Chapitre 7

Effets du projet sur les résultats des élèves et élaboration des instruments d'enquête

Ce chapitre présente les effets les plus pertinents de l'étude pilote du projet OCDE-CERI sur les résultats des élèves, et examine la validation des instruments d'enquête. Il offre également une vue d'ensemble des premières conclusions d'une analyse menée à l'échelle de la classe, en mettant l'accent sur les classes les plus performantes et sur les caractéristiques de leurs enseignants, élèves et activités pédagogiques adoptées. Il aborde enfin les leçons tirées de la phase pilote du projet, en les classant par thématique principale, avant de formuler des propositions d'amélioration relatives aux aspects opérationnels et instruments d'enquête en vue de la phase de validation.

Le projet OCDE-CERI

Le projet OCDE-CERI a réuni 13 équipes de 11 pays différents qui ont toutes reconnu l'importance de la créativité et de l'esprit critique pour l'avenir de leurs élèves. Elles ont travaillé en partenariat en vue de favoriser le développement de ces compétences, en s'appuyant sur des pratiques pédagogiques innovantes et une vision commune des tenants et aboutissants de la créativité et de l'esprit critique. Les équipes provenaient de pays membres de l'OCDE (Espagne [Communauté de Madrid], États-Unis [équipes Montessori et Vista], France [équipes CRI et Lamap], Hongrie, Pays-Bas, République slovaque et Royaume-Uni [Pays de Galles]) et d'économies non membres (Brésil, Fédération de Russie, Inde et Thaïlande).

Si le projet était essentiellement axé sur une expérience fondée sur l'élaboration de ressources pédagogiques, un second objectif consistait à mettre au point des instruments qui seraient utilisés dans le cadre d'une phase de validation du projet. À cette fin, l'OCDE a élaboré et testé sur le terrain des instruments dans le cadre d'un plan d'enquête quasi-expérimental, qui a consisté à sélectionner deux échantillons d'élèves, puis à leur administrer une série de questionnaires et de tests, en vue de mesurer plusieurs résultats clés (comme le potentiel créatif) et d'évaluer les variables explicatives pertinentes (par exemple le sexe, l'âge). Un groupe a bénéficié de pratiques pédagogiques visant à promouvoir le développement de la créativité et de l'esprit critique chez les élèves (le groupe expérimental), tandis que l'autre groupe a servi de point de référence (le groupe de contrôle). Pour les deux groupes d'élèves, l'administration des questionnaires et des tests a eu lieu à deux reprises : la première avant l'adoption des nouvelles pédagogies par le groupe expérimental (pré-expérience), et la seconde vers la fin de l'année scolaire (post-expérience).

Il convient d'insister sur ce que l'on entend exactement par groupe expérimental, étant donné que les points essentiels de l'analyse présentée ci-après reposent sur la comparaison entre ledit groupe et le groupe de contrôle. Dans le groupe expérimental, les enseignants avaient activement participé à des séances de formation professionnelle organisées dans le cadre du projet (voir le chapitre 5). Ces événements mettaient à leur disposition des outils communs décrivant les principales caractéristiques des activités pédagogiques qui permettraient de développer et d'évaluer la créativité et l'esprit critique (voir le chapitre 2). Les enseignants de certaines équipes jouissaient toutefois d'une totale autonomie quant aux aspects pratiques de mise en œuvre des activités sur le plan de la durée et du contenu. Ainsi, si l'ensemble des élèves dans le groupe expérimental ne suivaient pas exactement les mêmes activités pédagogiques, tous leurs enseignants participaient en revanche à des événements de formation professionnelle spécifiques et proposaient des activités s'inspirant des principes qui y étaient abordés.

L'OCDE a conçu la majorité des instruments et des tests spécialement pour ce projet, notamment un questionnaire « Élève », un test de performance en sciences et mathématiques

et un autre en arts visuels et musique. Tous les instruments étaient adaptés en fonction du niveau d'enseignement : ceux administrés aux élèves du primaire étant plus faciles que ceux destinés aux élèves du secondaire. En outre, le projet a également adopté le test EPoC, un test d'évaluation du potentiel créatif des enfants et adolescents dans une discipline spécifique mis au point par Todd Lubart, Maud Besançon et Baptiste Barbot (2011[1]). Dans ce chapitre, les conclusions au niveau de l'élève s'appuieront également sur certaines données obtenues par le biais du questionnaire « Enseignant ».

Le volet du projet OCDE-CERI impliquant les élèves s'est déroulé de novembre 2015 à juillet 2017, une période qui comprend deux années scolaires complètes tant pour l'hémisphère sud que l'hémisphère nord. Si les équipes locales ont fait preuve de divers degrés d'implication en fonction de l'année scolaire, du niveau d'enseignement et des instruments administrés, dans l'ensemble il a toutefois été possible de recueillir des réponses provenant de plus de 17 000 élèves. Ci-après, les dénominations « Vague 1 » et « Vague 2 » seront utilisées pour différencier les phases du projet qui incluent la collecte de données lors de la première et la seconde année, respectivement. Les évaluations pré- et post-expérience ont été menées à l'occasion de chacune des deux vagues, et si certains enseignants ont participé aux deux vagues (33 sur 380, dont 15 dans le groupe expérimental et 18 dans le groupe de contrôle), cela n'a pas été le cas pour la grande majorité des élèves. En effet, seuls 40 élèves se sont retrouvés dans cette situation, et leurs données tirées de la Vague 2 ont été exclues de l'analyse aux fins du présent rapport.

Ce chapitre présente les conclusions initiales décrivant l'effet des nouvelles activités pédagogiques sur les élèves. Ces conclusions ne constituent qu'une évaluation à court terme de l'efficacité de l'expérience auprès des élèves, conformément à l'objectif de cette phase pilote. L'étude des facteurs pertinents qui permettraient un examen plus approfondi des résultats (une confiance accrue des enseignants en leur capacité à adopter de nouvelles pédagogies au fil du temps, une capacité renforcée des élèves à exploiter et assimiler les nouveaux concepts et pédagogies sur de plus longues périodes, par exemple) ne pourra être menée qu'au moyen d'une évaluation des effets à plus long terme.

Avec en ligne de mire une éventuelle validation du projet, cette étude pilote a constitué un test important pour les différents instruments d'enquête, l'administration en situation réelle et la comparabilité ultérieure des données entre les pays. Afin de garantir une mise en œuvre cohérente des activités d'enquête, l'OCDE a diffusé à l'ensemble des équipes un protocole de recherche contenant des lignes directrices et des recommandations relatives à l'administration des instruments. Une fois l'étude pilote achevée, le protocole de recherche peaufiné a constitué l'un des résultats du projet sur le plan de l'élaboration des instruments. Concernant le choix des instruments à administrer à leurs participants, les équipes ont pris leurs décisions en toute autonomie en s'appuyant généralement sur la structure des expériences mises en œuvre (par exemple, dans les équipes qui se concentraient sur la créativité et l'esprit critique en mathématiques, le test de performance en sciences et mathématiques primait sur celui

en arts visuels et musique). Les équipes étaient ensuite chargées de traduire l'ensemble des instruments dans la ou les langues nationales, et de les administrer conformément au protocole de recherche ; tandis que la coordination de la collecte des données et la réalisation des analyses incombaient à l'OCDE.

Les questions de recherche

Dans le cadre de la phase pilote du projet OCDE-CERI, les questions de recherche abordées dans ce chapitre couvraient les deux grands domaines suivants : les enjeux liés à l'enquête et les résultats des élèves. Le premier domaine concernait l'ensemble des défis à relever pour une élaboration appropriée des instruments d'enquête et une mise en œuvre réussie de la collecte de données. Le second portait, quant à lui, sur l'efficacité de l'expérience à l'échelle des élèves, avec pour objectif de déterminer si les conclusions se révélaient encourageantes pour des équipes, niveaux d'enseignement ou thématiques déterminés ou pour toute combinaison de ces critères. À ce stade, la phase pilote devait constituer une étude de faisabilité : le fait de découvrir si l'expérience fonctionnait sous certaines conditions justifierait la réalisation d'une étude de validation ultérieure.

Les différents instruments conçus pour cette phase pilote ont représenté l'un des principaux résultats. Étant donné que la quasi-totalité des instruments était utilisée sur le terrain pour la première fois, plusieurs questions fondamentales ont nécessité des éclaircissements, et notamment les suivantes :

- Les instruments mesuraient-ils les concepts pour lesquels ils avaient été conçus ?
- Les instruments avaient-ils permis de mesurer une évolution significative de ces concepts, malgré le délai relativement court entre les évaluations pré- et post-expérience ?
- Les instruments recueillaient-ils toutes les informations requises pour une analyse pertinente ?
- Les instruments pouvaient-ils être affinés et améliorés ?

Sur le plan de la conception et de la gestion de l'enquête, les deux principaux défis résidaient dans l'hétérogénéité des parties prenantes participant au projet, et dans la période très courte durant laquelle l'enquête avait été organisée et conduite. Ces défis se sont traduits par des possibilités restreintes d'activités de formation et de soutien à l'utilisation des instruments et, dans certains cas, par l'attribution des responsabilités opérationnelles à des personnes qui manquaient d'expérience en matière de gestion d'enquêtes et de collecte de données. Il sera toutefois intéressant pour la phase de validation de déterminer si une telle conception complexe peut être assurée par du personnel interne plutôt que par des contractants externes. Plusieurs questions clés ont été posées à cet égard :

- Les aspects opérationnels de l'enquête ont-ils été menés conformément au protocole de recherche ?

- Les instruments d'enquête ont-ils été utilisés à bon escient ?
- La collecte de données s'est-elle déroulée conformément au protocole de recherche ?

Enfin, s'agissant des résultats des élèves, les questions soulevées étaient celles qui caractérisent la majorité des études expérimentales, à savoir :

- Quels sont les types d'effets qui peuvent être recensés ?
- Quel est le rôle joué par le contexte ?
- Observe-t-on des effets différents entre les divers sous-groupes d'élèves ?

Élaboration et validation des instruments

Comme mentionné précédemment, jusqu'à quatre instruments différents ont été administrés aux élèves : un questionnaire « Élève », le test EPoC d'évaluation de la créativité, un test de performance en sciences et mathématiques et un autre en arts visuels et musique. Tous les instruments administrés l'ont été à deux reprises : premièrement avant la mise en place de l'expérience (pré-expérience), puis deuxièmement une fois l'expérience achevée ou vers la fin de l'année scolaire (post-expérience), avec idéalement une période de six mois entre ces deux mesures. Les paragraphes suivants fournissent une brève description des caractéristiques et du contenu de chaque instrument.

Le questionnaire « Élève », dont les versions pré- et post-expérience ne variaient que très peu, contenait plusieurs batteries d'items permettant d'établir un ensemble d'indices pertinents, comme l'indice des sentiments positifs à l'égard de l'apprentissage (Dormann, Demerouti et Bakker, 2018[2] ; Schneider et al., 2016[3]) ou l'indice des dispositions à l'apprentissage liées à la créativité et à l'esprit critique (Carr et Claxton, 2002[4]). Le questionnaire comportait également des capsules d'ancrage relatives à ces deux compétences (King et al., 2004[5]), qui ont permis d'évaluer la compréhension que les élèves avaient de ces concepts et l'opinion qu'ils se faisaient de leur propre créativité et esprit critique. En outre, le questionnaire a recueilli certaines informations contextuelles sur les élèves et leur ménage (par exemple, le sexe, le niveau de formation des membres du ménage) ainsi que sur les activités des élèves dans le cadre scolaire et en dehors de celui-ci.

Le test EPoC a été conçu afin de mesurer le potentiel créatif des enfants et adolescents dans différents domaines de la pensée et de la production créatives : expression graphique-artistique, expression verbale et littéraire, résolution de problèmes de société, composition musicale et productions scientifiques et mathématiques (Lubart, Besançon et Barbot, 2011[1]). Dans ce test, les individus devaient produire une création (un dessin, un récit, une solution à un problème, par exemple) qui était ensuite évaluée de manière standardisée. Il existait dans chaque domaine deux types de tâches mettant en œuvre soit la pensée divergente-exploratoire, soit la pensée convergente-intégrative (synthèse créative).

La mesure finale du potentiel créatif englobait ces deux aspects de la créativité. Deux livrets équivalents ont été conçus pour chaque domaine, qualifiés de forme A et forme B, permettant ainsi de comparer les résultats pré- et post-expérience. Pour compléter le test EPoC, 40 à 50 minutes étaient nécessaires.

Le test de performance en sciences et mathématiques a été conçu par l'OCDE grâce à des items tirés des deux enquêtes à grande échelle suivantes : l'enquête TIMSS (*Trends in International Mathematics and Science Study*) menée par l'Association internationale pour l'évaluation du rendement scolaire (IEA) pour les élèves du primaire, et l'enquête PISA (Programme international pour le suivi des acquis des élèves) menée par l'OCDE pour les élèves du secondaire. Ce test comprenait des items de sciences et de mathématiques à réponse ouverte et fermée, des questions intégrées sur l'intérêt des élèves à l'égard de ces matières et des questions sur les pratiques pédagogiques utilisées dans leurs cours de sciences et de mathématiques. Comme pour le test EPoC, deux livrets d'un niveau de difficulté équivalent ont été conçus pour permettre une comparaison des résultats pré- et post-expérience. Chaque livret comportait 20 items à partir desquels était calculé le score final pour les élèves du primaire, tandis qu'il n'en comportait que 18 pour les élèves du secondaire. Les élèves disposaient de 45 minutes pour réaliser le test, indépendamment de leur niveau d'enseignement. Ci-après, ce test sera désigné sous le nom de « test de STIM » (sciences, technologie, ingénierie et mathématiques).

Le test de performance en arts visuels et musique avait été élaboré en interne par l'OCDE. Il comprenait des items d'arts visuels et de musique à réponse ouverte et fermée, des questions intégrées sur l'intérêt des élèves à l'égard de ces matières, et des questions sur les pratiques pédagogiques utilisées dans leurs cours d'arts visuels et de musique. Comme pour le test de STIM, deux tests différents ont été conçus pour les élèves de l'enseignement primaire et secondaire avec, dans chaque cas, l'élaboration de deux livrets équivalents pour permettre une comparaison des résultats pré- et post-expérience. Chaque livret comportait 53 items à partir desquels était calculé le score final pour les élèves de l'enseignement primaire, tandis qu'il en comportait 82 pour les élèves du secondaire. Les élèves disposaient de 30 minutes pour réaliser le test, indépendamment de leur niveau d'enseignement. Ci-après, ce test sera désigné sous le nom de « test d'AVM ».

L'ensemble des scores et indices abordés dans ce chapitre sont présentés par pays, niveau d'enseignement et discipline. Les scores ont été calculés en tant que scores simples pondérés, le coefficient de pondération appliqué dépendant du pourcentage d'élèves ayant correctement répondu à chaque item dans les différents pays, niveaux d'enseignement et batteries d'items. Pour les scores au test de STIM, ce pourcentage a été obtenu à partir des données tirées des enquêtes TIMSS et PISA (IEA, 2011[6] ; OCDE, 2006[7] ; OCDE, 2012[8]). Tandis que pour les scores au test d'AVM, le pourcentage pris en compte était celui des élèves ayant correctement répondu à chaque item dans chacun des pays. En l'absence de ces coefficients de pondération, on utilisait ceux des items internationaux propres au niveau

d'enseignement. Des méthodes plus complexes, comme les modèles inspirés de la théorie de la réponse d'item (TRI), ont également été utilisées pour le calcul des scores aux tests de performance. Toutefois, en raison de la forte corrélation entre les scores simples pondérés et les scores obtenus à l'aide de la théorie de la réponse d'item,[1] il a été jugé préférable de ne pas utiliser ces derniers afin de faciliter l'interprétation des résultats. La majorité des indices a été élaborée par le biais d'analyses factorielles distinctes par équipe et niveau d'enseignement, et les autres indices ont été obtenus en calculant la moyenne simple de deux items. Dans le cas des analyses factorielles, le respect de l'invariance de configuration entre l'ensemble des équipes et niveaux d'enseignement avait été garanti. De plus, les scores et les indices dépendaient de la discipline dans laquelle se déroulait l'expérience. Par exemple, si l'expérience avait lieu en cours de mathématiques, le score final de l'élève au test de STIM ne comportait que les scores qu'il avait obtenus aux items de mathématiques. Toutefois, si l'expérience avait lieu dans une autre matière que celle des mathématiques et des sciences, le score final de l'élève au test de STIM incluait les scores qu'il avait obtenus aux items de mathématiques et de sciences.

Entre la Vague 1 et la Vague 2, l'OCDE a mené une première évaluation des caractéristiques de chaque instrument. Si le questionnaire et le test de STIM n'ont que très peu évolué, plusieurs items ont été supprimés et remplacés dans le test d'AVM. L'annexe technique présente plus de précisions sur les procédures de sélection des items, les instruments, le calcul des scores et des indices ainsi que les contrôles de validité dont ils ont fait l'objet.

Le groupe visé par l'étude

Taille des populations visées par l'étude

L'échantillon initial pour la phase pilote du projet OCDE-CERI comprenait 20 273 élèves, dont 8 949 dans l'enseignement primaire et 11 324 dans l'enseignement secondaire. L'échantillon le plus petit était celui de l'équipe française (CRI), avec 354 élèves, et l'échantillon le plus grand celui de l'équipe thaïlandaise, avec 5 021 élèves. Très peu d'établissements et de classes ont abandonné le projet avant même sa mise en œuvre, conduisant à un échantillon d'élèves participants estimés à 19 129 (8 358 élèves dans le primaire et 10 771 dans le secondaire). Sur ces 19 129 élèves, 17 291 ont participé à au moins une évaluation comprise dans la collecte de données. À l'exception du taux de réponse affiché par l'équipe indienne, estimé à 64 %, celui de toutes les autres équipes s'élevait en moyenne à 95 %.

Sur la base des taux de réponse élevés et des informations disponibles concernant une part importante des mécanismes de non-réponse, l'analyse partira du principe que, pour tous les instruments, les mécanismes de réponse ont donné suite à une répartition des « données manquantes de façon complètement aléatoire » (MCAR) – pour plus de précisions,

veuillez consulter les travaux de Rubin (1976[9]). Cela revient à supposer que l'attrition (ou la non-réponse) n'a pas concerné certains groupes d'élèves plus que d'autres, ou qu'il n'y a eu aucun biais lié à la sélection. Dans la majorité des cas où des classes ou établissements avaient abandonné en cours de projet ou avant le lancement de la collecte de données, les équipes locales en avaient rapidement informé l'OCDE en fournissant également des explications. Par exemple, ces abandons de la part de classes ou d'établissements entiers s'expliquaient dans quelques équipes par le fait qu'ils s'étaient engagés dans de trop nombreux projets de recherche et que leur conseil d'administration leur demandait de se retirer de la majorité d'entre eux. Dans bien d'autres cas, les raisons de l'absence de données post-expérience étaient dues à des difficultés d'ordre opérationnel et à une mauvaise utilisation des instruments, et n'étaient donc pas liées aux caractéristiques des élèves.

S'agissant des tests de STIM et d'AVM, les scores obtenus par certains élèves ont été exclus de l'analyse car leur taux de réponse aux items était inférieur au seuil convenu. Cette décision avait pour but d'exclure les scores susceptibles d'être entièrement attribuables aux (faibles) efforts déployés pour passer le test plutôt qu'aux capacités des élèves. Fixé à 70 %, ce seuil a été établi en se fondant sur la valeur la plus faible possible tout en maintenant, dans des limites acceptables, la perte de données qui en résulte. Cela implique, par exemple, que si sur 20 items un élève en laissait 7 ou plus sans réponse, son score n'était pas considéré comme fiable et était donc exclu de l'analyse. La perte globale de données pour le test de STIM s'élevait à 8 % pour les élèves de l'enseignement primaire contre 6 % pour ceux de l'enseignement secondaire. Tandis que pour le test d'AVM, cette perte globale de données était estimée à 20 % pour les élèves du primaire contre 4 % pour ceux du secondaire. Toutefois, la majorité de ces pertes de données a été observée dans des équipes qui rencontraient également des problèmes d'ordre opérationnel dans ce domaine.

Sur les 17 291 élèves qui ont participé à au moins une évaluation, 12 265 d'entre eux ont complété au moins un instrument pré- et post-expérience. 5 703 élèves étaient scolarisés dans l'enseignement primaire et 6 562 élèves dans le secondaire (voir le tableau 7.1). Cela correspond à un taux global de réponse estimé à 71 %, avec une faible différence entre les deux niveaux d'enseignement : 75 % parmi les élèves du primaire contre 68 % parmi les élèves du secondaire. Le taux de réponse le plus élevé a été observé dans l'équipe française (CRI) et le plus faible dans l'équipe indienne, avec respectivement 98 et 28 %. En termes de données pré- et post-expérience disponibles pour chacun des instruments, les équipes ont collecté 8 986 questionnaires, 7 953 tests EPoC d'évaluation de la créativité, 7 376 tests de STIM et 1 500 tests d'AVM (avec pour ces deux tests de performance moins de 30 % de valeurs manquantes), et avec des taux de réponse estimés, respectivement, à 67, 75, 62 et 50 %.

Tableau 7.1. Nombre d'élèves ayant complété un instrument au début du projet et pourcentage de ceux qui ont également complété le même instrument à la fin du projet, selon l'équipe

	Questionnaires	Tests EPoC d'évaluation de la créativité	Tests de STIM	Tests d'AVM	Tout instrument
Équipe brésilienne	1 119 (51 %)	628 (90 %)	981 (31 %)	x	1 248 (62 %)
Équipe britannique (Pays de Galles)	791 (75 %)	821 (89 %)	725 (86 %)	x	852 (91 %)
Équipe néerlandaise	852 (69 %)	652 (56 %)	487 (63 %)	348 (75 %)	874 (73 %)
Équipe française (CRI)	325 (96 %)	204 (99 %)	319 (97 %)	x	345 (98 %)
Équipe française (Lamap)	207 (0 %)	361 (97 %)	201 (19 %)	x	364 (97 %)
Équipe hongroise	1 272 (89 %)	1 214 (62 %)	1 286 (87 %)	x	1 534 (85 %)
Équipe indienne	999 (31 %)	x	1 280 (25 %)	x	1 793 (28 %)
Équipe russe	860 (66 %)	1 310 (64 %)	1 547 (41 %)	740 (0 %)	2 122 (50 %)
Équipe slovaque	563 (63 %)	619 (90 %)	423 (61 %)	457 (64 %)	652 (88 %)
Équipe espagnole (Madrid)	467 (0 %)	x	361 (74 %)	x	670 (51 %)
Équipe thaïlandaise	4 333 (86 %)	3 645 (85 %)	3 426 (84 %)	456 (99 %)	4 590 (95 %)
Équipe américaine (Montessori)	90 0 %)	242 (38 %)	169 (53 %)	x	253 (37 %)
Équipe américaine (Vista)	1 621 51 %)	938 (41 %)	774 (30 %)	246 (45 %)	1 994 (58 %)
Total	13 499 (67 %)	10 634 (75 %)	11 979 (62 %)	2 247 (50 %)	17 291 (71 %)

Remarques : EPoC : évaluation du potentiel créatif ; STIM : sciences, technologie, ingénierie et mathématiques ; AVM : arts visuels et musique. Les données des tests de STIM et d'AVM sont uniquement celles des élèves qui ont répondu à au moins 70 % des items.

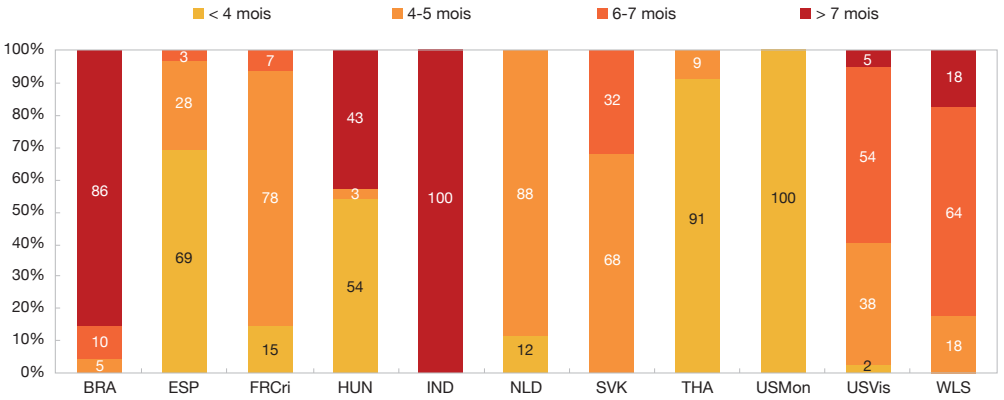
StatLink https://doi.org/10.1787/888934122171

L'expérience auprès des élèves

Le protocole de recherche recommandait que l'expérience auprès des élèves soit menée entre les mesures pré- et post-expérience, et que six à sept mois séparent ces deux mesures. Malheureusement, peu d'équipes sont parvenues à respecter cette recommandation du protocole de recherche. Comme l'illustre le graphique 7.1., seules quatre équipes ont réussi à établir une période de six mois ou plus entre les collectes de données pré- et post-expérience pour au moins 50 % de leurs élèves, tandis que d'autres équipes ne disposaient que d'une période de trois mois. Certaines équipes ne figurent pas dans le graphique 7.1. dans la mesure

où cette information n'était pas disponible. Aucune différence notable n'a été observée entre les différents niveaux d'enseignement.

Graphique 7.1. Pourcentage d'élèves selon l'équipe et la durée écoulée entre les collectes de données pré- et post-expérience

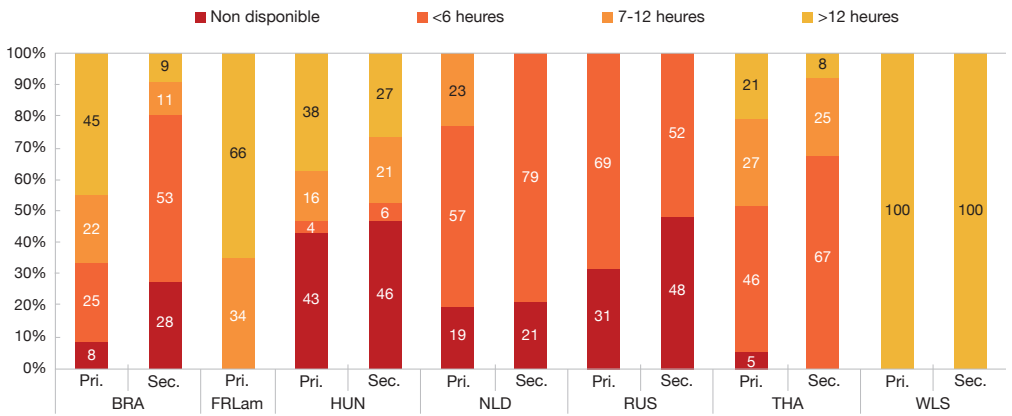


StatLink https://doi.org/10.1787/888934122190

Dans le cadre du présent chapitre, on entend par durée de l'expérience le nombre d'heures consacrées par les élèves aux nouvelles activités pédagogiques. Cette durée est très différente et bien plus courte que la durée du projet du point de vue des enseignants. En effet, pour les enseignants, l'expérience comportait également des réunions dans le cadre du plan de formation professionnelle et impliquait le fait de consacrer du temps en dehors des cours à réfléchir aux nouveaux instruments (comme le référentiel de compétences ; voir le chapitre 2) ainsi qu'à concevoir et mettre au point les nouvelles activités pédagogiques.

Le protocole de recherche n'émettait aucune recommandation explicite en termes de durée de l'expérience auprès des élèves, dans la mesure où les équipes locales devaient s'adapter à des contextes scolaires très différents. En effet, des différences notables ont pu être observées entre les équipes (voir le graphique 7.2.). Il est intéressant de constater qu'au niveau de l'enseignement primaire, l'expérience a duré plus longtemps pour la quasi-totalité des équipes pour lesquelles nous disposons de données. L'une des explications possibles pourrait reposer sur le nombre d'heures d'enseignement de chaque enseignant avec la même classe, qui est bien supérieur dans les établissements d'enseignement primaire que secondaire. Ainsi, plus d'heures de cours avec la même classe offrent aux enseignants davantage de flexibilité dans l'organisation de leurs activités pédagogiques, d'où le fait que les enseignants en poste dans le primaire soient plus susceptibles de consacrer davantage de temps au projet que leurs homologues du secondaire. Parmi les autres facteurs ayant pu jouer un rôle dans ce contexte figure le fait que les expériences se sont déroulées dans différentes disciplines (avec des temps d'instruction très variés) et qu'il existe diverses cultures pédagogiques dans l'enseignement primaire et secondaire.

Graphique 7.2. Pourcentage d'élèves selon la durée de l'expérience auprès des élèves, l'équipe et le niveau d'enseignement



StatLink https://doi.org/10.1787/888934122209

Selon les graphiques 7.1. et 7.2., et les données dont nous disposons, moins de quatre mois se sont écoulés entre les mesures pré- et post-expérience pour environ 24 % des élèves, et l'expérience a duré moins de six mois pour pratiquement 55 % des élèves. Si la prudence est de mise lors de l'interprétation de ces graphiques étant donné que les équipes n'ont pas toujours mesuré la durée des expériences auprès des élèves, ils nous permettent néanmoins d'affirmer que l'exposition des élèves aux nouvelles pédagogies a été relativement brève. (De prime abord, cette moyenne correspond à 1 % du temps moyen d'instruction sur une période de six mois dans les pays participants [données adaptées de l'OCDE (2018[10])]). Cela met en lumière le caractère pilote de l'étude et nous rappelle que l'objectif de cette phase ne consistait pas à évaluer l'efficacité de l'expérience, mais bien à concevoir des instruments et à les expérimenter sur le terrain. C'est pourquoi, même lorsque suffisamment de données ont été collectées, le véritable impact de l'expérience pourrait avoir été surestimé compte tenu du peu de temps qui s'est écoulé entre les deux mesures et de l'exposition limitée des élèves à l'expérience. En effet, même en permettant aux pratiques pédagogiques innovantes d'influer sur une partie du temps d'instruction restant et de s'y diffuser, au moins 90 à 95 % de ce temps d'instruction sera toujours consacré aux pratiques pédagogiques établies.

Caractéristiques de la population visée par l'étude

Cette section examine les principales caractéristiques de la population ayant participé au projet OCDE-CERI. Afin de mieux contextualiser les diverses réalités dans lesquelles chaque équipe a mené l'expérience, les données de l'enquête PISA 2015 (OCDE, 2015[11]) ont été utilisées comme valeur de référence. Si ces données ont été collectées en 2015, soit deux à trois ans avant la collecte de données du projet OCDE-CERI, elles fournissent néanmoins des estimations représentatives à l'échelon national de certaines variables également prises en compte dans cette phase pilote. Inclure ces données permet aux lecteurs de comparer les

résultats des échantillons participant au projet à ceux de leur population nationale respective. Cependant, ces comparaisons ne devraient être utilisées qu'à titre indicatif, dans la mesure où il n'était pas demandé aux équipes de travailler avec des échantillons représentatifs à l'échelle nationale.

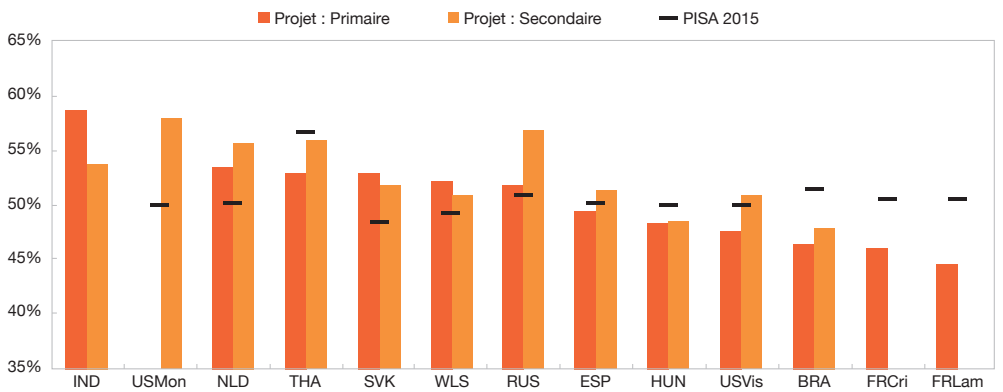
Âge

Le protocole de recherche recommandait aux équipes locales d'inclure dans leurs échantillons des classes de 3e année pour les élèves de l'enseignement primaire, et de 8e pour ceux de l'enseignement secondaire. Néanmoins, l'âge des élèves variait toujours quelque peu entre les différentes équipes. Pour les élèves du primaire, l'âge moyen dans les équipes était de 8.8 ans, avec un âge moyen minimum de 8 ans pour l'équipe américaine (Montessori) et maximum de 10.1 ans pour l'équipe brésilienne. Pour les élèves du secondaire, l'âge moyen dans les équipes était de 13.5 ans, avec un âge moyen minimum de 12.5 ans pour l'équipe indienne et maximum de 14.1 ans pour l'équipe russe. Un très faible nombre d'élèves présents dans l'échantillon étaient scolarisés dans le deuxième cycle de l'enseignement secondaire (174 dans l'équipe brésilienne et 42 dans l'équipe hongroise, soit 216 au total). En raison de la petite taille de ce groupe, ces élèves ont été intégrés dans l'analyse des élèves de l'enseignement secondaire.

Sexe

Dans les équipes participantes, le pourcentage de filles au sein des échantillons était relativement uniforme, allant de 45 % dans l'équipe française (Lamap) à 58 % dans l'équipe américaine (Montessori). Le graphique 7.3. présente le pourcentage de filles dans les équipes participantes selon le niveau d'enseignement. À l'exception des équipes indienne, russe et thaïlandaise, ce pourcentage était similaire dans les deux niveaux d'enseignement pour la majorité des équipes.

Graphique 7.3. Pourcentage de filles dans les différents échantillons et les données de référence de l'enquête PISA 2015, selon l'équipe et le niveau d'enseignement

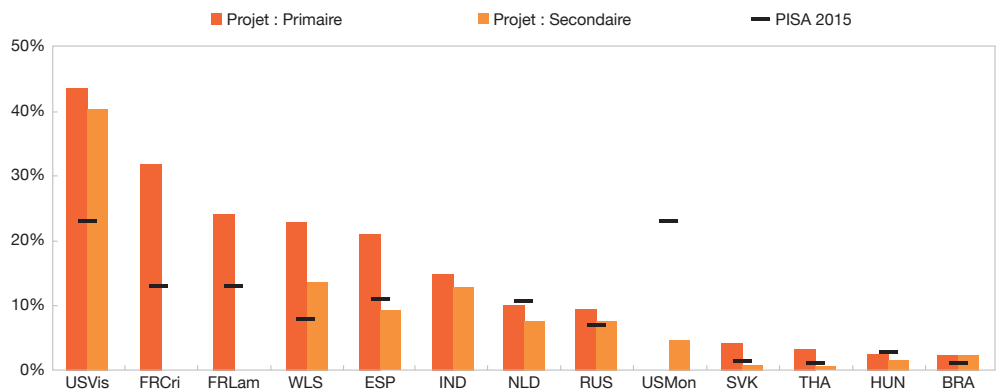


StatLink https://doi.org/10.1787/888934122228

Statut au regard de l'immigration

Conformément à la définition PISA (OCDE, 2015[12]), on entend par élève issu de l'immigration celui dont les deux parents sont nés à l'étranger (et ce, peu importe le pays de naissance de l'élève). Dans cette étude pilote, le pourcentage d'élèves issus de l'immigration variait grandement d'une équipe à l'autre, allant au minimum de 1 % dans l'équipe thaïlandaise à, au maximum, 44 % dans l'équipe américaine (Vista) (voir le graphique7.4.). Pour quelques équipes, ce pourcentage était sensiblement supérieur à celui des données tirées de l'enquête PISA 2015, mais aucune exigence n'avait été définie pour que les caractéristiques des populations échantillonnées soient similaires à celles des populations nationales respectives. Parmi les raisons pouvant expliquer ces différences figurent entre autres : le fait que les équipes participant à ce projet travaillaient dans des contextes plus diversifiés en termes de statut au regard de l'immigration que les établissements d'enseignement standards dans leur pays ; ou une véritable augmentation des effectifs d'élèves issus de l'immigration, qui ont été mesurés dans le cadre de ce projet trois à quatre ans après la collecte de données de l'enquête PISA 2015.

Graphique 7.4. Pourcentage d'élèves issus de l'immigration dans les différents échantillons et les données de référence de l'enquête PISA 2015, selon l'équipe et le niveau d'enseignement



StatLink https://doi.org/10.1787/888934122247

Le graphique 7.4. met en évidence une très faible présence d'élèves issus de l'immigration dans certaines équipes (inférieure à 2.5 %), du moins si l'on se réfère à la définition PISA. Toutefois, cette définition ne tient compte que du pays de naissance des parents de l'élève (en tant que couple). Lorsqu'on utilise les informations sur le pays de naissance des élèves et de chacun de leurs parents (obtenues grâce au questionnaire « Élève »), les données font ressortir plusieurs profils d'élèves différents, et ces différences sont d'autant plus grandes que sont prises en compte les données fournies par la variable relative à la langue principale parlée en famille par les élèves. Les options de réponse dans ce cas étaient les suivantes :

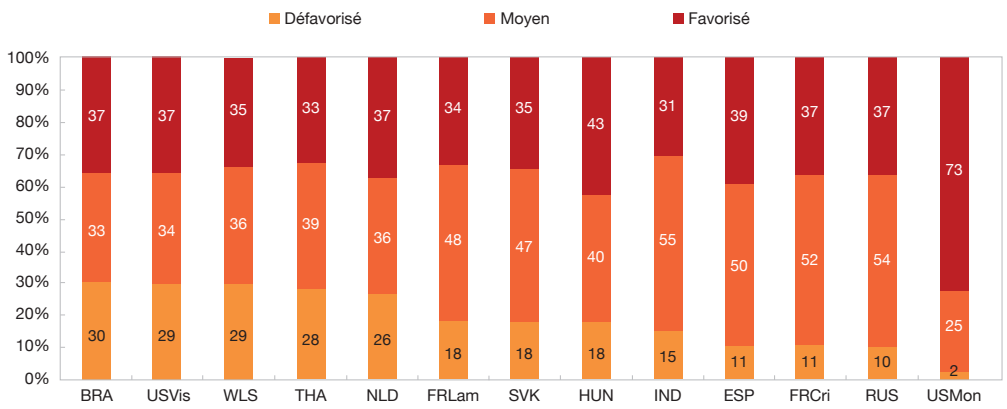
« Langue principale utilisée dans le pays », « Langue secondaire utilisée dans le pays » et « Langue étrangère ».

Afin de disposer de résultats pertinents pour l'ensemble des équipes, un nouvel indice décrivant le statut des élèves au regard de l'immigration a été créé grâce à toutes les informations fournies par les variables relatives au pays de naissance des élèves et de leurs parents, et à la langue parlée en famille par les élèves. L'annexe technique illustre les différences entre la variable de l'enquête PISA et le nouvel indice. Pour le reste de l'analyse, c'est ce nouvel indice qui sera utilisé comme variable pour décrire le statut des élèves au regard de l'immigration.

Milieu socio-économique

L'indice du milieu socio-économique a été conçu pour chaque pays et chaque niveau d'enseignement. Pour l'enseignement primaire, l'indice ne comportait que les informations relatives à la présence d'une bibliothèque familiale. Tandis que pour l'enseignement secondaire, il comportait également des données sur le niveau de formation le plus élevé des parents. L'indice répartissait les élèves en trois groupes en fonction de leur milieu socio-économique de base (défavorisé, moyen, favorisé), et avait comme objectif d'inclure au moins 15 % des élèves dans les catégories dites défavorisées et favorisées pour chacun des pays. Toutefois, cela n'a pas toujours été possible en raison de la nature profondément discrète de cette information et car les pourcentages d'élèves dans ces deux groupes variaient parfois considérablement d'une équipe à l'autre (voir le graphique 7.5.). La part d'élèves issus d'un milieu socio-économique défavorisé allait de 2 % dans l'équipe américaine (Montessori) à 30 % dans l'équipe brésilienne, tandis que pour ceux issus d'un milieu socio-économique favorisé cette part allait de 31 % dans l'équipe indienne à 73 % dans l'équipe américaine (Montessori).

Graphique 7.5. Pourcentage d'élèves issus d'un milieu socio-économique favorisé ou défavorisé, selon l'équipe



StatLink https://doi.org/10.1787/888934122266

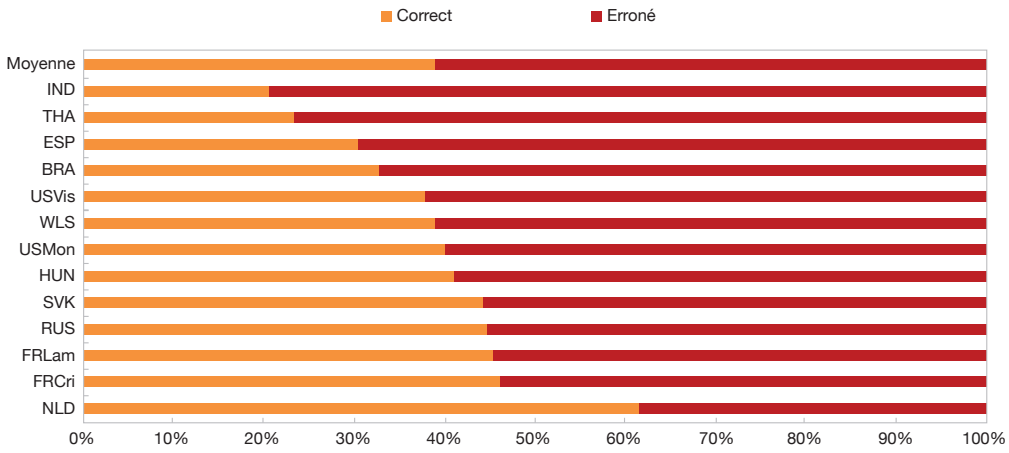
Compréhension initiale relative à la créativité et à l'esprit critique

Le questionnaire « Élève » contenait deux séries de capsules d'ancrage qui permettaient d'examiner l'opinion que les élèves avaient de leur propre créativité et esprit critique (le langage utilisé pour les élèves du primaire consistait en une version simplifiée de celui employé pour les élèves du secondaire). Pour chacune de ces compétences, les capsules décrivaient trois personnages affichant différents niveaux de créativité ou d'esprit critique. Les élèves ont premièrement été invités à évaluer le niveau de créativité ou d'esprit critique des personnages (allant de « Pas du tout » à « Extrêmement »), puis à s'identifier à l'un des personnages. Cet exercice a permis d'évaluer le degré de compréhension que les élèves avaient de ces compétences, en observant ceux qui avaient correctement classé les différentes capsules.[2] En outre, il a été possible d'évaluer l'opinion, tant relative qu'absolue, qu'ils avaient de leur propre créativité et esprit critique. On entendait par opinion relative, le niveau de créativité ou d'esprit critique que les élèves attribuaient au personnage auquel ils s'identifiaient. Par exemple, s'ils estimaient que le personnage auquel ils s'identifiaient était « Très créatif », leur opinion relative se traduirait alors par la catégorie de réponse « Très créatif ». En revanche, on entendait par opinion absolue le niveau de créativité ou esprit critique *a priori* de ce personnage, tel que défini lors de la conception des capsules. Si les élèves s'identifiaient au personnage dont la créativité était la plus faible, par exemple, leur opinion absolue se traduirait alors par la catégorie de réponse « Peu créatif », indépendamment du niveau de créativité qu'ils avaient attribué à ce personnage. Lorsque les élèves avaient une parfaite compréhension des notions de créativité et d'esprit critique, la corrélation entre l'opinion relative et l'opinion absolue devait être proche de 1.

Le graphique *7.6.* montre les divers degrés de compréhension qu'avaient les élèves concernant la créativité dans les différentes équipes en fonction de leur classement des capsules. En moyenne, le pourcentage d'élèves qui avaient réussi à classer correctement les trois capsules relatives à la créativité au début du projet s'élevait à environ 40 %, avec toutefois des différences substantielles d'une équipe à l'autre, allant de 61 % pour l'équipe néerlandaise à 21 % pour l'équipe indienne. À l'exception de la situation aux Pays-Bas, les pourcentages les plus élevés avoisinaient tous les 45 %, et aucune différence notable n'a été observée entre les élèves de l'enseignement primaire et ceux de l'enseignement secondaire.

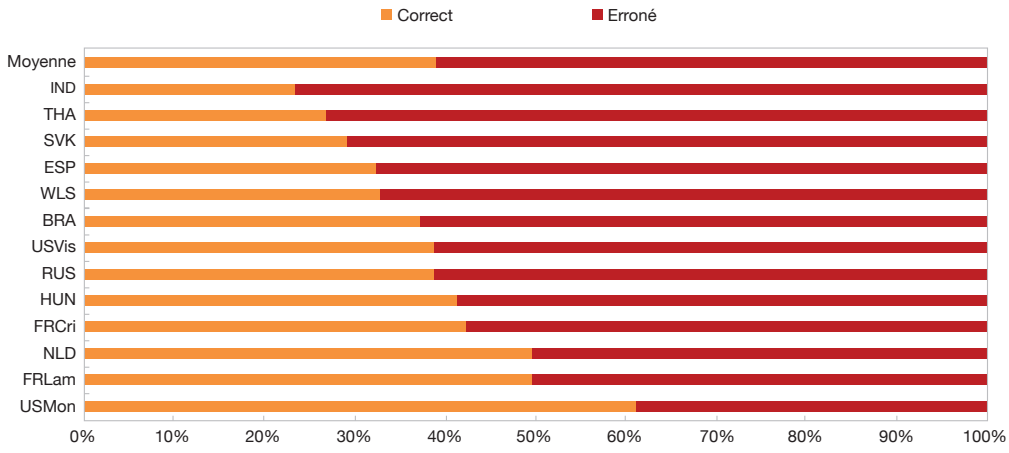
Le graphique *7.7.* montre, quant à lui, le pourcentage d'élèves ayant classé correctement les capsules relatives à l'esprit critique au début du projet. Le pourcentage le plus élevé a été observé dans l'équipe américaine (Montessori) et le plus faible dans l'équipe indienne, avec respectivement 61 et 23 %. En moyenne, le pourcentage s'élevait à environ 40 %, avec toutefois un écart très important entre les différents niveaux d'enseignement : la moyenne étant estimée à 47 % pour les élèves du secondaire contre seulement 30 % pour les élèves du primaire. Il ressort de cette différence et de toutes celles qui sont apparues lors de l'examen des plages de pourcentages (de 28 à 61 % pour les élèves du secondaire contre 17 à 50 % pour les élèves du primaire) que les élèves du primaire n'avaient pas une bonne compréhension des différents niveaux d'esprit critique présentés dans les capsules.

Graphique 7.6. Pourcentage d'élèves ayant classé correctement les capsules relatives à la créativité au début du projet, selon l'équipe



StatLink https://doi.org/10.1787/888934122285

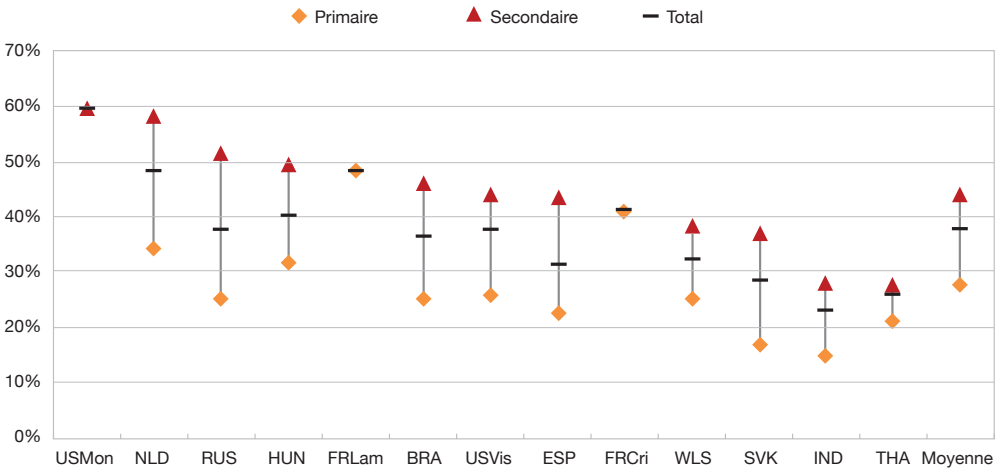
Graphique 7.7. Pourcentage d'élèves ayant classé correctement les capsules relatives à l'esprit critique au début du projet, selon l'équipe



StatLink https://doi.org/10.1787/888934122304

Le graphique 7.8. présente les mêmes informations que le graphique 7.7., mais ventile les données selon le niveau d'enseignement, afin de faire apparaître les différences entre les élèves de l'enseignement primaire et secondaire concernant le classement des capsules relatives à l'esprit critique.

Graphique 7.8. Pourcentage d'élèves ayant classé correctement les capsules relatives à l'esprit critique au début du projet, selon l'équipe et le niveau d'enseignement

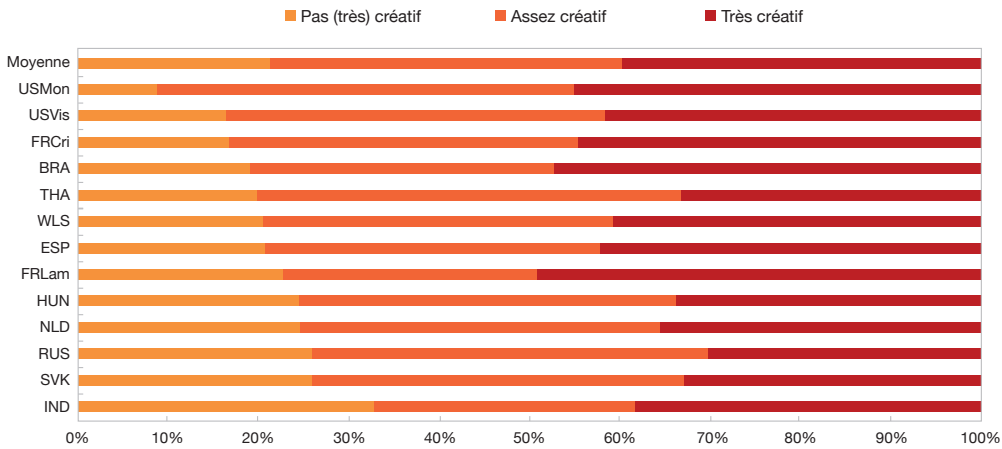


StatLink https://doi.org/10.1787/888934122323

Opinion que les élèves ont de leur propre créativité et esprit critique

Comme mentionné précédemment, les capsules peuvent également servir à analyser l'opinion que les élèves ont de leur propre créativité et esprit critique. S'agissant de l'opinion qu'ils avaient de leur créativité (voir le graphique 7.9.), le pourcentage le plus élevé d'élèves à s'être identifiés comme étant très créatifs a été enregistré dans l'équipe française (Lamap) (49 %), tandis que le pourcentage le plus faible a été observé dans l'équipe russe (30 %). Il est intéressant de noter que le pourcentage d'élèves à l'autre extrémité du spectre, c'est-à-dire ceux s'identifiant comme n'étant pas du tout créatifs, ou pas très créatifs, était similaire pour les deux équipes (23 et 26 %, respectivement), tandis qu'il variait considérablement entre toutes les autres équipes. Le pourcentage le plus faible a été observé dans l'équipe américaine (Montessori) et le plus élevé dans l'équipe indienne, avec respectivement 9 et 33 %. Toutefois, il semblerait que les équipes américaine (Montessori) et indienne faisaient figure d'exception, dans la mesure où le pourcentage d'élèves à s'identifier comme n'étant pas du tout créatifs, ou pas très créatifs, oscillait entre 17 et 26 % dans les autres équipes. La corrélation entre l'opinion relative et absolue qu'avaient les élèves de leur créativité était estimée à environ 0.3, tant au niveau de l'enseignement primaire que secondaire.

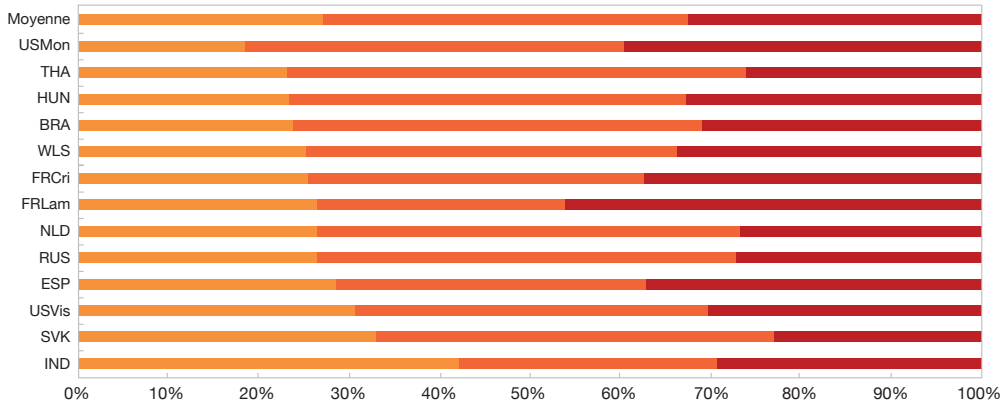
Graphique 7.9. Opinion relative qu'avaient les élèves de leur créativité au début du projet, selon l'équipe



StatLink https://doi.org/10.1787/888934122342

S'agissant de l'opinion relative qu'avaient les élèves de leur esprit critique (voir le graphique 7.10.), le pourcentage le plus élevé d'élèves à s'être identifiés comme faisant preuve d'un grand esprit critique a été enregistré dans l'équipe française (Lamap) (46 %), tandis que le pourcentage le plus faible a été observé dans l'équipe slovaque (23 %). Une variation similaire a été observée au niveau du pourcentage d'élèves à s'être identifiés comme ne faisant preuve d'aucun esprit critique, ou de peu d'esprit critique, allant de 18 % dans l'équipe américaine (Montessori) à 42 % dans l'équipe indienne. Dans le cas de l'esprit critique, l'examen de la corrélation entre l'opinion relative et absolue semble confirmer l'analyse avancée dans les paragraphes précédents, à savoir que les élèves du primaire n'avaient pas une bonne compréhension des différents niveaux d'esprit critique présentés dans les capsules. En effet, si la corrélation avoisinait 0.3 pour les élèves de l'enseignement secondaire (la même valeur que celle observée dans le cas de la créativité), elle était proche de 0 pour les élèves du primaire. Parmi les raisons pouvant expliquer cette différence pour les élèves du primaire figurent l'utilisation d'un langage simplifié pour leurs capsules (qui peut ne pas avoir été aussi efficace qu'escompté), le processus de développement naturel des enfants (la psychologie du développement ayant démontré que la pensée abstraite a tendance à se développer durant l'adolescence) ou l'existence d'une meilleure compréhension, commune aux enfants et adolescents, de la créativité que de l'esprit critique.

Graphique 7.10. Opinion relative qu'avaient les élèves de leur esprit critique au début du projet, selon l'équipe



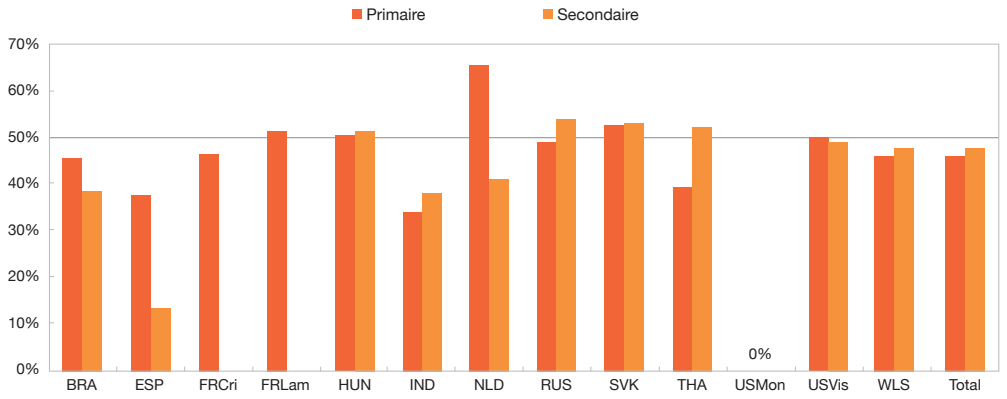
StatLink https://doi.org/10.1787/888934122361

Groupe expérimental et groupe de contrôle

Jusqu'ici, l'analyse descriptive a porté sur l'ensemble des élèves ayant pris part au projet. Toutefois, comme mentionné précédemment, le projet a été construit autour d'un modèle quasi-expérimental impliquant la répartition des élèves en deux groupes : un groupe de contrôle et un groupe expérimental. Le protocole de recherche recommandait la sélection de groupes de contrôle sensiblement comparables aux groupes expérimentaux, notamment sur le plan des résultats scolaires et du milieu socio-économique. Il demandait également que ces groupes soient de taille similaire.

Des différences significatives ont été observées entre ces groupes lors de l'analyse de leurs profils effectuée par équipe et niveau d'enseignement. Cela a notamment été le cas pour la répartition sous-jacente de certaines des principales variables socio-démographiques clés et pour la taille des échantillons des deux groupes. Dans l'ensemble, sur les 7 620 élèves du primaire participant au projet, ils étaient 4 134 à être répartis dans le groupe expérimental contre 3 486 dans le groupe de contrôle, ce dernier représentant donc, en moyenne, 46 % des élèves. Ce pourcentage oscillait toutefois entre 34 et 66 % dans les différentes équipes (voir le graphique 7.11.). Par ailleurs, sur les 9 657 élèves du secondaire participant au projet, 5 099 d'entre eux étaient dans le groupe expérimental contre 4 558 dans le groupe de contrôle, ce dernier représentant donc, en moyenne, 47 % des élèves. Comme pour les élèves de l'enseignement primaire, le pourcentage du groupe de contrôle variait considérablement d'une équipe à l'autre, allant de 13 % à 54 %. Dans la majorité des cas où on observait un déséquilibre important entre les deux groupes, cela était dû à divers types de problèmes rencontrés à l'occasion de la procédure de sélection des établissements d'enseignement ou des activités de collecte des données.

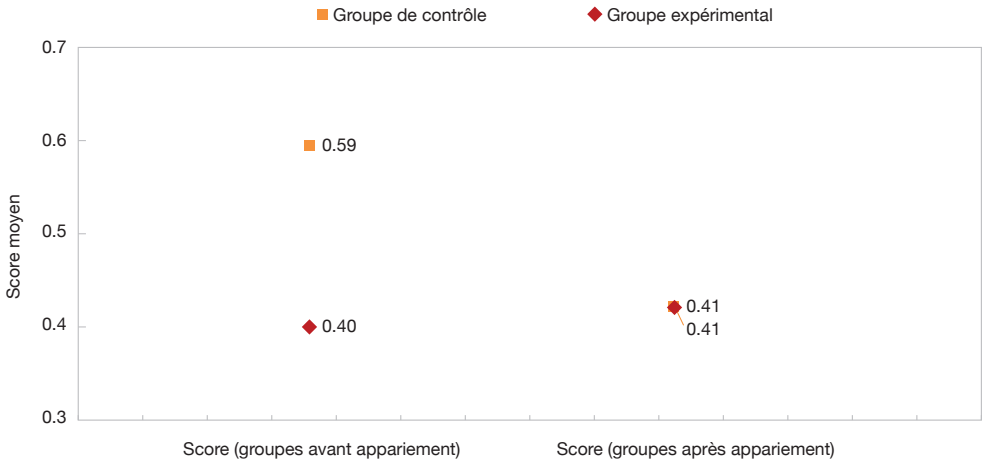
Graphique 7.11. Pourcentage d'élèves dans le groupe de contrôle, selon l'équipe et le niveau d'enseignement



StatLink https://doi.org/10.1787/888934122380

En termes de variables socio-démographiques, les équipes locales avaient été incitées à sélectionner des établissements, des enseignants et des élèves provenant, dans la mesure du possible, d'une pluralité de contextes (s'agissant par exemple de la taille des établissements, du milieu socio-économique, des résultats scolaires), et à garantir la comparabilité entre le groupe de contrôle et le groupe expérimental. Malgré les efforts déployés par les équipes, dans les faits il s'est avéré difficile de suivre ces recommandations, et on a constaté d'importantes différences entre les deux groupes. Afin de réduire au minimum l'impact de ces différences sur les conclusions du projet, la première étape de l'analyse s'est traduite par la mise en œuvre d'une méthode d'appariement des coefficients de propension (*propensity score matching*) (Rosenbaum et Rubin, 1983[13]). Cette technique consiste à réaligner les situations initiales des deux échantillons en attribuant aux élèves du groupe de contrôle différents coefficients de pondération.[3] Un ensemble spécifique de coefficients de pondération a été calculé pour chacun des instruments d'enquête. Pour le questionnaire, le calcul de ces coefficients visait à corriger des déséquilibres potentiels relatifs au sexe, au milieu socio-économique et à l'âge. Pour les tests de performance, l'objectif consistait également à corriger les déséquilibres potentiels dans les données initiales de la principale variable à l'étude (à savoir les scores au test EPoC d'évaluation de la créativité, au test de STIM et au test d'AVM). Le graphique 7.12. présente un exemple de l'effet de la méthode d'appariement des coefficients de propension.

Graphique 7.12. Exemple de l'effet de l'appariement des coefficients de propension sur les scores au test de STIM au début du projet pour les élèves du primaire de l'équipe thaïlandaise



StatLink https://doi.org/10.1787/888934122399

L'appariement des coefficients de propension visait à obtenir un groupe de contrôle et un groupe expérimental qui, après application de coefficients de pondération, pouvaient être considérés comme globalement équivalents (au début du projet). Le principal inconvénient de cette méthode est qu'elle implique une perte de données lorsque des élèves n'ont répondu à aucune des variables explicatives utilisées pour le calcul des coefficients de propension. Dans ce cas, il devient impossible de calculer le score de propension de ces élèves, qui ont donc été exclus de l'analyse. Cependant, cette exclusion se serait produite quoi qu'il en soit dans le cadre de toute analyse englobant les variables explicatives utilisées pour l'appariement. Afin de réduire au minimum la perte de données, l'appariement des coefficients de propension n'a pris en compte que les variables énumérées dans le paragraphe précédent (et dont la plupart sont disponibles pour l'ensemble des élèves), et a parfois privilégié des déséquilibres mineurs entre les groupes à l'exclusion d'un nombre significatif d'élèves de l'analyse. À titre indicatif, la perte de données imputable à cette méthode d'appariement s'élevait à 3 % de l'ensemble des données pour les questionnaires, à 11 % s'agissant des tests EPoC, à 6 % pour les tests de STIM et à 4 % pour les tests d'AVM. Veuillez consulter l'annexe technique pour de plus amples informations concernant cette analyse.

Mesure des effets de l'expérience auprès des élèves

L'objectif de cette phase pilote consistait principalement à concevoir des instruments pour une éventuelle validation de l'étude. Ce processus incluait la collecte effective des données et leur analyse afin, dans un premier temps, d'évaluer la validité des instruments puis de comprendre les effets de l'expérience, même ceux présentant une faible puissance statistique.

Les sections suivantes abordent les résultats associés à l'appartenance au groupe expérimental sur les élèves, et les facteurs qui ont influencé de manière uniforme ces résultats dans les différents pays. La première section présente les méthodes utilisées pour l'analyse, tandis que la deuxième porte sur les résultats observés de façon constante dans l'ensemble des pays pour tous les élèves. Enfin, une troisième section se concentre sur des sous-groupes d'élèves particuliers afin de révéler si l'expérience a eu des effets différents pour certaines sous-populations. Tous les résultats rendent compte des facteurs de confusion éventuels d'un ensemble de 13 variables explicatives.

Méthodologie

Les données ont été obtenues à partir d'un ensemble de modèles à plusieurs variables qui examinaient l'effet de l'expérience auprès des élèves en termes d'évolution pré- et post-expérience de plusieurs résultats d'intérêt. Tous les modèles ont été calculés avec des erreurs-types robustes groupées au niveau des établissements, en tenant compte par conséquent de la structure hiérarchique des données. De plus, ils incluaient tous un ensemble de variables de contrôle, telles que l'âge, le sexe, le milieu socio-économique, la discipline et la valeur des résultats d'intérêt au début du projet. Figuraient également parmi les variables de contrôle, le temps écoulé entre les collectes de données pré- et post-expérience ainsi que la durée de l'expérience auprès des élèves, lorsque ces informations étaient disponibles.

En raison de la forte disparité des travaux menés sur le terrain par les différentes équipes, l'analyse s'est limitée à un examen des résultats au regard des tendances positives ou négatives des conclusions. L'actuelle phase pilote a constitué une étude de faisabilité, de telle sorte que la véritable portée des conclusions n'ait qu'une importance limitée pour l'évaluation de l'efficacité de l'expérience auprès des élèves. Ainsi, au lieu d'être axée sur la taille des différents coefficients, l'analyse se concentrait davantage sur les tendances pouvant être observées dans les diverses équipes et aux différents niveaux d'enseignement.

Le seuil de signification statistique n'a pas été établi au niveau habituel estimé à 0.05. Proposé initialement par Fisher en 1926, ce seuil de 0.05 signifie que sur 100 essais le résultat d'intérêt sera observé à au moins 95 reprises. En d'autres termes, dans 1 essai sur 20 la conclusion s'avérerait erronée. Comme l'avait également fait remarquer ce même auteur, ce seuil avait été jugé approprié pour l'instauration de faits scientifiques « établis empiriquement » (Fisher,

1926[14]), ce qui ne constituait pas un des objectifs de la phase pilote. Par ailleurs, comme l'affirmait l'*American Statistical Association* en 2016, « les conclusions scientifiques, les orientations politiques ou décisions commerciales ne devraient pas être prises uniquement sur la base d'une valeur prédictive supérieure à un certain seuil » (Wasserstein et Lazar, 2016[15]).

Sur proposition de ces deux auteurs d'utiliser la signification statistique « comme un outil pour indiquer si un résultat doit faire l'objet d'un examen plus approfondi » et en vue de se conformer au caractère exploratoire et informatif – plutôt qu'évaluatif – du présent rapport, le seuil de signification statistique a été fixé à 0.2. Dans le cadre de la présente analyse, cela signifie que les résultats présentés comme statistiquement significatifs seront observés dans au moins quatre essais sur cinq. Toutes les considérations portant sur la nécessité d'une définition plus restrictive de la signification statistique seront abordées dans le cadre de la future étude de validation, le cas échéant, comme le seront les questions relatives à l'ampleur d'effets pris isolément. À titre d'information, la part des résultats présentés dans le tableau 7.2. affichant également une signification statistique à 0.1 s'élevait à pratiquement 80 %, tandis qu'elle était estimée à 65 % pour les résultats mentionnés dans le tableau 7.3.

Cette phase pilote a mis à disposition une myriade de données, en collectant jusqu'à plus de 2 000 variables pour chaque élève participant. Aux fins du présent chapitre, l'accent a été mis sur 36 résultats : 18 provenant du questionnaire (dont 8 de la section relative aux capsules) et les 18 autres étant tirés du test EPoC d'évaluation de la créativité, du test de STIM et du test d'AVM (6 résultats pour chaque test). Les effets de 29 variables sur les résultats d'intérêt ont fait l'objet d'une analyse initiale afin de sélectionner les variables explicatives les plus pertinentes. Sur ces 29 variables, 13 ont été retenues, dont une portait sur le temps écoulé entre les collectes de données pré- et post-expérience tandis que les autres étaient organisées en trois groupes principaux concernant, respectivement, le milieu d'origine des élèves, leurs réponses aux épreuves relatives aux capsules, et les pratiques et convictions de leurs enseignants.

L'analyse finale était composée de modèles par niveau d'enseignement et par équipe examinant : 1) l'effet de l'expérience auprès des élèves, après prise en compte de l'ensemble des variables de contrôle mentionnées précédemment ; et 2) l'effet de l'interaction des 13 variables explicatives à l'étude avec l'expérience (tout en conservant les variables de contrôle dans les modèles). Le nombre de modèles variait considérablement d'une équipe à l'autre, à cause des différences observées dans la disponibilité des données. Pour le point 1), le nombre de modèles allait de 34 pour l'équipe slovaque à 3 pour l'équipe française (Lamap), tandis que pour le point 2) il allait de 413 pour l'équipe thaïlandaise à 36 pour l'équipe française (Lamap). Les équipes américaine (Montessori) et espagnole ont été exclues de l'analyse à plusieurs variables en raison de problèmes de disponibilité des données. Le tableau 7.2. illustre les conclusions relatives au point 1).

Tableau 7.2. Résultats statistiquement significatifs (positifs et négatifs) associés à l'effet de l'expérience auprès des élèves

Instrument	Indice ou item	Modèles avec résultats positifs	Modèles avec résultats négatifs	Nombre total de modèles
Test de STIM	Pratiques pédagogiques en cours de STIM (P)	1	1	8
	Pratiques pédagogiques en cours de STIM (S)	3	0	6
	Intérêt pour les STIM (P)	2	3	9
	Intérêt pour les STIM (S)	2	2	7
	Score (P)	4	0	9
	Score (S)	1	1	7
Test d'AVM	Pratiques pédagogiques en cours d'AVM (P)	1	0	3
	Pratiques pédagogiques en cours d'AVM (S)	1	0	2
	Intérêt pour les AVM (P)	0	0	3
	Intérêt pour les AVM (S)	2	0	2
	Score (P)	2	0	3
	Score (S)	2	0	2
Capsules	Classement des capsules CR (P)	2	0	10
	Classement des capsules CR (S)	3	1	8
	Classement des capsules EC (P)	0	0	10
	Classement des capsules EC (S)	2	2	8
	Opinion relative sur la CR (P)	4	2	10
	Opinion relative sur la CR (S)	2	3	8
	Opinion relative sur l'EC (P)	1	2	10
	Opinion relative sur l'EC (S)	3	2	8
Test EPoC	Score total (P)	1	1	8
	Score total (S)	3	0	6
	Score en pensée convergente (P)	2	3	9
	Score en pensée convergente (S)	2	2	7
	Score en pensée divergente (P)	4	0	9
	Score en pensée divergente (S)	1	1	7
Questionnaire	Dispositions à l'apprentissage (P)	1	0	3
	Dispositions à l'apprentissage (S)	1	0	2
	Sentiments positifs (P)	0	0	3
	Sentiments positifs (S)	2	0	2
	Intérêt unique (P)	2	0	3
	Intérêt unique (S)	2	0	2
	Participation des parents (P)	2	0	10
	Participation des parents (S)	3	1	8
	Sentiment d'appartenance à l'école (S)	0	0	10
	Méthode d'apprentissage (S)	2	2	8
TOTAL	Élèves du primaire	34	24	130
	Élèves du secondaire	33	25	138
	Ensemble des élèves	67	49	268

Remarques : P = primaire ; S = secondaire ; CR = créativité ; EC = esprit critique. Tous les modèles incluaient un ensemble de variables de contrôle, telles que l'âge, le sexe, le milieu socio-économique, la discipline, la valeur des résultats d'intérêt au début du projet et, le cas échéant, le temps écoulé entre les collectes de données pré- et post-expérience ainsi que la durée de l'expérience auprès des élèves. Les colonnes des résultats positifs ou négatifs comprennent les modèles pour lesquels la signification statistique de l'expérience était inférieure à 0.20.

StatLink https://doi.org/10.1787/888934122418

Résultats globaux de l'expérience auprès des élèves

Cette phase pilote avait pour but la mise en œuvre de nouvelles activités pédagogiques qui profiteraient aux élèves sur le plan de la créativité et de l'esprit critique à plusieurs égards : leur potentiel créatif, la compréhension qu'ils ont de ces concepts, l'utilisation par leurs enseignants de pratiques pédagogiques en lien avec ces compétences, leurs dispositions à l'apprentissage de ces compétences, les méthodes d'apprentissage adoptées, etc. Par ailleurs, il apparaissait important de mesurer les possibles effets de cette expérience auprès des élèves par rapport à des indicateurs solidement établis, tels que les scores obtenus aux tests de performance axés sur les disciplines des STIM ou des AVM.

Pour que le but de cette phase pilote se concrétise, les élèves dans le groupe expérimental devraient avoir davantage progressé que leurs homologues dans le groupe de contrôle concernant les résultats d'intérêt. De plus, il serait souhaitable que ces constats puissent être observés dans tous les pays, ne serait-ce que pour quelques matières, thématiques, niveaux d'enseignement ou autres variables pertinentes.

L'expérience auprès des élèves semble avoir engendré des effets positifs : ainsi, sur l'ensemble des 268 modèles mis en œuvre, 25 % ont fait état d'un effet positif statistiquement significatif contre seulement 18 % ayant constaté un effet négatif statistiquement significatif, soit un total net estimé à 7 %. L'impact global de l'expérience était similaire entre les niveaux d'enseignement, avec un total net avoisinant les 7 % pour les élèves de l'enseignement primaire et secondaire pris isolément.

Pour les élèves du primaire, l'expérience semble avoir eu un effet particulièrement bénéfique au niveau des scores obtenus aux tests de performance. D'importants effets positifs ont notamment été observés pour :

- les scores au test de STIM (pour quatre équipes sur neuf)
- les scores au test d'AVM (pour deux équipes sur trois).

Dans les deux cas, aucun modèle n'a démontré d'effets négatifs importants de l'expérience.

Pour les élèves du secondaire, les résultats ont eu tendance à être moins réguliers s'agissant des différentes variables à l'étude. D'importants effets positifs ont toutefois été observés sur les points suivants :

- l'utilisation de pratiques pédagogiques en lien avec la créativité et l'esprit critique dans les cours de STIM (pour trois équipes sur six)
- l'intérêt des élèves pour les disciplines des AVM (pour deux équipes sur deux)
- les scores au test d'AVM (pour deux équipes sur deux).

Aucun effet négatif important n'a été observé pour l'une ou l'autre de ces variables.

S'agissant des scores obtenus au test EPoC d'évaluation de la créativité, l'expérience auprès des élèves a eu des effets plutôt contrastés. S'il y a eu, dans quatre équipes sur dix, des effets positifs pour les élèves du primaire, des effets négatifs ont également été enregistrés dans trois autres équipes. En

revanche, pour les élèves du secondaire, sur sept équipes une seule a constaté des effets positifs de l'expérience, contre trois ayant fait part d'effets négatifs. Cette disparité des effets s'est maintenue même au niveau des scores partiels au test EPoC, étant donné que les scores aux épreuves de pensée divergente et pensée convergente présentaient également des résultats contrastés.

Le temps écoulé entre les mesures pré- et post-expérience a également eu un impact positif sur le nombre de résultats positifs significatifs (variable non présentée dans le tableau). En effet, plus le délai entre les mesures pré- et post-expérience était important, plus les effets observés étaient positifs. Ainsi, sur les 227 modèles qui incluaient cette variable, 33 % ont constaté un effet positif engendré par ce délai plus important, contre 15 % ayant fait part d'un effet négatif, soit un total net de 17 %. Cet effet a été uniformément observé parmi les élèves de l'enseignement primaire et secondaire, avec un total net de 15 et 19 %, respectivement. Ce constat s'inscrit dans la continuité des recommandations du protocole de recherche ainsi que des données probantes issues des travaux de recherche-action dans le domaine de l'éducation qui soulignent la nécessité de disposer d'un délai suffisamment long entre les mesures pré- et post-expérience afin de pouvoir déterminer des évolutions significatives des résultats d'intérêt. Au niveau du primaire, les scores au test EPoC constituaient le résultat le plus positif (en termes de progression) lorsque le délai entre les collectes de données pré- et post-expérience était plus important (pour trois équipes sur huit). Ces délais plus importants semblent également avoir influencé positivement certains indices relatifs aux élèves du secondaire, dont les dispositions à l'apprentissage liées à la créativité et l'esprit critique, les sentiments positifs à l'égard de l'apprentissage et la capacité à classer correctement les capsules relatives à l'esprit critique. Dans tous les cas, ces effets ont été observés dans deux ou trois équipes sur sept, tandis qu'aucun impact négatif n'a été enregistré.

Résultats pour certains sous-groupes d'élèves

L'analyse des effets de l'expérience auprès des élèves sur différents groupes d'élèves (voir le tableau 7.3) nous a conduits à réorienter notre réflexion en vue de déterminer si certaines sous-populations (les filles par exemple) avaient tout particulièrement tiré profit de l'expérience. À cette fin, les interactions entre l'expérience et les différentes sous-populations à l'étude ont été intégrées une par une dans chacun des modèles présentés dans la section précédente. Après prise en compte de l'ensemble des éléments à l'étude, les résultats ont montré que dans les différents pays l'expérience semblait invariablement avoir été plus bénéfique pour les groupes suivants :

- les élèves dont les enseignants estimaient que la créativité pouvait être enseignée dans le cadre scolaire au début de l'expérience (total net évalué à 9 %)
- les élèves qui avaient classé correctement les capsules relatives à l'esprit critique au début du projet (avec des résultats différentiels positifs dans 18 % des modèles, et négatifs dans 11 % des modèles, soit un total net estimé à 7 %)
- les élèves qui n'étaient pas parvenus à classer correctement les capsules relatives à la créativité au début du projet (total net estimé à 6 %).

Tableau 7.3. Résultats significatifs (positifs et négatifs) associés à l'effet de l'expérience auprès des élèves pour les différents sous-groupes à l'étude

Variable	Résultats positifs	Résultats négatifs	Nombre de modèles étudiés	Pourcentage de résultats positifs	Pourcentage de résultats négatifs	Total net	Total net (primaire)	Total net (secondaire)
Sexe : fille	34	44	268	13%	16%	-4%	-5%	-2%
Milieu socio-économique défavorisé	41	43	265	15%	16%	-1%	-6%	6%
Milieu socio-économique favorisé	35	39	268	13%	15%	-1%	-4%	2%
Statut au regard de l'immigration (déf. du projet)	52	33	261	20%	13%	7%	2%	13%
Mauvaise opinion rel. sur la CR (pré-exp.)	51	29	268	19%	11%	8%	3%	15%
Bonne opinion rel. sur la CR (pré-exp.)	27	34	268	10%	13%	-3%	2%	-8%
Mauvaise opinion rel. sur l'EC (pré-exp.)	38	44	267	14%	16%	-2%	-8%	5%
Bonne opinion rel. sur l'EC (pré-exp.)	35	49	266	13%	18%	-5%	-4%	-7%
Bon classement des capsules CR (pré-exp.)	29	45	268	11%	17%	-6%	-6%	-6%
Bon classement des capsules EC (pré-exp.)	48	30	267	18%	11%	7%	6%	8%
Délai accru entre la collecte de données pré- et post-exp.	39	61	220	18%	28%	-10%	-13%	-7%
Indice plus élevé de l'évolution des pratiques	34	48	256	13%	19%	-5%	-3%	-8%
Bon class. des caps. CR par les enseignants (pré-exp.)	17	16	81	21%	20%	1%	-6%	13%
Bon class. des caps. EC par les enseignants (pré-exp.)	14	3	40	35%	8%	28%	42%	6%
Enseignant estimant que la CR est enseignable (pré-exp.)	15	10	56	27%	18%	9%	4%	13%
Enseignant estimant que l'EC est enseignable (pré-exp.)	7	12	34	21%	35%	-15%	-39%	13%
Discipline de l'enseignant : STIM (vs AVM)	6	18	59	10%	31%	-20%	9%	-27%
Discipline de l'enseignant : STIM (vs autre)	9	4	27	33%	15%	19%	x	19%
Discipline de l'enseignant : AVM (vs autre)	12	17	62	19%	27%	-8%	-19%	15%

Remarques : Opinion rel. = opinion relative ; pré-exp. = pré-expérience ; CR = créativité ; EC = esprit critique ; caps. = capsules ; class. = classement. Par « Nombre de modèles étudiés », on entend le nombre de cas dans lesquels il a été possible d'examiner l'effet de l'interaction entre l'expérience et chaque variable, dans les 13 équipes et pour les 36 résultats d'intérêt. Outre l'interaction entre l'expérience et chacune des variables, tous les modèles incluaient un ensemble de variables de contrôle, telles que l'âge, le sexe, le milieu socio-économique, la discipline, la valeur des résultats d'intérêt au début du projet et, le cas échéant, le temps écoulé entre les collectes de données pré- et post-expérience ainsi que la durée de l'expérience auprès des élèves. Le groupe de référence pour le « Milieu socio-économique » et pour les variables décrivant l'opinion relative des élèves sur leur créativité et leur esprit critique au début du projet, correspond au niveau « Moyen ». Dans le cas de la « Discipline de l'enseignant », la catégorie « Autre » regroupe toutes les matières autres que celles des domaines des STIM et des AVM.

StatLink https://doi.org/10.1787/888934122437

Par ailleurs, d'autres résultats positifs ont pu être observés lors de l'analyse de chaque niveau d'enseignement. Dans le secondaire, l'expérience semble notamment avoir été plus efficace pour les groupes d'élèves suivants :

- les élèves issus de l'immigration (total net estimé à 13 %)
- les élèves qui avaient une mauvaise opinion relative, puis une opinion relative moyenne, de leur créativité au début du projet (totaux nets estimés à 23 et 8 %, respectivement)
- les élèves qui avaient une mauvaise opinion relative, puis une opinion relative moyenne, de leur esprit critique au début du projet (totaux nets estimés à 12 et 7 %, respectivement)
- les élèves dont les enseignants avaient classé correctement les capsules relatives à la créativité au début du projet (total net estimé à 13 %).

En revanche, dans l'enseignement primaire, l'expérience semble avoir été plus efficace pour les groupes d'élèves suivants :

- les élèves qui avaient une opinion relative moyenne de leur esprit critique au début du projet (total net estimé à -9 % pour la mauvaise opinion et à -4 % pour la bonne opinion)
- les élèves dont les enseignants avaient correctement classé les capsules relatives à l'esprit critique au début du projet (total net estimé à 48 %).

Il est surprenant de constater que l'expérience semble avoir eu un effet négatif pour les élèves du primaire dont les enseignants estimaient au début du projet que l'esprit critique pouvait être enseigne dans le cadre scolaire (total net évalué à -39 %), tandis qu'à cet égard un effet positif a été enregistré pour les élèves du secondaire (total net estimé à 13 %).

En ce qui a trait à la discipline, les interactions entre la matière dans laquelle se déroulait l'expérience et l'expérience elle-même étaient souvent inexistantes en raison du cadre conceptuel de l'enquête adopté par les équipes locales (par exemple, tous les enseignants proposaient l'expérience dans les mêmes matières, les enseignants dans le groupe de contrôle enseignaient une discipline tandis que ceux dans le groupe expérimental en enseignaient une autre). 121 modèles ont toutefois pu être évalués, avec 53 dans l'enseignement primaire contre 68 dans l'enseignement secondaire.[4] D'après les observations, l'expérience a semblé particulièrement bien fonctionner dans les matières autres que celles des STIM et des AVM (principalement des expériences interdisciplinaires) pour les élèves du primaire, et dans les disciplines des AVM pour les élèves du secondaire, avec un total net estimé à 19 et 42 % respectivement.

Pour les élèves du primaire, l'effet positif de l'expérience dans les projets interdisciplinaires a principalement concerné les points suivants :

- la participation des parents (dans deux modèles sur quatre)
- les sentiments positifs des élèves à l'égard de l'apprentissage (dans deux modèles sur quatre)
- la compréhension que les élèves ont de la notion de créativité (capacité à classer correctement les capsules relatives à la créativité ; dans trois modèles sur quatre)

- la curiosité des élèves (pourcentage d'élèves à n'apprendre que ce qui les intéresse ; avec une baisse observée dans deux modèles sur trois).

En revanche, pour les élèves du secondaire, les effets positifs les plus fréquents associés à une expérience dans les matières des AVM ont concerné les points suivants :

- l'opinion relative des élèves sur leur créativité et leur esprit critique (dans deux modèles sur quatre pour les deux compétences)
- les dispositions des élèves à l'apprentissage en lien avec la créativité et l'esprit critique (dans deux modèles sur quatre)
- les méthodes d'apprentissage des élèves liées à la créativité et l'esprit critique (dans deux modèles sur quatre)
- le sentiment d'appartenance des élèves à l'école (dans deux modèles sur quatre).

Aperçu de l'analyse à l'échelle de la classe

Une autre façon d'examiner les données consiste à se concentrer, non pas sur les élèves, mais sur les classes. Ce faisant, il devient possible d'utiliser d'autres données provenant des questionnaires « Enseignant » relatives aux caractéristiques des enseignants et des environnements d'apprentissage, et d'en faire le sujet de nos analyses. La majorité de ces informations avait, par ailleurs, dû être exclue de l'analyse à l'échelle des élèves en raison du nombre limité de questionnaires « Enseignant ».

Le recours aux classes individuelles comme unités à l'étude dans l'analyse permet d'identifier celles dont les résultats sont les plus prometteurs et de relever les points qu'elles ont en commun. En outre, il a été possible dans certains cas de corréler ces données à celles relatives aux nouvelles activités pédagogiques spécifiques, offrant ainsi aux lecteurs des points de repère utiles en ce qui a trait aux caractéristiques des classes et aux expériences spécifiques auprès des élèves.

Méthodologie

L'analyse des classes ne s'est concentrée que sur un faible nombre de variables étudiées, à savoir : les scores obtenus au test EPoC d'évaluation de la créativité, au test de STIM et au test d'AVM ; l'intérêt pour les matières des domaines des STIM et des AVM ; l'utilisation de pratiques pédagogiques en lien avec la créativité et l'esprit critique ; le pourcentage d'élèves n'apprenant pas uniquement les sujets pour lesquels ils montraient déjà un intérêt ; la capacité des élèves à classer correctement les capsules relatives à la créativité et l'esprit critique ; et l'adoption de méthodes d'apprentissage liées à la créativité et l'esprit critique (uniquement pour les élèves du secondaire). L'analyse a isolé les 25 % de classes ayant enregistré la plus grande évolution de ces variables entre les mesures pré- et post-expérience

(analyse menée séparément pour le groupe de contrôle et le groupe expérimental, et par niveau d'enseignement), avant de les comparer avec le reste des classes afin de repérer leurs caractéristiques distinctives.

Deux différences majeures ont été observées entre l'analyse à l'échelle de la classe et celle au niveau des élèves présentée dans les sections précédentes. Le fait qu'il ait été impossible de mener séparément une analyse à l'échelle de la classe pour chaque équipe locale, en raison du nombre très varié de classes participantes dans chaque équipe, a constitué la première de ces différences. Aux fins de cet exercice, l'analyse conjointe de l'ensemble des équipes a tout de même permis de tirer des conclusions pertinentes à partir de ces données. La seconde différence portait sur le fait que l'analyse à l'échelle de la classe prenait en compte des variables qui ne pouvaient pas être intégrées à l'analyse axée sur les élèves. Parmi ces variables figurent notamment : le nombre d'heures d'enseignement par semaine avec la classe, le degré de préparation que l'enseignant estimait posséder pour favoriser le développement de la créativité et l'esprit critique chez ses élèves, l'ancienneté de l'enseignant et le climat en classe. La note 5 présente la liste exhaustive des variables explicatives utilisées pour l'analyse à l'échelle de la classe.[5]

Dans l'ensemble, 753 classes ont participé à cette phase pilote, mais en vue de garantir des estimateurs suffisamment fiables seules celles comportant au moins cinq élèves ont été intégrées à l'analyse. L'échantillon final était donc composé de 732 classes. Les données relatives à la classe ont soit été tirées du questionnaire « Enseignant », soit obtenues à partir de moyennes calculées à l'échelle de la classe sur la base des réponses apportées au questionnaire « Élève ». Dans ce dernier cas, les moyennes ont été calculées séparément pour chaque variable à l'étude, et seulement lorsque l'une des conditions suivantes était remplie : le taux de réponse observé dans la classe s'élevait à au moins 50 % ; ou la classe avait au minimum dix réponses valides.

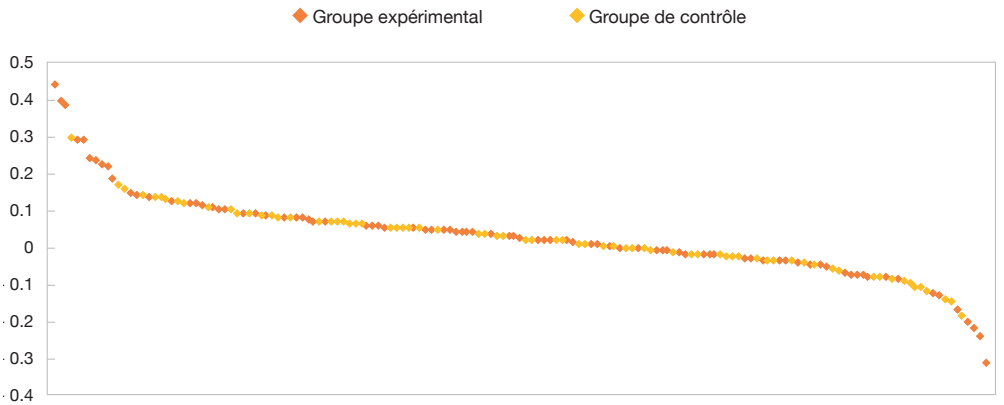
Effets de l'expérience auprès des élèves selon les résultats d'intérêt

L'examen de la répartition des classes du groupe de contrôle et du groupe expérimental en termes d'évolution des variables à l'étude entre les mesures pré- et post-expérience, a permis de mettre en évidence les variables pour lesquelles l'expérience auprès des élèves avait conduit aux résultats les plus satisfaisants. Cela a notamment été le cas pour les variables suivantes :

- les scores obtenus au test de STIM, dans l'enseignement primaire (présentés dans le graphique 7.13.)
- la capacité à classer correctement les capsules sur l'esprit critique, dans l'enseignement primaire
- l'utilisation de pratiques pédagogiques en lien avec la créativité et l'esprit critique, dans le primaire et le secondaire.

Dans le cas du graphique 7.13., par exemple, les classes ayant enregistré la plus grande évolution entre les mesures pré- et post-expérience sont situées à gauche, et il est possible de remarquer que le nombre de classes dans le groupe expérimental (représentées par un cercle) est supérieur à celui des classes dans le groupe de contrôle (représentées par une croix). Parmi les 10 premières classes, par exemple, seule 1 classe appartient au groupe de contrôle, et ce chiffre passe à 6 lorsque l'on tient compte des 20 premières classes. Il ressort de ce constat un effet positif imputable à l'expérience. Les conclusions pouvant être tirées de la simple observation d'une série de chiffres comme celle-ci corroboraient largement celles émises sur la base de l'analyse des données à l'échelle des élèves décrite dans les sections précédentes.

Graphique 7.13. Évolution des scores au test de STIM à l'échelle de la classe entre les mesures pré- et post-expérience, selon le groupe



Remarque : les classes sont triées par ordre décroissant de l'ampleur de l'évolution entre les mesures pré- et post-expérience.

StatLink https://doi.org/10.1787/888934122456

Effets de l'expérience auprès des élèves selon le profil de la classe

Élèves du primaire

Outre les conclusions précédemment citées, il est ressorti de l'analyse qu'au niveau du primaire l'expérience semblait avoir été bénéfique pour les classes qui présentaient un climat d'apprentissage difficile au début du projet.[6] L'examen des classes les plus performantes a permis de constater que la part des classes présentant au départ un climat d'apprentissage difficile était souvent plus de deux fois supérieure parmi les classes du groupe expérimental que celles du groupe de contrôle, avec une estimation supérieure à 78 % en moyenne.

On a également constaté, dans les classes les plus performantes du groupe expérimental, des scores légèrement meilleurs au test de STIM au début du projet ainsi qu'un milieu socio-économique plus favorisé par rapport à leurs équivalents dans le groupe de contrôle.

En ce qui concerne les profils des enseignants, on retrouvait dans les classes les plus performantes ceux avec un faible niveau de qualification, une moindre ancienneté et qui se sentaient moins préparés pour favoriser le développement des compétences en créativité et en esprit critique chez leurs élèves au début du projet.

Élèves du secondaire

S'agissant des élèves du secondaire, l'expérience semble avoir été la plus bénéfique pour les classes qui présentaient un climat d'apprentissage difficile au début du projet, même si la différence était moins marquée que pour les élèves du primaire (supérieure de 28 % en moyenne).

Un délai plus important entre les collectes de données pré- et post-expérience semble positivement corrélé à de meilleurs résultats. Par ailleurs, pour la quasi-totalité des résultats d'intérêt, il y a avait plus d'enseignants dans les classes les plus performantes du groupe expérimental, que du groupe de contrôle, à estimer, au début du projet, que la créativité et l'esprit critique pouvaient être enseignés dans le cadre scolaire.

Analyse au niveau des activités

Pour finir, le cadre conceptuel de l'étude permet d'observer dans les différentes équipes les effets spécifiques qui sembleraient avoir été obtenus grâce à certaines activités pédagogiques. Les activités énumérées dans le graphique 7.14. ont été sélectionnées en raison des résultats très positifs qu'elles affichaient (situés dans le quartile supérieur) dans plus de cinq classes (à l'exception de l'activité intitulée « *Secret of community* » qui n'a été utilisée que dans deux classes, mais qui a cependant fait état d'excellents résultats pour les variables disponibles à l'étude). Le graphique présente le profil global des activités, en incluant leurs principales caractéristiques, celles des classes dans lesquelles elles ont été mises en œuvre et leurs résultats les plus pertinents.

Dans certains cas, les équipes locales ont communiqué à l'OCDE des plans de cours et des descriptions détaillées de certaines activités pédagogiques déployées sur le terrain. Après avoir fait l'objet d'un examen par les pairs, la majorité de ce matériel a été intégré au référentiel de l'OCDE regroupant les plans de cours (voir le chapitre 4).

Graphique 7.14. Profils des activités pédagogiques les plus fructueuses

Secret de la communauté

Niveau	Primaire
Équipe à l'origine	Thaïlande
Pays de déploiement	Thaïlande
Durée totale de l'activité	3h 20m
Taille moyenne (et nombre) des classes	39.5 (2)
Part d'élèves peu performants (STEM)	Non disponible
Part d'élèves performants (STEM)	Non disponible
Performance moyenne de l'école	Haute
Performance moyenne de la classe	Haute
Climat en classe	Encourageant
Principaux résultats de l'activité	• Amélioration des scores EPoC • Usage accru de pratiques d'enseignements pertinentes

Œuvre géométrique

Niveau	Primaire et secondaire
Équipe à l'origine	Fédération de Russie
Pays de déploiement	Fédération de Russie et Thaïlande
Durée totale de l'activité	2h 30m
Taille moyenne (et nombre) des classes	26 (15)
Part d'élèves peu performants (STEM)	37%
Part d'élèves performants (STEM)	15%
Performance moyenne de l'école	Moyenne-haute
Performance moyenne de la classe	Moyenne-basse
Climat en classe	Moyen
Principaux résultats de l'activité	• Compréhension accrue de la créativité et de l'esprit critique • Intérêt accru (STEM) • Amélioration des dispositions pour l'apprentissage

Détective Pytha

Niveau	Secondaire
Équipe à l'origine	Thaïlande
Pays de déploiement	Thaïlande
Durée totale de l'activité	2h 30m
Taille moyenne (et nombre) des classes	37 (13)
Part d'élèves peu performants (STEM)	35%
Part d'élèves performants (STEM)	18%
Performance moyenne de l'école	Moyenne
Performance moyenne de la classe	Moyenne-basse
Climat en classe	Plutôt décourageant
Principaux résultats de l'activité	• Intérêt accru (STEM) • Usage accru de pratiques d'ens. pertinentes • Amélioration des dispositions pour l'apprentissage

Élevage animal

Niveau	Primaire et secondaire
Équipe à l'origine	Thaïlande
Pays de déploiement	Thaïlande
Durée totale de l'activité	3h 20m
Taille moyenne (et nombre) des classes	36 (10)
Part d'élèves peu performants (STEM)	19%
Part d'élèves performants (STEM)	26%
Performance moyenne de l'école	Moyenne
Performance moyenne de la classe	Moyenne
Climat en classe	Moyen
Principaux résultats de l'activité	• Intérêt accru (STEM) • Usage accru de pratiques d'ens. pertinentes • Amélioration des dispositions pour l'apprentissage

L'analyse à l'échelle de la classe est un type d'analyse intéressant qui permet une interprétation des résultats plus approfondie et offre une plus grande fiabilité en cas de données manquantes. En outre, les résultats obtenus dans le cadre de cette analyse ont coïncidé avec les principaux résultats de l'analyse au niveau de l'élève. Ce type d'analyse semble donc constituer une piste intéressante pour examiner sur le terrain les travaux de recherche menés dans le domaine de l'éducation. Toutefois, afin d'assurer sa portée informative, il est conseillé d'asseoir cette analyse sur un échantillon de classes de taille raisonnable permettant une analyse distincte par pays et niveau d'enseignement. Cet élément devait être pris en compte lors de la planification de la stratégie d'analyse de l'éventuelle validation du projet.

Conclusions

Au vu des résultats de l'analyse des données, du retour d'expérience des équipes locales et des éléments factuels collectés sur le comportement des instruments, cette phase pilote semble confirmer que les instruments adoptés, ainsi que la stratégie d'analyse, sont appropriés pour l'évaluation des effets de l'expérience auprès des élèves. Comme il a été démontré dans les sections précédentes, les instruments et la stratégie d'analyse ont permis de déterminer les effets, tant positifs que négatifs, et d'identifier ensuite les facteurs contextuels les influençant. Ils ont également autorisé une certaine latitude dans le choix des types d'analyse adoptés, selon que l'intérêt portait sur les élèves ou sur les classes.

Les principales conclusions tirées de l'analyse des données de cette phase pilote sont les suivantes :

- L'expérience auprès des élèves semble avoir engendré un effet positif global. En effet, sur l'ensemble des modèles évalué, 25 % ont fait état d'un effet positif statistiquement significatif contre seulement 18 % ayant constaté un effet négatif statistiquement significatif, soit un total net estimé à 7 %. Cet effet global était similaire entre les différents niveaux d'enseignement. Toutefois, dans le cadre de l'examen des effets spécifiques, les variations suivantes ont été observées :
    - pour les élèves du primaire, l'expérience semble avoir eu un effet particulièrement bénéfique sur les scores obtenus au test de STIM et au test d'AVM
    - pour les élèves du secondaire, l'expérience semble avoir été particulièrement bénéfique concernant l'utilisation de pratiques pédagogiques en lien avec la créativité et l'esprit critique dans les cours de STIM, l'intérêt des élèves pour les disciplines des AVM ainsi que les scores obtenus au test d'AVM.
- Il semblerait que l'expérience ait notamment profité de manière assez similaire entre les pays aux sous-groupes d'élèves suivants :
    - les élèves dont les enseignants estimaient, au début de l'expérience, que la créativité pouvait être enseignée dans le cadre scolaire
    - les élèves qui avaient classé correctement les capsules relatives à l'esprit critique au début du projet
    - les élèves qui n'étaient pas parvenus à classer correctement les capsules relatives à la créativité au début du projet.
- Enfin, d'autres effets positifs dont l'ampleur variait considérablement d'une équipe à l'autre ont pu être observés, traduisant ainsi la grande diversité des situations locales. Ces effets font l'objet d'une description détaillée au chapitre 8.

La présence de résultats communs laisse entendre qu'il est possible d'établir une collaboration avec les enseignants et les décideurs politiques pour une promotion active de la créativité et de l'esprit critique, et que cela peut se traduire par des améliorations considérables et reproductibles de plusieurs résultats présentant un intérêt pour les élèves. Par ailleurs, ces résultats sont particulièrement dignes d'attention en raison de la fiabilité des conclusions qui tenaient compte des éventuels facteurs de confusion des variables, comme le sexe et le milieu socio-économique.

L'expérience auprès des élèves semble surtout avoir impacté les scores aux tests de performance obtenus par les élèves du primaire, dont les résultats ont mis en évidence des tendances bien plus nettes que pour les élèves du secondaire. Deux facteurs clés peuvent expliquer ce constat : premièrement, il se peut que les élèves du primaire soient davantage réceptifs aux pédagogies plus ouvertes utilisées par les équipes, en raison de leur expérience relativement faible d'autres pratiques plus établies. Deuxièmement, les enseignants du primaire passent beaucoup plus de temps avec les mêmes élèves que leurs homologues du secondaire, et cette réalité peut faciliter une adoption plus généralisée des nouvelles pratiques pédagogiques au-delà du temps consacré à l'expérience à proprement parler.

Dans le cadre de l'examen de certains sous-groupes visés, les conclusions de cette phase pilote sont particulièrement encourageantes pour les élèves du secondaire. Ces sous-groupes appartenaient majoritairement à des populations pouvant être définies comme défavorisées, soit sur le plan des ressources culturelles (milieu socio-économique défavorisé, par exemple), soit d'un point de vue cognitif (par exemple, mauvaise compréhension des concepts de la créativité et de l'esprit critique). En outre, bon nombre d'expériences ont été conçues pour cibler ces sous-groupes spécifiques. Le fait que certaines expériences soient parvenues à combler les écarts qui existent entre ces sous-groupes et l'ensemble de la population concernant différents éléments à l'étude constitue un résultat hautement pertinent, et étayé par des données probantes, pour des politiques d'éducation ciblées.

Les résultats se sont également avérés positifs s'agissant de l'élaboration des instruments. Dans la majorité des cas, tous les items inclus dans le questionnaire ont été retenus et une invariance de configuration a été observée entre les équipes. Il ressort de ce constat que les instruments ont pu mesurer efficacement les concepts pour lesquels ils avaient été conçus (les différents indices visés, par exemple). Les résultats ont parfois uniquement concerné l'enseignement secondaire, mais on pouvait s'y attendre dans la mesure où il peut s'avérer difficile, au niveau du primaire, de cerner toute la complexité de certains indices. Par ailleurs, en raison des conditions très diverses dans lesquelles les différentes équipes ont mené les travaux de terrain, il n'a pas été possible de vérifier de manière rigoureuse la capacité des instruments à mesurer, entre les pays, une évolution importante des résultats d'intérêt. Cependant, étant donné que dans les faits ils ont mesuré certaines évolutions significatives de différents résultats dans l'ensemble des pays, les éléments factuels laissent entendre que ces instruments en ont bel et bien la capacité. Il serait toutefois conseillé de poursuivre les recherches dans ce sens.

Enfin, cette phase pilote a également démontré que la majorité des coûts engendrés par la gestion et la collecte de données de l'enquête pouvait être internalisée lorsque les établissements disposaient de personnel ayant préalablement acquis de l'expérience dans le domaine de la recherche-action. Dans le cas contraire, certaines équipes ont dû faire appel à des consultants externes (souvent trouvés grâce à des contacts avec le monde universitaire), ce qui a également garanti des résultats de bonne qualité tout en maintenant les coûts à un niveau relativement raisonnable. La majorité des équipes qui sont parvenues à suivre au plus près le protocole de recherche, à utiliser les instruments à bon escient et à communiquer rapidement à l'OCDE les informations essentielles concernant la collecte des données sont celles dans lesquelles la gestion des aspects opérationnels de l'enquête avait été confiée à des chercheurs expérimentés. Alors qu'actuellement les travaux de recherche-action dans le domaine de l'éducation reposent en grande partie sur des ressources limitées, les conclusions de cette phase pilote se révèlent être particulièrement pertinentes du point de vue de l'élaboration des politiques en la matière.

Notes

1) Pour les scores au test de STIM, la corrélation entre les scores simples pondérés et les scores obtenus à l'aide de la théorie de la réponse d'item s'élevait à 0.93 pour les élèves du primaire et à 0.84 pour ceux du secondaire. Pour les scores au test d'AVM, ces corrélations avaient été estimées à 0.77 et 0.69, respectivement.

2) L'ordre escompté des capsules relatives à la créativité a été défini comme suit : la capsule avec le meilleur niveau de créativité devait être classée au-dessus ou au même niveau que la capsule affichant un niveau de créativité moyen ; cette dernière devait être classée au-dessus ou au même niveau que la capsule avec le plus faible niveau de créativité ; la capsule affichant le meilleur niveau de créativité devait être évaluée comme « Assez créative » ou « Très créative » ; et celle affichant le plus faible niveau de compétences devait être évaluée comme « Pas très créative » ou « Pas du tout créative ». La même procédure a été appliquée pour les capsules relatives à l'esprit critique.

3) Le coefficient de pondération équivaut à la probabilité d'appartenir au groupe expérimental au vu des valeurs de chaque élève pour un ensemble de variables explicatives.

4) Dans le tableau 7.3, les 121 modèles peuvent être obtenus en additionnant les 59 modèles de la ligne « Discipline de l'enseignant : STIM (vs AVM) » et les 62 modèles de la ligne « Discipline de l'enseignant : AVM (vs autre) ». Les 27 modèles de la ligne « Discipline de l'enseignant : STIM (vs autre) » ne devraient pas être pris en compte dans le total, étant donné qu'ils sont déjà comptabilisés dans les 59 modèles cités précédemment.

5) Le nombre d'heures d'enseignement par semaine avec la classe ; la performance de la classe en fonction du pays selon les déclarations de l'enseignant ; le degré de préparation que l'enseignant estimait posséder pour développer la créativité et l'esprit critique chez ses élèves ; la conviction de l'enseignant quant à la possibilité d'enseigner la créativité et l'esprit critique dans le cadre scolaire ; l'ancienneté de l'enseignant ; le niveau de formation de l'enseignant ; la matière enseignée ; le milieu socio-économique moyen du ménage de l'élève tel que perçu par l'enseignant ; le pourcentage de filles dans la classe ; le statut des élèves au regard de l'immigration ; le climat en classe ; le temps écoulé entre les collectes de données pré- et post-expérience ; la durée de l'expérience auprès des élèves ; les scores obtenus au test EPoC d'évaluation de la créativité, au test de STIM et au test d'AVM ; le pourcentage d'élèves à n'apprendre que ce qui les intéresse ; et l'évolution du recours aux pratiques pédagogiques liées à la créativité et à l'esprit critique selon les élèves.

6) On entend par classes présentant un climat difficile en termes de participation celles dans lesquelles l'enseignant a déclaré être d'accord ou totalement d'accord avec au moins l'un des items suivants : « Quand le cours commence, je dois attendre un long moment avant que les élèves ne se calment » ou « Il est difficile de garder le groupe concentré pendant plus de quelques minutes ». Ont également été intégrées à ce groupe, les classes dans lesquelles l'enseignant avait déclaré être en désaccord ou en total désaccord avec au moins l'un des

items suivants : « Les élèves de cette classe veillent à instaurer un climat d'apprentissage agréable » ou « Les élèves de cette classe sont généralement actifs et enclins à participer aux activités et aux discussions en classe ».

Références

Carr, M. et G. Claxton (2002), « Tracking the Development of Learning Dispositions », *Assessment in Education: Principles, Policy & Practice*, doi:10.1080/09695940220119148, pp. 9-37, http://dx.doi.org/10.1080/09695940220119148. [4]

Dormann, C., E. Demerouti et A. Bakker (2018), « A model of positive and negative learning », dans Zlatkin-Troitschanskaia, O., G. Wittum et A. Dengel (éds.), *Positive Learning in the Age of Information: A Blessing or a Curse?*, Springer VS, Wiesbaden, http://dx.doi.org/10.1007/978-3-658-19567-0_19. [2]

Fisher, R. (1926), « The Arrangement of Field Experiments », *Journal of the Ministry of Agriculture of Great Britain*, Vol. 33, pp. 503-513. [14]

IEA (2011), *The TIMSS Assessments website*, https://timssandpirls.bc.edu/timss2011/international-released-items.html (consulté le 12 juin 2019). [6]

King, G., C. J. Murray, J. A. Salomon et A. Tandon (2004), « Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research », *American Political Science Review*, Vol. 98/1, pp. 191-207, http://dx.doi.org/DOI: 10.1017/S000305540400108X. [5]

Lubart, T., M. Besançon et B. Barbot (2011), *EPOC: Évaluation du potentiel créatif des enfants*, Éditions Hogrefe, Paris, France. [1]

OCDE (2018), *Regards sur l'éducation 2018 : Les indicateurs de l'OCDE*, Éditions OCDE, Paris, https://dx.doi.org/10.1787/eag-2018-fr. [10]

OCDE (2015), *Base de données PISA 2015*, https://www.oecd.org/pisa/data/2015database/ (consultée le 12 juin 2019). [11]

OCDE (2015), *PISA 2015 Technical Report*, Éditions OCDE, Paris. [12]

OCDE (2012), *Compendium for the cognitive item responses*, https://www.oecd.org/pisa/data/pisa2012database-downloadabledata.htm (consulté le 12 juin 2019). [8]

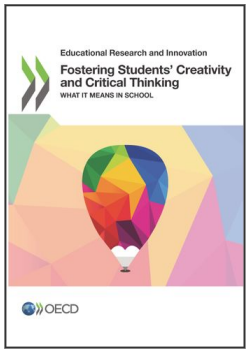
OCDE (2006), *Compendium for the cognitive item responses*, https://www.oecd.org/pisa/data/database-pisa2006.htm (consulté le 12 juin 2019). [7]

Rosenbaum, P. et D. Rubin (1983), « The central role of the propensity score in observational studies for causal effects », *Biometrika*, Vol. 70/1, pp. 41-55, http://dx.doi.org/10.1093/biomet/70.1.41. [13]

Rubin, D. (1976), « Inference and Missing Data », *Biometrika*, Vol. 63/3, pp. 581-592, http://dx.doi.org/10.2307/2335739. [9]

Schneider, B., J. Krajcik, J. Lavonen, K. Salmela-Aro, M. Broda, J. Spicer, J. Bruner, J. Moeller, J. Linnansaari, K. Juuti et J. Viljaranta (2016), « Investigating optimal learning moments in U.S. and finnish science classes », *Journal of Research in Science Teaching*, doi: 10.1002/tea.21306, pp. 400-421, http://dx.doi.org/10.1002/tea.21306. [3]

Wasserstein, R. et N. Lazar (2016), « The ASA Statement on p-Values: Context, Process, and Purpose », *The American Statistician*, doi: 10.1080/00031305.2016.1154108, pp. 129-133, http://dx.doi.org/10.1080/00031305.2016.1154108. [15]



Extrait de :

Fostering Students' Creativity and Critical Thinking

What it Means in School

Accéder à cette publication :
https://doi.org/10.1787/62212c37-en

Merci de citer ce chapitre comme suit :

Vincent-Lancrin, Stéphan*, et al.* (2020), « Effets du projet sur les résultats des élèves et élaboration des instruments d'enquête », dans Stéphan Vincent-Lancrin*, et al.*, *Fostering Students' Creativity and Critical Thinking : What it Means in School*, Éditions OCDE, Paris.

DOI: https://doi.org/10.1787/56e3eab3-fr