# Manuel des procédures du What Works Clearinghouse

Version 4.0 (Octobre 2017)

*Institute of Education Sciences*

# *What Works Clearinghouse Procedures Handbook*

Traduit en français par
Nathalie ROQUES (juillet 2019)

What Works Clearinghouse, Institute of Education Sciences, U.S. Department of Education. (2017, October). *What Works Clearinghouse: Standards Handbook (Version 4.0)*. Retrieved from http://whatworks.ed.gov

https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_handbook_v4.pdf

# Note concernant la traduction

Cette traduction est libre et n'engage en aucune manière le What Works Clearinghouse et ses organismes de tutelle.

Le document original, écrit par des américains, s'appuie sur des études effectuées aux USA. Pour faciliter la lecture par des français, le niveau des élèves a lui aussi été traduit (ainsi les élèves de 8ème grade américains sont devenus des élèves de 4ème).
La mise en forme de la version originale a été le plus souvent laissée inchangée (usage des guillemets, mots en italique, …). Certaines expressions américaines ont été conservées dans le texte, notamment les titres d'ouvrages et les noms des organismes américains. Dans quelques cas, une traduction les accompagne.

Aucune annexe n'a été traduite, et les passages non traduits dans le texte sont signalés par le symbole …/….

Vous trouverez également une traduction partielle du *What Works Clearinghouse Standards Handbook* (Version 4.0). Ces documents présentent à eux deux les règles et normes qui encadrent et soutiennent tous les textes et outils publiés par le WWC.

Ces traductions sont en accès libre sur le site www.mathadoc.fr

# Table des matières



## Tableaux



## Figures



## Liste des abréviations

# I. INTRODUCTION

Il est essentiel que les décideurs en matière d'éducation aient accès aux meilleures données probantes concernant l'efficacité des programmes, des politiques et des pratiques en matière d'enseignement. Cependant, il peut être difficile, long et coûteux pour ces décideurs d'avoir accès aux conclusions d'études pertinentes sur l'efficacité de ces interventions[1] et d'en tirer les conclusions qui s'imposent. Le centre de documentation *What Works Clearinghouse* (WWC) répond à ce besoin d'informations synthétiques et dignes de confiance en identifiant les recherches effectuées sur des interventions dans le domaine de l'enseignement, en évaluant la qualité de ces recherches, en résumant et en diffusant les preuves issues d'études respectant les normes édictées par le WWC.

Le WWC est une initiative de l'Institut des sciences de l'éducation (IES) du département de l'Éducation des États-Unis, créé en vertu de la loi de 2002 (Education Sciences Reform Act). Il s'agit d'un volet important de la stratégie de l'IES qui consiste à mettre à profit les recherches, les évaluations et les méthodes statistiques rigoureuses et pertinentes dans le but d'améliorer le système de l'éducation national. La mission du WWC est d'être **une source centrale et fiable de données scientifiques probantes sur « ce qui marche » (*What Works*) dans le domaine de l'enseignement**. Le WWC examine la recherche concernant les interventions visant à améliorer les résultats pertinents sur le plan éducatif, notamment celles qui concernent les élèves et le personnel enseignant.

Les examens systématiques menés par le WWC constituent la base de bon nombre des documents et des outils produits par cette institution. Ils appliquent et suivent des normes et des procédures cohérentes, objectives et transparentes, tout en assurant une couverture complète de la littérature pertinente existante. Les examens systématiques menée par le WWC suivent cinq étapes :

1. *Développement du protocole de l'examen*. Un protocole formel est élaboré pour chaque examen, y compris pour chaque centre d'intérêt du WWC (par exemple, l'alphabétisation des adolescents ou les mathématiques à l'école primaire) afin de définir les paramètres de la recherche qu'il convient d'inclure dans le champ de l'examen (par exemple les caractéristiques de la population et les types d'interventions) ; la recherche documentaire (par exemple les mots-clés utilisés pour interroger les bases de données) ; et toute application spécifique des normes à un sujet donné (par exemple, établissement des seuils acceptables d'attrition des échantillons et spécification des caractéristiques qui permettent de déterminer l'équivalence initiale des groupes étudiés).
2. *Identification de la littérature pertinente.* Les études sont rassemblées grâce à une recherche exhaustive de la littérature de recherche publiée et non publiée disponible. Cette recherche s'appuie entre autres sur des bases de données électroniques.
3. *Sélection des études*. Les manuscrits sont, dans un premier temps, examinés pour évaluer leur éligibilité afin de déterminer s'ils répondent à la question de recherche

[1] Une intervention est donc constituée par un ou plusieurs éléments caractérisés comme étant un programme, une politique ou une pratique.

initiale, fournissent des preuves potentiellement crédibles de l'efficacité d'une intervention et s'inscrivent dans le cadre du protocole d'examen.

4. *Examen des études*. Chaque étude éligible est examinée par rapport aux normes élaborées par le WWC. Le WWC utilise un processus d'examen structuré pour évaluer la validité des résultats présentés dans ces études en relation avec l'efficacité des interventions sur l'enseignement. Les normes développées par le WWC se concentrent sur la validité causale au sein de l'échantillon de l'étude (validité interne) plutôt que sur l'estimation de la réplication potentielle des résultats de l'étude et de son élargissement éventuel à d'autres contextes (validité externe) [2].
5. *Présentation des résultats*. Les détails de l'examen et ses conclusions sont résumés sur le site Web du WWC, et souvent dans une publication du WWC. Pour nombre de ses documents et outils[3] produits, le WWC combine les résultats d'études individuelles pour donner naissance à des mesures synthétisant l'efficacité d'une intervention, en présentant notamment l'ampleur des résultats et leur niveau de preuve.

En outre, le WWC examine certaines études en dehors du processus d'examen systématique, par exemple dans le cas où certaines études reçoivent une attention particulière de la part des médias. Ces examens sont également guidés par un protocole d'examen et utilisent les mêmes normes et procédures de présentation des résultats.

Ce *What Works Clearinghouse Procedures Handbook* (version 4.0) traduit ici par *Manuel des procédures WWC* fournit une description détaillée des procédures utilisées par le WWC lors du processus d'examen systématique, en particulier les étapes 1 à 3 et l'étape 5 décrites ci-dessus. L'étape 4 quant à elle est décrite dans un autre document intitulé *What Works Clearinghouse Standards Handbook[4]*, qui inclue la description des normes utilisées par le WWC pour examiner les études et attribuer l'un des classements suivants indiquant le niveau de preuve de l'étude : *conforme sans réserve aux normes WWC*, *conforme avec réserve aux normes WWC*, *non conforme aux normes WWC*. Depuis octobre 2017, ces deux documents remplacent le document unique utilisé antérieurement, le *What Works Clearinghouse Procedures and Standards Handbook (version 3.0, publiée en mars 2014)[5].* La figure I.1 montre comment les étapes du processus d'examen systématique du WWC sont réparties entre les deux manuels.

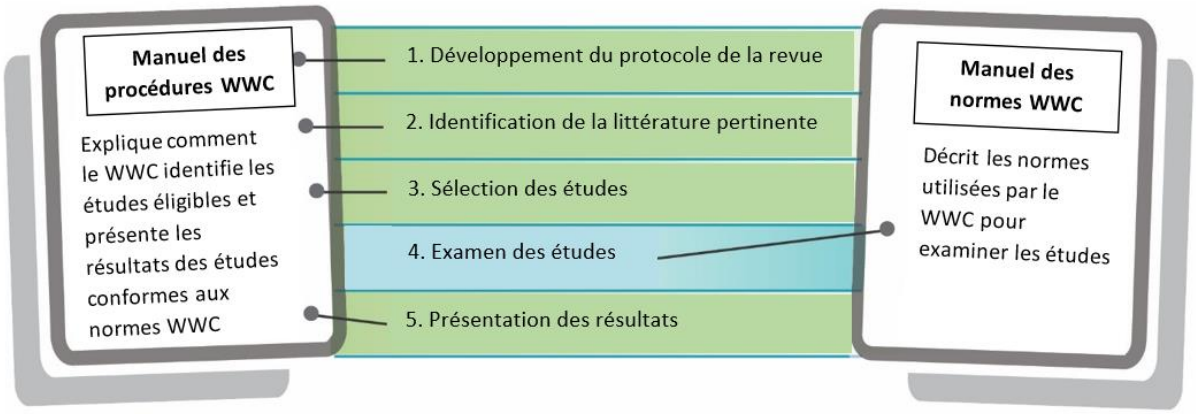
[2] Ces normes concernent le design (ou conception) des études qui comparent des groupes ; elles seront citées dans ce texte comme les « normes WWC ».

[3] Il s'agit notamment des *Rapports d'intervention*, des *Guides des pratiques* et du site internet Find What Works mis à disposition du grand publique librement.

[4] Dont la traduction partielle est intitulée *Manuel des normes WWC*

[5] Une nouvelle version des deux manuels est en cours de rédaction et sera effective à l'automne 2019.

**Figure I.1.** Étapes de l'examen systématique des études dans les manuels du WWC




En général, ce nouveau manuel des procédures contient les mêmes procédures que celles décrites dans le précédent et unique manuel sur les procédures et les normes (version 3.0). Toutefois, outre les modifications apportées à l'organisation (division en deux manuels), les mises à jour suivantes ont été apportées :

- **Le manuel contient des informations supplémentaires qui peuvent être utilisées lors des examens**. Ces informations supplémentaires peuvent s'avérer nécessaires en raison de mises à jour du *Manuel des normes WWC* ou pour la prise en compte de considérations méthodologiques pouvant varier selon les domaines.
- **Le manuel inclut des clarifications supplémentaires concernant le processus d'examen**. Ces clarifications visent à assurer la cohérence entre les examens. Elles incluent également de nouveaux commentaires sur la manière de définir ce qu'est une étude et de mener la correction indispensable au traitement des comparaisons multiples à l'intérieur d'une étude.
- **Le manuel comprend des formules mises à jour pour le calcul de la signification statistique des résultats.** La formule pour les résultats continus comprend un nouvel ajustement tenant compte de la faible taille de l'échantillon (le cas échéant) et une formule différente est proposée pour les résultats dichotomiques.
- **Le manuel inclut des procédures pour la présentation des résultats d'essais contrôlés randomisés calculant l'effet moyen du traitement sur les sujets traités (*complier average causal effects (CACE)*)**. Une nouvelle annexe décrit la manière dont le WWC présente les résultats et la signification statistique d'études calculant l'effet moyen du traitement sur les sujets traités dans le but de mesurer les effets de la participation à l'intervention plutôt que les effets de l'affectation à l'intervention.

Le reste du document est organisé comme suit : le chapitre II décrit les étapes suivies par le WWC pour développer un protocole d'examen. Le chapitre III décrit la manière dont le WWC identifie la littérature pertinente. Le chapitre IV décrit le processus de sélection permettant de déterminer si une étude est (ou non) éligible pour l'examen, et le chapitre V décrit les procédures utilisées pour examiner les études éligibles. Le chapitre VI décrit comment le WWC synthétise les preuves d'efficacité. Les procédures utilisées par le WWC pour assurer

un examen indépendant, systématique et objectif sont décrites dans les annexes (non traduites).

En cherchant à utiliser et appliquer les procédures décrites dans ce manuel, les examinateurs du WWC peuvent parfois ressentir le besoin d'obtenir des conseils supplémentaires. Si nécessaire, le WWC produira des documents à l'intention des examinateurs afin de leur fournir des éclaircissements et une interprétation des procédures assurant la cohérence des examens entre eux. Ces documents élaborés à l'intention des examinateurs préciseront comment ces procédures devraient être mises en œuvre dans les situations où le manuel de procédures actuel n'est pas suffisamment spécifique pour garantir la cohérence des examens.
Comme le WWC continue d'affiner et d'élaborer des procédures, le *Manuel des procédures WWC* sera révisé pour refléter ces changements. Les lecteurs qui souhaitent donner leur avis sur le *Manuel de procédures WWC*, ou sur le WWC en général, peuvent nous contacter à l'adresse http://ies.ed.gov/ncee/wwc/help.

# II. DÉVELOPPER LE PROTOCOLE D'EXAMEN

Avant de procéder à un examen systématique ou à tout autre analyse, le WWC élabore un protocole d'examen officiel pour servir de guide à ce dernier. Le WWC élabore un protocole d'examen à chaque fois qu'un nouveau sujet a été reconnu comme prioritaire (voir l'annexe A). Comme la recherche en éducation couvre un large éventail de sujets, d'interventions et de résultats, un protocole d'examen doit décrire les études pouvant faire l'objet d'un examen, la façon dont le WWC les recherchera et la manière dont elles seront examinées. Le protocole définit les types d'interventions entrant dans le champ de l'analyse, la population sur laquelle l'analyse est axée, les termes utilisés comme par exemple les mots-clés, les paramètres de la recherche documentaire et toute application des normes spécifique à l'examen. Plus précisément, les protocoles WWC incluent des instructions sur les questions suivantes :

• *Déclaration d'objectif*. Tous les protocoles d'examen de WWC commencent par une description de l'objectif général. Les protocoles relatifs à certains travaux d'analyses fournissent également des informations générales sur le sujet traité et décrivent les objectifs de l'examen.

• *Définitions clés*. Les protocoles définissent les termes et concepts clés spécifiques à la substance et à l'objectif de l'examen.

• *Procédures pour effectuer la recherche bibliographique.* Chaque protocole inclut une liste des mots-clés et des termes associés qui seront utilisés pour la recherche bibliographique et une liste des bases de données dans lesquelles effectuer cette recherche (voir l'a²nnexe B pour un exemple de liste de mots-clés et de bases de données). Un protocole peut également fournir des instructions particulières sur la prise en compte de la « littérature grise», y compris les soumissions publiques au WWC via le site Web ou ses membres, les recherches menées et diffusées par les distributeurs / concepteurs d'interventions, la littérature non publiée identifiée lors de revues et synthèses antérieures réalisées ou non par le WWC, la recherche non publiée identifiée par des listes de diffusion et les études publiées sur les sites Web de certaines organisations.

• *Critères d'éligibilité.* Les protocoles de tous les documents et outils diffusés par le WWC spécifient les critères permettant de déterminer si une étude est éligible pour être incluse ou non dans l'analyse. La direction du comité des examinateurs (méthodologiste en chef et expert, décrite plus en détail à l'annexe C) prend des décisions concernant les paramètres clés, tels que la définition des groupes de population éligibles, des types d'interventions, des caractéristiques de l'étude et des résultats digne d'intérêt. Voici des exemples de paramètres spécifiques à un examen habituellement définis dans les protocoles d'examen :

- *Populations éligibles*. Les protocoles spécifient les niveaux et les tranches d'âge (par exemple élèves du CP au CM2) et les caractéristiques de l'échantillon (par exemple plus de la moitié de l'échantillon est constituée d'élèves dont l'anglais n'est pas la langue maternelle) pour les populations d'élèves éligibles, ainsi que des sous-groupes présentant un intérêt pour l'examen. Le protocole peut spécifier une taille d'échantillon minimale requise pour que le WWC puisse tenir compte des résultats

d’une étude, cette taille pouvant dépendre de la population étudiée ou du type d'étude.

- *Interventions éligibles*. Les protocoles décrivent les types d’interventions qui entrent dans les limites de l’examen, y compris la nature de l’intervention (programmes basés sur des manuels, par exemple) ; l’environnement dans lesquels l’intervention est mise en place (par exemple, des cours ordinaires ou en complément de la journée scolaire) ; la durée minimum d’implémentation de l’intervention ; et si l'intervention est un produit issu du commerce.
- *Recherche éligible*. Les protocoles définissent la portée des recherches pouvant être incluses dans l’examen en fonction de caractéristiques telles que la date de publication, la langue et la localisation géographique de l’étude.
- *Résultats éligibles*. Les protocoles décrivent un ensemble de domaines dont les résultats présentent un intérêt pour l’examen (par exemple, les résultats en mathématiques ou sur les comportements scolaires posant problème). En fonction de la mesure des résultats, le protocole peut spécifier des normes plus élevées que celles requises dans le *Manuel des normes WWC* pour mener l’examen (par exemple concernant la fiabilité et la validité).

• *Normes de preuve*. Le WWC utilise les mêmes normes pour examiner toutes les études éligibles, comme indiqué dans le *Manuel des normes WWC*. Toutefois, au sein de ces normes, certains paramètres varient d’un examen à l’autre et doivent être spécifiés dans le protocole. Cela inclue le choix de la limite séparant les niveaux d'attrition d'échantillon acceptables et inacceptables, les mesures sur lesquelles les études doivent démontrer l'équivalence initiale (avant l’expérience), les propriétés psychométriques de la variable d’assignation dans les études de la régression par discontinuité et certains paramètres liés aux études dont l’unité d'affectation est le groupe – ce sont des études qui affectent des groupes (telles que des classes ou des établissements) à certaines conditions expérimentales plutôt que des individus (comme des élèves).

Chacun des éléments spécifiés doit être appliqué systématiquement à toutes les études entrant dans le champ d'application du protocole.

# III. IDENTIFIER LA LITTÉRATURE PERTINENTE

Après l'élaboration d'un protocole d'examen et la définition d'un sujet prioritaire pour l'examen systématique (voir l'annexe A), l'étape suivante du processus d'examen systématique consiste à effectuer une recherche systématique et exhaustive de la littérature pertinente. Une recherche documentaire est systématique lorsqu'elle utilise des termes et des processus de recherche bien spécifiés pour identifier les études pouvant être pertinentes. Elle est exhaustive lorsqu'un vaste éventail de bases de données, de sites Web et d'autres sources disponibles font l'objet d'une recherche sur les effets d'une intervention.

Après avoir établi un protocole d'examen pour un examen systématique WWC, les études sont rassemblées au moyen d'une recherche exhaustive de la littérature publiée et non publiée, y compris les soumissions des distributeurs / développeurs d'interventions, des chercheurs et du public par l'intermédiaire du WWC Help Desk. Seules les études rédigées en anglais et accessibles au public (accessibles sur le Web ou par le biais d'une publication, telle qu'une revue) au moment de la recherche documentaire sont éligibles pour l'examen mené par le WWC. De plus, les thèses de maîtrise et de spécialistes en éducation ne sont pas éligibles, alors que les thèses de doctorat peuvent l'être. Le WWC examine également certaines études individuelles en dehors du processus d'examen systématique, comme celles qui ont reçu une attention particulière de la part des médias (voir l'annexe A pour plus de détails).

Le personnel qualifié du WWC utilise les mots-clés définis dans le protocole d'examen pour effectuer une recherche dans un vaste ensemble de bases de données électroniques et de sites Web (voir l'annexe B). Les citations complètes et, le cas échéant, les résumés et les textes intégraux des études identifiées par ces recherches sont catalogués pour l'examen ultérieure de leur éligibilité. De plus, le WWC contacte les développeurs et les distributeurs d'interventions pour identifier d'autres recherches.

Toutes les citations recueillies au cours du processus de recherche font l'objet d'un examen préliminaire afin de déterminer si l'étude répond aux critères établis dans le protocole d'examen. Cette sélection est décrite au chapitre IV.

Le WWC exige également que les comités d'examinateurs en charge des examens identifient les études qui ont déjà été examinées par le WWC, soit qui ont été utilisées pour un autre document ou outil diffusé par le WWC, soit qui ont utilisé une version antérieure des normes WWC. Ces études seront réexaminées en utilisant les critères d'éligibilité du protocole spécifique utilisé et les normes mises à jour, si nécessaire.

# IV. SÉLECTIONNER LES ÉTUDES

Les études rassemblées au cours de la recherche bibliographique sont évaluées par rapport aux paramètres spécifiés dans le protocole d'examen afin d'identifier un ensemble d'études éligibles pour l'examen WWC. La sélection initiale de l'éligibilité est effectuée par un membre du personnel de WWC qui a été certifié comme sélectionneur (*screener*). Les études peuvent être considérées comme *inéligibles à l'examen WWC* pour l'une des raisons suivantes :

• *L'étude n'utilise pas de design[6] éligible*. Un design éligible est un design pour lequel le WWC dispose de normes (parfois pilotes) et qui met en œuvre une analyse primaire (et ne synthétise pas des résultats obtenus par d'autres études) pour examiner l'efficacité d'une intervention.

- *Designs éligibles*. Le WWC inclut les résultats d'essais contrôlés randomisés (ECR), d'études quasi expérimentale (EQE) et d'études de régression par discontinuité (ERD). Le WWC dispose également de normes pilotes pour les études de cas unique (ECU) qui peuvent également être examinées et décrites dans des rapports si cela est spécifié dans le protocole d'examen. Les études utilisant d'autres design ne sont pas éligibles à un examen.
- *Analyse primaire de l'efficacité d'une intervention*. De plus, certaines études ne sont pas des études primaires mesurant l'impact ou l'efficacité d'une intervention. Par exemple, les études sur la qualité de la mise en œuvre d'une intervention, les revues bibliographiques ou les méta-analyses ne peuvent pas être incluses dans une analyse WWC (mais peuvent néanmoins servir de sources supplémentaires pour l'analyse d'une étude éligible).

• *L'étude n'utilise pas un échantillon conforme au protocole*. Les caractéristiques des échantillons d'études éligibles pour l'examen sont répertoriées dans le protocole et peuvent inclure l'âge, le niveau d'étude, le sexe ou le niveau en anglais.

• *L'étude n'entre pas dans le cadre du protocole*. Chaque protocole identifie les caractéristiques des études éligibles dans le cadre de l'examen, et cela peut concerner les mesures de résultats, la période de publication, le terrain de l'étude et les types d'interventions.

- *Mesures des résultats*. Les études éligibles à l'examen doivent inclure au moins un résultat relevant des domaines identifiés par le protocole d'examen.
- *Date de publication*. Lorsque le WWC débute l'examen des études sur un nouveau sujet, une date limite est fixée pour l'inclusion de la recherche. En règle générale, cette limite est fixée à 20 ans avant le début de l'examen du sujet par le WWC. Cette période englobe généralement des recherches qui représentent de manière adéquate l'état actuel du domaine de recherche et évite d'inclure les recherches menées auprès de populations et dans des contextes très différents de ceux d'aujourd'hui.

[6] Le terme design utilisé dans ce texte désigne le type de conception d'une étude

- *Terrain d'étude*. Les protocoles d'examen peuvent limiter les études éligibles à celles réalisées dans certaines zones géographiques (comme les États-Unis), ou dans certains types d'établissements ou de classe.
- *Interventions*. Les protocoles d'examen décrivent les interventions éligibles pour l'examen et toutes les exigences d'éligibilité associées, telles que la réplicabilité (la possibilité pour les interventions d'être reproduites). En plus de satisfaire aux exigences spécifiques du protocole d'examen, pour être éligible, l'intervention doit être conforme à l'objectif de l'examen. Par exemple, chaque *Rapport d'intervention* est axé sur une intervention spécifique et seulement les études qui examinent l'efficacité de cette intervention sont éligibles pour l'examen associé. De plus, un *Rapport d'intervention* se concentre sur une seule intervention, alors qu'un *Guide des pratiques* peut se concentrer sur un plus grand nombre d'interventions. Lors de l'élaboration du protocole d'examen, il n'est pas possible d'anticiper l'objectif spécifique de tous les objectifs des examens à venir dont il pourra servir de guide, de sorte que tous les critères d'éligibilité pertinents ne peuvent pas être spécifiés dans le protocole d'examen. En particulier, si l'examen porte sur une intervention spécifique, mais que dans l'étude analysée, cette intervention est toujours proposée en association avec une deuxième intervention, l'étude ne sera pas éligible pour l'examen. Toutefois, si l'objectif de l'examen n'est pas spécifique à l'une des deux interventions et si les deux interventions sont individuellement éligibles pour être analysées selon le même protocole d'examen, le WWC considérera cette combinaison comme une seule intervention et l'étude sera éligible. Par exemple, si une application informatique fait l'objet d'un examen systématique, mais est toujours associée à des actions du type tutorat dans une étude, celle-ci pourra ne pas être éligible. Toutefois, si l'étude est examinée en dehors d'un examen systématique et si la combinaison de l'utilisation de l'application informatique et du tutorat est une intervention éligible au titre du protocole d'examen, l'étude pourra être éligible.

# V. EXAMINER LES ÉTUDES

## A. Définition d'une étude

L'examen des études éligibles par rapport aux normes WWC constitue le cœur du processus d'examen systématique. La définition de l'unité « étude » est importante, compte tenu de la manière dont le WWC rend compte et synthétise les preuves. Le niveau de preuve dans les *Guides des pratiques* et la synthèse des résultats d'un *Rapport d'intervention* dépendent du nombre d'études dont le design est conforme aux normes WWC. Par exemple, une classement *effets positifs* (*Rapport d'intervention*) nécessite au moins deux études dont le design est conforme aux normes WWC.

Une étude n'est pas nécessairement équivalente à un manuscrit, tel qu'un article de journal, un chapitre de livre ou un rapport. Une seule étude peut être décrite dans plusieurs manuscrits (par exemple, une intervention menée sur cinq années peut donner lieu à des publications intermédiaires). Alternativement, un manuscrit peut inclure plusieurs études (par exemple, de nombreux articles incluent plusieurs expériences distinctes). Dans le cas de plusieurs manuscrits faisant référence à une étude, le WWC sélectionne le premier manuscrit publiant des résultats pertinents et qui sera mentionné comme citation principale dans le document WWC, et répertorie les autres manuscrits décrivant l'étude comme sources connexes.

La question cruciale de la définition d'une étude par rapport à une analyse connexe est de savoir si elle fournit un test différent de l'intervention. Autrement dit, fournit-elle de nouveaux éléments de preuve distincts des éléments de preuve déjà disponibles ? Lorsque les analyses d'une même intervention partagent certaines caractéristiques, on peut craindre qu'elles ne fournissent pas des tests indépendants de l'intervention.

Souvent, la question de savoir s'il existe plus d'une étude découle de la présentation séparée de résultats qui partagent une ou plusieurs caractéristiques. Lorsque deux résultats partagent certaines caractéristiques, le WWC peut les considérer comme faisant partie de la même étude. Ces caractéristiques comprennent :

- **Les sujets de l'échantillon, tels que des enseignants ou des élèves.** Les résultats d'analyses qui incluent tout ou partie des mêmes enseignants ou élèves peuvent être liés.
- **Les procédures de formation de groupes, telles que les méthodes utilisées pour effectuer une affectation aléatoire ou un appariement.** Lorsque les auteurs utilisent des méthodes identiques (ou presque identiques) pour former les groupes et que ces méthodes sont utilisées dans plusieurs analyses, ou lorsqu'une seule procédure a été utilisée pour former les groupes, les résultats peuvent ne pas fournir de tests indépendants de l'intervention.
- **Les procédures de collecte et d'analyse de données.** De même que pour la formation de groupes, lorsque les auteurs utilisent des procédures identiques (ou presque identiques) pour collecter et analyser des données, les résultats peuvent être liés. Partager des procédures de collecte et d'analyse de données signifie collecter les mêmes mesures à partir des mêmes sources de données, les préparer pour une analyse à l'aide des mêmes règles et utiliser les mêmes méthodes analytiques avec les mêmes variables de contrôle.

•**L'équipe de recherche.** Lorsque les manuscrits partagent un ou plusieurs auteurs, les résultats rapportés dans ces manuscrits peuvent être liés.

Le WWC considère que les conclusions sur l'efficacité d'une même intervention ne constituent qu'une seule et même étude que si elles partagent au moins trois de ces quatre caractéristiques (voir des exemples à l'annexe D). En particulier, lorsque deux résultats remplissent ces conditions, ils présentent :
1. *Une similarité ou continuité dans les groupes d'intervention et de comparaison*[7] *utilisés pour produire les résultats.* Ils partagent des sujets de l'échantillon ou utilisent les mêmes procédures de formation de groupe, et
2. *Une similarité ou continuité dans les procédures utilisées pour produire les résultats.* Ils utilisent les mêmes procédures de collecte et d'analyse de données ou partagent les membres de l'équipe de recherche.

Quand on ne sait pas si les résultats répondent aux critères décrits ci-dessus, la direction du comité des examinateurs (méthodologiste en chef et expert, décrite plus en détail à l'annexe C) a le pouvoir discrétionnaire de déterminer en quoi consiste une ou plusieurs études, et la décision est clairement indiquée dans le document WWC qui fait état de l'examen.

## B. Le processus d'examen des études[8]

En 2011, après une évaluation du processus consistant à utiliser deux examinateurs pour examiner chaque étude (voir *Manuel*, version 2.1, p. 11), le WWC a mis en place un processus d'examen optimisé pour les essais contrôlés randomisés (ECR) et les études quasi expérimentales (EQE). Cette section décrit les étapes du processus d'examen de ces études. Une fois que les études éligibles ont été identifiées grâce à une recherche bibliographique exhaustive (décrite à l'annexe B), toutes les études sont examinées selon le processus suivant (également illustré figure V.1).

Chaque étude fait l'objet d'un **premier examen** dont les renseignements sont inscrits dans le guide d'examen de l'étude (*Study Review Guide, SRG*). Le fichier SRG et les instructions afférentes à sa complétion sont accessibles à l'adresse http://ies.ed.gov/ncee/wwc/StudyReviewGuide.aspx.

• Si le premier examinateur pense que l'étude ne respecte pas les normes WWC, un examinateur senior du comité des examinateurs menant l'examen analyse l'étude et détermine si les raisons indiquées par le premier examinateur pour justifier sa position semblent correctes.

- Si l'examinateur senior du comité des examinateurs est d'accord avec le classement du premier évaluateur, le fichier SRG est créé et complété.
- Si l'examinateur senior n'est pas d'accord, l'étude reçoit un deuxième examen.

[7] On parle aussi souvent de groupe de traitement et de groupe de contrôle.
[8] Des modifications notables ont été apportées à ce passage dans la version 4.1 (2019)

• Si le premier examinateur pense que l'étude est conforme aux normes WWC, ou pourrait respecter les normes avec davantage de données fournies par l'auteur de l'étude, celle-ci fait l'objet d'un deuxième examen.

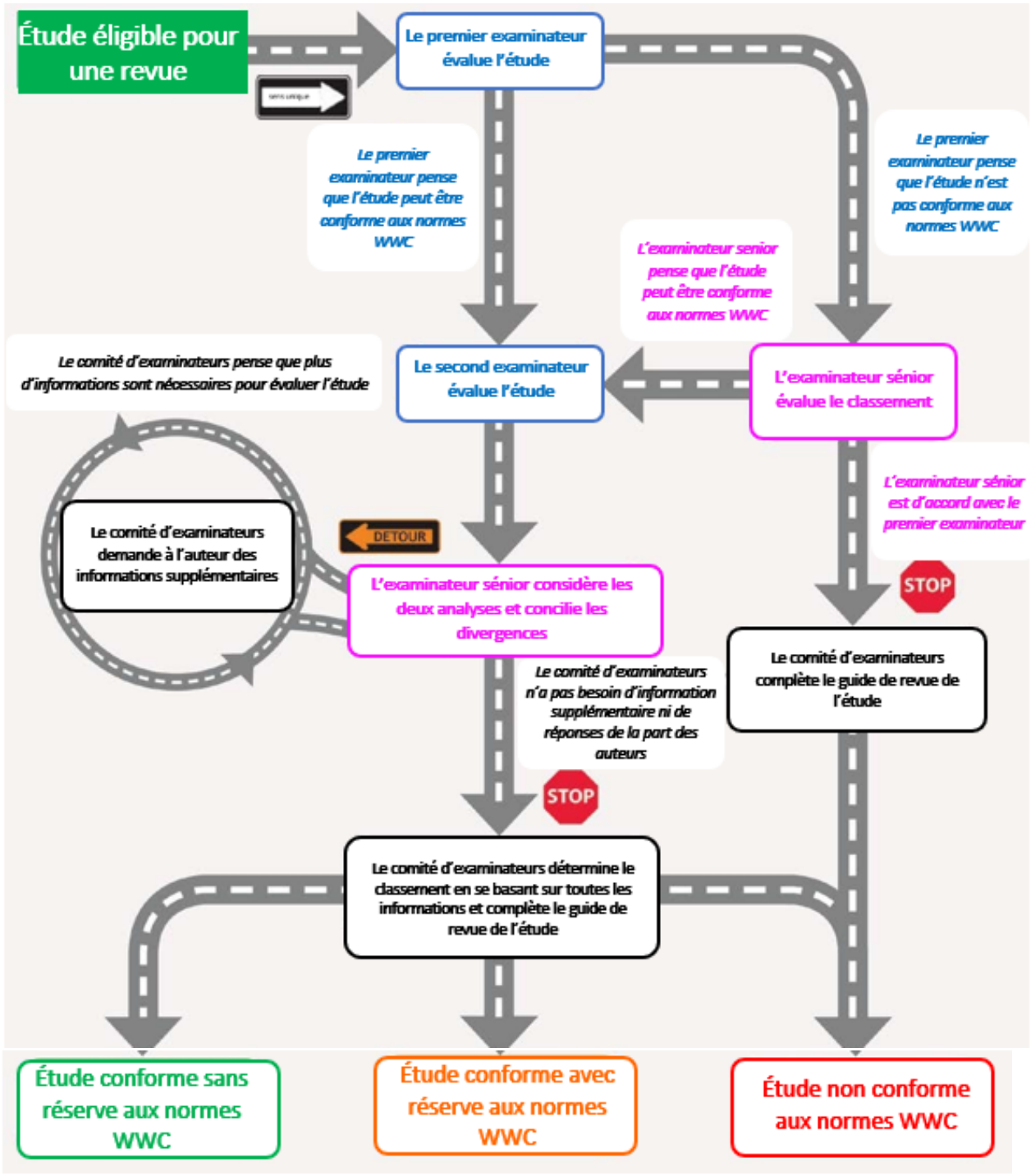
Si une étude fait l'objet d'un **deuxième examen**, celui-ci est mené sans connaissance de l'examen ou du classement précédent, de sorte qu'il ne puisse pas être influencé par les résultats précédents. Une fois le deuxième examen terminé, le coordonnateur demande au deuxième examinateur de comparer son classement à celle du premier (ou de l'examinateur senior, s'il n'a pas vérifié l'examen du premier).

• Si le deuxième examinateur et le premier examinateur (ou le senior) s'accordent sur le classement de l'étude et sur les éléments clés de l'examen, un fichier SRG est créé. Les éléments clés incluent le niveau d'attrition, l'établissement de l'équivalence initiale, les mesures à inclure, et les tailles d'effet. Des divergences mineures, telles que celles concernant la taille des échantillons, peuvent être résolues sans implication de la direction du comité des examinateurs.

• Si les examinateurs ne sont pas d'accord sur le classement final de l'étude, sur les motifs concernant ce classement ou sur d'autres éléments clés de l'examen, les divergences ou incertitudes sont soumises au méthodologiste en chef du comité pour résolution avant la création d'un fichier SRG.

Si nécessaire, le WWC contacte les auteurs de l'étude pour obtenir des informations permettant de compléter le SRG. Cette requête peut porter sur des informations relatives aux caractéristiques de l'échantillon, à la taille de l'échantillon, aux statistiques concernant le niveau initial, aux statistiques concernant les résultats, ou sur d'autres informations comme la formation des groupes, les facteurs de confusion et les mesures concourant à établir les conclusions. Le WWC peut également demander des informations sur les résultats d'analyses référencés dans l'article mais non présentés, y compris des données de synthèse telles que des moyennes, des écarts-types et des corrélations simples entre les mesures, mais le WWC ne demande pas que de nouvelles analyses soient conduites. Toutes les informations reçues via une requête et utilisées dans un rapport sont mises à la disposition du public et documentées dans le rapport final.

**Figure V.1**. Feuille de route pour l'examen des études dont le design est basé sur une comparaison de groupes



# VI. PRÉSENTER LES RÉSULTATS

Dans la mesure du possible, le WWC indique l'amplitude et la signification statistique des estimations de l'efficacité des interventions rapportées par les études, en utilisant des méthodes de calculs communes et en appliquant des corrections (par exemple pour les regroupements et les comparaisons multiples) susceptibles d'affecter les résultats rapportés par les études elles-mêmes. Ensuite, une heuristique est appliquée pour caractériser les résultats des études de manière à intégrer la direction, l'ampleur et la précision statistique des estimations d'effets. Enfin, dans certains de ses documents (*Rapports d'intervention* et *Guides des pratiques* par exemple), le WWC combine les résultats d'études individuelles en une mesure synthétique de l'efficacité, qui comprend les estimations globales des tailles d'effet, les évaluations globales de l'efficacité (classement de l'intervention) et une évaluation du niveau de preuve (étendue des preuves).

## A. Résultat issu d'une analyse individuelle

Le WWC définit le résultat d'une analyse comme étant la mesure de l'effet de l'intervention sur un échantillon, à un moment donné par rapport à l'introduction de l'intervention et en rapport à une condition de comparaison spécifiée.

### 1. Amplitude d'un résultat

Le WWC rend compte de l'amplitude des résultats de l'étude en calculant deux résultats : (a) la taille de l'effet (c.-à-d. la différence des moyennes standardisées) et (b) « l'indice d'amélioration » calculé par le WWC.

*Tailles d'effet*[9]
Pour toutes les études, le WWC enregistre les résultats de l'étude dans les unités déclarées par les auteurs de l'étude. De plus, le WWC calcule et enregistre l'ampleur de l'effet associé aux résultats de l'étude en utilisant les mesures de résultats pertinentes. En règle générale, pour améliorer la comparabilité des estimations de la taille de l'effet d'une étude à l'autre, le WWC utilise les écarts-types au niveau de l'élève pour calculer la taille de l'effet, indépendamment de l'unité d'affectation ou de l'unité d'intervention. Voir l'annexe E pour les calculs de tailles d'effet utilisés dans d'autres situations, telles que celles basées sur des tests $t$ au niveau de l'élève ou sur une affectation au niveau d'un groupe.

Quand les **résultats sont des variables continues**, le WWC a adopté l'indice de taille d'effet le plus couramment utilisé, la différence des moyennes standardisée connue sous le nom de $g$ de Hedges, avec un ajustement pour les petits échantillons. Cet indice est défini comme la différence entre le résultat moyen du groupe d'intervention et le résultat moyen du groupe de comparaison, divisé par l'écart-type groupé de la mesure du résultat. Si on définit pour les élèves des groupes d'intervention (i) et de comparaison (c) $y_i$ et $y_c$ comme les moyennes des résultats, $n_i$ et $n_c$ comme les tailles d'échantillon des élèves, $s_i$ et $s_c$ comme les écarts-

[9] Pour plus de détails sur ces calculs, se reporter au livre *Comparaison de deux groupes de données (Nathalie ROQUES, 2019).*

types au niveau de l'élève, et $w$ comme la correction pour les petites tailles d'échantillons, la taille de l'effet est donnée par (voir annexe E) :

$$g = \frac{w(y_i - y_c)}{\sqrt{\frac{(n_i - 1)s_i^2 + (n_c - 1)s_c^2}{n_i + n_c - 2}}}$$

avec $$w = [1 - 3/(4N - 9)]$$

Quand les **résultats sont des variables dichotomiques**, la différence des moyennes des groupes est calculée comme la différence de probabilité d'occurrence d'un événement. La mesure de la taille de l'effet choisie pour les résultats dichotomiques est l'indice de Cox, qui donne des valeurs de taille d'effet similaires aux valeurs du $g$ de Hedges que l'on obtiendrait si les moyennes des groupes, les écarts-types et les tailles d'échantillon étaient disponibles, en supposant que la mesure des résultats dichotomiques soit basée sur une distribution normale sous-jacente.

…./…

Le WWC suit également ces directives supplémentaires lors du calcul de la taille de l'effet:

• Si une étude rapporte à la fois des moyennes post-intervention non ajustées et ajustées, le WWC enregistre les moyennes ajustées et les écarts-types non ajustés et les utilise dans le calcul de la taille de l'effet.
• Lorsque seules les moyennes de groupe non ajustées sont publiées et que les informations sur la corrélation entre les tests ne sont pas disponibles, le WWC calcule la taille de l'effet avec la différence entre les deux groupes lors du prétest et la taille de l'effet avec la différence entre les deux groupes lors du posttest séparément, la taille de l'effet finale étant donnée par la différence de ces deux tailles d'effet. Le WWC considère que cet ajustement *post-hoc*[10] est un ajustement statistique acceptable pour les différences de base si le prétest et le post-test sont suffisamment liés, conformément aux exigences décrites dans le chapitre II.A du *Manuel des normes WWC*.
• Lorsque le WWC opère un ajustement et utilise la différence des différences à partir des résultats fournis par l'auteur de l'étude, il rapporte les niveaux de signification statistique des différences ajustées qui reflètent l'ajustement de la taille de l'effet. Par exemple, considérons une différence de 0,2 au prétest mesurant la réussite des élèves avant l'intervention. Si la différence après l'intervention était de 0,3, l'effet ajusté de la différence des différences serait de 0,1. Par la suite, la signification statistique rapportée par le WWC serait basée sur le résultat ajusté de 0,1 et non sur le résultat non ajusté de 0,3.
• Lorsque les tailles d'effet déclarées par l'auteur et celles calculées par le WWC sont différentes, le WWC tente d'identifier la source de la différence et explique la raison de cette divergence dans une note. En règle générale, lorsque cela se produit, le WWC communique la taille de l'effet calculée par le WWC car son calcul peut être vérifié, et l'utilisation des résultats calculés par le WWC facilite la comparabilité des résultats et des études. Cependant, le WWC signalera une taille d'effet proposée par l'auteur qui serait comparable au $g$ de Hedges si elle tient compte des différences initiales et que la taille de l'effet calculée

[10] Ajustement effectué a posteriori et ne reposant pas sur une hypothèse expérimentale préalablement définie.

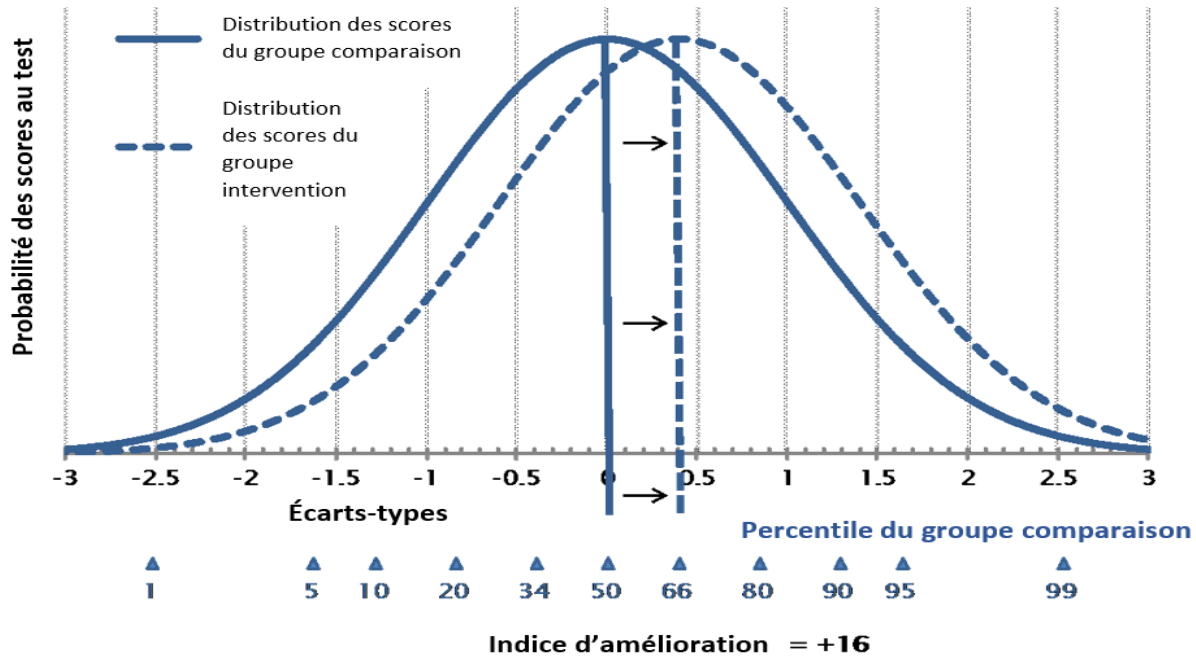
par le WWC ne le fait pas, ou bien si elle est basée sur l'ajustement post-hoc décrit ci-dessus. Pour le WWC, des tailles d'effet de 0,25 écart-type ou plus sont considérées comme étant **importante**. Des tailles d'effet au moins aussi importantes sont interprétées comme un effet positif (ou négatif) qualifié, même si elles peuvent ne pas atteindre la signification statistique dans une étude donnée[11].

*Indice d'amélioration*

Afin d'aider les lecteurs à juger de l'importance pratique de l'effet d'une intervention, le WWC traduit la taille de l'effet par un « indice d'amélioration ». L'indice d'amélioration pour une étude individuelle représente la différence entre le rang centile[12] correspondant à la moyenne du résultat pour le groupe d'intervention et le rang centile correspondant à la moyenne du résultat pour le groupe de comparaison (des détails sur le calcul de l'indice d'amélioration sont présentés à l'annexe E). L'indice d'amélioration peut être interprété comme le changement attendu du rang centile pour un groupe d'élèves du groupe de comparaison se situant dans la moyenne si celui-ci avait bénéficié de l'intervention.

La figure VI.1 illustre l'interprétation de l'indice d'amélioration. Dans cet exemple, l'effet moyen estimé de l'intervention est une amélioration de 0,4 écart-type des résultats aux tests de lecture. Ainsi, en moyenne, un élève du groupe témoin qui se situe au 50ème centile de l'échantillon de l'étude aurait obtenu un score supérieur de 0,4 écart-type s'il avait reçu l'intervention, autrement dit il se serait situé au 66 ème centile. L'indice d'amélioration obtenu est de +16, ce qui correspond à l'élévation des performances de l'élève moyen qui passerait du 50 ème au 66 ème centile de la distribution du groupe témoin. Pour plus de détails, voir l'annexe E.

**Figure VI.1**. Calcul de l'indice d'amélioration




[11] Dans la version 4.1 (2019) cette caractérisation disparait et seule la signification statistique du résultat est prise en compte.

[12] La proportion d'élève dont le score est inférieur à une valeur donnée.

## 2. Importance statistique d'un résultat

Pour évaluer correctement les effets d'une intervention, il est important de connaître la signification statistique des estimations des effets en plus de la différence moyenne, de la taille de l'effet ou de l'indice d'amélioration, comme décrit ci-dessus. Pour le WWC, une estimation statistiquement significative d'un effet correspond à une estimation pour laquelle la probabilité d'observer un effet au moins aussi grand que l'effet mesuré dans le cas où l'intervention n'a aucun impact est inférieure à 1 sur 20 (en utilisant un test $t$ avec $p = 0{,}05$), en supposant qu'il existe une seule mesure ou effet moyen dans chaque domaine. Le WWC accepte généralement les niveaux de signification statistique rapportés par le ou les auteurs de l'étude. Cependant, le WWC calculera les niveaux de signification statistique si l'étude ne comprend pas d'estimations de la signification statistique. Quand les résultats sont des variables continues, le WWC calcule leur signification statistique en calculant la statistique $t$ :

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

où $g$ est la taille de l'effet, et $n_i$ et $n_c$ les tailles moyennes des échantillons des groupes d'intervention et de comparaison, respectivement, pour un ensemble de résultats. Le deuxième terme du radical est une correction pour les échantillons de petite taille (Hedges et Olkin, 1985).

…/…

De plus, le WWC apportera des ajustements aux niveaux des signification statistique publiés dans l'étude si elles ne tiennent pas compte des regroupements en cas de divergence entre l'unité d'affectation et l'unité d'analyse. Ces estimations calculées ou recalculées par le WWC apparaissent dans les documents et outils publiés par le WWC avec une note décrivant la source des calculs et notant toute différence entre les résultats WWC et ceux rapportés par l'auteur.

*Correction tenant compte des regroupements*

Un problème de « désadéquation » se produit lorsque l'affectation est effectuée au niveau d'un groupe (par exemple, au niveau de la classe ou de l'établissement), tandis que l'analyse est réalisée au niveau individuel (au niveau de l'élève, par exemple). Ignorer la corrélation entre les résultats des individus appartenant aux mêmes groupes lors du calcul des erreurs types des estimations de l'effet, conduit à sous-estimer les erreurs types et à surestimer la significativité statistique des résultats. Les estimations ponctuelles des effets de l'intervention (comme la taille de l'effet et l'indice d'amélioration) obtenues à partir d'analyses de ce type ne sont, elles, pas affectées par cette situation particulière.

Pour évaluer la significativité statistique des effets d'une intervention dans les cas où les auteurs de l'étude n'ont pas effectués de correction liée au regroupement, le WWC calcule des estimations de la significativité statistique corrigées pour le regroupement sur la base des recommandations de Hedges (2007) (voir annexe F).

# B. Résultats issus de plusieurs analyses

Les études présentent souvent plusieurs résultats issus d'analyses qui font varier la condition de comparaison, la mesure du résultat, l'échantillon ou le moment de la mesure. Par exemple, les analyses peuvent inclure tous les participants à l'étude ou des sous-ensembles de cette population. De même, les analyses peuvent utiliser plusieurs mesures de résultats dans un même domaine, un même résultat mesuré à plusieurs moments dans le temps, ainsi qu'une mesure composite et ses composants.

Pour une étude comportant plusieurs analyses, toutes les analyses éligibles définies par le protocole sont examinées (comme décrit au chapitre II, les protocoles définissent les interventions, les populations et les résultats éligibles). Des requêtes sont adressées aux auteurs au besoin pour évaluer toutes les analyses éligibles présentées dans l'étude. Le classement de l'étude est alors le classement le plus élevé obtenu pour toutes les analyses éligibles.

## 1. Présentation des résultats issus de plusieurs analyses

Le WWC rend compte des résultats de toutes les analyses éligibles (telles que définies dans le protocole d'examen applicable) qui répondent aux normes WWC. Ces résultats sont divisés en résultats principaux et résultats complémentaires. Le classement des résultats de l'étude (décrit à la section C) repose sur l'analyse des résultats principaux. Le WWC utilise les critères suivants pour désigner, parmi tous les résultats répondant aux normes WWC, un résultat ou un ensemble de résultats comme étant le résultat principal : (1) ce résultat est basé sur l'échantillon complet ; (2) il utilise la mesure la plus agrégée (plutôt que des sous-échelles individuelles) ; et (3) il est mesuré à moment spécifié par le protocole (par exemple, la dernière période de suivi, la première période de suivi après la fin de l'intervention ou après un an d'exposition).

Les règles suivantes guident cette distinction faite parmi les résultats des analyses éligibles et conformes aux normes WWC, comme l'illustre l'exemple d'une étude portant sur deux cohortes d'élèves de quatrième. Dans cet exemple, l'étude inclut les analyses éligibles d'un échantillon groupé d'élèves ainsi que celles de chaque cohorte analysée séparément. La distinction qui sera faite entre les analyses qui donneront lieu aux résultats principaux et celles qui fourniront des résultats complémentaires dépend de la manière dont les analyses se conforment aux normes WWC.

• **Toutes les analyses éligibles répondent aux normes**. L'analyse groupée est présentée comme le résultat principal, alors que les autres analyses (les cohortes séparées) sont présentées comme des résultats complémentaires.
• **L'analyse groupée est conforme aux normes et l'une des analyses spécifiques à une cohorte est conforme aux normes**. L'analyse groupée est présentée comme donnant le résultat principal, tandis que seule l'analyse (associée à l'une des deux cohortes) qui respecte les normes est présentée comme fournissant un résultat complémentaire.

• **L'analyse groupée est conforme aux normes et aucune des analyses spécifiques à une cohorte n'est conforme aux normes**. L'analyse groupée est présentée comme donnant le résultat principal, sans résultat complémentaire.
• **L'analyse groupée ne répond pas aux normes, mais les deux analyses spécifiques à la cohorte respectent les normes.** Étant donné que les analyses spécifiques à une cohorte respectent les normes et combinent l'ensemble de l'échantillon, le WWC crée un échantillon groupé à partir des cohortes, qui donne lieu au résultat principal à l'aide des formules présentées ci-dessous à la section 2. Les résultats des analyses des cohortes considérées individuellement sont présentés en tant que résultats complémentaires. Toutefois, si les seuls résultats conformes aux normes de cet exemple étaient des résultats relatifs à des sous-échelles distinctes d'une mesure composite, les deux étant basées sur l'ensemble de l'échantillon, le WWC consignerait les résultats de chaque sous-échelle séparément comme résultats principaux (accompagnés de la moyenne non pondérée agrégeant les résultats, également décrite dans la section 2).
• **L'analyse groupée ne répond pas aux normes et une seule des analyses spécifiques à une cohorte est conforme aux normes**. Comme il n'existe pas d'ensemble d'analyses répondant aux normes couvrant l'ensemble de l'échantillon, l'analyse de la cohorte respectant les normes est présentée comme un résultat principal, sans résultat complémentaires. Toutefois, les examinateurs doivent également évaluer si le résultat calculé par le WWC basé sur la mise en commun des deux cohortes peut respecter les normes WWC, et proposer ce résultat groupé comme résultat principal le cas échéant[13]. Si le seul résultat respectant les normes dans cet exemple était plutôt une sous-échelle distincte d'une mesure composite, le WWC rendrait compte du résultat relatif à la sous-échelle qui respecte les normes WWC.

…/…

Voir l'annexe G pour connaître les procédures de rédaction des résultats des études faisant état des résultats des analyses en intention de traiter (ITT) ou calculant l'effet moyen du traitement sur les sujets traités (*complier average causal effects - CACE)*.

Enfin, le WWC procède à des ajustements en cas de comparaisons multiples (voir ci-dessous) mais cela ne concerne que les résultats principaux (les résultats complémentaires ne sont pas concernés).

## 2. Amplitude des résultats

Le WWC combine les résultats dans trois situations : plusieurs sous-échantillons sont concernés par une même mesure d'un résultat dans une étude, plusieurs mesures de résultats sont proposées dans une même étude et plusieurs études sont combinées.

[13] Les résultats calculés par le WWC peuvent respecter les normes WWC même lorsque les résultats rapportés par l'auteur issus de l'analyse groupée et de l'une des cohortes ne le font pas. Par exemple, l'analyse rapportée par l'auteur peut inclure une covariable endogène, alors que les résultats utilisés pour former les résultats regroupés calculés par le WWC ne tiennent pas compte de cette covariable endogène. En outre, l'attrition des résultats regroupés calculés par le WWC peut être faible, tandis qu'un seul des résultats spécifiques à une cohorte déclarés par l'auteur présente un taux d'attrition faible.

Certaines études présentent les résultats séparément pour plusieurs sous-échantillons de sujets sans présenter de résultat global (ou le résultat global peut ne pas correspondre aux normes WWC). On peut citer par exemple une étude sur les mathématiques au collège qui présente les effets séparément pour les élèves de sixième, cinquième et quatrième ; une étude sur l'alphabétisation des adolescents qui examine les élèves à risque élevé et faible ; et une étude d'initiation à la lecture qui considère les élèves ayant un niveau de compétence faible, moyen et élevé. Lorsque l'étude présente les résultats séparément pour des parties de l'échantillon sans présenter un résultat global, le WWC demande aux auteurs s'ils ont effectué une analyse sur l'échantillon complet (bien que le comité des examinateurs ait le pouvoir discrétionnaire de ne pas faire cette requête lorsqu'il y a lieu de croire qu'un échantillon complet n'a pas été analysé). L'analyse de l'étude est préférable car elle peut être plus précise que le calcul effectué par le WWC. Si le WWC ne parvient pas à obtenir les résultats globaux de l'auteur ou si le résultat global ne répond pas aux normes WWC, le WWC calcule la moyenne **de la mesure d'un même résultat pour les sous-échantillons à l'intérieur d'une étude**.

Plus concrètement, si une étude fournit des résultats pour $G$ sous-échantillons mutuellement exclusifs qui constituent l'ensemble de l'échantillon, mais pas de résultat global, le WWC calcule un résultat global. Pour les résultats continus, en définissant $n_{gj}$, $m_{gj}$ et $s_{gj}$ comme étant la taille, la moyenne des résultats et l'écart-type du sous-échantillon $g$ dans le groupe $j$[14], l'estimation $M_j$ de la moyenne du groupe $j$ et l'écart-type $S_j$ du groupe $j$ pour tous les sous-échantillons sont donnés par

$$M_j = \frac{\sum_{g=1}^{G} n_{gj} m_{gj}}{\sum_{g=1}^{G} n_{gj}}$$

$$S_j = \sqrt{\frac{\sum_{g=1}^{G}\left[(n_{gj}-1)s_{gj}^2 + n_{gj}(M_j - m_{gj})^2\right]}{\sum_{g=1}^{G} n_{gj} - 1}}$$

La taille de l'effet $g$ est alors donnée par

$$g = \frac{w(M_i - M_c)}{\sqrt{\frac{\left(\sum_{g=1}^{G} n_{gi} - 1\right)S_i^2 + \left(\sum_{g=1}^{G} n_{gc} - 1\right)S_c^2}{\sum_{g=1}^{G}(n_{gi} + n_{gc} - 2)}}}$$

avec[15] $w = [1 - 3/(4N - 9)]$

…/…

Si une étude présente des résultats conformes aux normes WWC pour plus d'une mesure de résultats dans un domaine, les tailles d'effet de tous les résultats de cette étude sont

[14] Groupe de comparaison ou groupe d'intervention, donc $j = i$ ou $j = c$

[15] Facteur correctif pour échantillon de petite taille, avec $N$ la taille totale de l'échantillon.

combinées en une **taille d'effet moyenne de l'étude** à l'aide de la moyenne non pondérée des tailles d'effets individuelles. **L'indice d'amélioration moyen de l'étude** est calculé directement à partir de la taille de l'effet moyenne de l'étude.

Pour les examens systématiques qui incluent plus d'une étude, si plusieurs études ont des résultats dans un domaine, les tailles d'effet moyennes de chaque étude sont combinées en une **taille d'effet moyenne du domaine** en utilisant la moyenne non pondérée des tailles d'effet de ces études[16]. **L'indice d'amélioration moyen du domaine** est calculé directement à partir de la taille de l'effet moyenne du domaine.

## 3. Signification statistique des résultats issus de plusieurs analyses

En tant que deuxième élément de synthèse des résultats de plusieurs analyses, le WWC évalue la signification statistique à l'aide des mêmes formules données à la section A pour le calcul de la statistique $t$. Pour les tailles d'effet moyennes de l'étude basées sur des mesures de résultats continues, $g$ est la taille de l'effet moyenne pour les résultats, et $n_i$et $n_c$ sont les tailles d'échantillon moyennes des groupes d'intervention et de comparaison respectivement, pour un ensemble de résultats.
…/…
Pour les résultats agrégés des mesures des résultats pour l'ensemble des sous-échantillons, le WWC calcule la statistique $t$ de la façon suivante pour les mesures continues

$$t = \frac{g}{\sqrt{\frac{N_i + N_c}{N_i N_c} + \frac{g^2}{2(N_i + N_c)}}}$$

où $g$ est la taille de l'effet basée sur $M_j$ et $S_j$ tels que définis ci-dessus, et $N_i$ et $N_c$ sont les tailles d'échantillon totales comprenant l'ensemble des sous-échantillons pour le groupe d'intervention et le groupe de comparaison, respectivement.

*Correction de Benjamini-Hochberg pour les comparaisons multiples*
Le WWC a adopté la correction de Benjamini-Hochberg (BH) pour prendre en compte les comparaisons multiples ou la « multiplicité » des résultats qui peuvent conduire à des estimations surestimées de la signification statistique des résultats (Benjamini & Hochberg, 1995). Des tests répétés sur des résultats fortement corrélés augmenteront la probabilité de conclure par erreur que les différences de moyennes des résultats d'intérêt entre les groupes d'intervention et de comparaison sont significativement différentes de zéro (erreur de type I dans les tests d'hypothèses). Ainsi, le WWC utilise la correction BH pour réduire la possibilité de commettre ce type d'erreur.

Si les $p$-valeurs exactes ne sont pas disponibles mais que les tailles d'effet le sont, le WWC convertit la taille de l'effet en statistiques $t$, puis obtient les $p$-valeurs correspondantes. Pour les résultats fondés sur des analyses où l'unité d'analyse est aussi l'unité d'affectation ou bien pour lesquelles les auteurs de l'étude ont effectué leur analyse de telle manière que leurs $p$-valeurs soient ajustées pour tenir compte de la divergence entre le niveau

[16] Des modifications notables ont été apportées à ce passage dans la version 4.1 (2019)

d'affectation et celui de l'analyse, les $p$-valeurs rapportées par les auteurs de l'étude sont utilisées pour la correction BH. Pour les résultats basés sur des analyses qui n'ont pas généré de $p$-valeurs tenant compte d'une telle divergence, le WWC utilise les $p$-valeurs corrigées pour la correction BH. Pour plus de détails, voir l'annexe F.

## C. Synthèses des résultats

Les documents publiés par le WWC, y compris les *Guides des pratiques* et les *Rapports d'intervention*, présentent des synthèses des données probantes pour les études individuelles et pour des ensembles d'études après leur analyse par des examens systématiques. Ces synthèses indiquent la direction (par exemple positive, négative ou indéterminée) et la force (par exemple, l'ampleur et la signification statistique) des résultats. Les synthèses sont basés sur des résultats conformes aux normes WWC et considérés par celui-ci comme étant les résultats principaux de l'étude.

### 1. Synthèse des preuves pour une étude individuelle[17]

En utilisant la taille de l'effet estimée et le niveau de signification statistique (en tenant compte du regroupement et des comparaisons multiples si nécessaire), le WWC classe les résultats de l'étude de chaque domaine dans l'une des cinq catégories suivantes : (a) effet positif (favorable) statistiquement significatif (b) effet positif important (c) effet indéterminé, (d) effet négatif (défavorable) important et (e) effet négatif statistiquement significatif. Pour les conclusions basées sur une seule mesure de résultat, les règles du tableau IV.1 sont utilisées pour déterminer la catégorie dans laquelle il convient de classer l'effet.

**Tableau IV.1.** Classement d'un effet basé sur l'unique mesure d'un résultat dans un domaine

| | |
|---|---|
| Effet positif statistiquement significatif | L'effet estimé est positif et statistiquement significatif (avec correction pour regroupement si nécessaire) |
| Effet positif important | L'effet estimé est positif et non statistiquement significatif mais important |
| Effet indéterminé | L'effet estimé n'est ni statistiquement significatif ni important |
| Effet négatif important | L'effet estimé est négatif et non statistiquement significatif mais important |
| Effet négatif statistiquement significatif | L'effet estimé est positif et statistiquement significatif (avec correction pour regroupement si nécessaire) |

Note : Une estimation statistiquement significative d'un effet est une estimation pour laquelle la probabilité d'observer un effet au moins aussi grand que l'effet mesuré sous l'hypothèse que l'intervention n'a eu aucun effet est inférieure à un sur 20 (en utilisant un test bilatéral, avec p = 0,05). Une analyse correctement menée est une analyse pour laquelle l'unité d'affectation et l'unité d'analyse sont identiques ou qui prend en compte la corrélation entre les résultats obtenus par des individus appartenant aux mêmes groupes. Une taille d'effet égale ou supérieure à 0,25 écart-type est considérée comme étant importante.

Si l'effet est basé sur plusieurs mesures de résultats dans un domaine, les règles du tableau IV.2 s'appliquent.

[17] Des modifications notables ont été apportées à ce passage dans la version 4.1 (2019)

**Tableau IV.2**. Classement d'un effet basé sur plusieurs mesures d'un résultat dans un domaine

| | |
|---|---|
| Effet positif statistiquement significatif | Lorsque l'un des éléments suivants est vrai :<br>1. Au moins un effet est positif et statistiquement significatif, et aucun n'est négatif et statistiquement significatif sur la base de tests statistiques univariés, rendant compte de multiples comparaisons (et apportant les corrections nécessaires en cas de regroupement).<br>2. La taille de l'effet moyenne des mesures de l'étude calculée par WWC est positive et statistiquement significative (en ayant apporté les corrections nécessaires en cas de regroupement).<br>3. L'étude indique que l'effet global pour l'ensemble des mesures est positif et statistiquement significatif sur la base d'un test statistique multivarié dans une analyse correctement menée. |
| Effet positif important | La taille de l'effet moyenne calculée par WWC est positive et non statistiquement significative mais important |
| Effet indéterminé | La taille de l'effet moyenne calculée par WWC n'est ni statistiquement significative ni importante |
| Effet négatif important | La taille de l'effet moyenne calculée par WWC est négative et non statistiquement significative mais importante |
| Effet négatif statistiquement significatif | Lorsque l'un des éléments suivants est vrai :<br>1. Au moins un effet est négatif et statistiquement significatif, et aucun n'est positif et statistiquement significatif sur la base de tests statistiques univariés, rendant compte de multiples comparaisons (et apportant les corrections nécessaires en cas de regroupement).<br>2. La taille de l'effet moyenne des mesures de l'étude calculée par WWC est négative et statistiquement significative (en ayant apporté les corrections nécessaires en cas de regroupement).<br>3. L'étude indique que l'effet global pour l'ensemble des mesures est négatif et statistiquement significatif sur la base d'un test statistique multivarié dans une analyse correctement menée. |

*Note* : Une estimation statistiquement significative d'un effet est une estimation pour laquelle la probabilité d'observer un effet au moins aussi grand que l'effet mesuré sous l'hypothèse que l'intervention n'a eu aucun effet est inférieure à un sur 20 (en utilisant un test bilatéral, avec p = 0,05). Une analyse correctement menée est une analyse pour laquelle l'unité d'affectation et l'unité d'analyse sont identiques ou qui prend en compte la corrélation entre les résultats obtenus par des individus appartenant aux mêmes groupes. Une taille d'effet égale ou supérieure à 0,25 écart-type est considérée comme étant importante.

Étant donné qu'elles ne sont pas directement comparables aux tailles d'effets au niveau individuel (par exemple au niveau des élèves), les résultats basés sur l'analyse de données au niveau des groupes, tels que les résultats au niveau d'un établissement, ne peuvent pas être pris en compte pour déterminer les effets importants lors de l'examen des interventions. (Cependant, les moyennes au niveau du groupe peuvent être utilisées pour calculer des tailles d'effet comparables aux tailles d'effet au niveau de l'élève, tant que le calcul utilise un écart-type basé sur des données au niveau individuel.) Par conséquent, dans les *Rapports d'intervention*, les tailles d'effet calculées au niveau des groupes sont exclues du calcul des tailles d'effet moyennes du domaine concerné et du calcul des indices d'amélioration. Cependant, la signification statistique et la direction (positive ou négative) des résultats au niveau du groupe sont prises en compte pour déterminer le classement des effets de l'étude.

En plus de classer les effets des études tel que décrits ci-dessus, le WWC identifie également sur le site Web *Trouver ce qui marche (Find What Works)* les études qui incluent un ou plusieurs résultats statistiquement significatifs et positifs. Cette mention, indiquée sur la page Web d'une étude, repose sur l'examen de tous les résultats de l'étude conformes aux

normes WWC, y compris les résultats complémentaires qui n'ont pas contribué au classement d'une intervention dans un *Rapport d'intervention*. Cette mention est indiquée quand l'étude inclut au moins un résultat principal ou complémentaire répondant aux normes WWC et qui est positif et statistiquement significatif après ajustements pour des regroupements et des comparaisons multiples, si nécessaire. Cette mention est spécifique à chaque examen de l'étude (par exemple, certains résultats peuvent être examinés au titre d'un protocole, mais pas d'un autre, ce qui pourrait affecter la manière dont l'ajustement pour des comparaisons multiples est appliqué).

Identifier les études dont les résultats sont statistiquement significatifs et positifs incite les décideurs à considérer des études qui fournissent des preuves de l'efficacité d'une intervention. Toutefois, cette sélection n'est pas fondée sur un examen systématique des données probantes, ainsi, d'autres études, ou même d'autres conclusions issues de la même étude, pourraient fournir des données contradictoires. En revanche, le classement des effets observés dans une étude et décrit ci-dessus par le WWC fournit une synthèse de toutes les preuves concernant une intervention analysée dans une étude conforme aux normes WWC, et le classement des effets observés à travers plusieurs études décrit ci-dessous fournit une synthèse de toutes les preuves concernant une intervention conforme aux normes suite à un examen systématique.

## 2. Synthèse des preuves pour un *Rapport d'intervention*[18]

Dans les *Rapports d'intervention*, le WWC attribue un classement aux effets de l'intervention dans chaque domaine de résultats et évalue l'étendue des données probantes en fonction de ce classement.

*Classement d'une intervention*
Le WWC combine les résultats concernant l'efficacité d'une intervention provenant de plusieurs études pour déterminer son classement. Le WWC utilise un ensemble de lignes directrices pour déterminer le classement d'une intervention tout comme il utilise des lignes directrices pour déterminer le classement des résultats d'une étude individuelle (tableau IV.3). Les critères du tableau IV.3 ne s'excluent pas mutuellement. Si plusieurs critères sont remplis, l'intervention se voit attribuer la note d'intervention la plus élevée à laquelle elle est éligible.

**Tableau IV.3**. Critères utilisés pour déterminer le classement attribué par le WWC à une intervention

| | |
|---|---|
| Effets positifs : preuve forte d'un effet positif sans preuve contraire | • Au moins deux études montrent des effets positifs statistiquement significatifs, dont au moins une est *conforme sans réserve aux normes WWC*, ET<br>• Aucune étude n'a montré d'effet négatif statistiquement significatif ou important. |
| Effets potentiellement positifs : preuve d'un effet positif sans preuve contraire majeure. | • Au moins une étude montre des effets positifs statistiquement significatifs ou importants, ET<br>• Moins ou le même nombre d'études montrent des effets indéterminés, ET<br>• Aucune étude n'a montré d'effet négatif statistiquement significatif ou important. |

[18] Des modifications notables ont été apportées à ce passage dans la version 4.1 (2019)

| | |
|---|---|
| Aucun effet discernable : aucune preuve affirmative d'effet | • Aucune des études n'a montré d'effet statistiquement significatif ou important, qu'il soit positif ou négatif. |
| Effets mixtes : preuve d'effets incohérents. | SOIT les deux points suivants sont vérifiés :<br>• Au moins une étude montre des effets positifs statistiquement significatifs ou importants, ET<br>• Au moins une étude montre des effets négatifs statistiquement significatifs ou importants, MAIS pas plus d'études de ce type que du type montrant des effets positifs statistiquement significatifs ou importants.<br>OU les deux suivants :<br>• Au moins une étude montre des effets statistiquement significatifs ou importants, ET<br>• Un nombre supérieur d'études montrent un effet indéterminé. |
| Effets potentiellement négatifs : preuve d'un effet négatif sans preuve contraire majeure. | SOIT les deux points suivants sont vérifiés :<br>● Une étude montre des effets négatifs statistiquement significatifs ou importants, ET<br>● Aucune étude n'a montré d'effet positif statistiquement significatif ou important.<br>OU les deux suivants :<br>• Deux ou plusieurs études montrent des effets négatifs statistiquement significatifs ou importants, au moins une étude montre des effets positifs statistiquement significatifs ou importants ET<br>• Plus d'études montrent des effets négatifs statistiquement significatifs ou importants que des effets positifs statistiquement significatifs ou importants. |
| Effets négatifs : preuves fortes d'un effet négatif sans preuve contraire. | • Deux ou plusieurs études montrent des effets négatifs statistiquement significatifs, dont au moins une est *conforme sans réserve aux normes WWC*, ET<br>• Aucune étude n'a montré d'effet positif statistiquement significatif ou important. |

*Note* : Une estimation statistiquement significative d'un effet est une estimation pour laquelle la probabilité d'observer un effet au moins aussi grand que l'effet mesuré sous l'hypothèse que l'intervention n'a eu aucun effet est inférieure à un sur 20 (en utilisant un test bilatéral, avec p = 0,05). Une taille d'effet égale ou supérieure à 0,25 écart-type est considérée comme étant importante. Un effet indéterminé est un effet tel que l'effet ou l'effet moyen n'est ni statistiquement significatif ni important.

*Étendue des preuves*

La dernière étape pour combiner les conclusions sur l'efficacité d'une intervention en provenance de plusieurs études est de rapporter l'étendue des preuves avec laquelle le classement de l'intervention a été déterminé. Les catégories de l'étendue des preuves ont été déterminées pour informer les lecteurs du nombre de preuves utilisées pour déterminer le classement de l'intervention, en combinant le nombre et la taille des études. Ce système se divise en deux catégories : a) moyen à important et b) faible (tableau IV.4).

**Tableau IV.4.** Critères utilisés pour déterminer l'étendue des preuves pour une intervention

| | |
|---|---|
| Moyenne à important | • Le domaine comprend plus d'une étude ET<br>• Le domaine comprend plusieurs terrains ET<br>• Les résultats du domaine sont basés sur un échantillon total d'au moins 350 élèves, OU d'au moins 14 classes (en considérant qu'il y a 25 élèves par classe) sur l'ensemble des études. |
| Faible | • Le domaine comprend une seule étude OU<br>• Le domaine comprend un seul terrain OU<br>• Les résultats du domaine sont basés sur un échantillon total de moins de 350 élèves ET sur moins de 14 classes (en considérant qu'il y a 25 élèves par classe). |

Avec une seule étude, il est possible que certaines caractéristiques de l'étude, telles que les mesures de résultats ou le moment choisi pour l'intervention, aient pu affecter les résultats. Plusieurs études réduisent le biais potentiel dû à une erreur d'échantillonnage. Par conséquent, le WWC considère que l'ampleur des preuves est limitée lorsque les conclusions ne reposent que sur une seule étude.

De même, avec l'étude d'un seul terrain (par exemple un établissement), il est possible que certaines caractéristiques de ce terrain (par exemple les données démographiques du directeur ou des élèves de l'établissement) aient pu affecter les résultats ou s'imbriquer aux résultats. Par conséquent, le WWC considère que l'ampleur des preuves est faible lorsque les conclusions ne reposent que sur l'étude d'un seul terrain.
La taille de l'échantillon de 350 a été choisie car il s'agit généralement de la plus petite taille nécessaire pour disposer d'une puissance statistique adéquate (par exemple, une probabilité de rejeter l'hypothèse nulle de 80% lorsqu'elle est fausse et une probabilité maximale de 5% de conclure à tort qu'il existe un effet), pour détecter les tailles d'effets significatives (par exemple 0,3 écart-type ou plus) pour un simple ECR (par exemple les élèves sont affectés aléatoirement au groupe traitement ou au groupe contrôle dans des proportions égales) sans utiliser de covariables dans l'analyse.

## 3. Synthèse des preuves pour un *Guide des pratiques*

En combinant les preuves pour chaque recommandation, le comité d'experts et les examinateurs du WWC prennent en compte les éléments suivants :

• le nombre d'études
• la qualité des études
• si les études représentent la gamme de participants, les terrains et les comparaisons sur lesquels la recommandation est basée
• si les résultats des études peuvent être attribués à la pratique recommandée
• si les résultats des études sont systématiquement positifs

Les comités d'experts qui rédigent les *Guides de pratiques* s'appuient sur un ensemble de définitions pour déterminer le niveau de preuve à l'appui de leurs recommandations (tableau IV.5).

**Table IV.5** Niveaux de preuve pour les *Guides des pratiques*

| Critères | Base de preuves solide | Base de preuve modérée | Base de preuve minimale |
|---|---|---|---|
| Validité | La recherche a une validité interne et une validité externe élevées basées sur des études qui répondent aux normes WWC. | La recherche a une validité interne élevée et une validité externe modérée ou une validité interne modérée et une validité externe élevée. | La recherche peut inclure des preuves en provenance d'études ne répondant pas aux critères pour preuves modérées ou fortes. |
| Effets sur résultats pertinents | La recherche montre des effets positifs constants sans éléments contradictoire dans des études avec haute validité interne. | La recherche montre une prépondérance de preuves d'effets positifs. Les preuves contradictoires doivent être discutées et examinées en fonction de leur pertinence par rapport à la portée du guide et de l'intensité de la recommandation en tant que composante de l'intervention évaluée. | Les preuves sont de niveau faible ou contradictoires. |
| Pertinence concernant l'objectif | La recherche est en lien direct avec l'objectif : contexte, échantillon, comparaison et résultats évalués. | La pertinence par rapport à l'objectif est variable. Au moins certaines recherches sont directement liées à l'objectif. | La recherche est hors du cadre du *Guide des pratiques*. |

| | | | |
|---|---|---|---|
| Relation entre recherche et recommandations | La recommandation est spécifiquement évaluée dans les études ou bien la recommandation est une composante majeure de l'intervention testée dans les études. | L'importance de la recommandation en tant que composante des interventions évaluées dans les études est variable. | L'importance de la recommandation en tant que composante des interventions évaluées dans les études est faible et / ou la recommandation reflète l'opinion d'un expert basée sur des extrapolations raisonnables à partir de la recherche. |
| Confiance du comité | Le comité a un haut niveau de confiance sur l'efficacité de la pratique | Le comité a déterminé que la recherche atteignait un niveau de preuve situé sous le niveau le plus élevé mais au-dessus du niveau minimal. Il se peut que le comité ne sache pas si la recherche a efficacement contrôlé d'autres explications ou si la pratique serait efficace dans la plupart ou dans tous les contextes. | De l'avis du comité, la recommandation doit être traitée dans le *Guide des pratiques* ; cependant, le comité ne peut pas mettre en évidence un corpus de recherches atteignant le niveau de preuve modéré ou élevé. |
| Opinion de l'expert | Non applicable | Non applicable | Opinion d'expert basée sur une interprétation défendable de la théorie. |
| Quand l'évaluation (des élèves) est au cœur de la recommandation | Les évaluations répondent aux normes de *The Standards for Educational and Psychological Testing* | Les évaluations répondent aux normes de *The Standards for Educational and Psychological Testing* mais les échantillons ne sont pas suffisamment représentatifs de la population. | Non applicable |

# Liste des abréviations

ECR : essai contrôlé randomisé (*randomized controled trials, RCT*)

ECU : étude de cas unique (*single-case design, SCD*)

EQE : étude quasi-expérimentale (*quasi-experimental design, QED*)

ERD : études de régression par discontinuité (*regression discontinuity design, RDD*)

ES : *Effect Size* (taille d'effet)

WWC : *What Works Clearinghouse*