation se fait au niveau des grappes (équation E.21 ; cependant, notez que l'équation E.28 doit être utilisée quand les tailles d'effet sont calculées à partir des écarts-types au niveau des grappes).

À titre d'exemple, dans un chapitre précédent on a calculé une taille d'effet $g = 0,40$ et une erreur standard $SE[g] = 0,22$ pour une étude avec affectation au niveau individuel et un échantillon total de 80 élèves (correspondant à un degré de liberté égal à 78). La statistique t pour le calcul de la valeur- p calculée par WWC serait donc $t = 0,40/0,22 = 1,82$. La valeur- p pourrait être calculée dans Microsoft Excel en utilisant la fonction T.DIST.2T(t , df), ce qui donnerait ici T.DIST.2T(1.82,78) = 0,07.

La valeur- p pourrait également être calculée en utilisant le logiciel R en saisissant $2*pt(-abs(t),df)$, ce qui donnerait ici $2*pt(-abs(1.82),78) = 0,07$. Le WWC ne considérerait pas ce résultat comme statistiquement significatif car la valeur- p est supérieure au seuil conventionnel de 0,05.

Le WWC n'applique pas de corrections de la significativité statistique pour les comparaisons multiples au niveau des résultats. Le WWC aborde plutôt la question des comparaisons multiples en focalisant son attention sur la taille d'effet calculée au niveau du domaine considéré. Cette taille d'effet permet de synthétiser plusieurs résultats principaux dans un domaine donné, comme cela est détaillé dans la partie suivante.

4. Calcul de l'indice d'amélioration

Le WWC peut traduire les tailles d'effet en indices d'amélioration et faciliter l'évaluation par les lecteurs de l'importance pratique à accorder à l'ampleur de l'effet d'une intervention. L'indice d'amélioration est le changement attendu du rang centile d'un élève moyen du groupe comparaison qui serait ensuite soumis à l'intervention (ou également la différence des rangs centiles entre un élève moyen du groupe intervention et un élève moyen du groupe comparaison). Si on reprend l'exemple donné au début de ce chapitre avec une taille d'effet telle que $g = 0,40$, cela se traduirait par un indice d'amélioration de +16 centiles.

Le calcul de l'indice d'amélioration se fait en deux étapes :

1. Convertir la taille de l'effet en indice U3 de Cohen. L'indice U3 de Cohen est la fraction des élèves du groupe comparaison surpassés par l'élève moyen du groupe intervention. Une taille d'effet de 0,40 correspond à un indice U3 de 0,66, ce qui signifie que l'élève moyen du groupe intervention obtient un score supérieur à 66 % des élèves du groupe comparaison (en supposant des données distribuées normalement⁵). Ou encore, l'élève moyen du groupe intervention se situe au 66^{ème} centile de la distribution du groupe comparaison. L'indice U3 est la proportion de la zone délimitée par la courbe normale standard et par l'axe vertical correspondant à la valeur de la taille de l'effet. Par exemple, l'indice U3 pour une taille d'effet de 0,40 peut être calculé à l'aide de la commande `pnorm(0,40)` dans le logiciel R ou de la fonction `NORM.S.DIST(0,4, VRAI)` dans Microsoft Excel.

2. Calcul de l'indice d'amélioration = U3 – 50 %. Le WWC calcule l'indice d'amélioration en soustrayant 50 % à l'indice U3. Cet indice représente donc la différence des rangs centiles entre un élève moyen du groupe intervention et un élève moyen du groupe comparaison (les centiles étant basés sur la distribution du groupe comparaison). Par exemple, l'indice d'amélioration pour une taille d'effet de 0,40 peut être calculé en points de pourcentage en utilisant `100*pnorm(0,40) – 50` dans R ou `100*NORM.S.DIST(0,4,TRUE) – 50` dans Microsoft Excel.

Le WWC peut également calculer des indices d'amélioration pour les tailles d'effet moyennes au niveau du domaine dans une étude (combinant potentiellement plusieurs résultats de l'étude ou mesures de résultats) et pour les tailles d'effet globales des méta-analyses qui synthétisent les résultats de plusieurs études. Le WWC synthétisera d'abord les tailles d'effet, puis calculera l'indice d'amélioration pour la taille d'effet globale, plutôt que de calculer la moyenne des indices d'amélioration. La partie suivante décrit les procédures mises en œuvre par le WWC pour synthétiser les tailles d'effet.

⁵ Les calculs de l'indice U3 supposent des variances de population égales pour les groupes intervention et comparaison. Bien que l'estimation U3 ne soit pas une estimation impartiale de la zone de la queue, le biais est généralement faible pour les études avec des tailles d'échantillon susceptibles d'être trouvées dans les revues WWC (Hedges & Olkin, 2016)

Partie2. Formules statistiques pour des résultats agrégés

Pour déterminer l'ampleur d'un effet global, le WWC combine les résultats dans trois situations : (a) les résultats de [plusieurs sous-échantillons](#) sont publiés au sein d'une étude en utilisant la même mesure, (b) [plusieurs résultats principaux](#) pour un domaine donné sont publiés au sein d'une même étude, et (c) les résultats de [plusieurs études](#) ont été calculés pour un domaine donné et doivent être synthétisés.

1. Synthèse de sous-échantillons au sein d'une étude

Certaines études présentent les résultats de plusieurs sous-échantillons de sujets sans publier de résultat global. Cela peut concerner par exemple, les résultats pour les élèves de 6^{ème}, 7^{ème} et 8^{ème} année ; ou les résultats des élèves à risque élevé et faible ; ou des sous-échantillons basés sur des caractéristiques démographiques (par exemple pour les garçons et les filles).

Dans ces situations, le What Works Clearinghouse (WWC) peut interroger les auteurs pour savoir s'ils ont effectué une analyse sur l'échantillon complet. L'analyse de l'étude est préférable, car elle peut être plus précise que le calcul du WWC. Si le WWC ne peut pas obtenir de résultats agrégés de la part de l'auteur, alors le WWC fait la moyenne des résultats de tous les sous-échantillons et calcule une seule mesure de résultat pour l'étude.

Le WWC utilisera les équations de ce chapitre si les résultats des sous-échantillon répondent à tous les critères suivants :

- Les sous-échantillons sont indépendants (sans chevauchement).
- Les résultats sont mesurés de la même façon.
- Les résultats sont collectés au même moment.
- Les procédures de notation des résultats sont les mêmes dans tous les sous-échantillons.
- Les moyennes non ajustées ou ajustées sont publiées.

Les équations suivantes permettent de calculer \bar{Y}_j , la moyenne du groupe j (groupe intervention ou comparaison) sur tous les G sous-échantillons mutuellement exclusifs ainsi que SD_j l'écart-type combiné :

[F.1]

$$\bar{Y}_j = \frac{\sum_{g=1}^G n_{gj} \bar{y}_{gj}}{\sum_{g=1}^G n_{gj}}$$
$$SD_j = \sqrt{\frac{\sum_{g=1}^G [(n_{gj} - 1)SD_{gj}^2 + n_{gj}(\bar{y}_j - \bar{y}_{gj})^2]}{\sum_{g=1}^G n_{gj} - 1}}$$

où n_{gj} , \bar{y}_{gj} et SD_{gj} sont respectivement la taille de l'échantillon, la moyenne des résultats et l'écart-type du sous-échantillon g du groupe j .

Le WWC publiera le résultat combiné comme résultat principal et les résultats de chaque sous-échantillon comme des résultats complémentaires. Les tailles d'effet et les erreurs standard seront calculées en utilisant les formules des chapitres précédents sur l'échantillon combiné. Par exemple, les

moyennes et des écarts-types combinés calculés à partir de l'équation F.1 seront utilisés dans l'équation E.5 qui permettra de calculer la taille de l'effet pour le résultat combiné dans les études avec affectation au niveau individuel.

2. Synthèse des résultats principaux d'une étude dans un domaine donné

Le WWC calculera une taille d'effet moyenne au niveau du domaine si les auteurs de l'étude publient plusieurs résultats principaux qui répondent aux normes WWC dans un domaine donné. Considérons une étude qui publie plusieurs mesures de résultats telles que deux mesures standardisées de la réussite en mathématiques. Le WWC fera la moyenne des tailles d'effet pour les deux mesures, à condition que les résultats répondent aux normes WWC et satisfassent aux critères pour être définis comme des résultats principaux. La présentation par le WWC des résultats d'une étude, en particulier pour la caractérisation de l'efficacité de l'intervention, se concentrera sur ces moyennes au niveau du domaine plutôt que sur les principaux résultats individuels qui composent les moyennes. Si une étude ne comprend qu'un seul résultat principal dans un domaine donné, alors l'interprétation au niveau du domaine sera l'interprétation faite sur ce seul résultat principal.

La taille d'effet moyenne au niveau du domaine est la moyenne arithmétique non pondérée des tailles d'effet calculées pour les résultats principaux individuels qui répondent aux normes WWC dans un domaine donné :

[F.3]

$$\bar{g} = \frac{1}{K} \sum_{i=1}^K g_i$$

où \bar{g} est la taille moyenne de l'effet au niveau du domaine, K est le nombre de résultats principaux satisfaisants aux normes WWC dans un domaine donné, et g_i est la taille d'effet du résultat i . Le WWC exclut les résultats qui ne répondent pas aux normes WWC du calcul de cette moyenne.

L'erreur standard pour cette taille d'effet moyenne au niveau du domaine est donnée par ce qui suit :

[F.4]

$$SE[\bar{g}] = \frac{1}{K} \sqrt{\sum_{i=1}^K SE[g_i]^2 + \rho \sum_{i \neq j} SE[g_i]SE[g_j]}$$

où ρ est le coefficient de corrélation moyen entre les mesures des résultats, et $SE[g_i]$ et $SE[g_j]$ sont les erreurs standard des tailles d'effet i et j ⁶. Cette formule est une approximation en partie parce qu'elle suppose que toutes les corrélations entre les résultats sont les mêmes⁷. Si le coefficient de corrélation ne peut être déterminé de façon satisfaisante, on lui attribuera la valeur $\rho = 1,0$. Cette formule générale s'applique à tous les estimateurs de taille d'effet et d'erreur standard des chapitres

⁶ La notation de sommation $i \neq j$ traite les paires comme ordonnées (par exemple le cas $i = 2$ et $j = 4$ est différent du cas $i = 4$ et $j = 2$), ce qui signifie que $\sum_{i \neq j} 1 = K(K-1)$.

⁷ Les corrélations entre les mesures des résultats sont utilisées à la place des corrélations entre les tailles d'effet. En général, les deux corrélations sont très similaires, surtout lorsque la corrélation entre les mesures est positive, ce qui est raisonnable dans ce contexte. Lorsqu'elles diffèrent, le coefficient de corrélation entre les mesures des résultats sera légèrement supérieur au coefficient de corrélation entre les tailles d'effet, ce qui se traduira par une estimation de l'erreur standard légèrement conservatrice, en limitant le risque de type I (Thompson et Becker, 2014).

précédents, pour les études avec affectation au niveau individuel comme les études avec affectation au niveau des grappes.

Le WWC détermine la significativité statistique au niveau du domaine en calculant d'abord t , le rapport de la taille de l'effet au niveau du domaine \bar{g} sur son erreur standard $SE[\bar{g}]$ à l'aide de la formule $t = \bar{g}/SE[\bar{g}]$. Le degré de liberté pour déterminer la valeur- p associée à cette statistique t est la moyenne des degrés de liberté calculés pour chacun des principaux résultats contributifs.

Le WWC utilise la statistique t et le degré de liberté au niveau du domaine pour calculer la valeur- p au niveau du domaine. On peut par exemple calculer cette valeur- p en utilisant Microsoft Excel avec la fonction suivante : $p = T.DIST.2T(t, df)$. On peut également utiliser le logiciel R de la manière suivante : $p = 2 * pt(-abs(t), df)$. Si la valeur- p d'un test t bilatéral est inférieure à 0,05, le résultat au niveau du domaine est statistiquement significatif. Aucune correction pour les comparaisons multiples n'est requise, car l'interprétation au niveau du domaine est le résultat d'un seul agrégat résumant plusieurs résultats principaux dans un domaine donné.

Exemple. Considérons un essai contrôlé randomisé (ECR) avec affectation au niveau individuel qui a publié deux résultats principaux dans le domaine de la réussite en mathématiques. Les résultats sont deux mesures standardisées distinctes avec comme coefficient de corrélation $\rho = 0,70$. Leurs tailles d'effet et leurs erreurs standards (tableau F.1) étaient basées sur des statistiques ajustées par régression à l'aide des formules pertinentes des chapitres précédents. Les deux résultats ont été obtenus sur des échantillons de même tailles : 50 élèves pour les groupes interventions et 50 élèves pour les groupes comparaisons, ce qui donne $df = 98$ comme degré de liberté au niveau du domaine.

Tableau F.1. Calculs au niveau du domaine pour un exemple d'étude avec deux résultats principaux

Résultats	g	SE	p
Réussite en mathématiques 1	0,446	0,153	N/A
Réussite en mathématiques 2	0,348	0,142	N/A
Moyenne au niveau du domaine	0,397	0,136	0,004

N/A : non applicable.

La taille d'effet au niveau du domaine est la moyenne des deux tailles d'effet. L'erreur standard au niveau du domaine est calculée en utilisant l'équation F.4 comme suit :

$$SE[\bar{g}] = \frac{1}{2} \sqrt{0.153^2 + 0.142^2 + 0.70(0.153 \times 0.142 + 0.142 \times 0.153)} = 0.136$$

La statistique t au niveau du domaine est donc $t = 0,397/0,136 = 2,919$ avec un degré de liberté égal à 98. La formule Excel $T.DIST.2T(2.919, 98)$ permet de calculer la valeur- p qui est ici égale à 0,004, ce qui indique un résultat statistiquement significatif au niveau du domaine. L'utilisation de la formule R de $2 * pt(-abs(2.919), 98)$ donne la même valeur- p .

3. Synthèse des résultats de plusieurs études pour d'un domaine donné

Le WWC combine les tailles d'effet de plusieurs études pour les publications du WWC qui considèrent plus d'une étude, comme les rapports d'intervention et les guides des pratiques. Le WWC calcule les tailles d'effet moyennes inter-études au niveau du domaine en conduisant une méta-analyse qui met en œuvre le modèle des effets fixes.

Le WWC a choisi le modèle des effets fixes parce que l'objectif du WWC est de proposer des conclusions à partir des études dans ses rapports d'intervention et ses guides de pratiques.

Contrairement au modèle de l'effet fixe (singulier), le modèle des effets fixes (pluriel) ne suppose pas que les études estiment un effet commun. Le modèle des effets fixes suppose au contraire que la variation observée entre les tailles d'effet dans une méta-analyse reflète la véritable variation des effets sur la population. En conséquence, les inférences à des populations plus larges sont limitées aux populations présentant des caractéristiques pertinentes qui sont les mêmes que les caractéristiques des populations des études, ces caractéristiques pertinentes étant celles associées au calcul de la taille d'effet.

La procédure utilisée par le WWC dans ses méta-analyses donne plus de poids aux études ayant estimés avec plus de précision les tailles d'effet. Par exemple, une expérience randomisée menée sur 300 élèves aura environ trois fois le poids d'une expérience randomisée menée sur 100 élèves. Cette approche est similaire à la façon dont les établissements calculent les moyennes pondérées cumulatives : une note obtenue dans un cours de trois heures aura trois fois le poids d'une note obtenue dans un cours d'une heure.

Pondération par l'inverse de la variance. Les tailles d'effet sont pondérées par l'inverse de leurs variances (qui sont largement déterminées par la taille de l'échantillon). La variance est le carré de l'erreur standard. Le poids W_s pour l'étude s dans une méta-analyse peut donc s'écrire comme suit :

[F.5]

$$W_s = \frac{1}{V[\bar{g}_s]} = \frac{1}{SE[\bar{g}_s]^2}$$

où $V[\bar{g}_s]$ et $SE[\bar{g}_s]$ sont respectivement la variance et l'erreur standard de \bar{g}_s qui est la taille d'effet moyenne au niveau du domaine de l'étude s . La taille de l'effet au niveau du domaine et l'erreur standard propre à l'étude proviennent des équations F.3 et F.4.

La moyenne pondérée \bar{G} pour une méta-analyses qui combine les tailles d'effet de J études est alors :

[F.6]

$$\bar{G} = \frac{\sum_{s=1}^J W_s \bar{g}_s}{\sum_{s=1}^J W_s}$$

Seuls les résultats au niveau du domaine qui satisfont aux normes WWC sont inclus dans ces moyennes. Le WWC effectue ces calculs séparément pour chacun des domaines.

L'erreur standard $SE[\bar{G}]$ de la moyenne pondérée \bar{G} est donnée par ce qui suit :

[F.7]

$$SE[\bar{G}] = \sqrt{\frac{1}{\sum_{s=1}^J W_s}}$$

Une moyenne calculée au niveau d'une méta-analyse statistiquement significative est une moyenne pour laquelle l'hypothèse nulle a été rejetée à l'aide d'un test z bilatéral pour une erreur de type I égale à $\alpha = 0,05$. La statistique du test z est égale au rapport de la taille d'effet de la méta-analyse sur son erreur standard : $z = \bar{G} / SE[\bar{G}]$.

Exemple. Considérons trois ECR qui satisfont tous aux *normes WWC sans réserve* et évaluent la même intervention en étudiant ses effets sur la réussite des élèves en mathématiques. Le tableau F.2 donne les tailles d'effet au niveau du domaine et leur erreur standard pour chaque étude. Les tailles d'effet au niveau du domaine pour chaque étude pourraient chacune synthétiser plusieurs résultats principaux d'une étude. Par exemple, la première ligne correspond à la moyenne au niveau du domaine calculée dans le tableau F.1.

Tableau F.2. Exemple d'une méta-analyse à effets fixes

Résultats	g	SE	p
Etude 1	0.397	0.136	0.004
Etude 2	0.413	0.178	0.022
Etude 3	0.156	0.054	0.004
Moyenne méta-analyse	0.205	0.048	<.001

L'erreur standard plus petite pour l'étude 3 indique que cette taille d'effet a été estimée avec plus de précision que pour les études 1 et 2, reflétant probablement une taille d'échantillon plus grande. Par conséquent, le poids de l'étude 3 est plus important que les poids des deux autres études. En utilisant l'équation F.5, les poids calculés pour les première, deuxième et troisième études sont respectivement de 54,066 ; 31,562 et 342,936.

La moyenne calculée par l'équation F.6 pour l'ensemble des études de la méta-analyse est égale à :

$$\bar{G} = \frac{54.066 \times 0.397 + 31.562 \times 0.413 + 342.936 \times 0.156}{54.066 + 31.562 + 342.936} = 0,205$$

L'erreur standard de cette moyenne est calculée en utilisant l'équation F.7 comme suit :

$$SE[\bar{G}] = \sqrt{\frac{1}{54.066 + 31.562 + 342.936}} = 0,048$$

Ces calculs ne dépendent pas directement de la taille des échantillons ou de l'unité d'affectation aléatoire des études, car les erreurs standards intègrent déjà ces caractéristiques. Par conséquent, l'erreur standard au niveau du domaine et spécifique à chaque étude agit comme une statistique récapitulative qui tient déjà compte d'une gamme variée de caractéristiques du design des études, et cela se répercute au niveau de la méta-analyse.

Enfin, la statistique du test z est égale à $z = 0,205 / 0,048 = 4,251$, ce qui correspond à une valeur p inférieure à 0,001, indiquant une moyenne statistiquement significative au niveau de la méta-analyse.

Repondération par la qualité associée à chaque résultat. Pour limiter le biais des résultats qui sont *conformes aux normes WWC avec réserves*, le WWC utilisera une procédure visant à pondérer le poids des études dans une méta-analyses à effets fixes quand les poids calculés pour les résultats *conformes aux normes WWC sans réserve* ne représentent pas la majorité des poids des études sélectionnées dans la méta-analyse. Cette procédure garantira que la moyenne calculée au niveau de la méta-analyse soit principalement basée sur les tailles d'effet de résultats dont la qualité permet d'inférer avec le plus de confiance possible.

Le WWC met en œuvre cette procédure de repondération uniquement lorsque toutes les conditions suivantes sont remplies :

1. La synthèse comprend un ou plusieurs résultats *conforme aux normes WWC avec réserves* et un ou plusieurs résultats *conforme aux normes WWC sans réserve*.
2. Les résultats *conformes aux normes WWC avec réserves* représentent plus de 50 % des poids (inverses des variances) calculés pour la méta-analyse.
3. La somme totale des poids calculés pour les résultats *conformes aux normes WWC sans réserve* est égale à au moins 87,2, ce qui est le seuil nécessaire pour détecter une taille d'effet égale à 0,30 avec une puissance de 80 %.

Si l'un de ces critères n'est pas rempli, le WWC synthétisera les tailles d'effet à partir des résultats correspondants en utilisant uniquement les poids calculés comme égaux à l'inverse de la variance.

Si tous ces critères sont remplis, la formule suivante est utilisée pour repondérer la moyenne méta-analytique :

[F.8]

$$\bar{G}^* = \frac{\sum_{s=1}^J W_s^* \bar{g}_s}{\sum_{s=1}^J W_s^*}$$

La formule suivante fournit la variance de la moyenne méta-analytique repondérée :

[F.9]

$$SE[\bar{G}^*] = \frac{\sqrt{\sum_{s=1}^J W_s^{*2} SE[\bar{g}_s]^2}}{\sum_{s=1}^J W_s^*}$$

où W_i^* est le poids de l'étude i calculé de façon à ce que la somme des poids des études *conformes aux normes WWC avec réserves* soit égal à 0,49 et que la somme des poids des études *conformes aux normes WWC sans réserves* soit égal à 0,51.

Exemple de repondération. Considérons que les premier et deuxième résultats du tableau F.2 soient *Conformes aux normes WWC sans réserves*, mais que le troisième résultat soit *Conforme aux normes WWC avec réserves*. Bien que le troisième résultat ait le niveau de qualité le plus bas, il est associé au poids dont la proportion dans la méta-analyse est la plus importante, cette proportion étant égale à $342,936 / (54,066 + 31,562 + 342,936) = 80,0\%$.

Dans ce cas, le premier et le deuxième résultat pourront être repondérés et contribuer de façon prépondérante à la taille d'effet calculée pour la méta-analyse⁸ :

$$W_1^* = 0.51 \frac{54.066}{54.066 + 31.562} = 0.322$$

$$W_2^* = 0.51 \frac{31.562}{54.066 + 31.562} = 0.188$$

$$W_3^* = 0.49 \frac{342.936}{342.936} = 0.490$$

La moyenne et l'erreur standard au niveau de la méta-analyse seront calculées comme suit :

⁸ Remarque de la traductrice : la somme des poids est égale à 85,6 calculés initialement et est donc inférieure au seuil requis de 87,2 pour procéder à une repondération.

$$\bar{G}^* = \frac{0.322 \times 0.397 + 0.188 \times 0.413 + 0.490 \times 0.156}{0.322 + 0.188 + 0.490} = 0.282$$

$$SE[\bar{G}^*] = \frac{\sqrt{0.322^2 \times 0.136^2 + 0.188^2 \times 0.178^2 + 0.490^2 \times 0.054^2}}{0.322 + 0.188 + 0.490} = 0.061$$

Analyse de modérateurs. Dans tous les cas où les tailles d'effet des résultats *conformes aux normes WWC avec réserves* sont combinées à des tailles d'effet de résultats *conforme aux normes WWC sans réserve*, le WWC effectuera un test en considérant le niveau de qualité du résultat comme un modérateur de la taille d'effet calculée au niveau de la méta-analyse.

Lorsque la méta-analyse comprend une combinaison d'études qui ont utilisé à la fois des mesures de résultats indépendantes et non indépendantes, le WWC effectuera un test pour prendre en compte la composition des mesures (la proportion de mesures indépendantes dans une étude) et les tailles d'effet.

Ces deux ensembles de tests utiliseront les pondérations par l'inverse de la variance (spécifiées dans l'équation F.5) pour maximiser la puissance statistique et mieux détecter les différences dans les tailles d'effet.

Glossaire

Symbole	Description (français)	Description (américain)	Equation
β	Coefficient de régression standardisé pour estimer l'effet d'une intervention	Standardized regression coefficient for intervention impact estimate	E15, E16
η	Augmentation de la variance que l'on peut attribuer au fait d'avoir regroupé les élèves en grappes	Design effect for a cluster-level assignment study	E22, E23, E25, E26, E27, E29, E30, E31, E33
γ	Ajustement pour les petits nombres de grappes pour les études avec affectation au niveau des grappes	Small numbers of clusters correction for a cluster-level assignment study	E20, E24, E32, E33
ρ_{cor}	Coefficient de corrélation entre les scores prétests et scores posttests	Correlation between the pretest and posttest measures	E19
ρ	Coefficient de corrélation moyen entre les mesures des résultats	Average correlation among outcome measure	F4
ρ_{ICC}	Coefficient de corrélation intraclasse	Intraclass correlation coefficient	E20, E21, E22
ω	Facteur de correction pour petits échantillons	Small sample bias correction term	E1, E3, E5, E6, E7, E8, E9, E11, E12, E13, E16, E20, E23, E24, E25, E26, E27, E29, E30, E31, E32
b	Différence des moyennes non standardisée entre les groupes comparaison et contrôle	Unstandardized intervention-comparison group mean difference of the outcome	E1, E14, E15, E20, E27
df	Degré de liberté	Degrees of freedom	E3, E4, E21, E23, E24 E25, E26, E28, E29, E30
F	Résultat du test ANOVA (statistique F)	F statistic	E8, E13
G	Nombre de sous-échantillons mutuellement exclusifs	Number of mutually exclusive subgroups	F1, F2
\bar{G}	Taille d'effet moyenne pour une méta-analyse à effets fixes	Fixed-effects meta-analytic average effect size	F6
\bar{G}^*	Taille d'effet moyenne pondérée pour une méta-analyses de façon à ce que les études <i>conformes sans réserve aux normes WWC</i> reçoivent la majeure partie du poids analytique	Rewighted meta-analytic average effect size such that studies rated Meets WWC Standards Without Reservations receive most of the meta-analytic weight	F8
g	Taille d'effet standardisée, g de Hedges	Hedge's g standardized effect size	E1, E5, E6, E7, E8, E9, E11, E12, E13, E16, E20, E23, E24, E 25, E26, E27, E29, E30, E31, E32
g_i	Taille d'effet du résultat principal i	Effect size for the i th main finding in a study	F3
\bar{g}	Taille moyenne de l'effet au niveau du domaine pour une étude	Domain-level composite effect size for a study	F3
\bar{g}_s	Taille d'effet moyenne au niveau du domaine pour l'étude s	Domain-level composite effect size for study s	F6, F8
J	Nombre total d'études	Total number of studies	F6, F7 F,8, F9
K	Nombre total de résultats principaux conformes aux normes	Total number of main findings meeting WWC standards for an outcome domain in a single study	F3, F4

	WWC pour un domaine dans une étude		
M	Nombre total de grappes dans une étude avec affectation au niveau des grappes	Total number of clusters in a cluster-level assignment study	E20, E21, E22, E27, E28, E30
N	Nombre total de sujets dans l'échantillon analytique	Total number of individuals in the analytic sample	E4, E14, E16, E20, E21, E22, E27, E30
n_c	Nombre de sujets dans l'échantillon analytique du groupe comparaison	Number of individuals in the comparison group analytic sample	E2, E5, E6, E7, E8, E9, E11, E12, E13, E14, E16, E17, E18, E23, E26, E29, E31, E32, E33
n_{gj}	Nombre de sujets du sous-échantillon g du groupe j	Number of individuals in subsample g in group j	F1
n_i	Nombre de sujets dans l'échantillon analytique du groupe contrôle	Number of individuals in the intervention group analytic sample	E2, E5, E6, E7, E8, E9, E11, E12, E13, E14, E16, E17, E18, E23, E26, E29, E31, E32, E33
R^2	Coefficient de corrélation du modèle de régression linéaire multiple	Multiple correlation between the covariates and the outcome	E11, E12, E13, E18, E26
SD	Ecart-type total au niveau individuel des scores	Total individual-level standard deviation of the outcome	E14
SD_B	Écart-type groupé au niveau des grappes des scores	Pooled within-group, cluster-level standard deviation of the outcome	E27, E30
SD_c	Écart-type au niveau individuel du groupe comparaison	Individual-level standard deviation of the comparison group	E2, E5, E17, E18, E33, E34
SD_{gj}	Écart-type des scores du sous-échantillon g du groupe j	Outcome standard deviation for subsample g in assignment group j	F1
SD_i	Écart-type au niveau individuel du groupe intervention	Individual-level standard deviation of the intervention group	E2, E5, E17, E18, E33, E34
SD_j	Écart-type des scores combiné pour tous les sous-échantillons du groupe j	Outcome standard deviation combined across all subsamples for assignment group j	F1
SD_p	Écart-type groupé des scores au niveau individuel	Pooled within-group, individual-level standard deviation of the outcome	E1, E2, E9, E14, E19, E20, E24, E25
$SD_{p.gain}$	Écarts-types au niveau individuel des gains de scores	Pooled within-group, individual-level standard deviation of the pre-post gain score	E19
SD_x	Écart-type total de la mesure du niveau initial	Standard deviation of the baseline measure	E15
SD_y	Écart-type total des scores post-test	Standard deviation of the outcome measure	E15
$SE[b]$	Erreur standard de la différence des moyennes non standardisée	Standard error of the unstandardized mean difference	E9, E10, E30
$SE_{CC}[b]$	Erreur standard tenant compte du regroupement pour la différence des moyennes non standardisée dans les études avec affectation au niveau des grappes	Cluster-corrected standard error of the unstandardized mean difference in a cluster-level assignment study	E24
$SE_{UC}[b]$	Erreur standard ne tenant pas compte du regroupement pour la différence des moyennes non standardisée dans les études avec affectation au niveau des grappes	Standard error uncorrected for clustering of the unstandardized mean difference in a cluster-level assignment study	E25
$SE[g]$	Erreur standard de la taille d'effet (le g de Hedges)	Standard error of the Hedges' g effect size	E6, E9, E11, E23, E24, E25, E26, E29, E30

$SE[\bar{g}]$	Erreur standard de la taille d'effet moyenne au niveau du domaine	Standard error of the domain-level average effect size in a single study	F4
$SE[g_i]$	Erreur standard de la taille d'effet i dans une étude	Standard error of the i th effect size in a study	F4
$SE[\bar{g}_s]$	Erreur standard de la taille d'effet moyenne au niveau du domaine pour l'étude s	Standard error of the domain-level average effect size for study s	F5, F9
$SE[\bar{G}]$	Erreur standard de la taille d'effet moyenne pour une méta-analyse à effets fixes	Standard error of the fixed-effects meta-analytic average	F7
$SE[\bar{G}^*]$	Erreur standard de la taille d'effet moyenne pondérée pour une méta-analyses de façon à ce que les études <i>conformes sans réserve aux normes WWC</i> reçoivent la majeure partie du poids analytique	Standard error of the reweighted meta-analytic average such that studies rated Meets WWC Standards Without Reservations receive most of the meta-analytic weight	F9
$SE[y_i]$	Erreur standard de la moyenne du groupe intervention	Standard error of the intervention group mean	E10, E17, E18
$SE[y_c]$	Erreur standard de la moyenne du groupe comparaison	Standard error of the comparison group mean	E10, E17, E18
$SE_{CC}[y_i]$	Erreur standard des moyennes du groupe intervention tenant compte des regroupements dans les études avec affectation au niveau des grappes.	Cluster-corrected standard error of the intervention group mean in cluster-level assignment study	E33
$SE_{CC}[y_c]$	Erreur standard des moyennes du groupe comparaison tenant compte des regroupements dans les études avec affectation au niveau des grappes	Cluster-corrected standard error of the comparison group mean in a cluster-level assignment study	E33
$SE_{UC}[y_i]$	Erreur standard des moyennes du groupe intervention ne tenant pas compte des regroupements dans les études avec affectation au niveau des grappes	Standard error uncorrected for clustering of the intervention group mean in a cluster-level assignment study	E34
$SE_{UC}[y_c]$	Erreur standard des moyennes du groupe comparaison ne tenant pas compte des regroupements dans les études avec affectation au niveau des grappes	Standard error uncorrected for clustering of the comparison group mean in a cluster-level assignment study	E34
t	Résultat d'un test t pour la différence des moyennes des groupes	test statistic for group mean difference	E7, E12, E43
t_{CC}	Résultat d'un test t pour la différence des moyennes des groupes corrigée en tenant compte du regroupement en grappe	Cluster-corrected test statistic for group mean difference in cluster-level assignment study	E31
t_{UC}	Résultat d'un test t pour la différence des moyennes des groupes corrigée ne tenant pas compte du regroupement en grappe	test statistic uncorrected for clustering for group mean difference in cluster-level assignment study	E32
$V[\bar{g}_s]$	Variance de la taille d'effet moyenne au niveau du domaine pour l'étude s	Variance of the domain-level average effect size for study s	F5

W_s	Poids (inverse de la variance) pour l'étude s dans une méta-analyse	Meta-analytic inverse variance weight for study s	F5, F6, F7
W_s^*	Poids d'une étude pondéré dans une méta-analyse de façon à ce que les études <i>conformes sans réserve aux normes WWC</i> reçoivent la majeure partie du poids analytique	Rescaled meta-analytic weight for study s that ensures that most the meta-analytic weight goes to studies rated Meets WWC Standards Without Reservations	F8, F9
\bar{Y}_j	Moyenne combinée dans tous les sous-groupes affectés à la condition j	Combined group mean across all subsamples in assignment group j	F1
\bar{y}_c	Moyenne du groupe comparaison	Comparison group mean	E5
\bar{y}_{gj}	Moyenne des scores du sous-échantillon g du groupe j	Outcome mean for subsample g in assignment group j	F1
\bar{y}_i	Moyenne du groupe intervention	Intervention group mean	E5