bonnes pages

L'article qui suit de Gary A. Troia (Université de Maryland, Collège Park, Etats-Unis) paru dans *Reading Research Quarterly* (Vol. 34, n°1, 1er trimestre 1999) s'insère dans notre rubrique Bonnes Pages, **composée de textes et de documents d'origines diverses qui, à propos de l'écrit, de la lecture et de l'écriture et de leurs apprentissages, ont un intérêt documentaire et informatif certain parce qu'ils abordent des sujets pas ou peu traités dans notre revue ou encore parce qu'ils exposent une réflexion, une position, un point de vue originaux.**

ANALYSE CRITIQUE DE LA METHODOLOGIE EXPERIMENTALE DES RECHERCHES SUR LA CONSCIENCE PHONOLOGIQUE[1]

Gary A. Troia

Il a été montré que la conscience phonologique, c'est à dire la conscience de et la capacité à manipuler la structure phonologique des mots, détermine largement la réussite en lecture et en orthographe (ex. Adams, 1990 ; Blachman, 1984, 1989 ; Bradley et Bryant, 1978 ; Lundberg, Olofsson et Wall, 1980 ; Mann, 1984, 1993 ; Stanovich, 1986 ; Treiman, 1991 ; Wagner et Torgesen, 1987). Les enfants qui risquent de subir un échec en lecture et ceux identifiés comme dyslexiques réussissent souvent mal aux épreuves de mesure de la conscience phonologique (Clafee, Lindamood et Lindamood, 1973 ; Fox et Routh, 1980 ; Rosner et Simon, 1971 ; Zifcak, 1981). De plus, bien que ces enfants puissent faire de nets progrès en lecture et en écriture s'ils suivent des cours de phonétique, ces progrès sont souvent associés à une meilleure compréhension et reconnaissance visuelle du mot plutôt qu'à une amélioration de leur capacité de décodage analytique (voir Alexander, Anderson, Heilman, Voeller et Torgesen, 1991). La capacité de décodage alphabétique ne s'améliore donc pas toujours suite aux interventions basées uniquement sur les correspondances graphème-phonème et sur les caractéristiques orthographiques (ex. Ball et Blachman, 1988, 1991).

Par conséquent, des programmes d'entraînement de la conscience phonologique ont été mis en place et évalués pour savoir s'ils peuvent améliorer les capacités métaphonologiques et les performances en lecture d'enfants avec ou sans difficultés. Les investigateurs ont rapporté que l'entraînement des capacités de segmentation phonémique conduit à une amélioration significative d'un point de vue statistique de la conscience phonologique (ex. Alexander et al., 1991 ; Elkonin, 1973 ; Skjelfjord, 1976 ; Zhurova, 1973). De plus, il s'est avéré que les enfants qui apprennent à segmenter ou à enchaîner des sons réussissent statistiquement beaucoup mieux dans les activités d'attaques de mots que les enfants qui n'ont pas reçu de formation à la synthèse ou à l'analyse de phonèmes (Fox et Routh, 1976 ; Treiman et Baron, 1983 ; Wallach et Wallach, 1976 ; Williams, 1980).

Le but de cette analyse est d'examiner la qualité méthodologique d'études ayant entraîné la conscience phonologique chez des enfants. Après une présentation d'ensemble des études qui ont été sélectionnées, nous traiterons de leurs qualités et de leurs limites méthodologiques en utilisant les critères conventionnels d'évaluation en recherche quantitative. Une telle critique est nécessaire car la rigueur méthodologique avec laquelle les études sur les interventions relatives à la conscience phonologique sont conduites est directement liée à l'interprétation et à la possibilité de généraliser les conclusions de ces recherches. De plus, ce domaine de recherche s'est rapidement développé ces dix dernières années et il est probable qu'il continue à attirer beaucoup d'attention dans la communauté éducative ; pourtant, l'analyse critique de la recherche expérimentale n'a pas encore été faite. Il faudrait clairement démontrer la viabilité et l'efficacité des programmes d'entraînement de la conscience phonologique avant que les praticiens ne les adoptent pour prévenir ou remédier aux difficultés en lecture de différentes populations. Enfin, il

[1] Titre original : Phonological awareness intervention research : a critical review of the experimental methodology.

y a de nettes différences entre ces études quant à la nature des interventions employées, les méthodes utilisées pour évaluer l'efficacité des traitements, et les conclusions spécifiques obtenues. Une étude de ce type aidera les éducateurs et les chercheurs à comprendre l'impact d'une telle diversité sur notre connaissance dans ce domaine.

◆ MÉTHODE

Les procédures de recherche

Nous avons tout d'abord mené une recherche automatique dans deux bases de données, *Educational Resources Information Center* (ERIC) et *Psychological Literature Abstracts* (PsycLit), afin d'identifier les études qui pouvaient être intégrées à cette analyse. Les descriptifs utilisés lors de la recherche informatique étaient : métalinguistique, métaphonologie, phonologie, phonème, conscience phonologique, conscience phonémique et capacités phonologiques, couplés avec : entraînement, instruction, enseignement ou intervention. Nous avons ensuite localisé d'autres études citées en référence dans les articles obtenus suite à la recherche sur les bases de données. Enfin, nous avons effectué une recherche manuelle dans une sélection de journaux afin d'être sûrs que nous avions identifié toutes les recherches publiées sur ce sujet : *Annals of Dyslexia, Journal of Educational Psychology, Journal of Experimental Child Psychology, Journal of Learning Disabilities, Learning Disabilities Research and Practice, Reading Research Quarterly et Remedial and Special Education*. Ces journaux ont été sélectionnés car ils sont souvent mentionnés dans les articles relatifs à la conscience phonologique. Grâce à ces procédures de recherche, nous avons sélectionné 68 études qu'il était possible d'intégrer dans notre analyse.

Les critères de sélection

Pour être sélectionnée dans notre corpus, chaque étude devait répondre à plusieurs critères. Tout d'abord, l'étude devait être publiée dans un journal à comité de lecture. Évaluer les travaux publiés dans ce domaine est une sorte de test décisif remettant en cause la croyance répandue mais naïve que seules les meilleures recherches sont publiées. Si la plupart des expériences publiées ne répondent pas aux exigences de rigueur méthodologique attendues dans les recherches quantitatives, on peut se demander si la procédure d'examen par les pairs permet de dépister les études mal réalisées ou mal construites.

Deuxièmement, les enfants du ou des groupe(s) expérimentaux devaient être comparés à des enfants affectés à une sorte de groupe contrôle. Les recherches conduites sans groupe contrôle ne peuvent pratiquement pas être interprétées car l'amélioration des performances ne peut pas être imputée uniquement au traitement subi : l'histoire, la maturation et les tests sont des sources de confusion possibles.

Troisièmement, les conditions expérimentales devaient inclure un entraînement à l'analyse ou à la synthèse auditive soit isolément soit conjointement avec un entraînement phonétique. C'est pourquoi nous avons exclu de notre corpus les études qui évaluaient uniquement les effets d'un entraînement phonétique car elles ne sont pas vraiment représentatives de la manière dont se construit la conscience phonologique qui est censée être un précurseur de la réussite en lecture.

Quatrièmement, nous avons éliminé les études qui utilisaient des interventions extrêmement brèves (ex. : un entraînement conduit sur une ou deux séances). Un tel entraînement est plus représentatif de paradigmes d'évaluation dynamique que de pratiques pédagogiques types qu'on estime adéquates à l'apprentissage de la lecture et de l'écriture.

En appliquant ces quatre critères, nous avons identifié 39 études pouvant être incluses dans notre analyse (elles sont précédées d'un astérisque dans notre bibliographie). Il faut noter qu'un certain nombre d'auteurs ont conduit des études longitudinales pour examiner les effets à long terme de leurs interventions. Nous n'avons examiné ces études ultérieures que pour inclure les mesures de maintien et de généralisation des acquis, afin de ne pas tomber dans le piège consistant à évaluer de multiples rapports portant sur une même étude.

Fiabilité de l'évaluation

Nous avons sélectionné au hasard dix des 39 études afin qu'un deuxième lecteur les évalue pour établir la fiabilité du codage. Un étudiant diplômé en sciences de l'éducation qui avait suivi des cours en recherche méthodologique a codé de façon indépendante cet échantillon issu du corpus, en évaluant les critères de validités interne et externe. Les différents points s'accordaient à 91%. Les contradictions dans le codage ont été résolues après discussion ; par contre, les jugements relatifs au codage rapportés ici sont ceux de l'auteur.

Généralités

La plupart (22) des études sélectionnées ont été conduites aux Etats-Unis. Les autres se sont déroulées au Canada, au Portugal, en Israël, en Australie, en Nouvelle-Zélande, en

Belgique, au Royaume-Uni et en Scandinavie. Toutes les études ont utilisé une méthodologie de recherche quantitative et des plans expérimentaux ou quasi expérimentaux. Un plan longitudinal fut utilisé dans 10 des études originales (Bentin et Leshem, 1993 ; Brady, Fozler, Stone et Winbury, 1994 ; Hatcher, Hulme et Ellis, 1994 ; Kennedy et Backman, 1993 ; Korkman et Peltomaa, 1993 ; Kozminsky et Kozminsky, 1995 ; Lie, 1991 ; Lundberg, Frost et Peterson, 1988, Uhry et Shepherd, 1993 ; Warrick, Rubin et Rowe-Walsh, 1993).

Les effets des programmes d'interventions en classe ont été évalués dans 12 études (Blachman, Ball, Black et Tangel, 1994 ; Brady et al., 1994 ; Haddock, 1976 ; Kennedy et Backman, 1993 ; Kozminsky et Kozminsky, 1995 ; Lie, 1991 ; Lundberg et al., 1988 ; McGuinness, McGuinness et Donohue, 1995 ; Olofsson et Lundberg, 1983 ; Rosner, 1974 ; Tangel et Blachman, 1992 ; Williams, 1980). Les participants de 3 études ont reçu un enseignement basé sur l'informatique (Foster, Erickson, Foster, Brinkman et Torgesen, 1994, expériences 1 et 2 ; Wise et Olson, 1995). La longueur de l'intervention variait beaucoup selon les études, allant de 2 semaines (Content, Morais, Alegria et Bertelson, 1982 ; Hohn et Ehri, 1983 ; Vellutino et Scanlon, 1987) à 2 ans (Bradley et Bryant, 1983) avec une durée moyenne d'environ 11 semaines. De même, le nombre de séances de traitement suivies par les élèves allait de 5 (Slocum, O'Connor et Jenkins, 1993) à plus de 100 (Lundberg et al., 1988 ; MacGuinness et al., 1995), avec une moyenne d'environ 32 séances. Rosner (1974) n'a pas indiqué, dans son étude, combien de séances d'entraînement les élèves avaient suivi.

Des élèves identifiés grâce à des tests de réussite comme étant non-lecteurs participaient à 13 des études (Ball et Blachman, 1988 ; Blachman et al., 1994 ; Bradley et Bryant, 1983 ; Brady et al. 1994 ; Cary et Verhaeghe, 1994, expériences 1 et 2 ; Hohn et Ehri, 1983 ; Lundberg et al., 1988 ; O'Connor, Jenkins et Slocum, 1995 ; Tangel et Blachman, 1992 ; Torgesen, Morgan et Davis, 1992 ; Warrick et al., 1993). Les chercheurs de 7 études (Kennedy et Backman, 1993, Korkman et Peltomaa, 1993 ; O'Connor et al., 1993, 1995 ; Vellutino et Scanlon, 1987 ; Warrick et al., 1993 ; Williams, 1980) ont étudié les effets d'un entraînement métaphonologique avec des enfants souffrant d'un handicap documenté.

Dans toutes les études sauf 4, les participants étaient des enfants âgés de 4 à 7 ans. Williams (1980) a sélectionné des enfants avec un handicap âgés de 7 à 12 ans, alors que Kennedy et Backman (1993) ont sélectionné des enfants connaissant des problèmes d'apprentissage âgés de 11 à 15 ans. Vellutino et Scanlon (1987) se sont concentrés sur de mauvais lecteurs ou des lecteurs moyens dans des classes du deuxième au sixième grade (approximativement équivalant en France du CE1 à la 6ème, NDT) tandis que Wise et Olson (1995) ont choisi des mauvais lecteurs dans des classes du deuxième au cinquième grade (approximativement équivalant en France du CE1 au CM2, NDT).

Toutes les études du corpus sauf 2 (Bradley et Bryant, 1983 ; Haddock, 1976) ont pris comme critère de mesure au moins une activité d'analyse ou de synthèse de phonèmes. Toutes les études sauf 9 incluaient une évaluation de la réussite en lecture suite au traitement, portant par exemple sur la performance à un test standard d'identification de mots (ex. Ball et Blachman, 1980), un test expérimental de décodage (ex. Williams, 1980) ou une tâche comparable à la lecture (ex. Torgesen et al., 1992). Les effets de l'entraînement sur les performances orthographiques ont été évalués dans environ la moitié des études. Les 39 études ont montré de façon empirique que l'entraînement métaphonologique conduisait à une amélioration importante de la conscience phonologique et des performances en lecture et/ou en orthographe (lorsqu'elles étaient mesurées). Cela était prévisible puisque les conclusions où l'effet est nul ne sont habituellement pas publiées dans les journaux de recherche (un point à revoir).

◆ RÉSULTATS

L'ensemble des critères d'évaluation de la validité interne et externe de ces études sont présentés dans le tableau 1 *(à consulter sur le site AFL http://www.lecture.org)* et sont basés sur les travaux de Campbell et Stanley (1966), Cook et Campbell (1979) et Huck, Cormier et Bounds (1974). Chaque critère s'accompagne d'une brève définition et d'un coefficient proportionnel à son importance dans les interprétations causales et dans la possibilité de généraliser l'étude. Plus précisément, un coefficient 1 est affecté à chaque critère pour lequel une non-satisfaction serait regrettable mais ne menacerait pas sérieusement le contrôle expérimental ou la validité externe. Un coefficient 2 est affecté à chaque critère pour lequel une non-satisfaction serait considérée comme un défaut méthodologique nuisible mais pas crucial. Les critères considérés comme essentiels pour éviter toute interprétation alternative de la causalité et pour permettre la généralisation des conclusions à d'autres populations et à dans d'autres circonstances sont affectés d'un coefficient 3.

Trois grandes catégories de critères de validité interne ont été définies : les caractéristiques générales du plan expérimental, les mesures et le traitement statistique.

De même, trois catégories de critères de validité externe ont été établies : les hypothèses de recherche, la sélection et la description des participants et les mesures de transfert et de maintien des acquis.

Dans les tableaux 2 et 3 *(à consulter sur le site AFL)*, le codage pour chaque critère de validité interne et externe sont présentées pour chaque étude du corpus. D'habitude, cette information n'est pas rapportée, mais nous avons estimé possible de le faire ici étant donné le petit nombre d'études analysées. Nous n'avons pas l'intention de montrer du doigt des études ou des chercheurs en particulier. Mais cette information devrait faciliter la discussion sur la qualité d'un groupe varié d'études. La proportion (et le pourcentage) d'études qui ont satisfait chaque critère est résumée à la fin de chaque tableau. Chaque point pour lequel l'information n'était pas suffisante a été évalué comme étant négatif. Cette façon de faire a surtout affecté deux critères : l'effet des maîtres au sein des groupes et le transfert dans une activité comparable. Enfin, le tableau 4 résume la rigueur méthodologique relative des investigations en présentant pour chaque étude le pourcentage des critères de validité interne et externe satisfaits, le nombre de critères de validité interne et externe cruciaux violés, une note pondérée (décrite ci-après), et un classement général. La note pondérée a été calculée en additionnant les facteurs de pondération de chaque point évalué négativement. Une note pondérée basse indique donc que l'étude a suivi une méthodologie assez bonne. Les classements sont effectués à partir de trois facteurs : la note pondérée, le nombre total de violations de critères affecté du coefficient 3 et la proportion totale des critères d'évaluation satisfaits.

La validité interne

Les caractéristiques générales des plans expérimentaux. Les individus ont été distribués de manière aléatoire dans 21 études (54%). Dans 14 de ces études (67%), les chercheurs ont rassemblé les enfants en fonction de mesures de pré-testage et/ou de leurs taux de QI, avant de les affecter de manière aléatoire à leur groupe respectif. Bien entendu, dans les études où l'enseignement était dispensé à des classes entières par les enseignants, les élèves ne pouvaient pas être répartis de manière aléatoire dans des groupes, même si dans quatre études (Haddock, 1976 ; Kozminsky et Kozminsky, 1995 ; Lie, 1991 ; Williams, 1980), les classes avaient été affectées de manière aléatoire aux groupes expérimentaux et de contrôle. Dans les 8 autres études effectuées dans des classes, soit l'affectation aléatoire des classes aux groupes n'a pas été vérifiée, soit il y a eu confusion entre traitement et école (les élèves des groupes expérimentaux allaient dans des écoles différentes de ceux des groupes contrôle).

Bien que des groupes contrôle aient été utilisés dans toutes les investigations, dans environ la moitié du corpus (51%), les individus ont continué de suivre leur programme d'enseignement habituel et n'ont pas fait l'objet d'interventions spécifiques. L'interprétation des résultats du traitement présentée dans ces études n'est pas fiable car une amélioration des performances suite à un traitement peut être simplement due à la nouveauté de l'intervention expérimentale plutôt qu'aux caractéristiques spécifiques de l'entraînement. Les participants du groupe contrôle dans les autres études avaient reçu un enseignement différent.

Par exemple, Bradley et Bryant (1983) et Byrne et Fielding-Barnsley (1991) apprenaient aux enfants de leurs groupes contrôle à trier des images en fonction d'associations sémantiques, alors que les enfants des groupes expérimentaux apprenaient à classer ces mêmes images par catégories regroupant les mêmes sons. Dans l'étude de Cunningham (1990) ainsi que dans l'étude conduite par Torgesen et ses collègues (1992), les élèves du groupe contrôle écoutaient des histoires qu'on leur lisait et posaient des questions alors que les élèves dans les conditions expérimentales apprenaient à segmenter ou à enchaîner des sons.

Haddock (1976) et O'Connor et al. (1995) ont tous les deux enseigné la correspondance lettre-son aux enfants de leurs groupes contrôle tandis qu'ils enseignaient la conscience phonologique, soit de façon isolée soit combinée à un entraînement lettre-son, aux autres enfants. (Certaines des études où le groupe contrôle suivait un autre type d'intervention comprenaient également un autre groupe contrôle qui ne recevait, lui, aucun traitement spécial. Ex. : Ball et Blachman, 1988 ; Bentin et Leshem, 1993 ; Bradley et Bryant, 1983 ; Castle, Riach et Nicholson, 1994, expérience 2 ; Kozminsky et Kozminsky, 1995 ; Olofsson et Lundberg, 1983 ; Vellutino et Scanlon, 1987). Ces groupes contrôle sans traitement servaient à valider l'efficacité des interventions expérimentales.

Dans moins de la moitié (40%) des 20 études où le groupe contrôle recevait un autre type d'intervention, le matériel éducatif était le même que celui du groupe expérimental (ex. Bradley et Bryant, 1983 ; Byrne et Fielding-Barnsley, 1991). On peut plus facilement attribuer les effets du traitement à la seule intervention expérimentale si les participants des différents groupes utilisent le même matériel. Dans les autres cas, ces effets pourraient être attribués à un matériel éducatif expérimental probablement créatif et motivant.

Dans seulement le tiers (36%) des études, les expérimentateurs ont rapporté soit que tous les élèves étaient suivis par un seul enseignant (Castle et al., 1994, expériences 1 et 2 ; Cunningham, 1990 ; Foster et al., 1994, expériences 1 et 2 ; Fox et Routh, 1984 ; Warrick et al., 1993), soit que les enseignants étaient répartis dans les différents groupes (Haddock, 1976 ; Hatcher et al., 1994 ; O'Connor et al., 1993, 1995 ; Slocum et al., 1993 ; Torgesen et al., 0992 ; Uhry et Shepherd, 1993) afin d'éviter la confusion enseignant/condition. La plupart des autres études n'ont pas fourni une information adéquate sur l'affectation des enseignants aux différents groupes classes.

Bien entendu, les auteurs des 12 études limitées à une classe ne pouvaient pas éliminer la possibilité d'un biais introduit par l'enseignant. Cependant, Haddock (1976) a bien tenté de le faire : dans chaque classe participante, les élèves ont été affectés soit au groupe expérimental, soit au groupe contrôle. Dans chaque classe, l'enseignement du professeur variait selon le groupe auquel les élèves appartenaient. Malheureusement, la validité du traitement n'a pas été évaluée et il est donc possible qu'il y ait eu biais de l'enseignant et/ou propagation des traitements d'un groupe à l'autre.

Dans un peu plus de la moitié (56%) des investigations, les conditions du traitement étaient décrites de façon suffisamment détaillée. Bien sûr, la place disponible dans les journaux limite souvent les auteurs quant à la quantité d'information qu'ils peuvent fournir. Par conséquent, un certain nombre de chercheurs ont indiqué qu'une description plus complète de leurs programmes d'entraînement était disponible sur demande. Il n'en reste pas moins que les lecteurs des recherches, et surtout les praticiens qui souhaitent utiliser des traitements innovants avec leurs élèves, ont besoin de descriptions détaillées des programmes d'interventions.

Onze (28%) des expériences analysées ont utilisé un enseignement basé sur un objectif. Parmi elles, deux études avaient lieu dans une seule classe (Haddock, 1976 ; Rosner, 1974). Dans les 28 autres études, on ne demandait pas aux élèves d'atteindre une performance précise à chaque étape de l'enseignement. L'utilisation d'un enseignement basé sur un objectif fournit une information de qualité sur l'efficacité du traitement et sur les interactions possibles entre une progression normale et une progression liée à cet objectif. De plus, des différences de performance entre les groupes peuvent surgir (en faveur du traitement non expérimental ou du groupe contrôle) si un nombre important des élèves d'un groupe ne maîtrise pas suffisamment une compétence particulière. Inversement, il n'est souvent pas possible d'utiliser un enseignement basé sur un objectif dans une recherche en classe entière car l'enseignant ne souhaite pas arrêter la progression du traitement lorsque certains élèves ne parviennent pas à satisfaire l'objectif. En outre, comme la plupart des élèves atteignent des niveaux de performance similaires, un entraînement basé sur un critère est susceptible de réduire la variation des mesures, limitant ainsi l'efficacité des analyses statistiques.

Dans 5 (25%) des 20 études dans lesquelles le groupe contrôle a suivi un autre type d'intervention, les investigateurs n'ont pas équilibré les temps d'enseignement entre les différents groupes. En particulier, ni Bradley et Bryant (1983), ni Olofsson et Lundberg (1983) n'ont fourni une information satisfaisante permettant de déterminer si les participants suivant un traitement expérimental et les ceux des groupes contrôle avaient passé le même temps sur des activités similaires. O'Connor et ses collègues (1993, 1995) ont fait faire aux participants du groupe contrôle des exercices comprenant des parties isolées des programmes expérimentaux. Mais les élèves des groupes expérimentaux continuaient à suivre un entraînement plus long tout en participant à l'intégralité des programmes. Enfin, Torgesen et al. (1992) ont fait remarqué que le groupe qui suivait un traitement expérimental avec analyse et synthèse de phonèmes avait une semaine d'entraînement supplémentaire par rapport au groupe suivant un traitement avec synthèse de phonèmes et par rapport au groupe contrôle. Les résultats de ces 5 études doivent être interprétés avec prudence car il est possible que les effets positifs du traitement annoncés soient un artefact d'une interaction prolongée avec l'enseignant.

La perte de participants était problématique dans 10 (26%) des études, car elle n'était pas la même selon les groupes. La perte est en général due à des déménagements, à trop d'absences, à la disparition de données ou à des raisons non spécifiées. Seuls Lundberg et ses collègues (1988) ont rapporté que les enfants perdus au cours de leur investigation avaient les mêmes caractéristiques que les autres élèves de leur étude d'après les mesures des critères au pré-testage (capacités de réception du langage, conscience métaphonologique, aptitudes en lecture, ...). Il est par conséquent peu probable qu'il y ait eu des différences fondamentales entre les participants restés dans l'étude et ceux qui en étaient sortis.

Les mesures. La plupart des investigateurs (87%) ont effectué leurs mesures dépendantes de manière satisfaisante. Ball et Blachman (1988, 1991) et Weiner (1994) ont fourni des détails précis sur leurs mesures dépendantes (ex. tâches

requises, listes de stimuli utilisés lors des test, durées) et sur leurs méthodes de notation. Ces auteurs donnent de remarquables exemples de descriptions de mesures d'objectifs.

Lorsque les analyses de la fiabilité ne sont pas fournies, il est impossible de savoir si une erreur de mesure n'a pas faussé les conclusions. Malheureusement, seul un tiers (33%) des études attestait de la fiabilité d'au moins une majorité de mesure des objectifs. Il semble que la plupart des chercheurs soit ne contrôlaient pas leurs évaluations, soit n'avaient pas les moyens nécessaires d'obtenir des données de fiabilité au cours de leurs investigations.

Des contrôles de réalisation n'ont été effectués que dans 5 des études analysées (13%), afin d'être sûr que les conditions de traitement étaient mises en œuvre correctement. O'Connor et ses associés (1993) ont observé spécifiquement chaque enseignant au moins une fois par semaine dans chaque condition de traitement et ont enregistré au hasard différentes sessions d'intervention pour vérifier la fidélité du traitement. Dans une étude ultérieure, O'connor et al. (1995) ont observé régulièrement des leçons dans chaque condition de traitement et ont assuré une formation continue aux enseignants afin d'éviter une dérive de l'enseignement. Hatcher et ses collègues (1994) ont assuré l'intégrité du traitement en demandant aux enseignants de remplir une note par écrit après chaque séance de traitement et en tenant des réunions régulières avec eux. Slocum et al. (1993) ont rapporté que chacun de leurs enseignants avait reçu une formation durant laquelle ils devaient enseigner à la fois aux auteurs et aux enfants qui ne participaient pas à leur étude, afin qu'ils soient à 100% fiables. Les auteurs ont ensuite observé des séances d'enseignement par intermittence tout au long de l'étude. Williams (1980) déclarait avoir effectué des listes de contrôle des plans des leçons et des observations de classes, bien que ces observations n'aient pas été fréquentes (une fois toutes les 2 ou 3 semaines). Seuls Williams (1980) et O'Connor et al. (1995) ont fourni des indices relatifs à la fidélité du traitement dans leurs études (ex. pourcentage de précision dans la mise en place).

Dans ces recherches en interventions, une question importante concernant les mesures est la sensibilité des mesures dépendantes sélectionnées pour évaluer la réponse au traitement. Une mauvaise sensibilité des mesures (ex. l'épreuve est trop facile ou trop difficile) peut conduire à des effets plafond ou plancher qui restreignent les marges de variation observées et limitent l'efficacité des analyses statistiques. Il n'y avait aucun effet plafond ou plancher dans seulement 14 (36%) des études analysées. Ainsi, dans la plupart des expériences, les différences de performances entre les groupes expérimentaux et les groupes contrôle peuvent être atténuées ou masquées.

Le traitement statistique. Le nombre de participants dans une expérience affecte considérablement la force des statistiques. Dans plus des trois quarts (77%) des études, la taille des groupes a été jugée suffisante pour effectuer des analyses statistiques. Dans les études où les groupes étaient estimés trop petits, les investigateurs n'ont pas effectué d'analyses de puissance pour déterminer si les taux d'erreur de type II restaient dans des limites acceptables. Dans près des deux tiers (61%) des études, les investigateurs ont contrôlé la probabilité d'erreur de type I, généralement en ajustant les taux d'erreur expérimentale, en effectuant des analyses multivariées suivies de tests univariés, ou en utilisant des comparaisons planifiées. Les autres études ont effectué un grand nombre d'analyses de variance et/ou ont classifié un grand nombre de coefficients de corrélation sans ajuster les niveaux de probabilité.

L'unité de l'analyse correspondait à l'unité du traitement dans 13 (33%) des études. Plus précisément, dans 6 des 32 expériences où l'enseignement était donné à des petits groupes d'élèves (en général entre 3 et 6 enfants dans chaque groupe) ou à la classe entière plutôt qu'à chaque élève individuellement, la moyenne du groupe représentait l'unité dans l'analyse des données. Dans les 7 autres études où l'unité de l'analyse était correcte, on enseignait individuellement à chaque enfant. Dans une des études où l'unité d'analyse n'était pas l'unité du traitement (Torgesen et Davis, 1996), un système de modèle hiérarchique linéaire, c'est à dire une méthode statistique pour évaluer les courbes de croissance individuelle, fut utilisé.

Toutes les études sauf 3 (92%) ont utilisé des analyses statistiques acceptables (même si elles n'étaient pas forcément les meilleures), comme les tests t et F, les analyses de variance et de covariance et les mesures répétées. Olofsson et Lundberg (1983) n'ont effectué aucun test statistique, disant que les résultats des mesures des objectifs étaient distribués de façon bi modale. Cependant, des tests non paramétriques auraient été possibles et adaptés. O'Connor et al. (1995) et Williams (1980) ont calculé les coefficients de corrélation entre chaque condition expérimentale et non à l'intérieur de celles-ci, ceci ayant pour conséquence de gonfler les corrélations à cause des différences entre les groupes. Sinon, leurs analyses statistiques étaient satisfaisantes.

Seules 4 expériences (10%) ont fourni les tailles des effets. Dans le quart (23%) des autres études (Bradley et Bryant, 1983 ; Byrne et Fielding-Barnsley, 1991 ; Hohn et Ehri,

1983 ; McGuinness et al., 1995 ; Olofsson et Lundberg, 1983 ; Rosner, 1974 ; Tornéus, 1984 ; Wise et Olson, 1995), il serait impossible de calculer les tailles des effets à partir des données disponibles. Ceci est assez surprenant car il suffit de statistiques descriptifs de base pour calculer la taille d'un effet. La communication des tailles des effets faciliterait l'évaluation de l'efficacité relative des traitements expérimentaux utilisés dans ces études d'intervention.

Pour résumer la question de la validité interne du corpus, environ la moitié des études n'ont pas effectué de répartition aléatoire, ce qui est un sérieux défaut méthodologique. Dans environ la moitié des études, les chercheurs n'ont pas contrôlé les effets Hawthorne[2], ce qui remet en question l'efficacité démontrée des interventions relatives à la conscience phonologique. De même, la plupart des études n'ont pas rendu compte de la fidélité du traitement et il n'est donc pas évident que les résultats d'un traitement, rapportés de façon fiable, puissent vraiment être attribués à l'intervention. Dans environ deux tiers des investigations, la mesure de la sensibilité n'était pas adaptée.

Au contraire, les mesures des critères ont été clairement effectuées dans la plupart des expériences. Dans la majorité des études, le temps passé en traitement par chaque participant a été uniformisé entre les groupes (le cas échéant). Dans la quasi-totalité des études retenues, des analyses statistiques adaptées ont été effectuées, bien que beaucoup d'auteurs n'aient pas utilisé une unité d'analyse appropriée. Seules 9 (23%) des expériences analysées satisfaisaient environ les deux tiers ou plus des critères de validité interne applicables (Ball et Blachman, 1988 ; Byrne et Fielding-Barnsley, 1991 ; Castle et al., 1994, expérience 2 ; Foster et al., 1994, expériences 1 et 2 ; Hatcher et al., 1994 ; Slocum et al., 1993 ; Torgesen et al., 1992 ; Uhry et Shepherd, 1993). Parmi celles-ci, six ne respectaient pas deux ou plus des critères de validité considérés comme indispensables au contrôle expérimental (c'est à dire avec un coefficient de pondération égal à 3). Il est intéressant de noter que deux autres investigations (Bentin et Leshem, 1993 ; Castle et al., 1994, expérience 1) violaient une seule exigence fondamentale (dans les deux cas, il s'agissait de la fidélité du traitement), bien qu'elles ne satisfassent pas une majorité des critères de validité interne.

Traduction : ***Magali FRANÇOIS, Léna COÏC, Denis FOUCAMBERT***

Les tableaux cités et la bibliographie sont consultables sur le site Internet de l'AFL : http://www.lecture.org

La deuxième partie de l'étude de Gary A. Trioa, qui traite des critères de validité externe, paraîtra dans le n°70 de juin 2000.

Rigueur méthodologique relative des études

Études	% de critères de validité interne	Nombre de défauts cruciaux
Ball & Blanchman, 1988,1991	65	2
Bentin & Leshem, 1993	47	1
Blachman, Ball, Black & Tangel, 1994	47	4
Bradley & Bryant, 1983, 1985	47	4
Brady, Fowler, Stone & Winbury, 1994	27	4
Byrne & Fielding-Barnsley, 1991, 1993, 1995	65	2
Cary & Verhaeghe, 1994 (Expérience 1)	20	4
Cary & Verhaeghe, 1994 (Expérience 2)	35	3
Castle, Riach & Nicholson, 1994 (Expérience 1)	59	1
Castle et al., 1994 (Expérience 2)	65	1
Content, Morais, Alegria & Bertelson, 1982	41	3
Cunningham, 1990	53	2
Foster, Erickson, Foster, Brinkman & Torgesen, 1994 (Expérience 1)	80	2
Foster et al ., 1994 (Expérience 2)	73	3
Fox & Routh, 1984	53	4
Haddock, 1976	53	3
Hatcher, Hulme & Ellis, 1994	67	1
Hohn & Ehri, 1983	60	2
Kenney & Backman, 1993	33	3
Korkman & Peltomaa, 1993	40	4
Kozminsky & Kozminsky, 1995	47	3
Lie, 1991	47	3
Lundberg, Frost & Peterson, 1988	27	4
McGuinness, McGuinness & Donohue, 1995	40	3
O'Connor, Jenkins, Leicester & Slocum, 1993	53	2
O'Connor, Jenkins & Slocum, 1995	59	2
Olofsson & Lundberg, 1983, 1985	31	4
Rosner, 1974	40	4
Slocum, O'Connor & Jenkins, 1993	71	1
Tangel & Blachman, 1992	47	4
Torgesen & Davis, 1996	40	3
Torgesen, Morgan & Davis, 1992	82	2
Tornéus, 1984	40	3
Uhry & Shepherd, 1993	65	2
Vellutino & Scanlon, 1987	53	3
Warrick, Rubin & Rowe-Walsh, 1993	27	5
Weiner, 1994	47	3
Williams, 1980	40	4
Wise & Olson, 1995	53	2

[2] On appelle effet Hawthorne les résultats qui ne sont pas dus aux facteurs expérimentaux, mais à l'effet psychologique de participer à une recherche et d'être l'objet d'une attention spéciale (NdT).