



HAL
open science

L'interdépendance positive au sens de la classe Puzzle : examen détaillé et approche expérimentale numériquement assistée à grande échelle

Anaïs Robert

► To cite this version:

Anaïs Robert. L'interdépendance positive au sens de la classe Puzzle : examen détaillé et approche expérimentale numériquement assistée à grande échelle. Psychologie. Université Clermont Auvergne, 2022. Français. NNT : 2022UCFAL002 . tel-04675988

HAL Id: tel-04675988

<https://theses.hal.science/tel-04675988v1>

Submitted on 23 Aug 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Clermont Auvergne
U. F. R. Psychologie, Sciences Sociales et Sciences de l'Éducation
Laboratoire de Psychologie Sociale et Cognitive (LAPSCO) - CNRS, UMR 6024
École Doctorale Lettres, Langues, Sciences Humaines et Sociales



L'interdépendance positive au sens de la classe Puzzle :
examen détaillé et approche expérimentale
numériquement assistée à grande échelle

Thèse de Doctorat présentée par Anaïs Robert
En vue de l'obtention du titre de Docteur en Psychologie (spécialité Psychologie Sociale
Expérimentale) sous la direction de Pascal Huguet et la co-direction de Céline Darnon
29 Septembre 2022

TOME I

Membres du jury

Céline DARNON	Professeure des universités, Université Clermont Auvergne – LAPSCO UMR 6024	Co-directrice de la thèse
Pascal HUGUET	Directeur de recherche CNRS, Université Clermont Auvergne – LAPSCO UMR 6024	Directeur de la thèse
Nicolas MICHINOV	Professeur des universités, Université Rennes 2 – LP3C UR 1285	Examinateur & Président du jury
Pascal PANSU	Professeur des universités, Université Grenoble Alpes – LaRAC	Rapporteur
Isabelle RÉGNER	Professeure des universités, Aix-Marseille Université – LPC UMR 7290	Rapporteure

Remerciements

Je tenais tout d'abord à remercier l'action « Innovation numérique pour l'excellence éducative » du Programme d'investissements d'avenir (PIA 2) et la mission Monteil pour le numérique éducatif, sans qui l'étude ProFan et le financement que j'ai obtenu dans ce cadre n'aurait pas vu le jour.

Pascal, merci de m'avoir fait confiance. Travailler à vos côtés a été riche d'enseignement, aussi bien professionnellement qu'humainement. Je sors grandi de cette expérience et j'espère que notre collaboration se poursuivra encore longtemps (et qui sait, peut-être qu'un jour j'arriverai à vous tutoyer :P). Céline, je tenais à te remercier d'avoir toujours répondu présente lorsque j'en ai eu besoin, merci pour ta bienveillance et pour ton humanité.

Je remercie tout particulièrement Marie pour son aide précieuse tout au long de ma thèse. Tu as été un soutien moral et logistique important (cf. entre autres les 598 pages d'annexe mouhahaha), je t'en suis extrêmement reconnaissante.

Je tenais également à remercier chaleureusement toute l'équipe ProFan et particulièrement celles et ceux avec qui j'ai étroitement travaillé ces dernières années (Arnaud, Céline, Eva, Isabelle, Pascal B.), ainsi que tous les acteur·rices qui ont rendu ce projet possible.

Merci à tous les collègues doctorant·es, post-doctorant·es, enseignant·es et personnels du LAPSCO que j'ai côtoyé ces dernières années. J'espère que vous ne m'en voudrez pas de ne pas toutes et tous vous citer. Sachez tout de même que je vous remercie pour votre accueil chaleureux, votre bienveillance ainsi que pour tous les bons moments que nous avons passé ensemble.

Une mention spéciale au « *Club des affreux* », Jordan et Medhi. Merci pour votre soutien ces dernières semaines, et pour toutes nos conversations aussi riches que fantasques. Nos éclats

de rire ont été un exécutoire précieux ces dernières années. Je suis fière de pouvoir vous compter parmi mes plus proches amis (#Su*****).

Je remercie également mes ami·e·s qui de près ou de loin ont participé à la réalisation de cette thèse. Je garde précieusement en mémoire les souvenirs de nos discussions, de nos éclats de rire et de nos bamboches endiablées. Merci de m'avoir soutenu et épaulé, je mesure la chance que j'ai d'être entourée par des personnes aussi formidables que vous.

C'est avec une pensée émue que je remercie également ma famille. Merci pour votre soutien sans faille. Un célèbre chanteur français a dit un jour « *On ne choisit pas ses parents, on ne choisit pas sa famille* »¹. Avec un peu de recul je suis convaincue que nous sommes extrêmement bien tombé·e·s, merci la vie ! J'ai une pensée toute particulière pour mes parents, Odile et Michel, qui n'ont jamais cessé de croire en moi et qui m'ont enseigné la résilience et la pugnacité, c'est avec beaucoup de fierté que je vous dédie cette thèse.

Et comment ne pas adresser quelques mots à la personne qui m'a choyé, accompagné et soutenue ces 5 dernières années. Guillaume, tu es la plus belle personne que la vie m'a donné la chance de rencontrer, grandir et vieillir à tes côtés est un cadeau que je nous promets de chérir sans répit.

¹ Spéciale dédicace à ma mère qui m'a transmis cette culture musicale qui me rend presque imbattable aux « blind test » ou tests à l'aveugle (JC si tu passes par là :P).

Résumé

Développée par Aronson et collaborateurs dans les années 1970, la classe Puzzle est une méthode d'apprentissage coopératif fondée sur la mise en œuvre d'un mécanisme d'interdépendance positive entre les élèves. Malgré sa popularité, les fondements théoriques et empiriques de cette méthode suscitent aujourd'hui de nombreuses interrogations. Notre synthèse narrative et quantitative des travaux expérimentaux publiés ces quarantes dernières années sur la classe puzzle et ses conséquences sur les performances scolaires montre qu'ils présentent de nombreuses limites méthodologiques (e.g., petits échantillons, groupes contrôles insuffisants) et des tailles d'effets anormalement élevées, notamment pour les études jugées les plus fragiles après réévaluation de leurs qualités méthodologiques. Ce premier constat rend difficile, voire impossible, toute conclusion ferme s'agissant de l'efficacité de la méthode promue par Aronson. Nos travaux dans le cadre de l'expérimentation ProFan permettent de tester cette efficacité à grande échelle, avec des groupes contrôles appropriés et une attention particulière à la question de la qualité de l'opérationnalisation du dispositif par les enseignant·es. Nos résultats suggèrent l'importance d'un entraînement de ces dernier·es à la méthode proposée pour en saisir les bénéfices auprès de leurs élèves. Ils invitent plus généralement à poursuivre l'effort engagé pour tester les méthodes pédagogiques coopératives à grande échelle, en prêtant une attention particulière aux conditions à réunir sur le terrain pour que l'interdépendance positive attachée à la coopération entre élèves puisse exprimer toute son efficacité.

Mots clés apprentissage coopératif ; classe Puzzle ; interdépendance positive ; qualités méthodologiques

Abstract

Developed by Aronson and colleagues in the 1970s, the Jigsaw Classroom is a cooperative learning method based on positive interdependence between students. Despite its popularity, the theoretical and empirical underpinnings of this method raise many questions. Our narrative and quantitative synthesis of studies published in the last 40 years on the effects of the Jigsaw Classroom method on school achievement reveals numerous methodological weaknesses (e.g., insufficient sample sizes, lack of control groups) as well as abnormally high effect sizes, especially for studies rated as the most fragile after reassessment of their methodological qualities. These results prevent from drawing any firm conclusion on the effectiveness of Aronson's method. The large-scale ProFan experiment allowed us to test its effectiveness on a large sample, with suitable control groups and accounting for the role of the quality of its implementation by the teachers. Our results suggest how important it is for teachers to be trained in the proposed method in order to understand the benefits for their students. More generally, they suggest that we should continue to test cooperative learning methods at large scale, paying more attention to the conditions that need to be met under which cooperation can fully express its efficacy.

Keywords Jigsaw classroom ; cooperative learning ; positive interdependence ;
methodological qualities

Table des matières

TABLE DES MATIERES	5
TABLE DES TABLEAUX	7
TABLE DES FIGURES	8
INTRODUCTION GENERALE	13
CHAPITRE 1 : LA CLASSE PUZZLE, HISTORIQUE ET ENJEUX ACTUELS	20
1 LA CLASSE PUZZLE, CONTEXTE HISTORIQUE	20
1.1 L’AFFAIRE « BROWN VS. BOARD OF EDUCATION »	20
1.2 L’HYPOTHESE DU CONTACT (ALLPORT, 1954)	21
1.3 LA COMPETITION A L’ECOLE	23
2 LA CLASSE PUZZLE, UNE METHODE D’APPRENTISSAGE COOPERATIF STRUCTUREE 24	
2.1 DEROULEMENT D’UNE CLASSE PUZZLE	24
2.2 LES PRINCIPES D’INTERDEPENDANCE POSITIVE ET DE RESPONSABILITE INDIVIDUELLE	26
2.3 LES DANGERS DU TRAVAIL DE GROUPE : PARESSE SOCIALE, CAVALIER SEUL ET BONNE-POIRE	27
3 LA CLASSE PUZZLE : ETAT DES TRAVAUX DE RECHERCHE	29
4 QUELQUES DONNEES DESCRIPTIVES	31
4.1 NOMBRE ANNUEL D’ARTICLES PUBLIES SUR LA CLASSE PUZZLE (1976 – 2019)	31
4.2 NOMBRE D’ARTICLES PUBLIES SUR LA CLASSE PUZZLE EN FONCTION DU PAYS.....	32
4.3 NOMBRE D’ARTICLES PUBLIEE SUR LA CLASSE PUZZLE EN FONCTION DE LA POPULATION ETUDIEE ...	34
4.4 EXPERIENCES PRINCEPS	35
4.5 NOMBRE D’ARTICLE PUBLIES EN FONCTION DE LA METHODE PUZZLE EMPLOYEE	38
4.6 CLASSE PUZZLE ET USAGE DES TECHNOLOGIES DE L’INFORMATION ET DE LA COMMUNICATION	40
5 DEVELOPPER LES COMPETENCES SOCIO-COGNITIVES PAR LE BIAIS DE LA METHODE PUZZLE ?	41
5.1 VARIABLES DEPENDANTES ETUDIEES DANS LE CADRE DES TRAVAUX REALISES SUR LA METHODE PUZZLE 41	
5.2 RELATIONS INTERGROUPEES ET DYNAMIQUE DE GROUPE	44
5.3 SENTIMENT D’AUTO-EFFICACITE.....	44
5.4 PERFORMANCES ACADEMIQUES	46
CHAPITRE 2 : LA CLASSE PUZZLE, ELEMENTS META-ANALYTIQUES	50
1 SYNTHESE QUANTITATIVE : CRITERES D’INCLUSION ET D’EXCLUSION DES ETUDES 50	
1.1 TAILLE D’EFFET	51
1.2 CRITERES METHODOLOGIQUES	55
1.2.1 <i>Construction des critères méthodologiques</i>	55
1.2.2 <i>Les critères et leur pondération</i>	57
1.2.3 <i>Le coefficient de concordance W de Kendall et son interprétation</i>	60
1.2.4 <i>Résultats</i>	61
1.3 ANALYSE DE LA REPARTITION METHODOLOGIQUE DES 41 ETUDES CONSIDEREES	62
2 EN CONCLUSION LES EFFETS DE LA METHODE PUZZLE AURAIENT-ILS ETE SURESTIMES PAR SES CONCEPTEURS ?	70
CHAPITRE 3 : LE DISPOSITIF PROFAN	73
1 VUE D’ENSEMBLE	73
2 PROCEDURE	75
2.1 PARTICIPANT·ES	75

3	METHODE.....	78
3.1	ASPECTS GENERAUX	78
3.2	DEROULEMENT CHRONOLOGIQUE DE LA PROCEDURE POUR LES ELEVES.....	80
3.3	DETAILS DE L'INDUCTION EXPERIMENTALE.....	84
4	MESURES.....	86
4.1	QUESTIONNAIRE Q1	86
4.2	QUESTIONNAIRE Q2	87
4.3	QUESTIONNAIRE Q3	88
4.4	MESURE DES PERFORMANCES ACADEMIQUES	88
4.5	GRILLE D'OBSERVATION POST-SEQUENCE PEDAGOGIQUE.....	90
4.6	MESURE DES COMPETENCES SOCIALES	93
CHAPITRE 4 : PROBLEMATIQUE GENERALE, HYPOTHESES ET PREMIERES ANALYSES.....		94
1	PROBLEMATIQUE GENERALE ET HYPOTHESES.....	94
2	QUALITE DE L'OPERATIONNALISATION DU DISPOSITIF PROFAN.....	102
2.1	TENDANCES CENTRALES.....	102
2.2	RESULTATS.....	103
2.3	CATEGORISATION DE L'IMPLEMENTATION DES SEQUENCES PROFAN	104
3	EFFET DE L'INDUCTION EXPERIMENTALE SUR LA SATISFACTION EXPRIMEE PAR LES ENSEIGNANT-E-S A L'EGARD DU DISPOSITIF.....	113
3.1	RESULTATS.....	114
3.2	SEQUENCES PLUTOT REUSSIES	114
3.2.1	<i>En moyenne</i>	114
3.2.2	<i>Séquence par séquence</i>	116
3.3	SEQUENCES DEGRADEES.....	116
3.3.1	<i>En moyenne</i>	116
3.3.2	<i>Séquence par séquence</i>	117
3.4	SEQUENCES ECHOUÉES.....	118
3.4.1	<i>En moyenne</i>	118
3.4.2	<i>Séquence par séquence</i>	119
4	EN CONCLUSION.....	119
CHAPITRE 5 : INTERDEPENDANCE POSITIVE, NIVEAU DE SATISFACTION DES ENSEIGNANT-E-S ET PERFORMANCE DES ELEVES.....		120
1	PRECISIONS SUR LES ANALYSES STATISTIQUES.....	121
2	RESULTATS PRINCIPAUX.....	123
2.1	COHORTE 2018 ASSP MATHÉMATIQUES SEQUENCE 3	124
2.2	COHORTE 2018 COMMERCE MATHÉMATIQUES SEQUENCE 2	127
2.3	COHORTE 2017 COMMERCE MATHÉMATIQUES SEQUENCE 3	130
2.4	COHORTE 2018 MELEC MATHÉMATIQUES SEQUENCE 3	133
2.5	COHORTE 2018 ASSP FRANÇAIS SEQUENCE 1.....	136
2.6	COHORTE 2017 MELEC FRANÇAIS SEQUENCE 3	139
3	RESUME ET DISCUSSION.....	142
CHAPITRE 6 : DISCUSSION GENERALE		149
1	LA LITTERATURE PUZZLE ET SES FAIBLESSES.....	150
2	PROFAN : UNE ETUDE PLUS AMBITIEUSE.....	152
3	LES RESULTATS DU DISPOSITIF PROFAN	154
4	DISCUSSION ET LIMITES DES RESULTATS DE L'ETUDE PROFAN.....	158
BIBLIOGRAPHIE		167

Table des tableaux

Tableau 1. Comparaison des quatre premières versions de la méthode Puzzle

Tableau 2. Formules de calcul des estimations de taille d'effet moyenne (source : Stanczak et al., 2020)

Tableau 3. Tailles d'effets moyennes estimées (g de Hedge) pour les performances académiques post-test (IC 95% [1.02, 1.76]) et présentées par ordre alphabétique des premiers auteurs

Tableau 4. Coefficient de concordance W de Kendall avant et après discussion entre les trois

Tableau 5. Répartition des études qui ont testées les performances académiques selon leur qualité méthodologique et la valeur estimée du g de Hedge

Tableau 6. Nombre de participant·e·s dans les conditions expérimentales par cohorte selon leur filière et leur sexe

Tableau 7. Corrélations entre perception par les élèves de leurs compétences en Mathématiques et en Français et les autres variables du questionnaire Q1

Tableau 8. Extrait de la grille d'observation fournies aux enseignant·e·s des établissements G1, G2 et G3

Tableau 9. Catégorisation de l'implémentation des séquences ProFan, cohorte par cohorte, filière par filière pour le Français, les Mathématiques et les Enseignements professionnels

Table des figures

Figure 1. Déroulement d'une classe Puzzle

Figure 2. Procédure de sélection des articles selon le diagramme de flux PRISMA

Figure 3. Nombre annuel d'articles publiés sur la classe Puzzle (1976 – 2019)

Figure 4. Nombre d'articles publiés sur la classe Puzzle en fonction du pays

Figure 5. Nombre d'articles publiés en fonction de la population étudiée

Figure 6. Nombre d'articles publiés en fonction de la méthode Puzzle employée

Figure 7. Nombre de variables dépendantes étudiées en fonction des articles publiés

Figure 8. Nombre d'études classées selon le critère méthodologique attribué

Figure 9. Nombre d'élèves par catégorie d'âge selon la cohorte

Figure 10. Vue d'ensemble du déroulement de la procédure pour l'année de Première (i.e., Temps 1-Année scolaire 2017-2018 et Temps 2-Année scolaire 2018-2019)

Figure 11. Vue d'ensemble du déroulement de la procédure pour l'année de Terminale (i.e., Temps 2-Année scolaire 2018-2019 et Temps 3-Année scolaire 2019-2020)

Figure 12. Extrait du questionnaire élève

Figure 13. Pourcentage de temps déclaré dans les modalités d'organisation de classe pour les conditions G1, G2 et G3 en MELEC pour la Séquence 2 de Mathématiques pour la cohorte 2018

Figure 14. Pourcentage de temps déclaré dans les modalités d'organisation de classe pour les conditions G1, G2 et G3 en COMMERCE pour la Séquence 2 de Français pour la cohorte 2018

Figure 15. Pourcentage de temps déclaré dans les modalités d'organisation de classe pour les conditions G1, G2 et G3 en COMMERCE pour la Séquence 4 de Français pour la cohorte 2018

Figure 16. Nombre de séquences « plutôt réussie » vs. « dégradée » vs. « échouée » par cohorte (tout enseignement confondue), en Français, en Mathématiques et pour les Enseignements professionnels par filière

Figure 17. Niveau de satisfaction exprimé par les enseignant·e·s pour les conditions Collectif structuré (G1), Collectif non structuré (G2) et Libre (G3) toutes cohortes et toutes filières confondues en Français et en Mathématiques pour les séquences jugées plutôt réussies

Figure 18. Niveau de satisfaction exprimé par les enseignant·e·s pour les conditions Collectif structuré (G1), Collectif non structuré (G2) et Libre (G3) toutes cohortes et toutes filières confondues en Français et en Mathématiques pour les séquences jugées dégradées

Figure 19. Niveau de satisfaction exprimé par les enseignant·e·s pour les conditions Collectif structuré (G1), Collectif non structuré (G2) et Libre (G3) toutes cohortes et toutes filières confondues en Français et en Mathématiques pour les séquences jugées échouées

Figure 20. Pourcentage de temps déclaré dans les modalités d'organisation de classe pour les conditions G1, G2 et G3 en ASSP pour la Séquence 3 de Mathématiques pour la cohorte 2018

Figure 21. Score au test de Mathématiques Séquence 3 en ASSP pour la cohorte 2018 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) et du niveau initial (Faible vs. Moyen vs. Fort) chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un haut niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

Figure 22. Score au test de Mathématiques Séquence 3 en ASSP pour la cohorte 2018 en fonction du contexte (G1, Collectif Structuré vs. G2 vs. Collectif non structuré) et du niveau initial (Faible vs. Moyen vs. Fort) chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un bas niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

Figure 23. Pourcentage de temps déclaré dans les modalités d'organisation de classe pour les conditions G1, G2 et G3 en COMMERCE pour la Séquence 2 de Mathématiques pour la cohorte 2018

Figure 24. Score au test de Mathématiques Séquence 2 en COMMERCE pour la cohorte 2018 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) et du niveau initial (Faible vs. Moyen vs. Fort) chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un haut niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

Figure 25. Score au test de Mathématiques Séquence 2 en COMMERCE pour la cohorte 2018 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) et du niveau initial (Faible vs. Moyen vs. Fort) chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un bas niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

Figure 26. Pourcentage de temps déclaré dans les modalités d'organisation de classe pour les conditions G1, G2 et G3 en COMMERCE pour la Séquence 3 de Mathématiques pour la cohorte 2017

Figure 27. Score au test de Mathématiques Séquence 3 en COMMERCE pour la cohorte 2017 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) et du niveau initial (Faible vs. Moyen vs. Fort) chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un haut niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

Figure 28. Score au test de Mathématiques Séquence 3 en COMMERCE pour la cohorte 2017 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) et du niveau initial (Faible vs. Moyen vs. Fort) chez les élèves exposés à des enseignant·e·s

caractérisé-e-s par un bas niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

Figure 29. Pourcentage de temps déclaré dans les modalités d'organisation de classe pour les conditions G1, G2 et G3 en MELEC pour la Séquence 3 de Mathématiques pour la cohorte 2018

Figure 30. Score au test de Mathématiques Séquence 3 en MELEC pour la cohorte 2018 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) et du niveau initial (Faible vs. Moyen vs. Fort) chez les élèves exposés à des enseignant-e-s caractérisé-e-s par un haut niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

Figure 31. Score au test de Mathématiques Séquence 3 en MELEC pour la cohorte 2018 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) et du niveau initial (Faible vs. Moyen vs. Fort) chez les élèves exposés à des enseignant-e-s caractérisé-e-s par un bas niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne.

Figure 32. Pourcentage de temps déclaré dans les modalités d'organisation de classe pour les conditions G1, G2 et G3 en ASSP pour la Séquence 1 de Français pour la cohorte 2018

Figure 33. Score au test de Français Séquence 1 en ASSP pour la cohorte 2018 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) chez les élèves exposés à des enseignant-e-s caractérisé-e-s par un haut niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

Figure 34. Score au test de Français Séquence 1 en ASSP pour la cohorte 2018 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) chez les élèves exposés à des enseignant-e-s caractérisé-e-s par un bas niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

Figure 35. Pourcentage de temps déclaré dans les modalités d'organisation de classe pour les conditions G1, G2 et G3 en MELEC pour la Séquence 3 de Français pour la cohorte 2017

Figure 36. Score au test de Français Séquence 3 en MELEC pour la cohorte 2017 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) chez les élèves exposés à des enseignant-e-s caractérisé-e-s par un haut niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

Figure 37. Score au test de Français Séquence 3 en MELEC pour la cohorte 2017 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) chez les élèves exposés à des enseignant-e-s caractérisé-e-s par un bas niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

Introduction générale

L'introduction du numérique dans la plupart des activités professionnelles entraîne des modifications comportementales profondes dans le monde du travail. Le changement dans les relations hiérarchiques traditionnelles au sein des organisations remet ainsi en question les modes classiques de communication et de collaboration. De même, en favorisant les échanges en réseau, le numérique invite à la polyvalence et à des compétences à interagir augmentées. Pourvoir les futures générations d'un répertoire de compétences qui leur permettront de faire face aux transformations du travail dans l'économie du futur apparaît alors comme un enjeu éducatif majeur. En France, le recours à l'apprentissage par petits groupes d'élèves est une modalité pédagogique fortement encouragée par le Ministère de l'Éducation nationale. À titre d'illustration, l'article L111-1 du Code de l'éducation datant du 24 Août 2021 déclare que : « *Le service public de l'éducation fait acquérir à tous les élèves le respect de l'égalité des êtres humains, de la liberté de conscience et de la laïcité. Par son organisation et ses méthodes, comme par la formation des maîtres qui y enseignent, il favorise la coopération entre les élèves.* ». Néanmoins, selon un rapport récent de l'Enquête PériODique sur l'Enseignement (EPODE, 2018), le travail des élèves en petits groupes semble peu répandu dans les écoles françaises : seul·e un·e enseignant·e sur deux indique y avoir « souvent » ou « toujours » recours, contre deux sur trois en Angleterre, en Flandre ou en Espagne.

Les méthodes pédagogiques par le biais desquelles les enseignant·e·s réunissent les élèves en petits groupes, lesquels travaillent conjointement en vue d'atteindre un objectif commun, sont désignées sous le terme d'apprentissage coopératif (Slavin, 2011). Issus de la Psychologie et des Sciences de l'éducation, les travaux sur l'apprentissage coopératif sont consignés dans plusieurs centaines d'études réalisées de l'école primaire à l'université, qui

attestent des bénéfices sociaux, motivationnels et cognitifs des travaux de groupes coopératifs (Hattie, 2009 ; Johnson & Johnson, 2009 ; Johnson et al., 2000 ; Kyndt et al., 2013).

Une méthode fréquemment évoquée par la littérature scientifique concernant l'apprentissage coopératif est la méthode de la « classe Puzzle » (*Jigsaw classroom*). Développée par le psychologue social Elliott Aronson et ses collaborateurs (1976, 1978), cette méthode a été historiquement élaborée dans le but de favoriser l'intégration des enfants issus de minorités ethniques dans le contexte de la déségrégation scolaire aux États-Unis. En effet, de par sa capacité supposée à rendre indispensables les contributions de tous les élèves à l'atteinte d'un objectif commun, la classe Puzzle devrait valoriser ceux issus des minorités en question, à leurs propres yeux et/ou aux yeux des pairs issus de la majorité, et favoriser ainsi leur intégration. Outre les apprentissages, la méthode Puzzle est également supposée efficace pour promouvoir le plaisir d'apprendre, la capacité des élèves à travailler en groupe ou encore leur motivation (Topping et al., 2017). La question des progrès en matière de travail en groupe, impliquant de développer la capacité des élèves à raisonner, à se décentrer et à coordonner leurs points de vue, est aujourd'hui rendue encore plus nécessaire par la transition numérique. En effet, cette dernière offre désormais les conditions technologiques d'une montée en puissance du travail collectif qui n'est plus strictement contrainte par des contingences spatiales et/ou temporelles.

Néanmoins, les fondements théoriques et mêmes empiriques de la classe Puzzle suscitent aujourd'hui de nombreuses interrogations (e.g., Roseth et al., 2018, Stanczack et al., 2021). Au niveau théorique, une question importante est de savoir si et dans quelle mesure l'interdépendance positive au fondement de la méthode Puzzle est bien de nature à faciliter l'intégration scolaire et par quels mécanismes exactement. Une autre question est de savoir si cette interdépendance est de nature également à améliorer les performances scolaires des élèves les plus faibles, quelle que soit leur origine sociale ou ethnique. En effet, les difficultés de ces

élèves peuvent tenir précisément à ce que leurs échecs répétés dans l'espace public de la classe, donc en présence de leurs pairs, produisent avec le temps des dynamiques de comparaison et d'auto-évaluation peu propices à l'apprentissage (Monteil & Huguet, 1999). Or, la valorisation théoriquement associée à l'interdépendance positive au sens de la classe Puzzle pourrait installer les élèves les plus en difficulté dans des dynamiques plus favorables aux apprentissages scolaires. Nous verrons que cette autre question, centrale dans nos travaux, demeure cependant à investiguer en dépit d'un intérêt dominant pour tester les effets de la classe Puzzle sur les performances des élèves. Contrairement aux objectifs de départ, davantage centrés autour de la question du climat de classe, ceux le plus souvent exprimés dans la littérature sur la classe Puzzle correspondent en effet davantage à la question des bénéfices de l'interdépendance positive sur les apprentissages des élèves. Même si évidemment les deux questions sont étroitement liées (climat scolaire et performances scolaires), nous verrons que dans les travaux (quasi-)expérimentaux conduits dans ce cadre, la seconde question traitant des effets sur les performances scolaires apparaît dominante, sans examen systématique de la question de l'intégration scolaire proprement dite pourtant à l'origine de la méthode proposée. Au niveau empirique, nous verrons également que les résultats produits sur cette question dominante n'offrent pas à ce jour de conclusion très claire, notamment en raison des nombreuses faiblesses méthodologiques des travaux en question (échantillons de petite taille, faiblesse des groupes contrôles, etc.).

Dans les travaux présentés au sein de ce manuscrit de thèse, nous avons testé à grande échelle la question des bénéfices éventuels de l'interdépendance positive entendue au sens de la classe Puzzle sur les performances des élèves à l'aide d'une méthodologie, à notre sens, plus conforme aux standards scientifiques actuels que celle exprimée dans la littérature spécialisée. Afin de bien saisir l'ampleur des enjeux au cœur de ce travail de thèse, nous avons opté pour une organisation en six chapitres. Dans le chapitre 1, nous présentons le contexte historique

dans lequel la méthode de la classe Puzzle (Aronson et al., 1978, Aronson & Patnoe, 2011) a été développée ainsi que la méthodologie associée visant à engager un mécanisme d'interdépendance positive et de responsabilité individuelle. Dans le but d'identifier précisément les forces et les faiblesses de la littérature en question, nous proposons une synthèse narrative, de 1976 à nos jours, dans laquelle nous décrivons les caractéristiques des études publiées (e.g., nombre d'articles publiés en fonction du pays, de la population étudiée, du type de méthode employée). Comme nous le verrons, cette synthèse, adossée à la méthode PRISMA (Liberati et al., 2009), révèle une incohérence importante entre l'efficacité supposée de la méthode Puzzle et la qualité des preuves empiriques. En effet, les travaux consacrés à cette méthode sont peu nombreux, majoritairement focalisés sur la question des performances académiques (plutôt que sur le climat de classe), chez les étudiant·e·s de l'enseignement supérieur (plutôt que chez les élèves plus jeunes comme ceux des premières études impulsées par Aronson). Mais surtout, ces travaux présentent de nombreuses limites méthodologiques rendant difficile, voire impossible, toute conclusion ferme s'agissant de l'efficacité de méthode puzzle.

L'application de cette méthode pédagogique dans le primaire, le secondaire ou encore à l'université requiert pourtant en toute rigueur d'en estimer précisément l'efficacité, c'est pourquoi nous avons réalisé dans le chapitre 2 une synthèse quantitative des effets de l'interdépendance positive au sens de la classe Puzzle sur la performance académique, puisque cette variable dépendante est de loin la plus étudiée dans le cadre des travaux réalisés au sujet de cette méthode. Dans ce but, nous avons dans un premier temps calculé la taille de l'effet classe Puzzle sur les performances académiques pour les études sélectionnées dans notre synthèse. Dans un second temps, et compte tenu des faiblesses déjà évoquées en référence aux travaux dans ce domaine, nous avons construit un indice de qualité méthodologique basé sur 6 critères. Ce travail permet de classer les études retenues pour notre synthèse selon cinq

catégories : « Médiocre », « Plutôt faible », « Intermédiaire », « Plutôt Forte » et « Excellente ». Comme nous le verrons, cette catégorisation révèle que très peu d'études s'avèrent satisfaisantes et qu'elles présentent des tailles d'effets anormalement élevées. Les études les plus solides présentent quant à elles une taille d'effet plus proche des estimations relatives aux effets attribués à l'apprentissage coopératif qu'il s'agisse ou non de la méthode Puzzle (Hattie, 2009). Les résultats de la catégorisation présentée au chapitre 2, tout en invitant à la prudence s'agissant des vertus encore souvent conférées à la méthode Puzzle, suggèrent de poursuivre les travaux à l'aide de dispositifs plus ambitieux que les précédents, s'agissant en particulier de la puissance statistique de test ou encore des standards de qualité méthodologiques à réunir.

Sans être restreinte à cet unique objectif, l'étude ProFan, dont la description fait l'objet du chapitre 3, permet précisément de tester l'efficacité de la méthode Puzzle dans des conditions méthodologiques et statistiques plus satisfaisantes que celles des études précédentes. Au niveau le plus général, cette étude, inscrite dans l'action « Innovation numérique pour l'excellence éducative » du Programme d'investissements d'avenir (PIA3), avait pour objectif d'analyser et de tester des modes d'enseignement et d'apprentissage propres à faire émerger de nouvelles compétences induites par la transformation digitale du travail et de son environnement social. Conçue dans le cadre de la seconde mission interministérielle confiée au recteur Monteil, ProFan s'adressait à des élèves de lycées professionnels issus de trois filières de formation : métiers de l'électricité et de leur environnement connecté (MELEC), commerce (COMMERCE) et accompagnement, soins et services à la personne (ASSP). Ce dispositif a été déployé dans chacune de ces 3 filières, en classe de première et en classe de terminale, dans 109 établissements de dix académies (Bordeaux, Poitiers, Limoges, Rennes, Nantes, Strasbourg, Nancy-Metz, Reims, Montpellier, Toulouse) couvrant 5 régions de France métropolitaine. L'étude était réalisée sur un ensemble de 3 groupes d'établissement distincts

répondant à des modes d'organisation pédagogiques spécifiques dans les enseignements de français, de mathématiques et professionnels selon un calendrier commun (72 séquences pédagogiques au total). Ainsi, plus de 10 000 élèves ont participé à cette opération avec un suivi longitudinal sur 2 promotions. Nous en détaillons au chapitre 3 les éléments principaux, échantillons, induction expérimentale (fondée sur des consignes favorisant une interdépendance positive), groupes contrôles (fondés soit sur des consignes plus classiques de travail en groupe moins susceptibles de garantir une interdépendance positive, soit sur un travail individuel en classe), avec comme précédemment (chapitre 2) une centration sur les performances académiques en tant que variable dépendante.

Le chapitre 4 et le chapitre 5 sont consacrés à la présentation des résultats de l'expérimentation ProFan, que nous limitons donc dans nos travaux à la question des effets de l'interdépendance positive sur les performances académiques. Notre analyse des données tient compte de la qualité de l'opérationnalisation du dispositif ProFan sur le terrain, qualité estimée à travers différents indicateurs renseignant à la fois sur la nature et sur l'ampleur des écarts entre les consignes fournies aux enseignant·e·s et la conduite de leurs séquences pédagogiques. Ce travail préalable révèle une implémentation très hétérogène du dispositif, d'où notre classement des 72 séquences pédagogiques selon 3 catégories : 1) Opérationnalisation plutôt réussie, 2) Opérationnalisation dégradée, et 3) Opérationnalisation échouée. Les séquences pédagogiques correspondant à une opérationnalisation dégradée ou échouée n'étant par définition pas ou peu exploitables/interprétables, nous focalisons pour l'essentiel dans ce chapitre sur les effets éventuels de l'interdépendance dans le cadre des séquences jugées plutôt réussies (environ un tiers de toutes les séquences). Ces effets attendus dans les séquences réussies sont examinés à la lumière de la satisfaction ou insatisfaction éprouvée par les enseignant·e·s à l'égard des consignes à respecter pour leurs enseignements ; une contrainte rendu nécessaire pour la conduite de cette grande expérimentation mais en effet susceptible de

réduire leur liberté pédagogique. La question est alors de savoir si le bénéfice attendu de l'interdépendance positive s'exprime plus ou moins selon que les enseignant·e·s se déclarent satisfait·e·s ou non des modalités d'organisation pédagogiques à respecter dans l'étude. Enfin, il s'agit aussi de savoir si ce bénéfice de l'interdépendance positive dépend de l'évaluation à laquelle les élèves procèdent s'agissant de leurs performances passées. En effet, et comme suggérée antérieurement dans cette introduction, la valorisation théoriquement associée à l'interdépendance positive au sens de la classe Puzzle est susceptible d'installer les élèves les plus en difficulté dans des dynamiques favorables aux apprentissages scolaires. Il n'est pas exclu que cette valorisation ait davantage de sens pour ces dernier·e·s que pour ceux plus en réussite (ou moins en échec) si l'on considère que la réussite est valorisante en tant que telle dans nos univers scolaires, ce qui est bien le cas. Aussi peut-on attendre légitimement au niveau statistique un effet d'interaction entre l'interdépendance positive et l'auto-évaluation des élèves s'agissant de leurs performances actuelles : le bénéfice de cette interdépendance, s'il peut s'exprimer chez tous les élèves, pourrait s'exprimer davantage encore chez les plus faibles. Pour bien comprendre l'enjeu de notre approche, il faut bien voir que la question du rôle des auto-évaluations comme celles de la satisfaction/insatisfaction des enseignant·e·s à l'égard des méthodes pédagogiques qui leurs sont proposées (à des fins expérimentales ou à d'autres fins) ne sont jamais pris en compte dans la littérature sur la classe Puzzle, et rarement pris en compte dans la littérature sur l'apprentissage coopératif d'une manière plus générale.

Enfin, les résultats de nos travaux (ceux présentés aux chapitres 4 et 5) sont discutés et mis en perspective dans un sixième et dernier chapitre. Cette discussion générale offre comme nous le verrons une vision à la fois optimiste et pessimiste des possibilités offertes par une approche scientifique des méthodes pédagogiques et de leur acceptation sur le terrain.

Chapitre 1 : La classe Puzzle, historique et enjeux actuels

1 La classe Puzzle, contexte historique

1.1 L'affaire « Brown vs. Board of Education »

Il y a plus de soixante ans, la plus haute juridiction des États Unis, la Cour suprême, s'apprêtait à juger une affaire dont les répercussions seraient historiques : « *Brown vs. Board of Education* » (Brown contre le bureau de l'éducation). En 1951, Linda Brown, une élève noire résidant à Topeka au Kansas se voit refuser l'inscription dans une école blanche de son quartier, l'obligeant à s'inscrire dans l'établissement réservé aux élèves noirs situé à plus d'un kilomètre du domicile familial. À cette époque, la loi de l'État du Kansas autorisait le principe des écoles séparées pour l'enseignement primaire dans les villes de plus de 15 000 habitants. En signe d'opposition, le père de Linda rassembla d'autres parents d'élèves victimes de cette ségrégation afin d'intenter une *class action*, c'est à dire un recours collectif contre l'administration, et cela avec le soutien de la *NACCP* (National Association for the Advancement of Colored People — l'association nationale pour la promotion des gens de couleur), une organisation de défense des droits civiques. Le 17 mai 1954, à l'unanimité des neuf juges, la Cour suprême rend son arrêt 347 et décrète inconstitutionnelle la ségrégation dans les écoles publiques des États-Unis d'Amérique.

D'après le *Chief Justice* (i.e., le président de la Cour suprême), cette décision fut en partie fondée sur les recherches en psychologie sociale (e.g., Clark & Clark, 1950) présentées par les plaignants au cours des différents procès. Ces recherches soutenaient l'idée selon laquelle la ségrégation des écoles faisait naître chez les enfants des groupes minoritaires un sentiment d'infériorité de nature à endommager leur estime de soi. Même si les bâtiments, ou encore la qualité de l'enseignement des écoles accueillant les enfants des minorités étaient très

similaires à celles de leurs homologues blancs, ces derniers ne bénéficiaient tout de même pas des mêmes chances en matière d'éducation.

À l'époque, les experts en psychologie sociale invités à témoigner durant l'affaire Brown étaient convaincus que la déségrégation des écoles allait non seulement réduire les préjugés, mais augmenterait également l'estime de soi et les performances académiques des enfants des groupes minoritaires. Dans certains états comme le Texas, la déségrégation des écoles avait été mise en œuvre par le biais du *busing*, un programme de transport scolaire visant à promouvoir la mixité sociale dans les établissements scolaires publiques. À cette époque, la ségrégation était à la fois raciale et résidentielle. Ainsi, et pour la première fois de leur vie, des enfants issus de groupes ethniques et raciaux différents se retrouvaient en contact les uns avec les autres.

Cependant, les effets attendus de cette opération furent tout autre. En 1978, lorsque Walter Stephan et David Rosenfield passèrent en revue des dizaines d'études réalisées à la suite de la déségrégation, aucune d'entre elles ne démontraient une augmentation significative de l'estime de soi des élèves issus des minorités. Au contraire, dans 25% des cas, l'estime de soi de ces derniers avait même diminué. De plus, Stephan et Rosenfield (1978) rapportèrent que la déségrégation engendrait une réduction des préjugés dans seulement 13% des études. Le constat était donc sans appel.

1.2 L'hypothèse du contact (Allport, 1954)

Néanmoins, les experts invités à témoigner durant le procès Brown n'avaient pas certifié que les bénéfices liés à la déségrégation allaient être constatés sur-le-champ. En effet, certaines conditions devaient être observées. Ces dernières avaient notamment été spécifiées par Gordon Allport dans son ouvrage « *The nature of prejudice* », publié l'année de la parution de l'arrêt *Brown v. Board*. L'hypothèse dite du « contact » d'Allport (1954) affirmait que la

réduction des préjugés et de la discrimination ne pouvait s'opérer que sous certaines conditions qui n'étaient alors pas respectées. Premièrement, le contact devait être approuvé et appuyé officiellement par les autorités. Néanmoins, et ce malgré la décision de la Cour suprême, dans certains quartiers les autorités locales n'avaient pas formellement accepté et appliqué la loi, tandis que d'autres étaient même en plein mépris de cette dernière. Pettigrew (1961; tel que cité par Aronson & Bridgeman, 1979) observa que la déségrégation se déroulait sans heurt dans les quartiers où les autorités locales approuvaient cette dernière. Toutefois, il était clair que ce n'était pas suffisant dans le but d'observer un effet sur la réduction des préjugés. Deuxièmement, Allport affirmait que le contact devait avoir lieu entre des groupes de statuts égaux. Cependant, avant 1954, la fameuse doctrine « *Separate but equal* » faisait loi et autorisait les états à imposer des mesures de ségrégation raciale, pourvu que les conditions offertes aux différents groupes soient égales. Malheureusement il existait beaucoup de disparités et très peu d'égalités. Les écoles des quartiers qui accueillaient la plupart des minorités ethniques n'offraient pas la même qualité d'éducation que celle accordée à leurs homologues blancs. De ce fait, il était impossible que le contact ait lieu entre membres de groupes de statuts égaux. Enfin, Allport suggérait que le contact devait se dérouler dans un contexte de coopération dans le but d'atteindre un but commun. Cependant en classe, et ce encore à l'heure actuelle, les élèves n'étaient d'ordinaire jamais engagés dans la poursuite d'un but commun. En effet, le processus éducatif était hautement compétitif. La compétition s'exprimait non seulement durant les heures de classe ou pendant les évaluations, mais aussi lors des échanges informels ou, typiquement, les enfants apprennent à lever la main en réponse aux questions de l'enseignant.e. Cette atmosphère de compétition permanente entraînait les élèves à se considérer les uns et les autres comme des adversaires à vaincre et exacerbait non seulement les préjugés et la discrimination déjà existante, mais également les différences en termes de performance scolaire entre les enfants issus des minorités et leurs homologues blancs.

1.3 La compétition à l'école

Les effets délétères de la compétition furent notamment illustrés par Sherif et al. (1961) dans la célèbre expérience de « la caverne des voleurs ». Dans cette expérience qui se déroulait lors d'un camp d'été, un conflit fut généré par le biais d'activités concurrentielles (e.g., des tournois) entre deux groupes de garçons âgés d'une dizaine d'années. Ces activités créèrent un terrain fertile à la colère et l'hostilité, augmentant considérablement l'antagonisme entre les deux groupes. C'est seulement après que les deux groupes aient été amenés à travailler en coopération dans le but de résoudre un problème commun que les conflits se dissipèrent, démontrant ainsi que la présence d'un objectif commun (« but supra-ordonné ») était un moyen de résoudre les conflits entre les groupes.

C'est précisément ce constat qui mena Elliott Aronson et ses collègues (1978) à considérer qu'il était nécessaire de développer une méthode pédagogique permettant de modifier l'atmosphère dans les salles de classe afin que les élèves ne se considèrent plus comme des concurrent·e·s mais comme des ressources les un·e·s pour les autres. Les systèmes éducatifs occidentaux tendant en effet à être organisés autour de valeurs qui favorisent la compétition (e.g., les notes ou les classements ; Butera et al., 2021).

L'objectif d'Aronson et de ses collaborateurs était donc de développer un programme de recherche dont le but n'était pas seulement de comparer les effets de la coopération et de la compétition en classe. En effet, il s'agissait de concevoir une méthode d'apprentissage coopérative qui pourrait être facilement utilisée à long terme par les enseignant·e·s, et d'évaluer les effets de cette intervention *via* une série d'expérimentations sur le terrain. C'est ainsi que la méthode dite de « la classe *Puzzle* » (« Jigsaw Classroom ») fût créée.

2 La classe Puzzle, une méthode d'apprentissage coopératif structurée

2.1 Déroulement d'une classe Puzzle

Dans une salle de classe traditionnelle, les élèves sont souvent récompensés lorsqu'ils réussissent à attirer l'attention de l'enseignant·e en surpassant leurs concurrent·es qui ne sont autre que leurs camarades de classe. À contrario dans la *classe Puzzle*, les élèves atteignent leurs objectifs d'apprentissage en étant attentifs à leurs pair·es, en posant les bonnes questions, autrement dit en s'entraidant. Dans l'expérience princeps, Aronson et ses collègues se rendirent dans une classe de cinquième année (i.e., CM2) durant laquelle les élèves étudiaient la biographie de Joseph Pulitzer. Afin d'exposer au mieux le déroulement de cette méthode, prenons l'exemple d'une classe composée d'élèves de CM2 à qui l'enseignant·e propose un apprentissage en histoire sur la période du début des temps modernes (cf. Figure 1).

- 1) *Phase individuelle* : Dans un premier temps, la classe est divisée par l'enseignant·e en sous-groupes de 4 à 6 élèves qui s'efforcent individuellement de se familiariser avec une partie du matériel à apprendre. Préalablement organisé en plusieurs parties distinctes mais complémentaires au regard de la globalité de l'apprentissage proposé, le matériel proposé en phase 1 à chaque élève se présente donc comme une pièce d'un puzzle à reconstituer ultérieurement. Dans notre exemple, le matériel distribué par l'enseignant·e peut se présenter comme une synthèse d'informations sur les grandes découvertes (une première partie du matériel), la monarchie absolue en France (une deuxième partie), ou encore la révolution française (une troisième partie). Chaque élève ne disposant à ce stade que d'une partie du matériel délivré par l'enseignant·e (une seule pièce du puzzle), il lui est donc impossible de chercher à maîtriser l'ensemble des informations requises pour atteindre l'objectif d'apprentissage dans son intégralité.

- 2) *Phase « expert »* : Dans un deuxième temps, les élèves quittent momentanément leur groupe de la phase 1 pour interagir dans le cadre d'un nouveau groupe de travail avec celles et ceux disposant du même matériel ou corpus d'informations qu'eux (la même pièce du puzzle). Sur la base de cette interaction, il est attendu que les élèves deviennent en quelque sorte « experts » du matériel qui leur a été spécifiquement attribué en phase 1 et conditionne au moins en partie la mise en place d'un mécanisme d'interdépendance positive en phase 3.

Cette étape du dispositif est particulièrement importante car elle fournit en principe aux élèves les moins chevronnés l'occasion de prendre appui sur leurs camarades plus expérimentés afin de maîtriser au mieux le corpus d'informations qui leur a été attribué. Les groupes « experts » donnent à tous les élèves l'occasion de se faire une idée précise de la façon de présenter le matériel à leurs camarades (ceux de la phase 1) et ce, sans tenir compte des inégalités antérieures en matière de compétences ou de préparation.

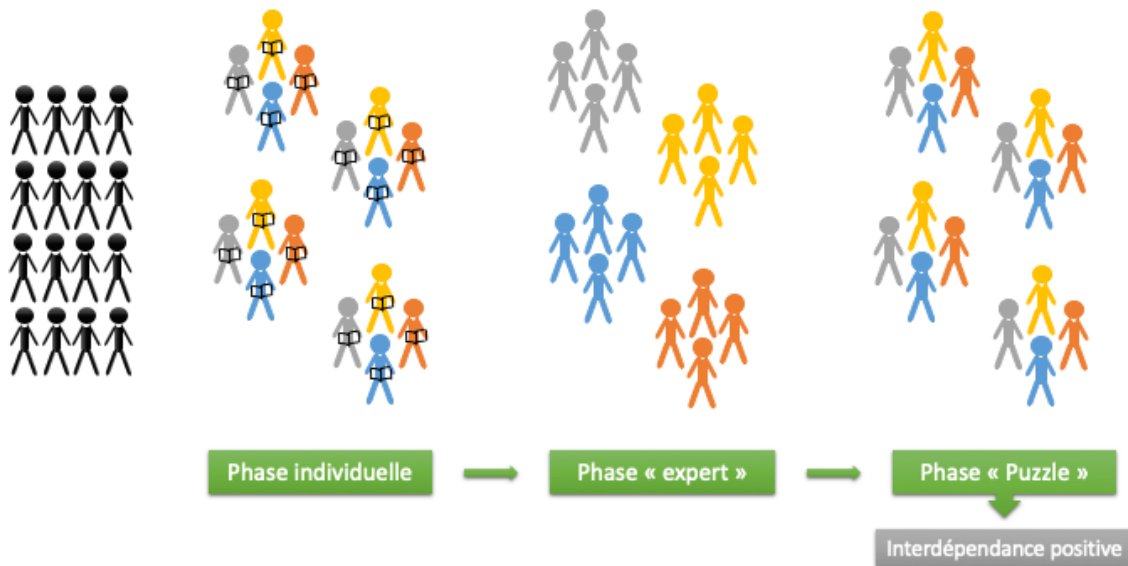
- 3) *Phase « puzzle »* : Dans cette troisième phase, les élèves retournent dans leurs groupes d'origine (phase 1) avec l'objectif d'exposer à leurs camarades les informations sur lesquelles ils sont en principe devenus experts. Chacun possédant une pièce du puzzle, cette pièce une fois partagée avec le reste du groupe permet en principe à tous d'accéder à la totalité des informations requises pour atteindre l'objectif d'apprentissage dans son intégralité.

Durant cette étape il est également précisé aux élèves, que par la suite, chacun sera évalué individuellement sur l'entièreté de la leçon. La procédure évoquée précédemment rappelle donc fortement la technique du puzzle, chaque élève possédant une seule pièce essentielle afin de reconstituer l'intégralité de la leçon. C'est cette ressemblance qui mena Aronson et ses collaborateurs à désigner cette méthode sous le nom de classe « puzzle » (Aronson & Patnoe, 2011).

Figure 1

Déroulement d'une classe Puzzle

□



2.2 Les principes d'interdépendance positive et de responsabilité individuelle

Afin de prévenir les phénomènes délétères susceptibles de se produire lorsque les individus travaillent en groupe (e.g., paresse sociale, phénomène de cavalier seul/profiteur, « bonne-poire », etc.), ce dispositif repose sur deux éléments centraux qui vont structurer le travail de groupe (Buchs, 2017, Topping et al., 2017) :

- 1) *L'interdépendance positive*, qui implique que la réussite de chacun est conditionnée par la contribution et la réussite de chaque membre du groupe. Elle permet aux élèves d'un groupe coopératif de percevoir que leur réussite dépend de celle des autres élèves avec qui ils sont associés et cela, en vue d'atteindre un objectif commun.

Cette interdépendance peut reposer sur les *buts* (i.e., la nécessité d'entreprendre des efforts collectifs afin d'atteindre un but commun), mais également sur les *ressources* (e.g., chaque

membre du groupe dispose d'une partie du matériel à apprendre et la solution exige l'articulation des différentes parties) ou encore les *rôles* (i.e., chaque membre a un rôle spécifique et indispensable au bon fonctionnement du groupe).

- 2) *La responsabilité individuelle*, qui implique que chaque membre du groupe doit apporter sa contribution à travers la tâche qui lui est assignée. Elle est présente lorsque les élèves se sentent responsables de leur apprentissage et perçoivent que leur propre effort, participation ou engagement dans la tâche, est essentiel pour atteindre les objectifs fixés pour le groupe.

2.3 Les dangers du travail de groupe : paresse sociale, cavalier seul et bonne-poire

Le principe de *responsabilité individuelle* est un des éléments qui permet de modérer les phénomènes délétères susceptibles de se produire lors d'une activité de groupe. En effet, la « paresse sociale » (i.e., « *social loafing* ») réfère à un phénomène de réduction de l'effort individuel en situation de travail collectif dans laquelle seule la performance du groupe est identifiable : dans cette situation, les individus fournissent typiquement moins d'effort relativement à ce dont ils se montrent capables lorsqu'ils travaillent seuls sur la tâche proposée. Mis en évidence par Ringelmann (1913), ce phénomène de paresse sociale, reproduit et théorisé par Latané et al. (1979) a été ensuite répliqué de nombreuses fois (Huguet, 1995 ; Huguet & Monteil, 2001 ; pour une méta-analyse, cf. Karau & Williams, 1993). Il résulte largement de l'intégration d'une idéologie individualiste conduisant les individus à rechercher les différences plutôt que les similitudes avec leurs congénères autrement dit, à affirmer l'unicité du « soi » par une différenciation interpersonnelle permanente. D'où cette tendance, bien repérée dans les cultures d'Europe de l'Ouest et d'Amérique du Nord, à valoriser davantage la performance individuelle que la performance collective (pour des travaux montrant le lien direct entre la recherche d'unicité du soi et la paresse sociale, cf. Charbonnier et al., 1998 ; Huguet et al., 1999). Dans le même registre, Karau et Williams (1993) ont également démontré que la paresse

sociale est le fait des hommes plus que des femmes et des occidentaux plus que des orientaux, autrement dit des catégories sociales qui ont le plus intégré les normes et valeurs de l'individualisme.

Le phénomène de « cavalier seul » (i.e., *free-rider*), mis en évidence par Olson (1965) et appliqué quelques années plus tard au travail de groupe par Kerr (1983 ; voir aussi Kerr & Bruun, 1981), désigne quant à lui une tendance pour les membres d'un groupe à profiter du bénéfice d'une action collective sans s'être acquittés de leur part du travail. En effet, en règle générale, plus un groupe compte de membres au service d'une cause commune et plus il est difficile d'identifier la contribution de chacun. Par conséquent, les individus perçoivent leur participation comme infime par rapport à l'ensemble de la production du groupe et ne parviennent plus à trouver la motivation suffisante pour s'impliquer dans la tâche. En retour, le phénomène de « bonne-poire » (i.e., *sucker effect* ; Kerr, 1983) se produit lorsque les membres les plus productifs d'un groupe réalisent que certain·es de leurs coéquipier·es font « cavalier seul ». Refusant de soutenir leur pair·es qui ne contribuent pas à l'effort collectif, les membres les plus performants vont progressivement ajuster leurs contributions individuelles à la baisse.

Néanmoins, plusieurs travaux ont montré que ces différents phénomènes se dissipent dès lors que la production collective demeure facilement identifiable (Latané et al., 1979), lorsque la tâche à accomplir collectivement est perçue comme intéressante et stimulante (Brickner et al., 1986 ; Jackson & Williams, 1985), ou encore lorsque les membres du groupe ne se perçoivent pas supérieurs à leurs congénères (absence d'un biais de supériorité généralisée/*above average effect* ; cf. Huguet et al., 1999).

En réduisant en principe la probabilité que ne surviennent de tels phénomènes (paresse sociale, etc.), notamment grâce à l'interdépendance positive et à la responsabilité individuelle, la classe Puzzle est une méthode hautement structurée. Selon Aronson et Patnoe (2011), c'est en particulier l'interdépendance requise entre les élèves qui en fait une méthode de travail

unique incitant les élèves à participer activement à leur apprentissage. De cette façon, chaque élève devient une ressource précieuse pour ses camarades. D'autre part, apprendre les uns des autres diminue petit à petit le besoin de se distinguer dans la mesure où l'apprentissage d'un élève améliore la performance des autres élèves au lieu de l'inhiber, comme c'est généralement le cas dans la plupart des classes traditionnelles. Dans ce dispositif coopératif, l'enseignante devient « un facilitateur d'apprentissage » et partage le processus d'apprentissage et d'enseignement avec ses élèves au lieu d'en être l'unique source. Mais pour bien saisir à la fois la nature et la quantité des travaux attachés à la classe Puzzle, et la réalité des bénéfices attendus dans ce cadre pour les élèves, nous avons jugé indispensable d'en produire une synthèse avec une focalisation sur la question des gains éventuellement associés à cette méthode en matière de performances scolaires. Cette synthèse fait aussi l'objet d'un article en langue française en cours de préparation.

3 La classe Puzzle : état des travaux de recherche

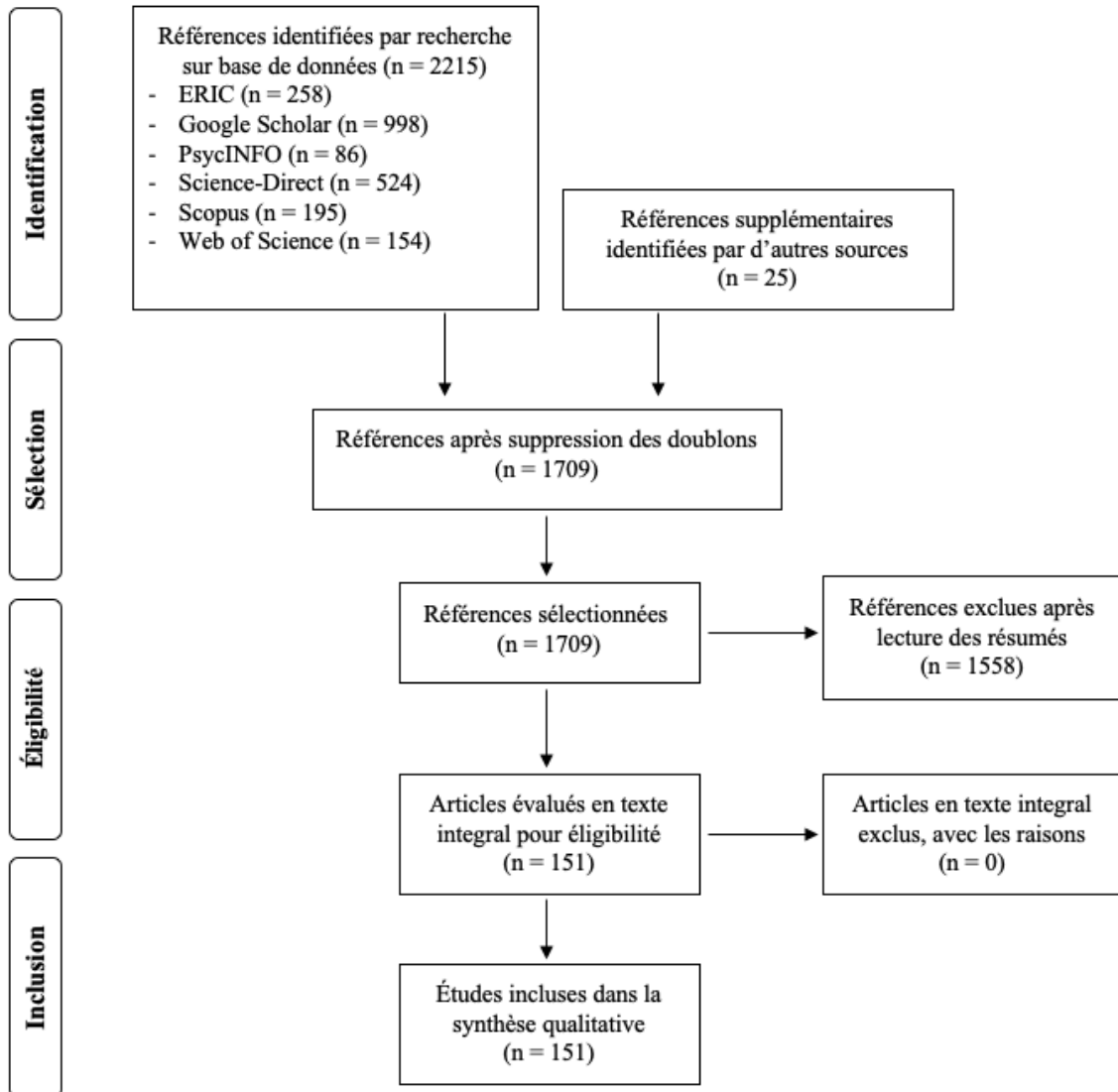
Notre synthèse des travaux sur la classe Puzzle recouvre plus de quarante années de recherche (depuis 1976 à nos jours) à partir des bases de données « ERIC », « Google Scholar », « PsycINFO », « Science-Direct », « Scopus » et « Web of Science ». Ces dernières ont été explorées entre Novembre 2017 et Janvier 2021 à partir des mots-clés suivants : « jigsaw » AND « cooperative learning ». Adossée à la méthode PRISMA (Liberati et al., 2009), notre recherche porte sur 2215 références extraites des bases de données auxquelles 25 références supplémentaires identifiées par d'autres sources ont été ajoutées. Sur cette base élargie, 1709 références ont été conservées après avoir retiré les doublons et les documents non exploitables. Une lecture des titres et des résumés de chacune de ces occurrences nous a permis *in fine* de retenir 151 publications faisant explicitement mention de la classe Puzzle. Enfin, dans la mesure où certains des articles sélectionnés rapportaient plusieurs études (n = 4 articles au total), le

nombre total de références s'élève à 156. La Figure 2 reprend l'ensemble du processus de sélection et d'exclusion des articles.

Figure 2

Procédure de sélection des articles selon le diagramme de flux PRISMA

□



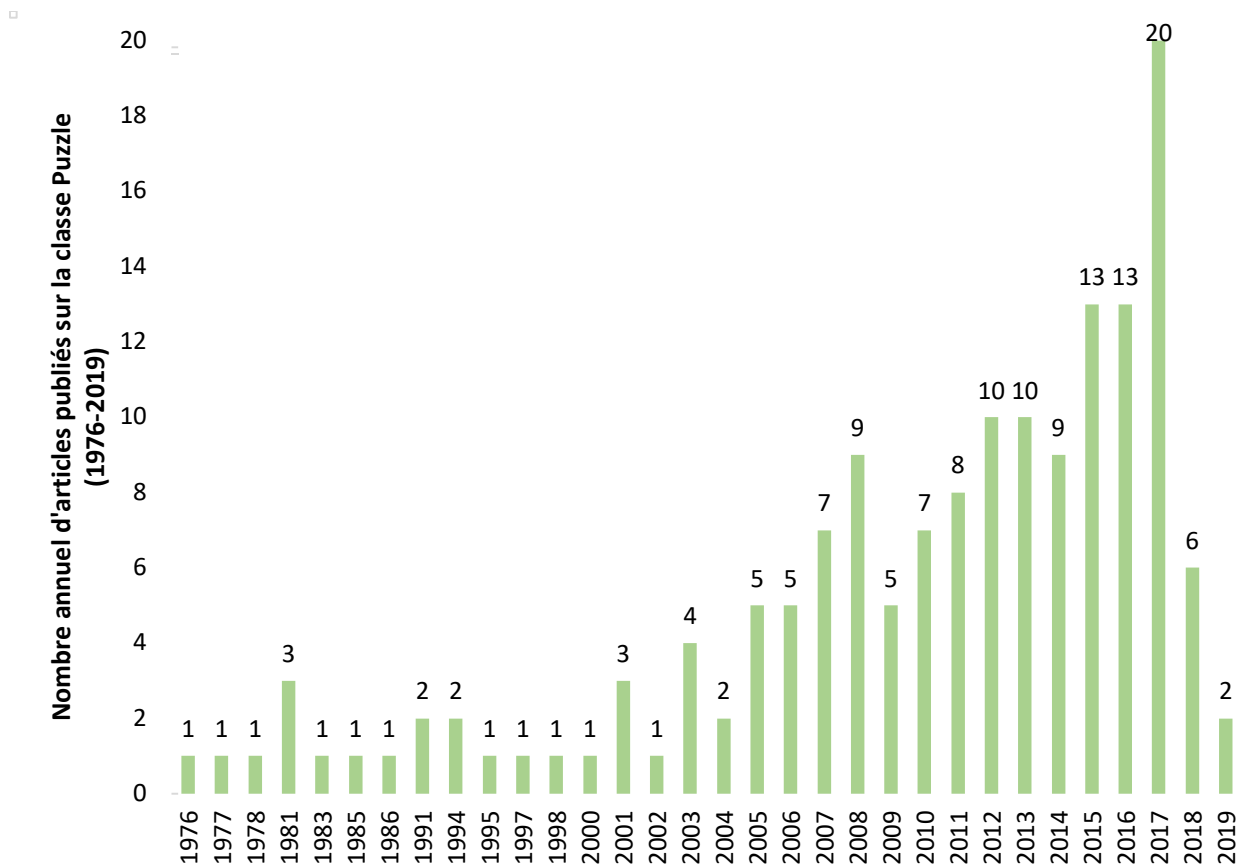
4 Quelques données descriptives

4.1 Nombre annuel d'articles publiés sur la classe Puzzle (1976 – 2019)

La Figure 3 ci-dessous répertorie le nombre annuel d'articles réalisés sur la classe Puzzle depuis la première publication à ce sujet en 1976. En dépit d'une augmentation depuis les années 2000, qui atteint son pic en 2017, le nombre de publications impliquant la classe Puzzle est assez faible en comparaison à d'autres champs de recherche sur le travail en groupe pour lesquels ce sont des centaines d'articles qui sont publiés chaque année (e.g., les travaux sur la paresse sociale ont généré plus de 1300 publications depuis les années 70). Par ailleurs, et même si les méthodes de travail collaboratives sont considérées comme parmi les plus grandes innovations éducatives de la période récente (Gillies, 2014), les données de la Figure 3 suggèrent qu'à l'échelle internationale, l'intérêt porté à la classe Puzzle demeure modeste et disparate d'un point de vue géographique (cf. Figure 4).

Figure 3

Nombre annuel d'articles publiés sur la classe Puzzle (1976 – 2019)



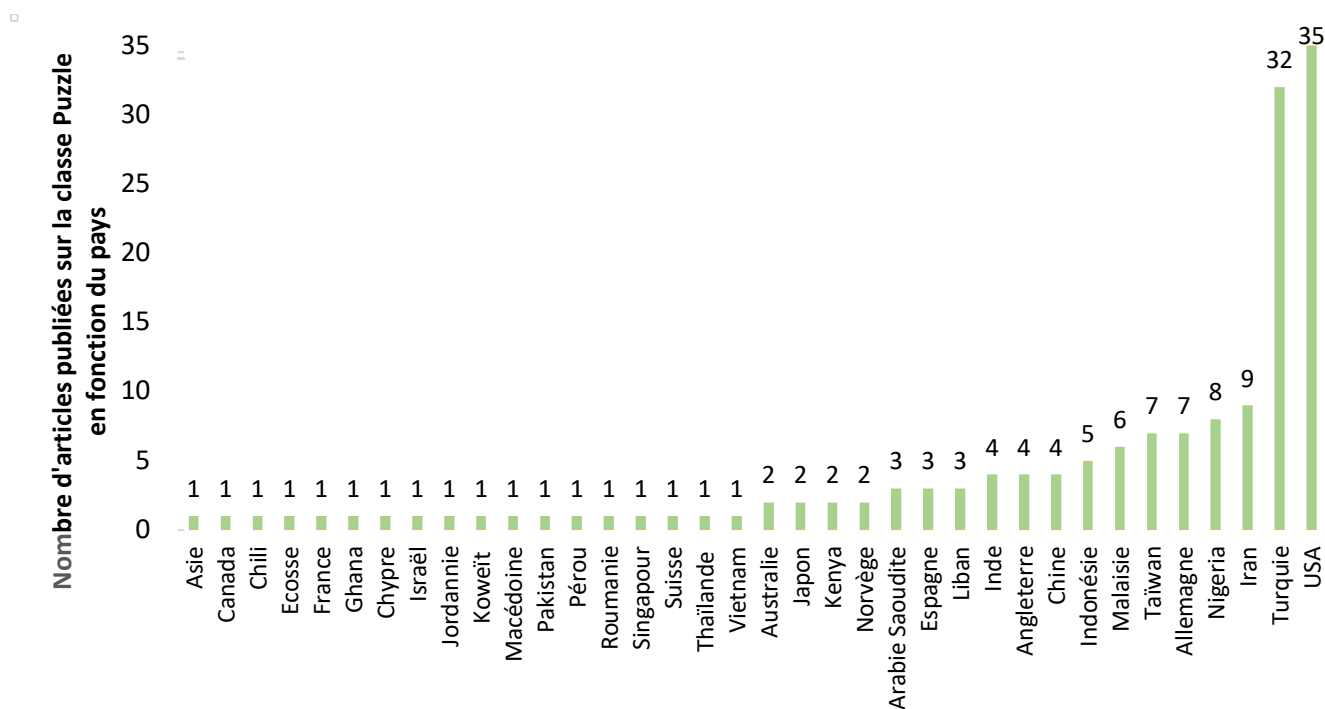
4.2 Nombre d'articles publiés sur la classe Puzzle en fonction du pays

La Figure 4 recense le nombre d'articles publiés sur la classe Puzzle selon le pays dans lequel ces études ont été conduites. Sans surprise, les États-Unis ($n = 35$), berceau de la classe Puzzle, arrivent en tête. Plus curieusement, la Turquie ($n = 32$) occupe la deuxième place. Cette observation néanmoins est assez cohérente avec un mouvement plus large en faveur de l'apprentissage coopératif dans ce même pays depuis les années 2000 (Dirlikli et al., 2016), suite au vaste programme de rénovation de son système éducatif en 2004 pour répondre aux standards de l'Union Européenne. Le nombre d'articles publiés partout ailleurs est extrêmement faible, pour la plupart au voisinage de zéro.

Ces disparités pourraient découler des caractéristiques qui structurent les systèmes éducatifs. En effet, des différences en termes de conception de l'éducation pourraient valoriser certaines attitudes plutôt que d'autres et par conséquent, l'intérêt porté à certains modes d'apprentissage. À l'exemple des travaux de Mons et al. (2012) qui, en s'appuyant sur les données collectées par l'enquête PISA 2000, définissent trois grands modèles éducatifs et leurs effets sur les attitudes des élèves. Le premier modèle qualifié « *d'éducation totale* », développé principalement dans les pays nordiques et anglo-saxons, serait notamment caractérisé par la recherche d'un bagage culturel commun dans la scolarité qui s'étend au-delà des disciplines académiques classiques, et se traduirait par des relations de proximité entre élèves et enseignant·e·s, un climat de discipline plus détendu et un suivi personnalisé des apprenant·e·s. Un deuxième modèle qualifié de « *producteur* » et développé en Belgique, en Suisse ou encore en Autriche, valoriserait le lien entre éducation et marché du travail par l'introduction dès le secondaire d'enseignements préprofessionnels mais cela, dans l'unique but de promouvoir la hiérarchisation des filières. Le troisième modèle qualifié « *d'éducation académique* » développé en France, en Italie ou en encore au Portugal, érigerait quant à lui l'école en forteresse dispensatrice de savoirs universels, et se caractériserait par des contenus disciplinaires théoriques détachés du monde professionnel. Selon Mons et al. (2012), cette typologie pourrait également être mise en relation avec l'expression de certaines attitudes chez les élèves. Les résultats de leurs travaux indiquent que le type « *éducation totale* » se traduirait chez les élèves par des attitudes plus favorables envers la compétition et la coopération que chez les élèves de type « *producteur* ». De plus et comme le soulignent Mons et al. (2012), la typologie proposée ne représente qu'un fragment de la variété des modèles éducatifs qui existent dans notre société. *In fine*, ces différentes caractéristiques pourraient tout de même affecter l'attention portée aux méthodes de travail collaboratives et dans le cas présent, le nombre de travaux réalisés sur la méthode Puzzle.

Figure 4

Nombre d'articles publiés sur la classe Puzzle en fonction du pays

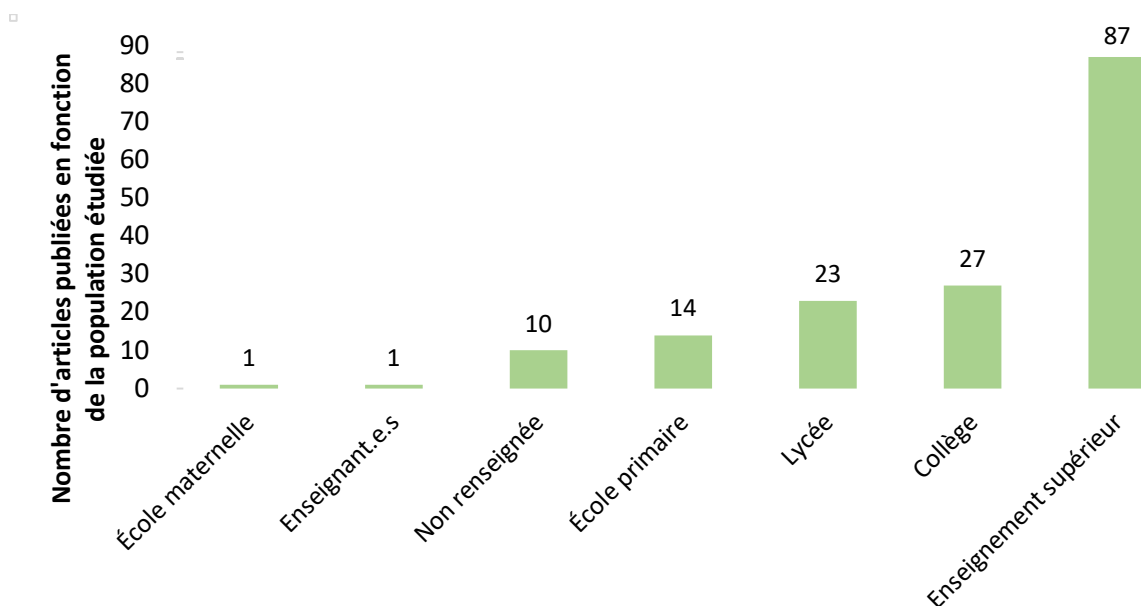


4.3 Nombre d'articles publiée sur la classe Puzzle en fonction de la population étudiée

La Figure 5 recense les populations étudiées dans le cadre des travaux réalisés sur la classe Puzzle. La population la plus étudiée est celle de l'enseignement supérieur ($n = 87$; pour plus de détails, voir Figure 5), sans surprise dans la mesure où les étudiant·e·s inscrits à l'université constituent aussi la population à laquelle les chercheur·e·s, pour la plupart des universitaires, ont le plus facilement accès. Seules les études de Darnon et al. (2012) et de Lai et al. (2015) portent sur la population des lycéen·e·s. Quasiment aucune étude n'intègre des élèves du niveau primaire (ou maternel), alors que l'enseignement primaire était l'objet des premiers travaux conduits en référence à la classe Puzzle.

Figure 5

Nombre d'articles publiés en fonction de la population étudiée



4.4 Expériences princeps

Ainsi, la première expérience qui examina les effets de la méthode Puzzle fut conduite par Blaney et al. (1977). Cette dernière se déroula dans dix classes de 5^{ème} (i.e., CM2) de sept écoles primaires à Austin (Texas). Trois classes furent également sélectionnées dans ces écoles en tant que conditions contrôle. Ces dernières étaient animées par des enseignant.e.s utilisant une pédagogie traditionnelle et qui étaient évalués par leurs pair.e.s comme étant très compétent.e.s. Les élèves des classes exposées à la méthode Puzzle se retrouvèrent en groupe environ 45 minutes par jour, trois fois par semaine pendant six semaines. Le programme scolaire était similaire dans les classes exposées à la méthode Puzzle et dans celles pilotées par une pédagogie plus traditionnelle. Les résultats de ces travaux révélèrent une augmentation significative de l'estime de soi chez les élèves exposés à la méthode Puzzle, et cela aussi bien chez les élèves issus de minorités que chez leurs homologues blancs. Ces derniers rapportèrent également des attitudes générales plus positives vis-à-vis à la fois de leur propre groupe et d'autres groupes ethniques. De même, les élèves exposés à la méthode Puzzle rapportèrent aimer plus l'école

que ceux exposés à une pédagogie plus traditionnelle. Cependant, ce ne fut pas le cas pour les élèves d'origine mexicaine. Selon Aronson (1978), ce résultat était dû au fait que ces enfants avaient eu des difficultés en anglais, créant un certain embarras avec le fait de travailler dans un groupe dominé par des anglophones. En effet dans une salle de classe traditionnelle, il est relativement facile pour les élèves de devenir « invisible » en restant silencieux ou en refusant de participer. Cette conduite est en revanche difficilement envisageable dans une classe Puzzle.

Lucker et al. (1976) s'intéressèrent quant à eux à l'impact de la méthode Puzzle sur le développement des compétences académiques des élèves. Ils réalisèrent leurs travaux sur 303 élèves de 5^{ème} et 6^{ème} années (i.e., CM2 et 6^{ème}). Six classes furent exposées à la méthode des classes Puzzle et cinq à une pédagogie plus traditionnelle. Pendant deux semaines les élèves étudièrent l'Amérique coloniale, puis furent évalués à l'aide d'un test standardisé. Les résultats montrèrent que le pourcentage de bonnes réponses des élèves issus des minorités était significativement meilleur lorsque ces derniers étaient exposés à la méthode de la classe Puzzle plutôt qu'à une pédagogie traditionnelle. Chez leurs homologues blancs, cependant, aucune différence n'était observée entre les deux conditions de l'étude. Une exposition de deux semaines à la classe Puzzle avait donc suffi à réduire l'écart de performance entre les enfants issus des minorités et leurs homologues blancs.

Bridgeman (1981) s'est quant à elle intéressée aux effets de la méthode Puzzle sur le développement de l'empathie. Elle présenta une série de « cartoons » à des enfants de dix ans dont la moitié étaient exposés pendant huit semaines à la méthode des classes puzzle. Ces cartoons étaient conçus dans le but de mesurer la capacité des enfants à faire preuve d'empathie. Sur la première vignette, on pouvait voir un petit garçon qui semblait triste quand il disait au revoir à son père à l'aéroport. Sur la suivante l'enfant recevait un colis d'un facteur. Après avoir ouvert le colis et avoir découvert qu'il contenait une maquette d'avion, l'enfant éclatait en sanglots. Il était ensuite demandé aux enfants d'expliquer la raison pour laquelle le petit garçon

pleurait. La majorité d'entre eux répondirent que l'avion rappelait au petit garçon le fait qu'il était séparé de son père et que c'était pour cette raison qu'il était triste. Pour finir, Bridgeman demanda aux enfants ce que pensait le facteur qui livrait le colis. La plupart des enfants de cet âge firent la même erreur basée sur l'hypothèse que leurs propres connaissances sont universelles. Ils supposèrent à tort que le facteur saurait que le garçon était triste car le cadeau lui rappelait le départ de son père. Toutefois, les réponses des enfants exposés à la méthode Puzzle étaient différentes. En effet, ces derniers étaient capables de prendre en compte le point de vue du facteur en comprenant qu'il allait éprouver une certaine confusion en voyant un petit garçon pleurer en recevant un beau cadeau car il n'avait pas connaissance de la scène qui s'était déroulée à l'aéroport. Ces premiers résultats fondamentaux furent ainsi répliqués dans différents états comme à Watsonville en Californie (e.g., Geffner, 1978). Cependant, de tels effets ne sont pas observés lorsque l'empathie est mesurée au moyen de dilemmes moraux.

Moskowitz et al. (1983, 1985) sont les premiers à nuancer les évaluations antérieures de la méthode Puzzle qui leurs apparaissent fragiles. Les résultats de leurs travaux réalisés sur 384 élèves de 5^{ème} année (i.e., CM2) ne révélaient d'effets significatifs de la méthode Puzzle ni sur les performances académiques, ni sur l'estime de soi, ni sur les attitudes à l'égard de l'école (ni sur le locus de contrôle interne vs externe). De même, dans la revue de Newman et Thompson (1987), la méthode Puzzle était la méthode d'apprentissage coopérative la moins efficace lorsqu'il s'agissait de favoriser les performances académiques des élèves dans des disciplines telles que les sciences ou les langues. Selon Moskowitz et collaborateurs (1983, 1985), cette absence d'effets positifs pourrait être dû à un problème dans la structure même de la méthode Puzzle. Ainsi, l'absence de toute production ou récompense commune aurait pour effet de réduire drastiquement la mise en œuvre d'une interdépendance positive entre les élèves alors confrontés à un système de récompense plutôt individualiste voire même compétitif. C'est aussi ce constat qui amena Slavin (1980) à apporter les premières modifications à la version initiale

de la méthode Puzzle. Les modifications et autres variantes de la méthode puzzle (cf. Tableau 1 et Annexe A) n'ont cependant pas donné lieu à beaucoup de travaux pour en valider l'efficacité. C'est pourquoi plutôt que de les présenter dans le corps du texte nous les présentons en annexe de manière néanmoins assez détaillée (cf. Annexe A).

4.5 Nombre d'article publiés en fonction de la méthode Puzzle employée

La Figure 6 répertorie le nombre d'articles publiés en fonction de la méthode Puzzle employée. Le recours à la version originelle de cette méthode est largement majoritaire (n = 94). La classe Puzzle a également fait l'objet de différents développements : *Jigsaw II*, *Jigsaw III* et *Jigsaw IV*, dont les éléments de bases restent assez similaires (pour plus de détails, voir Tableau 1). Néanmoins, très peu d'études ont été réalisées sur la 3^{ème} (n = 1) et la 4^{ème} version (n = 5) de la classe Puzzle ainsi que sur ses variantes, les *Reverse* et *Subject Jigsaw* (n = 1 et n = 5 respectivement).

Figure 6

Nombre d'articles publiés en fonction de la méthode Puzzle employée

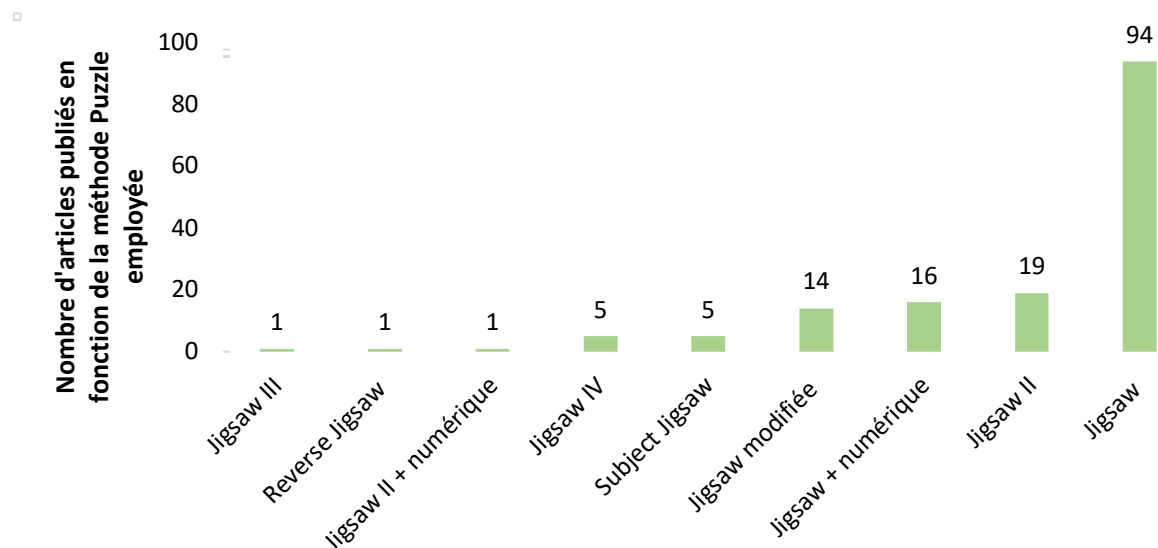


Tableau 1

Comparaison des quatre premières versions de la méthode Puzzle

	Jigsaw I (Aronson et al, 1978)	Jigsaw II (Slavin,1986)	Jigsaw III (Stahl, 1994)	Jigsaw IV (Holliday, 1995)
	Formation des			Introduction au matériel
1.	groupes d'élèves et travail individuel	<i>Idem à Jigsaw I</i>	<i>Idem à Jigsaw I</i>	
	Travail en groupe expert			Formation des groupes d'élèves et travail individuel
2.		<i>Idem à Jigsaw I</i>	<i>Idem à Jigsaw I</i>	
	Retour des élèves dans leur groupe d'origine et partage des connaissances			Travail en groupe expert
3.		<i>Idem à Jigsaw I</i>	<i>Idem à Jigsaw I</i>	
	Évaluation et score individuel	Évaluation individuelle et score du groupe	Évaluation des connaissances e.g. quiz bowl	Quizz sur le matériel dans les groupes experts
4.				Retour des élèves dans leur groupe d'origine et partage des connaissances
5.			Évaluation individuelle et score du groupe	

6.	Quizz sur le matériel partagé dans les groupes
7.	Évaluation des connaissances (e.g., quiz bowl)
8.	Évaluation individuelle et score du groupe
9.	Approfondissement du matériel si nécessaire

4.6 Classe Puzzle et usage des technologies de l'information et de la communication

L'engouement pour l'usage des technologies de l'information et de la communication (TIC) dans le domaine du travail collaboratif (ou *computer-supported collaborative learning*, CSCL) n'a pas échappé à la communauté des chercheurs intéressés par la classe Puzzle. En effet, 17 études implémentent la classe Puzzle par le biais d'un outil numérique (i.e., ordinateur portable, tablette, programme informatique, environnement numérique, etc.). Dans ces études, la technologie est utilisée comme un outil de médiation dans le but de soutenir le processus coopératif en facilitant notamment les interactions entre les pair-es. C'est par exemple le cas des travaux de Huang et al. (2014) réalisés sur 63 étudiant-es qui, pendant une activité qui portait sur l'écologie fluviale de la région de Taïwan, étaient exposés soit à la méthode Puzzle soit à une pédagogie plus traditionnelle (i.e., travail individuel). Pour ce faire, ces dernier-es

avaient tous à leur disposition des tablettes qui leurs permettaient d'accéder à une plateforme d'apprentissage Google. Néanmoins, les étudiant·e·s du groupe expérimental (i.e., méthode Puzzle) étaient les seul·e·s à pouvoir utiliser la fonction de messagerie instantanée ou celle de vidéoconférence pour participer à des discussions en temps réel avec leurs pair·e·s. Les résultats de ces travaux montrent une augmentation significative des performances post-test chez les étudiant·e·s exposé·e·s à la méthode Puzzle. De plus, ces derniers rapportaient être plus satisfaits de l'utilisation de la plateforme d'apprentissage en ligne que leurs homologues du groupe contrôle (i.e., travail individuel). Cependant il est à noter que ces études, plutôt que d'évaluer l'efficacité de la méthode Puzzle, s'inscrivent surtout dans des perspectives centrées sur l'utilisation et l'acceptabilité des outils numériques.

Ainsi, les éléments de base des quatre premières versions de la méthode Puzzle restent assez similaires (pour plus de détails, voir Annexe A). Nous constatons que les versions I et II sont celles les plus couramment étudiées dans la littérature scientifique ($N_{Jigsaw\ I} = 94$, $N_{Jigsaw\ II} = 19$; cf. Figure 6 et Annexe A) et qu'à l'inverse, très peu d'études ont été réalisées sur la 3^{ème} et la 4^{ème} version de la méthode Puzzle, ainsi que sur ses variantes, les *Reverse* et *Subject Jigsaw*. Il est également à noter que les travaux consacrés à la méthode Puzzle et ses extensions ont mesuré essentiellement les performances académiques des élèves.

5 Développer les compétences socio-cognitives par le biais de la méthode Puzzle ?

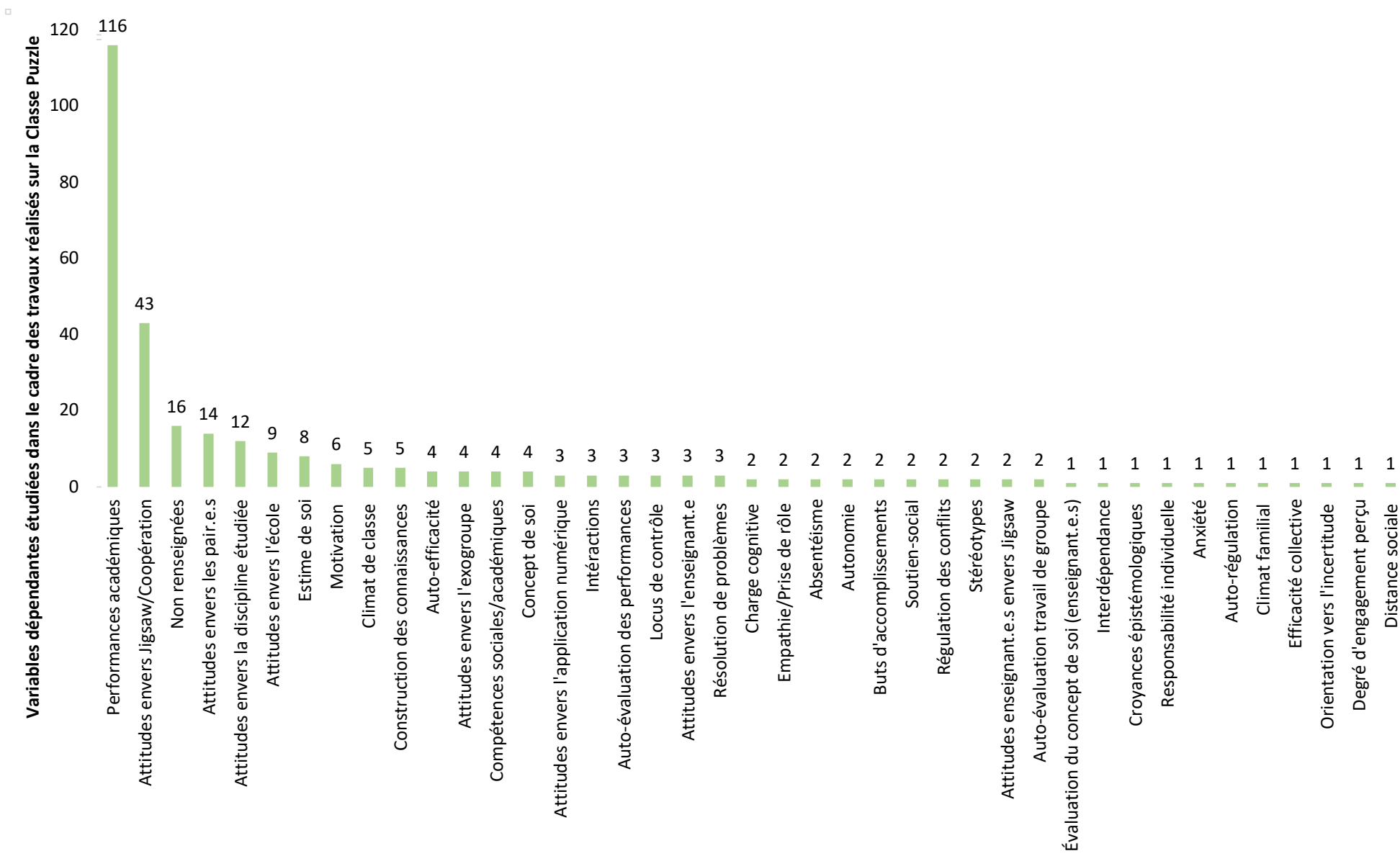
5.1 Variables dépendantes étudiées dans le cadre des travaux réalisés sur la méthode Puzzle

La Figure 7 recense les variables dépendantes étudiées dans les travaux sur la classe Puzzle. La performance académique est de très loin la variable la plus étudiée ($n = 116$). Bien que cette méthode ait été conçue à l'origine dans le but de favoriser l'intégration scolaire des minorités ethniques, les variables de nature à capturer tout changement dans les relations

intergroupes ont été manifestement délaissées. Par ailleurs, les données rapportées dans la Figure 7 suggèrent que les compétences dites « sociales » (i.e., collaboration, communication, résolution de problème, créativité, etc.) ont fait elles-aussi l'objet d'un nombre très restreint d'études, alors que la littérature suggère que la classe Puzzle, comme d'autres méthodes d'apprentissage coopératif, contribue à leur développement. Les travaux réalisés dans ces différents domaines sont néanmoins rapidement évoqués ci-dessous.

Figure 7

Nombre de variables dépendantes étudiées en fonction des articles publiés



5.2 Relations intergroupes et dynamique de groupe

Développée à la suite de la déségrégation des écoles aux USA (cf. supra), la classe Puzzle fut présentée à de nombreuses reprises comme une méthode de choix lorsqu'il s'agissait entre autres de favoriser les relations intergroupes ou de diminuer les préjugés. En effet, le paradigme de la classe Puzzle répondrait aux conditions prescrites par l'hypothèse du contact d'Allport (1954) qui, comme nous l'avons vu précédemment, suggère que la coopération interethnique se traduit par une réduction des préjugés. Néanmoins, les recherches réalisées à ce propos rapportent des résultats en demi-teinte. Les travaux de Blaney et al. (1977), Bridgeman (1981), Walker et Crogan (1998) ou encore plus récemment Rego et Moledo (2005) rapportent des effets positifs de la méthode Puzzle sur les relations interethniques (e.g., diminution des traits négatifs attribués aux membres de l'exogroupe). Néanmoins, les travaux de Moskowitz et al. (1983, 1985) et Bratt (2008) ne révèlent quant à eux aucun effet de la méthode Puzzle sur les relations intergroupes. Selon Bratt (2008), les travaux rapportant des effets positifs de la méthode Puzzle sur les relations intergroupes présenteraient en réalité un certain nombre de limites méthodologiques (i.e., tests statistiques non appropriés, non-équivalence des groupes, plan quasi-expérimental, effet de l'enseignant·e et non de l'intervention, etc.) et remettraient en question l'efficacité de la méthode concernant la réduction des préjugés interethniques.

5.3 Sentiment d'auto-efficacité

Selon Hänze et Berger (2007), la méthode Puzzle permettrait de satisfaire les besoins fondamentaux mentionnés dans la théorie de l'autodétermination (i.e., besoin de compétence, d'autonomie et d'appartenance sociale ; Deci & Ryan, 1985 ; 2000) et par conséquent, favoriserait un apprentissage profond, intrinsèquement motivé. En tant que méthode

d'apprentissage coopérative, la classe Puzzle devrait donc satisfaire le besoin d'être en relation avec autrui. En comparaison avec une méthode pédagogique plus traditionnelle, les élèves devraient également se sentir plus autonomes car une plus grande liberté leur serait accordée afin de structurer le processus d'apprentissage (i.e., autonomie dans la phase individuelle). Enfin, l'attribution claire des rôles et des tâches dans la classe Puzzle (i.e., le fait que chaque élève soit experte d'un segment d'informations) fournirait aux élèves l'occasion de se sentir plus compétents que dans une classe traditionnelle. Pour ce faire, 137 élèves de Terminale durant un cours de physique-chimie étaient exposés soit à la méthode de la classe Puzzle, soit à une méthode pédagogique plus traditionnelle (i.e., travail individuel). Conformément aux hypothèses de Hänze et Berger (2007), les élèves exposés à la méthode Puzzle se percevaient plus compétents, plus autonomes et plus socialement liés à leurs camarades de classe que ceux du groupe contrôle. Néanmoins, la hausse des performances en physique-chimie n'était observée que chez les étudiant·es qui rapportaient une hausse dans ces besoins fondamentaux. Pour les autres, les effets de la classe Puzzle étaient nuls, voire même négatifs.

Les travaux de Darnon et al. (2012) étaient quant à eux réalisés sur 33 élèves d'un lycée professionnel qui étaient exposés durant quatre cours soit à la méthode Puzzle, soit à une méthode pédagogique plus traditionnelle (i.e., travail individuel). Le sentiment d'auto-efficacité des élèves en français et en mathématiques était mesuré avant et après l'expérimentation par le biais d'une échelle de Likert en 5 points. Les résultats de ces travaux indiquent que le sentiment d'auto-efficacité en français et en mathématiques augmentait uniquement chez les élèves exposés à la méthode Puzzle. Selon ces auteur·es, les méthodes d'apprentissage comme la classe Puzzle amélioreraient le sentiment d'auto-efficacité des élèves car elles leur fourniraient une plus grande expérience de maîtrise (i.e., chaque élève a la possibilité de devenir « expert·e » et d'expliquer le corpus d'informations qui lui a été attribué aux membres de son groupe) ainsi qu'une plus grande expérience vicariante (i.e., dans la

phase « puzzle » les sources d'information ne sont autres que les pair-es). Cependant la taille de l'échantillon, là encore, invite une nouvelle fois à la prudence pour conclure fermement au sujet des bénéfices attachés à la classe Puzzle.

Des résultats similaires sont rapportés par Crone et Portillo (2013) chez 70 étudiant-e-s pendant un cours de psychologie qui étaient exposés soit à la méthode Puzzle, soit à une version réduite de la classe Puzzle, soit à une méthode pédagogique plus traditionnelle (i.e., travail individuel). Les résultats de cette étude indique que la confiance des étudiant-e-s en leur capacité à communiquer oralement des explications ainsi que leur sentiment d'auto-efficacité académique augmentaient chez ceux exposés à la méthode Puzzle. Ces auteurs n'observaient néanmoins pas d'effet de la classe Puzzle sur les performances académiques des étudiant-e-s.

5.4 Performances académiques

En comparaison des variables présentées précédemment, les performances académiques sont celles les plus largement testées dans le cadre des travaux réalisés au sujet de la méthode Puzzle (n = 116 études). Néanmoins, les preuves de l'efficacité de cette dernière sur les performances académiques et cela, en comparaison au travail individuel, demeurent pour le moins ambiguës. La méthode Puzzle est en effet supposée favoriser la mise en œuvre d'interactions dites « facilitatrices » qui peuvent prendre la forme de comportements d'assistance, d'aide, et d'échange de ressources et d'informations (Johnson & Johnson, 1989 ; 2002), susceptibles de permettre une compréhension plus approfondie du corpus d'informations étudié et l'utilisation de compétences cognitives élevées (i.e., poser des questions, synthétiser, expliquer, etc.). Cependant, un tel effet n'est pas toujours observé. Ainsi, les études qui testent l'effet de la classe Puzzle sur les performances académiques rapportent à la fois des effets positifs (e.g., Dori et al., 1995 ; Lazarowitz et al., 1994 ; Luckier et al, 1976 ; Tarhan et al., 2013)

et nuls (e.g., Berger & Hänze, 2009 ; Hänze & Berger, 2007 ; Moskowitz et al., 1985, Souvignier & Kronenberg, 2007).

Selon Moskowitz et al. (1983 ; 1985), qui sont parmi les premiers à remettre en question l'efficacité de la méthode Puzzle, l'absence de récompense commune aurait pour effet de fortement diminuer l'interdépendance positive entre les élèves, pourtant indispensable au bon déroulement de la classe Puzzle. Pour rappel, c'est aussi ce constat qui conduisit Slavin (1978 ; 1980) à développer une 2^{ème} version de la méthode Puzzle dans laquelle il proposa d'ajouter une récompense de groupe afin d'accroître l'interdépendance positive. Cependant, aucune étude n'a pour le moment cherché à comparer l'efficacité de la 2^{ème} version de la classe Puzzle comparativement à la 1^{ère}. Plus récemment, Roseth et al., (2018) font valoir de nouveaux éléments susceptibles d'expliquer les effets contrastés de la méthode Puzzle s'agissant des performances académiques. Au-delà des objections habituelles qui consistent à affirmer que le principal inconvénient de la méthode Puzzle réside dans le fait que les élèves sont exposés de manière limitée aux contenus pédagogiques qu'ils n'ont pas étudiés eux-mêmes (e.g., Nolan et al., 2018), Roseth et al. (2018) suggèrent que la procédure employée dans la méthode Puzzle (i.e., Phase « expert » et Phase « puzzle ») pourrait en réalité engager différents types de relations entre les élèves qui provoqueraient des effets opposés sur l'apprentissage. En effet dans la phase « expert », le fait de travailler avec d'autres élèves sur le même matériel engendrerait une indépendance des ressources, qui aurait pour conséquence d'orienter les élèves vers la compétition. Dans la phase « puzzle », le fait de travailler avec des pair-e-s sur un corpus d'informations complémentaire impliquerait non seulement des relations d'interdépendance positive, mais également des relations compétitives dès que les membres du groupe ont accès aux ressources des un-e-s et des autres. Les résultats de l'étude de Roseth et al. (2018) réalisée sur 258 étudiant-e-s durant un cours d'anatomie humaine montrent de faibles effets de la méthode Puzzle sur les performances académiques.

En conclusion, les données descriptives présentées précédemment semblent assez peu compatibles avec les affirmations suggérant un vaste intérêt de recherche pour la classe Puzzle, dont l'efficacité aurait été démontrée par une vingtaine d'années d'expérimentations rigoureusement menées dans le monde entier, pour reprendre les mots d'Aronson (2011). Quarante ans après la naissance de cette méthode, les publications scientifiques à son sujet demeurent peu nombreuses, très disparates géographiquement (concentrées pour l'essentiel sur les États-Unis et la Turquie), et majoritairement focalisées sur la question des performances académiques chez les étudiant·e·s de l'enseignement supérieur en dépit d'un intérêt premier pour des mesures plus en rapport avec la tolérance et la réduction des discriminations raciales parmi les élèves de l'enseignement primaire. S'il est vrai que les travaux princeps sur la classe Puzzle ont bien intégré des considérations et mesures en rapport avec les dynamiques intergroupes, l'estime de soi, ou encore l'empathie, ces dernières sont désormais totalement délaissées au profit de mesures des performances académiques. Aronson (2011) assure par exemple que les effets de la classe Puzzle sur l'empathie sont bien documentés alors qu'elles ne sont en réalité intégrées qu'à deux études (Bratt, 2008 et Bridgeman, 1981). Par ailleurs, très peu de facteurs médiateurs ou modérateurs ont été testés pour identifier les processus supposés à l'œuvre dans la classe Puzzle. Enfin, et peut-être surtout, les travaux sur cette méthode présentent de nombreuses limites méthodologiques (cf. Chapitre 2 de ce manuscrit) qui rendent difficile, voire impossible, toute conclusion ferme s'agissant de son efficacité.

C'est pourquoi la suite de notre développement a pour objectif de réexaminer les études en rapport avec la classe Puzzle sur la base de leurs forces et faiblesses méthodologiques, pour *in fine* conclure sur la taille des effets obtenus dans les études jugées les meilleures, comparativement à celles plus faibles. Les travaux qui évaluent les effets de l'apprentissage coopératif sur la réussite globale des élèves sont rassemblés depuis les années 1980 sous la forme de méta-analyses regroupant en général les résultats de plusieurs dizaines voire centaines

d'études (e.g., Huddy, 2013 ; Hilk, 2013 ; Johnson & Johnson, 2009 ; Johnson et al., 2000 ; Newman & Thompson, 1987 ; Slavin, 1980, 1996). Néanmoins, et même si certains ont testé les effets de la classe Puzzle (e.g., Johnson et al., 2000 ; Kyndt et al., 2013 ; Stanczak, 2020), aucun de ces travaux n'a spécifiquement ciblé les effets de la classe Puzzle au cours des quarante dernières années. L'application de cette méthode pédagogique en milieu scolaire ou universitaire requiert pourtant en toute rigueur d'en estimer précisément l'efficacité. La performance académique étant de très loin la variable dépendante la plus étudiée (cf. Figure 7), les éléments méta-analytiques fournis dans le chapitre suivant concernent exclusivement cette partie majoritaire de la littérature consacrée à la méthode Puzzle.

Chapitre 2 : La classe Puzzle, éléments méta-analytiques

La suite de notre développement a pour objectif d'estimer précisément les soutiens empiriques à la méthode Puzzle. Dans ce but, nous proposons une synthèse quantitative des résultats produits en référence à cette méthode, avec une focalisation sur la question de son influence sur les performances académiques (variable dépendante très majoritaire, cf. Chapitre 1). Il ne s'agit pas néanmoins d'une méta-analyse quantitative au sens strict du terme, en raison du caractère souvent incomplet des informations méthodologiques et/ou statistiques consignées dans la littérature de référence. Dans un premier temps, nous calculerons la taille de l'effet Puzzle sur les performances académiques à partir des paramètres disponibles dans les études sélectionnées dans notre synthèse. Dans un second temps, nous construirons un indice de « qualité méthodologique » sur la base de 6 critères pour permettre à terme un classement de ces mêmes études selon 5 degrés de qualité méthodologique : « Médiocre », « Plutôt faible », « Intermédiaire », « Plutôt Forte » et « Excellente ». Enfin, nous discuterons de la répartition des études qui ont testées les performances académiques en fonction de leur qualité méthodologique ainsi que des facteurs susceptibles d'expliquer la grande hétérogénéité s'agissant de la taille des effets Puzzle que nous avons calculé.

1 Synthèse quantitative : Critères d'inclusion et d'exclusion des études

Parmi les 151 références incluses dans notre synthèse qualitative de la littérature (cf. graphique PRISMA du Chapitre 1), nous avons exclu de notre examen plus quantitatif 110 d'entre-elles sur la base des critères suivants : échantillon inférieur à 20 participant·e·s (N = 7 études), variantes de la classe Puzzle jugées trop éloignées de la version princeps (i.e., *Reverse Jigsaw* et *Subject Jigsaw* ; cf. Annexe A ; N = 6 études), absence de groupe contrôle (N = 38

études), absence des éléments statistiques (i.e., moyenne et écart-type) nécessaires au calcul d'une taille d'effet ($N = 59$ études). Ce taux élevé d'exclusion, soit plus de 72% des travaux testant majoritairement l'influence de la classe Puzzle sur les performances académiques, exprime d'emblée la faiblesse méthodologique de ce champ de recherche et constitue en soi un premier résultat—assez surprenant—de notre approche plus quantitative de la littérature Puzzle.

1.1 Taille d'effet

La taille des effets observés dans chacune des 41 études incluses a été calculée avec la procédure classique du d de Cohen, correspondant à la différence entre les moyennes des deux groupes (Puzzle vs. Contrôle), divisée par l'écart-type intra-étude. Cet indicateur d , ainsi que la variance associée, ont été ensuite systématiquement corrigés par un facteur J pour estimer le paramètre « g » de Hedge permettant d'éviter toute surestimation de la taille d'effet pour les petits échantillons (Borenstein et al., 2011, Stanczak, 2020). Dans le cas des études qui rapportaient plusieurs groupes expérimentaux (Souvignier & Kronenberger, 2007 ; Law, 2011 ; Iweka, 2017 pour l'utilisation de trois groupes expérimentaux) nous avons seulement estimé la taille d'effet de la différence entre le groupe exposé à la méthode Puzzle et le groupe contrôle. De plus, et dans la mesure où une part importante des 41 études considérées dans notre synthèse rapportaient plusieurs mesures d'apprentissages ($N = 36$), nous avons combiné les scores obtenus sur ces différentes mesures puis nous en avons fait la moyenne. Sur la base des recommandations de Borenstein et collaborateurs (2010), nous avons également estimé un coefficient de corrélation inter-mesures fixé à $r = .71$, dans le but d'appliquer une correction assez forte et homogène sur ces mesures (Stanczak, 2020). Les formules utilisées pour calculer la taille d'effet moyenne sont présentées dans le Tableau 2 et le g de Hedge pour chacune des 41 études dans le Tableau 3.

Tableau 2

Formules de calcul des estimations de taille d'effet moyenne (Borenstein et al., 2010)

<i>d</i> de Cohen	$\frac{\text{Moyenne (Jigsaw)} - \text{Moyenne (Contrôle)}}{\text{Écart - type intra étude (S)}}$
<i>S</i> (Écart-type intra-étude)	$\sqrt{\frac{(n1 - 1)SD1^2 + n2(-1)SD2^2}{n1 + n2 - 2}}$
Correction <i>J</i>	$J = 1 - \frac{3}{4(df-1)} \text{ où } df = (n1+n2)-2$
<i>g</i> de Hedge	$g = d * J$
Intervalle de confiance	$IC = g \pm (1.96 * \sqrt{\text{Variance de } g})$

Notes. Une interprétation communément admise consiste à désigner la taille de l'effet comme étant petite ($g = 0.2$), moyenne ($g = 0.5$) et grande ($g = 0.8$) sur la base des repères suggérés par Cohen (1988, 1992).

Tableau 3

Tailles d'effets moyennes estimées (g de Hedge) pour les performances académiques (IC 95% [1.00, 1.73]) et présentées par ordre alphabétique des auteurs cités en premier nom dans les articles considérés

Auteurs	VD	Performances académiques	Variance	IC 95%
Akçay 2016 ²		$g = 1.96$	0.15	[1.21, 2.72]
Alamri 2018 ^{1 2}		$g = .81$	0.12	[0.11, 1.50]
Al-Salkhi 2015 ^{1 2}		$g = .36$	0.07	[-0.17, 0.89]
Artut et Tarim 2007 ^{1 2}		$g = 1.11$	0.05	[0.64, 1.57]
Aydin et Biyikli 2017 ^{1 2}		$g = 1.31$	0.07	[0.77, 1.84]
Basyah et al. 2018 ²		$g = 3.17$	0.14	[2.42, 3.92]
Çagatay et Demircioglu 2013 ^{1 2}		$g = 0.74$	0.09	[0.16, 1.33]

Chu 2014 ²	$g = .62$	0.03	[0.26, 0.98]
Dori et al. 1995 ²	$g = 1.52$	0.05	[1.09, 1.95]
Doymus 2008 ^{1 2}	$g = 1.73$	0.05	[1.26, 2.18]
Evcim et Ipek 2012 ²	$g = 1.30$	0.09	[0.69, 1.91]
Farahnaz et al. 2013 ²	$g = 1.06$	0.13	[0.33, 1.78]
Ghaith et Abd El-Malak 2004	$g = .25$	0.06	[-0.23, 0.74]
Göcer 2010 ^{1 2}	$g = 4.29$	0.21	[3.38, 5.20]
Gömleksi'z 2007 ²	$g = 2.42$	0.11	[1.79, 3.04]
Hornby 2009 ²	$g = .75$	0.09	[0.15, 1.35]
Hosseini et al. 2014 ^{1 2}	$g = .91$	0.09	[0.30, 1.52]
Huang et al. 2014 ²	$g = 2.59$	0.11	[1.93, 3.25]
Iweka 2017 ^{1 2}	$g = 0.96$	0.05	[0.50, 1.42]
Jafariyan et al. 2017 ^{1 2}	$g = 1.55$	0.14	[0.80, 2.29]
Karacop et Diken 2017 ^{1 2}	$g = 1.43$	0.10	[0.80 ; 2.05]
Kumar et al. 2017 ^{1 2}	$g = 1.71$	0.10	[1.07 ; 2.35]
Law 2011 ¹	$g = .044$	0.02	[0.15 ; 0.74]
Lazarowitz et al. 1994 ¹	$g = .36$	0.05	[-0.08, 0.80]
Marhamah et Mulyadi 2013 ^{1 2}	$g = 5.70$	0.37	[4.49, 6.90]
Mari et Gumel 2015 ²	$g = 1.11$	0.05	[0.65, 1.56]
Mattingly et VanSickle 1991	$g = 0.01$	0.08	[-0.56, 0.58]
Mutlu 2018 ²	$g = 0.86$	0.08	[0.29, 1.42]
Nebel et al. 2017 ¹	$g = 1.08$	0.08	[0.52, 1.64]
Özdemir et Arslan 2016 ²	$g = 2.37$	0.16	[1.57, 3.16]
Roseth et al. 2019	$g = 0.26$	0.10	[0.07, 0.45]
Sahin 2010 ^{1 2}	$g = .96$	0.05	[0.50, 1.42]
Sahin 2011 ²	$g = .85$	0.06	[0.37, 1.33]
Souvignier et Kronenberger 2007 ²	$g = -0.48$	0.02	[-0.78, -0.18]
Tarhan et al. 2013 ¹	$g = 2.57$	0.11	[1.89, 3.23]
Tarhan et Sesen 2012 ¹	$g = 1.51$	0.12	[0.80, 2.21]
Tran et Lewis 2012 ^{1 2}	$g = .66$	0.05	[0.21, 1.10]
Ural et al. 2017 ^{1 2}	$g = 1.38$	0.09	[0.76, 1.99]
Yapici 2016 ²	$g = 2.64$	0.13	[1.92, 3.37]
Yoruk 2016 ¹	$g = .61$	0.07	[0.08, 1.12]

Zahra 2014 ^{1 2}	$g = .68$	0.10	[0.05, 1.30]
	$M_{g\ de\ Hedge} = 1.36$	$M_{variance} =$	
		0.11	

Notes. ¹ Implémentation de la méthode Puzzle (e.g., modifications mineures ou majeures apportées à la version initiale ou pas d'indications concernant l'implémentation de la méthode) ; ² Limites méthodologiques (e.g., mesure des performances pré- et post-test similaires, utilisation de tests statistiques non adaptés)

Nous constatons dans le Tableau 3 que dans plus de 80 % des cas, les effets Puzzle que nous avons calculés sont de taille moyenne voire de grande taille (i.e., $g > 0.5$). Cette observation est néanmoins à interpréter avec beaucoup de précaution. En effet, la grande majorité des références mentionnées dans le Tableau 3, et de façon générale la littérature consacrée à la classe Puzzle, comportent des biais méthodologiques plus ou moins importants. Soit nous constatons que des modifications sont apportées à la version initiale de la méthode Puzzle (sans pour autant correspondre à l'une de ses variantes identifiées, cf. Annexe A), rendant son implémentation probablement non fidèle, soit nous ne disposons pas assez d'informations permettant d'attester de l'implémentation de cette dernière. À titre d'exemple, une version caricaturale mais bien réelle de ce déficit d'information est présente notamment dans les travaux de Mari et Gumel (2015) : « *Le groupe expérimental a été exposé à la méthode d'apprentissage coopérative de la classe Puzzle, tandis que le groupe contrôle a été exposé à un cours magistral* » (Mari et Gumel, 2015, p.198, notre traduction). Encore faudrait-il décrire la manière dont la méthode évoquée a été opérationnalisée et implémentée *in situ*.

Or, la fidélité avec laquelle une intervention est implémentée peut évidemment affecter ses résultats (Caroll et al., 2007 ; Lortie-Forgues & Inglis, 2019). Par conséquent, pour conclure à l'efficacité d'une méthode ou intervention quelle qu'elle soit, il est primordial de procéder en premier lieu à l'évaluation de la fidélité avec laquelle elle a été mise en œuvre. Selon Caroll et

al. (2007), l'absence d'une telle évaluation rend toute conclusion ferme impossible si bien que la présence d'un effet, ou au contraire son absence, ne peut pas être attribuée à une mauvaise implémentation ou à des insuffisances inhérentes à l'intervention en question. Or, une confiance relativement forte dans les données expérimentales est particulièrement important pour les travaux scientifiques qui ont pour objectif de valider empiriquement des méthodes ou des interventions à destination des enseignant-e-s (Johnson & Johnson, 2000).

Plus généralement, un examen détaillé des études intégrées au Tableau 3 fait apparaître des difficultés de différentes natures pour la plupart d'entre-elles. Ces difficultés sont en rapport avec le groupe contrôle, l'implémentation de l'intervention, le plan expérimental lui-même, la taille de l'échantillon, la durée de l'expérimentation, la qualité et la mesure de la variable dépendante. Pris dans leur ensemble, ces éléments viennent non seulement questionner la qualité méthodologique des études empiriques conduites au sujet de la classe Puzzle, mais également la crédibilité scientifique que nous pouvons accorder à leurs résultats.

1.2 Critères méthodologiques

Les faiblesses méthodologiques évoquées plus haut font qu'il est difficile de conclure directement sur la taille moyenne de l'effet attaché à la classe Puzzle. Un travail supplémentaire est donc nécessaire (cf. infra) et consiste à estimer de manière la plus objective possible la qualité méthodologique de chacune des études publiées. Ce travail et sa raison d'être sont présentés ci-dessous.

1.2.1 Construction des critères méthodologiques

Pour chacune des 41 études présentées dans le Tableau 3 (celles focalisant sur les performances académiques en variables dépendantes), nous avons calculé un score (méthode des juges, cf. infra) sur la base des 6 critères méthodologiques suivants : a) la nature du groupe

contrôle, b) l'implémentation de l'intervention, c) le « design » expérimental, d) la taille de l'échantillon, e) la qualité et la mesure de la variable dépendante et f) la durée de l'expérimentation. Ces critères tiennent compte des faiblesses méthodologiques mentionnées antérieurement (cf. Point 2 du présent Chapitre). La qualité méthodologique de chaque étude a donc été réévaluée—la première évaluation incombant en principe aux comités de lecture eux-mêmes—sur ces six critères. Chaque critère donne lieu à un score, positif ou négatif, dont le cumul permet d'apprécier quantitativement la qualité méthodologique de chacune des études considérées. Cette approche permet à terme d'ordonner ces études, de celles jugées « Médiocres » à celles jugées « Excellentes », et ainsi de disposer d'une estimation moyenne de la taille de l'effet attaché à la classe Puzzle qui tient compte de la qualité des réalisations expérimentales ou quasi-expérimentales sous-jacentes.

Qu'elles soient narratives ou plus quantitatives, les méta-analyses en psychologie (comme en neurosciences) négligent trop souvent l'examen préalable de la qualité méthodologique des études rassemblées dans la perspective d'une vision synthétique d'un champ donné de recherche. Or, pour qui fréquente le monde de la recherche, il est bien évident que toutes les recherches publiées ne sont pas de même niveau méthodologique, et que toutes ne sont pas aussi crédibles les unes que les autres. Afin de juger de la crédibilité de travaux scientifiques, la communauté scientifique se fonde d'une façon générale sur la qualité des supports de publication, à savoir leur facteur d'impact. Sans disqualifier cette pratique (toutes les revues en effet ne se valent pas en termes de qualité scientifique), nous invitons à ne pas oublier pour autant que les travaux publiés dans de grandes revues scientifiques ne correspondent pas toujours à la rigueur méthodologique attendue. De même, nombre de travaux plus en correspondance avec les canons de la science sont pourtant publiés dans des revues comparativement plus modestes (Pansu et al., 2013). Pour cette raison, nous n'avons pas opté pour un classement des études qui serait fondé sur le facteur d'impact et/ou la renommée de

leurs supports de publication. Plutôt que d'adopter cette démarche, nous avons choisi de tenter d'objectiver les qualités méthodologiques de chacune des études considérées pour ensuite les ordonner dans cette dimension et calculer la taille de l'effet attaché à la classe Puzzle par catégorie d'études : celles jugées « Médiocres », « Plutôt faibles », « Intermédiaires », « Plutôt fortes », et celles jugées « Excellentes ».

Le nombre de points attribués (points positifs) ou retirés (points négatifs) selon les critères considérés est évidemment discutable du fait de sa nature arbitraire. Il n'y a en effet pas de méthode absolue pour décider par exemple d'attribuer 4 points aux études dotées d'échantillons conformes aux calculs de puissance de test (l'un des critères retenus), et d'en retirer 4 aux études avec un échantillon de taille insuffisante. Il reste qu'ajouter ou retirer des points selon que le critère en question est ou n'est pas satisfait fait globalement sens. De même, il n'existe pas de méthode absolue pour la pondération des critères elle-même, par exemple l'attribution ou le retrait de 4 points, plutôt que 2 par exemple, à tel ou tel des critères considérés. Il reste que, même relativement arbitraire, cette pondération permet d'éviter d'une part les excès d'une approche méta-analytique qui ne tiendrait pas compte dans le détail de la qualité méthodologique des travaux dans le détail, et d'autre part, un calcul de la taille d'effet à l'aveugle de la qualité méthodologique des études considérées.

1.2.2 Les critères et leur pondération

a) Groupes contrôles :

- Deux groupes contrôles en parallèle (3 études sur 41 soit 7 %) : + 6 points
 - Un groupe fondé sur une configuration de travail individuel excluant *de facto* toute interdépendance positive.
 - Un groupe fondé sur une configuration de travail collectif mais sans consigne d'interdépendance positive, qui n'exclut pas nécessairement la possibilité de

cette interdépendance en mode spontané néanmoins peu probable en l'absence de consigne explicite

- Un seul groupe contrôle fondé sur une configuration de travail individuel excluant *de facto* toute interdépendance positive (33 études sur 41 soit 81 %) : + 4 points
- Un seul groupe contrôle fondé sur une configuration de travail collectif mais sans consigne d'interdépendance positive, qui n'exclut pas nécessairement la possibilité de cette interdépendance en mode spontané néanmoins peu probable en l'absence de consigne explicite (5 études sur 41 soit 12%) : + 2 points

b) Implémentation de l'intervention :

- Pas ou peu d'informations sur l'implémentation de la méthode Puzzle (4 études sur 41 soit 10 %) : - 6 points
- Modifications problématiques : le retrait de la phase « expert », le mélange de certaines étapes, l'absence d'interdépendance positive ou de responsabilité individuelle (2 études sur 41 soit 5 %) : - 4 points
- Modifications mineures donc moins problématiques : les 3 étapes sont respectées mais une étape est ajoutée telle qu'un retour dans les groupes experts, une présentation devant toute la classe après la phase puzzle ou une récompense collective comme dans *Jigsaw II* (25 études sur 41 soit 61 %) : + 2 points
- Implémentation fidèle à la version originale de la méthode Puzzle (10 études sur 41 soit 24 %) : + 4 points

c) Design expérimental :

- Quasi-expérimentale, pas de randomisation (23 études sur 41 soit 56 %) : + 2 points
- Expérimentale, randomisation (18 études sur 41 soit 44 %) : + 4 points

d) Taille de l'échantillon :

Si la taille de l'échantillon indiquée dans l'étude est :

- Supérieure ou égale aux prédictions de G*power (3 études sur 41 soit 7 %) : + 4 points
- Inférieure ou égale aux prédictions de G*power (38 études sur 41 soit 93 %) : - 4 points

e) Qualité et mesure de la variable dépendante :

- Variable dépendante existante mais non mentionnée très explicitement, pas ou peu d'informations (6 études sur 41 soit 14 %) : - 4 points
- Évaluation pré- et post-test similaires en tout point (26 études sur 41 soit 63 %) : - 2 points
- Évaluation post-test uniquement, pas de mesure pré-test (2 études sur 41 soit 4 %) : + 2 points
- Évaluation pré- et post-test différentes (8 études sur 41 soit 19 %) : + 4 points

Pour des raisons évidentes, nous avons opté pour une pondération qui pénalise les publications ne mentionnant pas explicitement la mesure utilisée pour évaluer les performances. Nous appliquons aussi une pénalité, plus faible, aux publications décrivant des études dans lesquelles la même mesure des performances est utilisée avant et après l'intervention. En effet pour ces dernières, les élèves peuvent s'attendre aux questions du post-test (connaissant celles du pré-test). Dans ce cas, il n'est donc pas possible de déterminer avec certitude si les gains éventuels en termes d'apprentissage sont causés par l'intervention ou/et par un effet de familiarité avec le matériel (puisque les questions support de l'évaluation ont déjà été vues au pré-test), si ce n'est même par un travail plus ou moins important à la maison sur ces questions entre les deux temps de mesure dans le cas des

études impliquant au moins une journée entre ces deux moments clefs. Nous avons choisi, en revanche, de ne pas pénaliser les publications décrivant des études dans lesquelles les apprentissages des élèves sur le thème traité n'étaient évalués qu'une fois l'intervention terminée. Nous ajoutons donc des points à ces études, mais nous en ajoutons encore davantage à celles qui mesurent les performances avant et après l'intervention avec des évaluations de nature différente.

f) Durée de l'expérimentation :

- Non mentionnée (7 études sur 41 soit 17 %) : + 1 point
- Inférieure ou égale à 1 journée (5 études sur 41 soit 12 %) : + 2 points
- Supérieure ou égale à 1 journée (29 études sur 41 soit 71 %) : + 3 points

Ce critère n'est pas considéré majeur dans notre analyse, aussi avons-nous opté pour une pondération qui ne pénalise pas les publications n'en faisant pas mention. Plutôt que d'enlever un point, nous en ajoutons davantage aux études mentionnant ce critère avec des durées plus ou moins longues.

De cette façon, l'intérêt scientifique de chaque étude peut être évalué par un score variant de -24 à + 25 points. Dans le cas présent, trois juges ont évalué l'intérêt scientifique des 41 études répertoriées dans le Tableau 3, en attribuant à chacune d'elle un score calculé à partir des six critères mentionnés précédemment. Par la suite, et dans le but de mesurer quantitativement le degré de consensus entre ces trois juges, nous avons eu recours au coefficient de concordance W de Kendall.

1.2.3 Le coefficient de concordance W de Kendall et son interprétation

Le coefficient de concordance W de Kendall est utilisé afin de mesurer le degré de concordance entre plusieurs classements (trois ou plus) d'un même ensemble d'individus ou

d'objets sur une échelle ordinale (i.e., attribution à chaque élément de l'ensemble d'un rang compris entre 1 et n , n désignant le nombre total d'éléments à classer). Ce dernier est notamment employé afin d'évaluer le degré *d'accord (de concordance)* entre plusieurs juges. Plus la valeur du coefficient W est proche de 1 et plus le degré de concordance est élevé :

$$W = \frac{s}{\frac{l}{12} k^2 (n^3 - n)}$$

$$s = \sum_j \left(R_i - \sum_i \frac{R_i}{n} \right)^2$$

Où i désigne les éléments (individus ou objets) à classer ($i = 1$ à n) ; j désigne les juges ou les critères de classement ($j = 1$ à k) et R_i la somme des rangs attribués à l'élément générique i sur les différents classements.

La valeur du coefficient de concordance W de Kendall est comprise entre 0 et 1. Plus la valeur du coefficient est élevée et plus l'association est forte. En règle générale, les coefficients de Kendall d'une valeur supérieure ou égale à 0.9 sont considérés comme très bons. Un coefficient de Kendall élevé ou significatif indique ainsi que les évaluateurs appliquent globalement les mêmes standards pour évaluer les échantillons. Typiquement, le coefficient de Kendall est calculé une première fois sur la base des scores communiqués par chacun des juges avant toute discussion entre eux, et une seconde fois après discussion pour résolution des divergences de point de vue éventuelles.

1.2.4 Résultats

Le coefficient de concordance W de Kendall apparaît d'emblée assez élevé (i.e., avant toute discussion entre les 3 juges ; $W = 0.89$, $p < .001$), attestant ainsi de la clarté et de l'applicabilité des critères retenus pour évaluer la qualité méthodologique des 41 études considérées. Après discussion, le coefficient atteint une valeur de $W = 0.91$, $p < .001$, qui

confirme que les trois juges appliquent les mêmes standards afin de classer chaque étude selon leur qualité méthodologique (cf. Tableau 4 ci-dessous).

Tableau 4

Coefficient de concordance W de Kendall avant et après discussion entre les trois juges

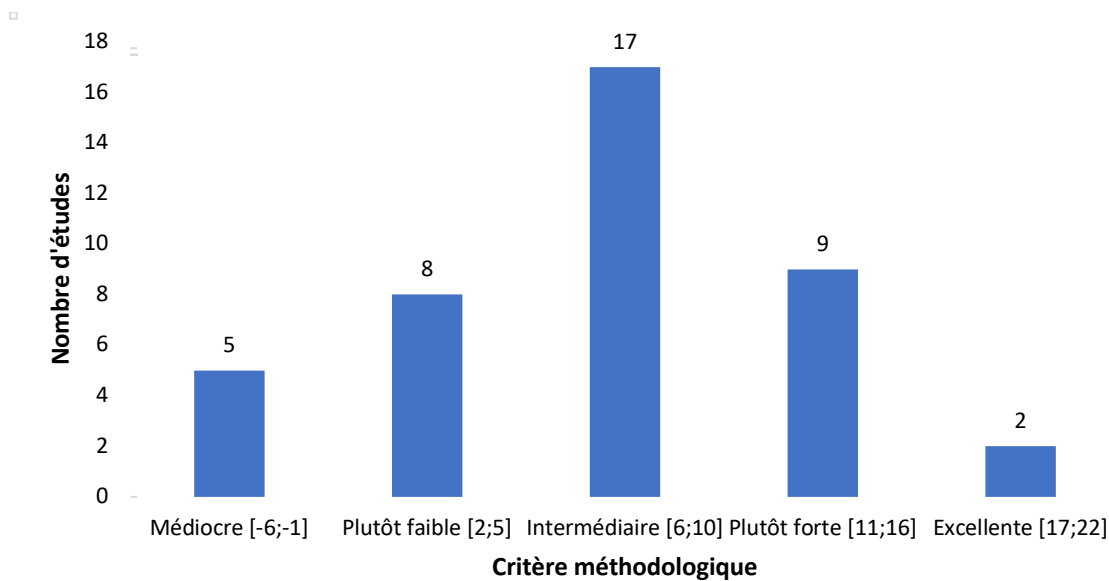
	Avant discussion	Après discussion
N	3	3
W de Kendall	.89	.91
Khi-deux	104.30	106.90
Ddl	39	39
Sig (<i>p</i>)	.001	.001

1.3 Analyse de la répartition méthodologique des 41 études considérées

Le coefficient de concordance final (i.e., après discussion) étant satisfaisant, nous avons ensuite fractionné la distribution des 41 scores (étendue de -6 à 22 points ; $M = 6.65$, $ET = 5.75$) en quintiles, et compté le nombre d'études relevant de chacun. Comme le montre la Figure 8, seules 2 études peuvent être qualifiées de qualité « Excellente » [17 ; 22], 9 de qualité « Plutôt fortes » [11 ; 16], 17 de qualité « Intermédiaire » [6 ; 10], 8 de qualité « Plutôt faible » [2 ; 5] et 5 de « Médiocre » qualité [-6 ; -1].

Figure 8

Nombre d'études classées en fonction de chaque niveau de qualité méthodologique



Ce premier résultat est frappant : sur les 41 études considérées, 30 (soit 73%) sont de piètre qualité (« Médiocre », « Plutôt faible » ou « Intermédiaire »), alors même que n'ont été intégrées à notre analyse que des études publiées dans des revues avec comité de lecture. La nécessité de ne pas calculer à l'aveugle la taille de l'effet associé à la classe Puzzle apparaît donc clairement. Compte tenu de la proportion très élevée d'études qui, bien que publiées, ne montrent pas le niveau attendu de qualité méthodologique d'une étude scientifique que l'on pourrait qualifier de crédible (i.e., quintiles 4 et 5 : qualité « Plutôt forte » et « Excellente »), un calcul de taille d'effet sans tenir compte de cette réalité pourrait mener à des erreurs d'interprétation qui auraient pour conséquence de surestimer les effets associés à la méthode Puzzle. D'autant plus que, comme le montre le Tableau 5 ci-dessous, le g de Hedge varie considérablement entre les études classées dans un même quintile et entre les études relevant de quintiles différents (cf. Tableau 5).

Ainsi, les deux études jugées de qualité « Excellente » d'un point de vue méthodologique montrent un effet moyen de petite taille ($M_{g\ de\ Hedge} = .35$), les études de qualité

« Plutôt forte » donnent lieu à un effet moyen de grande taille ($M_{g \text{ de Hedge}} = 1.27$). De leurs côtés, les études de qualité « Intermédiaire », « Plutôt faible » ou « Médiocre » suggèrent dans 60 % des cas (18 études sur les 30 rapportées dans le Tableau 5) un effet de grande taille ($g > .80$), très supérieur à l'effet moyen rapporté par Johnson et al. (2000) dans leur méta-analyse ($d = 0.20$), et à l'effet de l'apprentissage coopératif sur la performance académique, en comparaison à un apprentissage individuel ($d = 0.40$, Hattie, 2009). Estimées à l'aveugle de leurs qualités méthodologiques, les 41 études considérées donnent lieu à un effet moyen de très grande taille ($M_{g \text{ de Hedge}} \text{ à l'aveugle des quintiles} = 1.36$), effet en réalité peu informatif. En effet, les deux g de Hedge les plus élevés, qui par conséquent gonflent la taille de l'effet Puzzle, sont aussi associés aux études dont les qualités méthodologiques sont les plus faibles. La prudence s'impose par conséquent s'agissant du g de Hedge toutes catégories méthodologiques confondues, prudence d'autant plus légitime que le g de Hedge est beaucoup plus modeste s'agissant des deux seules études classées « Excellentes » ($g = .35$). Dans ces conditions, il est difficile de conclure fermement sur la taille de l'effet lié à la classe Puzzle en matière de performances académiques.

Tableau 5

Répartition des études qui ont testé les performances académiques selon leur qualité méthodologique et la valeur estimée du g de Hedge

Qualité méthodologique	Auteurs	Année	g de Hedge
Excellente	Law	2011	0.44
	Roseth et al.	2019	0.26
			$M_{g \text{ de Hedge}} = 0.35$
Plutôt forte	Yoruk	2016	0.61
	Souvignier et Kronenberger	2007	-0.48
	Ghaith et Abd El-Malak	2004	0.25

	Hosseini et al.	2014	0.91
	Huang et al.	2014	2.59
	Tarhan et al.	2013	2.57
	Evcim et Ipek	2012	1.30
	Doymus	2008	1.73
	Tarhan et Sesen	2012	1.51
			$M_g \text{ de Hedge} = 1.27$
	<hr/>		
	Iweka	2017	0.96
	Göcer	2010	4.29
	Sahin	2010	0.96
	Sahin	2011	0.85
	Dori et al.	1995	1.52
	Farahnaz et al.	2013	1.06
	Kumar et al.	2017	1.71
	Nebel et al.	2017	1.08
Intermédiaire	Akçay	2016	1.96
	Chu	2014	0.62
	Gömleksiz	2007	2.42
	Jafariyan et al.	2017	1.55
	Karacop et Diken	2017	1.43
	Lazarowitz et al.	1994	0.39
	Mattingly et VanSickle	1991	0.01
	Tran et Lewis	2012	0.66
	Ural et al.	2017	1.38
			$M_g \text{ de Hedge} = 1.27$
	<hr/>		
	Artut et Tarim	2007	1.11
	Hornby	2009	0.75
	Alamri	2018	0.80
Plutôt faible	Marhamah et Mulyadi	2013	5.70
	Yapici	2016	2.64
	Al-Salkhi	2015	0.36
	Mari et Gumel	2015	1.11
	Özdemir et Arslan	2016	2.37

			$M_g \text{ de Hedge} = 1.85$
	Aydin et Biyikli	2017	1.31
	Basyah et al.	2018	3.17
Médiocre	Çagatay et Demircioglu	2013	0.74
	Mutlu	2018	0.86
	Zahra	2014	0.68
			$M_g \text{ de Hedge} = 1.35$

Autre point important : les études répertoriées dans le Tableau 5 ont été réalisées dans plus de 80% des cas (i.e., 33 études sur les 41 répertoriées dans le Tableau 5) sur des échantillons de moins de 80 participant·es. Par exemple, Hosseini et al. (2014) ont testé l'efficacité de la méthode *Jigsaw II* en expression écrite sur un échantillon total de 40 participant·es. Tarhan et Sesen (2012) ont comparé l'efficacité de la méthode Puzzle à une pédagogie plus traditionnelle (i.e., travail individuel) en chimie sur seulement 38 participant·es. Les valeurs estimées de la taille des effets sont respectivement de $g = .91$ et $g = 1.51$, autrement dit très largement supérieures à la taille d'effet moyenne des études catégorisées comme « excellentes » ($g = .35$).

En accord avec nos conclusions invitant à la prudence, la présence de tailles d'effet anormalement élevées dans les études menées sur des échantillons de petite taille est l'indicateur d'un probable biais de publication en faveur des résultats positifs (effet tiroir ou « *file drawer effect* » ; Funder & Ozer, 2019, Kühberger, 2014, Rosenthal, 1979). L'effet tiroir, aussi appelé biais de publication, désigne le phénomène selon lequel les résultats positifs (i.e., dans le sens des hypothèses) et statistiquement significatifs ont davantage de chance d'être publiés dans des revues à comité de lecture (Kühberger, 2014, Rosenthal, 1979). La méthode la plus couramment utilisée pour estimer ce biais consiste à représenter les tailles d'effets de chaque étude en fonction de la taille de leurs échantillons sous la forme d'une représentation

graphique en entonnoir désignée sous le nom de « *funnel plot* » (pour plus de détails concernant cette procédure, voir Stern et al., 2011).

L'effet tiroir a été mis en évidence dans plusieurs disciplines telles les sciences politiques (e.g., Gerber et al., 2008), la sociologie (e.g., Gerber & Malhotra, 2008) ou encore la psychologie (e.g., Kühberger, 2014, Fergusson & Heene, 2012, Franco et al., 2016). Kühberger et collaborateurs (2014) ont par exemple observé une corrélation de $r = -.45$ entre la taille de l'effet et la taille d'échantillon de 341 études publiées dans différents domaines de la psychologie, qui suggère une tendance de ce champ à rapporter sélectivement les résultats positifs et à surestimer les tailles d'effets provenant des échantillons de petite taille. Ce processus de sélection biaisé, qui semble être par ailleurs une pratique relativement répandue en psychologie, résulterait du fait qu'une trop grande importance est accordée à la significativité des différences observées comme condition préalable à la publication. Auteurs, experts et éditeurs privilégient les résultats positifs et statistiquement significatifs, supposés être plus concluants. Ce phénomène se traduit à long terme par un faible taux de publication des travaux qui rapportent des résultats non significatifs ou négatifs (i.e., qui vont dans le sens inverse des hypothèses ou qui suggèrent des hypothèses plus complexes ; Fanelli, 2012). Le biais consistant à ne publier que les résultats statistiquement significatifs peut également conduire les chercheurs à avoir recours à des pratiques contre-productives pour la cumulation de connaissances en sciences : falsification de données, plagiat, choix méthodologiques peu rigoureux pour que les données soutiennent les hypothèses, etc. Nous ne pouvons que spéculer sur l'ampleur de ce phénomène dans la littérature consacrée à la classe Puzzle puisqu'il est impossible (comme dans beaucoup d'autres domaines) de quantifier précisément le nombre d'études non publiées en raison de résultats non significatifs. L'hypothèse d'un biais de publication dans la littérature consacrée à la classe Puzzle est soutenu par les récents travaux de Stanczak (2020). Sans pour autant distinguer la qualité méthodologique des études, l'auteur

a réalisé une méta-analyse sur un échantillon de 20 études publiées entre 2000 et 2020 qui testent l'efficacité de la méthode Puzzle sur les performances académiques. Les résultats de cette méta-analyse indiquent qu'il existe une relation négative entre la taille d'échantillon des études et leur taille d'effet ($\tau = -.43, p = .007$), suggérant la présence d'un biais de publication et de biais méthodologiques.

Le biais de publication consistant à ne rapporter que les résultats statistiquement significatifs peut naturellement fausser les connaissances dont nous disposons dans un champ disciplinaire donné. En effet avec le développement de ces connaissances, il devient indispensable d'avoir recours à l'utilisation de méta-analyses dans le but de synthétiser le nombre important de données produites à propos d'un phénomène donné. Néanmoins, si les résultats non significatifs sont systématiquement omis du processus de publication, alors l'estimation présentée comme la « taille de l'effet moyenne » d'un ensemble de travaux pourrait, dans une certaine mesure, représenter en réalité la taille d'effet moyenne de la portion de la distribution qui n'est constituée que de surestimations. Selon Ferguson et Heene (2012), environ 25% des méta-analyses publiées dans les 10 revues de psychologie les plus prestigieuses (au regard de leur facteur d'impact) présentent un risque de biais de publication. Les auteurs rapportent l'existence d'une corrélation négative entre la taille d'échantillon et l'estimation de la taille d'effet dans 80% des cas, ce qui suggère que la plupart des méta-analyses se basent sur des études dont les tailles d'effet sont exagérément élevées. Par conséquent, la surestimation de la taille de l'effet dans les analyses de puissance peut conduire à une sous-estimation de la taille de l'échantillon nécessaire pour détecter l'effet recherché, conduisant à 1) mener des études en sous puissance au regard de la véritable taille d'effet et 2) augmenter la probabilité de commettre une erreur de types I et II (e.g., Christley, 2010).

Bien qu'un biais de publication puisse être à l'origine de l'augmentation de la taille des effets observés, d'autres explications alternatives doivent également être envisagées (Stern et

al., 2011). La littérature fait ainsi mention de travaux qui 1), révèlent que, de manière générale, les études réalisées sur des échantillons de petite taille ont tendance à rapporter des tailles d'effet beaucoup plus importantes que les études réalisées avec de larges échantillons (e.g., Slavin et Smith, 2009 ; Schäfer et Schwarz, 2019) et 2) permettent par ailleurs de spécifier les facteurs susceptibles d'être à l'origine de l'augmentation de la taille des effets observés (e.g., Baker et al., 2019 ; Cheung et Slavin, 2016 ; Kraft, 2020). Ainsi, la grande hétérogénéité des tailles d'effet observées entre les études pourrait résulter de la qualité méthodologique souvent médiocre des études de petite envergure (i.e., mauvaise conception méthodologique, analyses inadéquates, etc.), provoquant ainsi des effets surestimés.

Il existe également d'autres éléments susceptibles d'influencer la taille de l'effet observé qui reposent sur les particularités méthodologiques intrinsèques aux études. Par exemple, les études randomisées produisent généralement des effets de plus faible amplitude que les études non randomisées (Baye et al., 2018 ; Cheung et Slavin, 2016). Cheung et Slavin (2016) constatent par exemple que la taille de l'effet moyenne dans les interventions pédagogiques à visée expérimentale est de $d = 0.16$ lorsque ces dernières sont randomisées et de $d = 0.23$ dans les études non randomisées. Dans le cas des 41 études attachées à notre synthèse, la taille de l'effet moyenne pour les 23 études qui présentent un design quasi-expérimental est de $g = 1.39$, et de $g = 1.35$ pour les 18 études qui présentent un design expérimental. Néanmoins, d'autres travaux ne rapportent pas de différences significatives entre les études selon qu'elles soient randomisées ou non, ni selon la durée de l'intervention (e.g., deBoer et al., 2014, Gersten et al., 2009). De la même façon, les études qui utilisent des mesures telles que des tests standardisés ont tendance à produire des effets moins importants que les études utilisant des mesures qui n'ont pas fait l'objet d'une procédure de validation (deBoer et al., 2014, Pellegrini et al., 2019). Les études qui comparent un groupe expérimental à un groupe contrôle, qui utilisent des analyses de données dites conservatrices ou qui intègrent des

populations importantes et hétérogènes ont également tendance à produire des tailles d'effets de faible amplitude (e.g., Cheung & Slavin, 2016 ; Karlsson & Bergmark, 2015). Les travaux menés par deBoer et al. (2014) révèlent également que les études dans lesquelles l'intervention est menée par le chercheur-e ou ses collaborateurs, plutôt que par des enseignant-e-s, obtiennent des tailles d'effets de plus forte amplitude. De même, les interventions axées sur les élèves travaillant individuellement ou en petits groupes produisent des tailles d'effet de plus forte amplitude que celles axées sur les élèves travaillant en classe entière (e.g., Lipsey et al., 2012). Enfin, la méta-analyse conduite par Wolf et collaborateurs (2020) indiquent que les études financées par des commanditaires produisent des tailles d'effet moyennes 1,7 fois plus importantes que les études non commanditées. Au final, la qualité intrinsèque, souvent médiocre, des études investiguant les effets de la méthode Puzzle sur l'amélioration des performances académiques est très probablement responsable de l'hétérogénéité ainsi que de l'ampleur des tailles d'effets observées et par conséquent, contribuent à fausser les estimations de l'efficacité de la méthode Puzzle.

2 En conclusion les effets de la méthode Puzzle auraient-ils été surestimés par ses concepteurs ?

Dans son ouvrage co-écrit par Shelley Patnoe et réédité en 2011, Eliott Aronson déclarait que la méthode Puzzle atténue certains des aspects indésirables provoqués par la compétition entre les élèves, en favorisant leur intérêt pour la coopération :

« In any classroom situation, the jigsaw method curbs some of the undesirable aspects of excessive competition and increases the interest children have in cooperating with one another. Thus, the research demonstrated that what seemed to be a deeply ingrained kind of behavior–competitiveness–can be modified. » (Aronson & Patnoe, 2011, p.13)

Néanmoins, nos observations suggèrent une réalité bien moins évidente. Nous avons en effet constaté que les publications au sujet de la méthode Puzzle se sont majoritairement focalisées sur la question des performances académiques en dépit d'un intérêt premier pour des mesures plus en rapport avec la tolérance et la réduction des discriminations raciales. Les publications concernant ces mesures étant peu nombreuses, nous nous sommes focalisées sur la variable dépendante dominante de cette littérature, à savoir les performances académiques. Notre catégorisation des études dans ce cadre selon leurs qualités méthodologiques révèle que très peu s'avèrent satisfaisantes, avec des tailles d'effets anormalement élevées sur les études jugées les plus faibles ($g > .80$). Les deux études jugées les plus solides d'un point de vue méthodologique suggèrent une taille d'effet ($M_g \text{ de Hedge} = 0.35$) plus cohérente avec les estimations rapportées précédemment par Johnson et collaborateurs (2000, $d = 0.20$). En bref, notre examen de la partie dominante de la littérature consacrée à la classe Puzzle révèle non seulement un intérêt pour des mesures différentes de celles envisagées au départ, mais aussi des faiblesses méthodologiques importantes et nombreuses qui interdisent toute conclusion ferme quant à la taille de l'effet Puzzle.

Cette conclusion justifie pleinement la poursuite des travaux sur les effets de l'interdépendance positive au sens de la classe Puzzle à l'aide de dispositifs plus ambitieux que les précédents, s'agissant en particulier de la puissance de test et des qualités méthodologiques à réunir. L'étude « ProFan » présentée dans le chapitre suivant correspond à cette ambition, avec plus généralement pour objectif de valider des pédagogies fondées sur une interdépendance positive encore très peu utilisées dans nos systèmes d'enseignements, qu'il s'agisse de la voie générale, technologique ou professionnelle. En effet, il ne suffit pas de faire travailler les élèves en groupe pour leur assurer la mise en place d'une quelconque interdépendance positive, par hypothèse valorisante pour ceux les plus faibles. L'étude « ProFan » permet précisément de comparer les effets de l'interdépendance positive au sens de

la classe Puzzle relativement à un travail de groupe plus classique et moins structuré, et à un travail individuel en classe.

Chapitre 3 : Le dispositif ProFan

1 Vue d'ensemble

Ce chapitre est consacré à la description de l'étude « ProFan ». Au niveau le plus général, cette étude, inscrite dans l'action « Innovation numérique pour l'excellence éducative » du Programme d'investissements d'avenir (PI3), avait pour objectif d'analyser et de tester des modes d'enseignement et d'apprentissage propres à faire émerger de nouvelles compétences induites par la transformation digitale du travail et de son environnement social. Conçue dans le cadre de la seconde mission interministérielle confiée au recteur Monteil, ProFan s'adressait à des élèves de lycées professionnels issus de trois filières de formation : métiers de l'électricité et de leur environnement connecté (MELEC), commerce (COMMERCE) et accompagnement, soins et services à la personne (ASSP). Ce dispositif a été déployé dans chacune de ces 3 filières, en classe de première et en classe de terminale, dans 109 établissements de dix académies (Bordeaux, Poitiers, Limoges, Rennes, Nantes, Strasbourg, Nancy-Metz, Reims, Montpellier, Toulouse) couvrant 5 régions de France métropolitaine. L'étude était réalisée sur un ensemble de 3 groupes d'établissement (G1, G2, G3 ; pour plus de détails, voir Procédure) selon un calendrier commun. Ainsi, plus de 10 000 élèves ont participé à cette opération avec un suivi longitudinal sur 2 promotions (pour plus de détails, voir Participant·e·s).

La transition numérique offrant les conditions technologiques d'une montée en puissance du travail collectif—cette modalité quelle que soit ses formes n'étant plus strictement contrainte par des contingences spatiales et/ou temporelles—il s'agissait plus spécifiquement avec le dispositif ProFan 1- de développer la capacité des élèves à travailler en groupe, à se décentrer et à coordonner leurs points de vue, autant de compétences sociales rendues encore plus nécessaires par la transition numérique ; et 2- d'étudier les effets du travail collectif avec ou

sans consignes susceptibles de favoriser une interdépendance positive entre les membres des groupes considérés (interdépendance entendue au sens de la méthode dite de la *classe puzzle*, cf. chapitres précédents). Les observables (variables dépendantes) étaient de trois types. Premièrement, les performances individuelles des élèves (tests standardisés construits pour l'étude) sur des contenus développés dans les séquences pédagogiques—en présentiel—impliquant ou non un travail collectif (l'observable central dans nos travaux). Deuxièmement, les performances et autres comportements des élèves en situation de résolution collectives de problèmes non issus des séquences pédagogiques (en référence à ce que nous nommerons plus loin « la boîte à outils » ou BAO) et donc non directement en rapport avec les compétences professionnelles à développer dans les différentes filières. Et enfin troisièmement, la réponse des élèves à plusieurs questionnaires permettant de sonder leurs représentations et auto-évaluations dans différents domaines (perception de compétences, estime de soi, etc. ; cf. Annexes G, H, I et J) et à différents moments de l'étude (pour en saisir l'évolution dans le temps et en relation avec les modalités de travail considérées).

Le groupement de chercheur·e·s (15 statutaires, 3 post-doctorant·e·s et 6 doctorant·e·s) qui, animé par la mission Monteil, a conçu le dispositif et en a exploité les résultats, était composé au total de sept équipes de recherche (5 françaises et 2 suisses) inscrites dans le spectre des sciences de la cognition et du comportement (cf. Annexe B). En interaction permanente avec des inspecteurs et inspectrices de l'éducation nationale, ces équipes sont aussi intervenues dans la structuration des séquences pédagogiques proposées par l'inspection générale pour les trois filières, dans le but d'en co-construire la compatibilité avec les objectifs scientifiques de l'étude eux-mêmes co-construits avec l'inspection. Pour bien comprendre les efforts déployés, il faut rappeler que les réunions dans ce cadre ont été nombreuses (plusieurs centaines d'heures d'interaction au total entre chercheurs et corps d'inspection) en collaboration étroite avec la mission Monteil. Ces réunions avaient à la fois pour but d'assurer une compréhension mutuelle

des objectifs, des étapes et des procédures à mettre en œuvre sur le terrain, et de construire les séquences pédagogiques et une plateforme numérique (conçue par l'une des sept équipes du consortium) support de l'étude (contenant notamment toutes les ressources nécessaires à la réalisation des activités des enseignants) dans le respect des contraintes imposées par cette dernière.

Aussi, le dispositif ProFan ne traduit pas simplement l'activité d'équipes de recherche sur le terrain scolaire au sens le plus habituel d'une mise à leur disposition d'élèves pour tester des hypothèses élaborées par et pour une communauté donnée de chercheurs, avec la possibilité parfois offerte *in fine* aux enseignant·es et cadres d'en consulter les principaux résultats sous une forme simplifiée pour en dégager les plus-values pratiques éventuelles. Il s'agit bien davantage d'un authentique processus de co-construction entre chercheur·es et cadres de l'éducation nationale pour l'atteinte d'un objectif collectivement partagé et susceptible de retombées à la fois scientifiques et pédagogiques, dans le cas présent en relation avec la question centrale du travail collectif à l'heure de la transition numérique.

En dehors notamment du nombre d'établissements impliqués dans l'étude, et du temps consacré à sa préparation, l'ampleur des efforts déployés dans ce cadre peut être appréciée également à la lumière de la quantité de données produites et centralisées sur la plateforme numérique support de l'étude, soit plusieurs millions au total avec les traces laissées par les utilisateurs sur ladite plateforme.

2 Procédure

2.1 Participant·es

L'expérimentation se déroulait sur deux promotions consécutives de bachelier·es professionnels issus de trois filières de formation (ASSP, COMMERCE et MELEC), selon le calendrier suivant :

a. Temps 1 – Année scolaire 2017-2018 :

Entrée des classes de Première dans le dispositif.

b. Temps 2 – Année scolaire 2018-2019 :

Classes de Terminale (i.e., les élèves de 1ère en 2017/2018) et entrée des nouvelles classes de Première dans le dispositif.

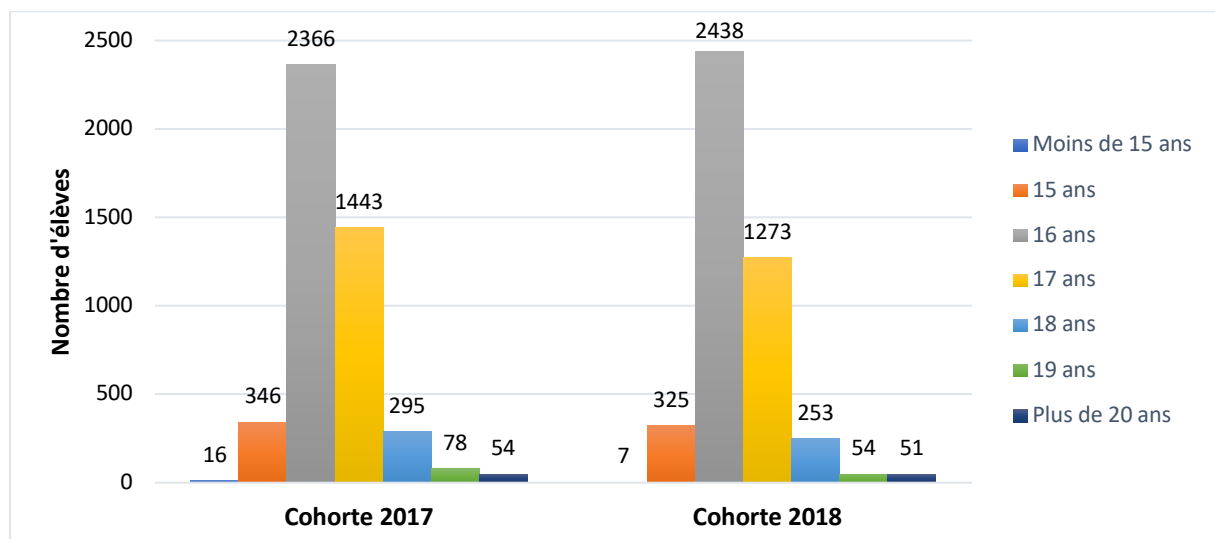
c. Temps 3 – Année scolaire 2019-2020 :

Classes de Terminale (i.e., les élèves de Première en 2018/2019)

Au total, 10395 élèves (918 classes issues de 109 lycée professionnels) ont participé volontairement à cette expérience avec l'autorisation de leurs parents. Sur l'ensemble des élèves des deux cohortes, 5221 étaient des filles et 4227 étaient des garçons. Dans la mesure où l'âge des participant·e·s a été mesuré avec une échelle en 7 points allant de 0 (*moins de 15 ans*) à 6 (*20 ans et plus*), nous ne disposons pas de l'âge exact de chaque participant·e (cf. Figure 9 ci-dessous montrant une répartition du nombre d'élèves par catégorie d'âge selon la cohorte).

Figure 9

Nombre d'élèves par catégorie d'âge selon la cohorte



Nous constatons que la catégorie majoritairement représentée est celle des élèves de 16 ans et cela, quelle que soit la cohorte ($N_{\text{Cohorte 2017}} = 2366$, $N_{\text{Cohorte 2018}} = 2438$). La répartition des

participant·e·s dans les différentes conditions selon leur filière et leur sexe est présentée dans le Tableau 1.

Tableau 7

Nombre de participant·e·s dans les conditions expérimentales par cohorte selon leur filière et leur sexe

Filière	Sexe	Condition			Total
		G1	G2	G3	
Cohorte 2017					
<i>ASSP</i>					
	Fille	454	530	497	1481
	Garçon	25	45	25	95
	Total	552	628	587	1767
<i>COMMERCE</i>					
	Fille	351	357	308	1016
	Garçon	268	271	296	835
	Total	715	691	791	2197
<i>MELEC</i>					
	Fille	8	8	7	23
	Garçon	454	409	254	1117
	Total	526	449	287	1262
<i>TOTAL</i>					
	Fille	814	895	812	2521
	Garçon	749	725	575	2049
	Total	1934	1786	1665	5385
Cohorte 2018					
<i>ASSP</i>					
	Fille	928	1085	1056	3069
	Garçon	58	79	48	185

Total	1060	1222	1172	3454
<i>COMMERCE</i>				
Fille	728	721	659	2108
Garçon	574	586	612	1772
Total	1429	1374	1465	4268
<i>MELEC</i>				
Fille	18	15	10	43
Garçon	934	803	525	2262
Total	1018	850	573	2441
<i>TOTAL</i>				
Fille	1675	1821	1725	5221
Garçon	1574	1468	1185	4227
Total	3704	3481	3210	10395

Note. Il existe une asymétrie de sexe inversée en ASSP et en MELEC, les filles ne représentant au total que 66 participant·e·s en MELEC contre 4510 participant·e·s en ASSP, et les garçons ne représentant au total que 280 participants en ASSP contre 3379 participants en MELEC. Par conséquent, seule la filière COMMERCE permet l'examen éventuel d'effets de sexe dans le dispositif ProFan et cela, quelle que soit la cohorte considérée.

3 Méthode

3.1 Aspects généraux

Quelle que soit la filière de référence, tous les contenus enseignés aux élèves dans le cadre de l'étude ProFan étaient issus des programmes officiels avec un tronc commun (enseignement de mathématiques et enseignement de français), et des enseignements professionnels nécessairement spécifiques à chaque filière. Seule la consigne fournie aux

enseignant·e·s s'agissant de la modalité pédagogique à mettre en œuvre pour ces différents contenus (généraux ou spécifiques) différait selon les établissements.

Dans un premier groupe d'établissements (N = 39), ci-après dénommé « G1- Interdépendance positive », les enseignant·e·s—tous volontaires—avaient pour consigne de respecter la méthode dite de la « classe puzzle », favorisant la mise en œuvre auprès de leurs élèves d'un travail collectif structuré par une forte interdépendance positive (cf. plus bas « induction expérimentale »).

Dans un deuxième groupe d'établissements (N = 36), ci-après dénommé « G2 – Coopération non structurée », d'autres enseignant·e·s eux-mêmes volontaires avaient simplement pour consigne de faire travailler leurs élèves en groupe (sans mention ni de la méthode dite d'interdépendance positive ni d'aucune autre méthode de travail collectif).

Dans un troisième groupe d'établissements (N = 34), ci-après dénommé « G3 - Contrôle », d'autres enseignant·e·s encore ne recevaient aucune consigne particulière. Ils étaient informés des contenus d'enseignement ciblés par l'étude ProFan dans leur filière de référence, mais n'avaient à ce titre accès qu'à seulement certains des documents disponibles sur la plateforme accessible en ligne.

Les 10 académies couvrant les 109 établissements mobilisés au total pour « G1 », « G2 » et « G3 » étaient choisies en fonction de leur position géographique dans le but de couvrir de manière assez équilibrée la France métropolitaine : nord-est /nord-ouest, sud-est/sud-ouest à l'exclusion de l'Ile de France en raison d'une disparité ingérable entre les académies de Paris-Versailles et de Créteil sur l'enseignement professionnel. Pour plus de détails concernant le nombre d'établissements par académie et leur nombre dans chaque groupe, voir Annexe B. Les établissements de ces dix académies étaient eux-mêmes sélectionnés en fonction principalement de la présence des trois filières considérées (ASSP, COMMERCE, MELEC) et de leur équilibre démographique, avec dans les trois groupes G1, G2 et G3 des établissements

citadins et ruraux. Présentes dans un maximum d'établissements tout en offrant une diversité de formations, la sélection des trois filières impliquées dans ProFan était elle-même assez stratégique. La procédure a d'abord consisté à choisir des établissements capables, pour diverses raisons, de soutenir la lourdeur du dispositif à mettre en place dans le groupe G1, pour ensuite choisir des établissements comparables au titre de groupes contrôles (G2 et G3). Dans chacune des 10 académies concernées, un "réfèrent académique ProFan » était désigné par la mission Monteil et avait objectif de faciliter le suivi de l'étude ProFan sur le terrain en relation avec leurs référents ProFan à l'échelle de chaque établissement. Enfin, tous les enseignant·e·s impliqué·e·s dans ProFan à l'échelle des trois groupes d'établissements bénéficiaient pour leur participation d'une indemnité pour mission particulière (IMP).

3.2 Déroulement chronologique de la procédure pour les élèves

En classe de première, chaque élève remplissait en début d'année un questionnaire (ci-après dénommé « Q1 ») pendant une première séance d'1 heure dans des conditions de passation préservant l'anonymat, c'est pourquoi la passation se déroulait en ligne et supposait un ordinateur par élève. Ensuite pendant une autre séance de 2 heures, chaque élève réalisait les tâches de la boîte à outils (i.e., « BAO 1 »). Chacune de ces activités étaient réalisées par classe ou demi-classe en fonction du nombre de postes de travail disponibles), en se connectant à la plateforme ProFan selon l'agenda qui était fixé par le réfèrent d'établissement dans le cadre des horaires de l'enseignement professionnel. Le réfèrent d'établissement était également responsable de la gestion informatique et du suivi des activités de la BAO au sein de l'établissement. Par la suite, les élèves participaient à deux séquences pédagogiques en français, en mathématiques et en enseignement professionnel (cf. Séquence 1 et Séquence 2 ; pour plus de détails concernant le contenu des séquences, voir Annexe D, E et F), qui étaient suivies chacune d'une évaluation individuelle. En fin d'année scolaire, chaque élève répondait de

nouveau au questionnaire (i.e., Q2) et réalisait des versions différentes des tâches de la boîte à outils (i.e., BAO 2), selon les mêmes modalités qu'en début d'année. En classe de terminale, les élèves ne passaient ni de questionnaire ni de boîte à outils en début d'année scolaire. Ils participaient néanmoins à deux séquences pédagogiques en français, en mathématiques et en enseignement professionnel (cf. Séquence 3 et Séquence 4, pour plus de détails concernant le contenu des séquences voir ; Annexe D, E et F), qui étaient chacune suivies d'une évaluation individuelle. En fin d'année scolaire, chaque élève répondait au questionnaire (i.e., Q3) et réalisait les tâches de la boîte à outils (i.e., BAO 3), selon les mêmes modalités qu'en classe de première. La procédure était similaire pour les trois années scolaires quelle que soit la condition expérimentale à laquelle les élèves étaient assignés. Pour une vue d'ensemble de la procédure en classe de première et en classe de terminale, voir Figure 10 et Figure 11 ci-dessous.

Figure 10

Vue d'ensemble du déroulement de la procédure pour l'année de Première (i.e., Temps 1-Année scolaire 2017-2018 et Temps 2-Année scolaire 2018-2019)

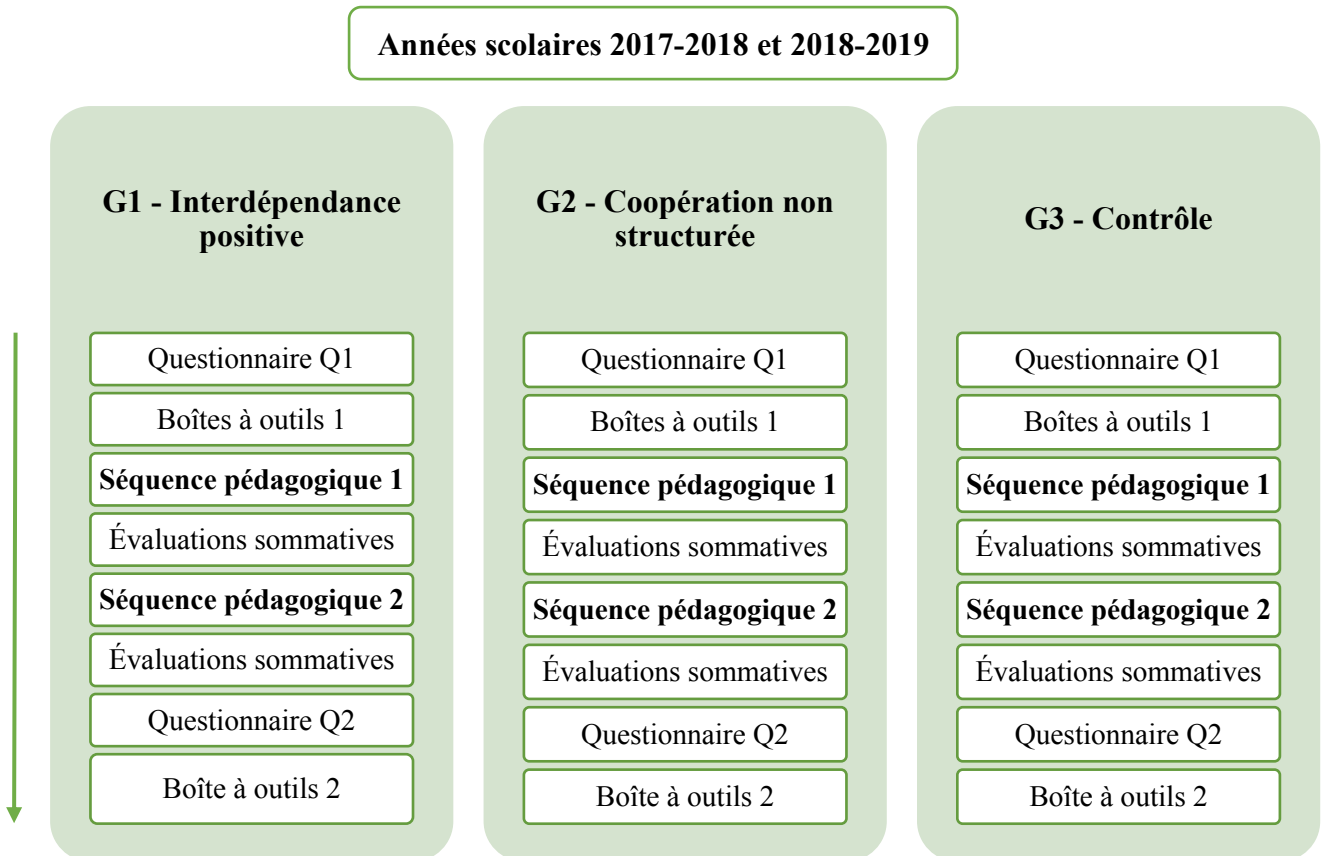
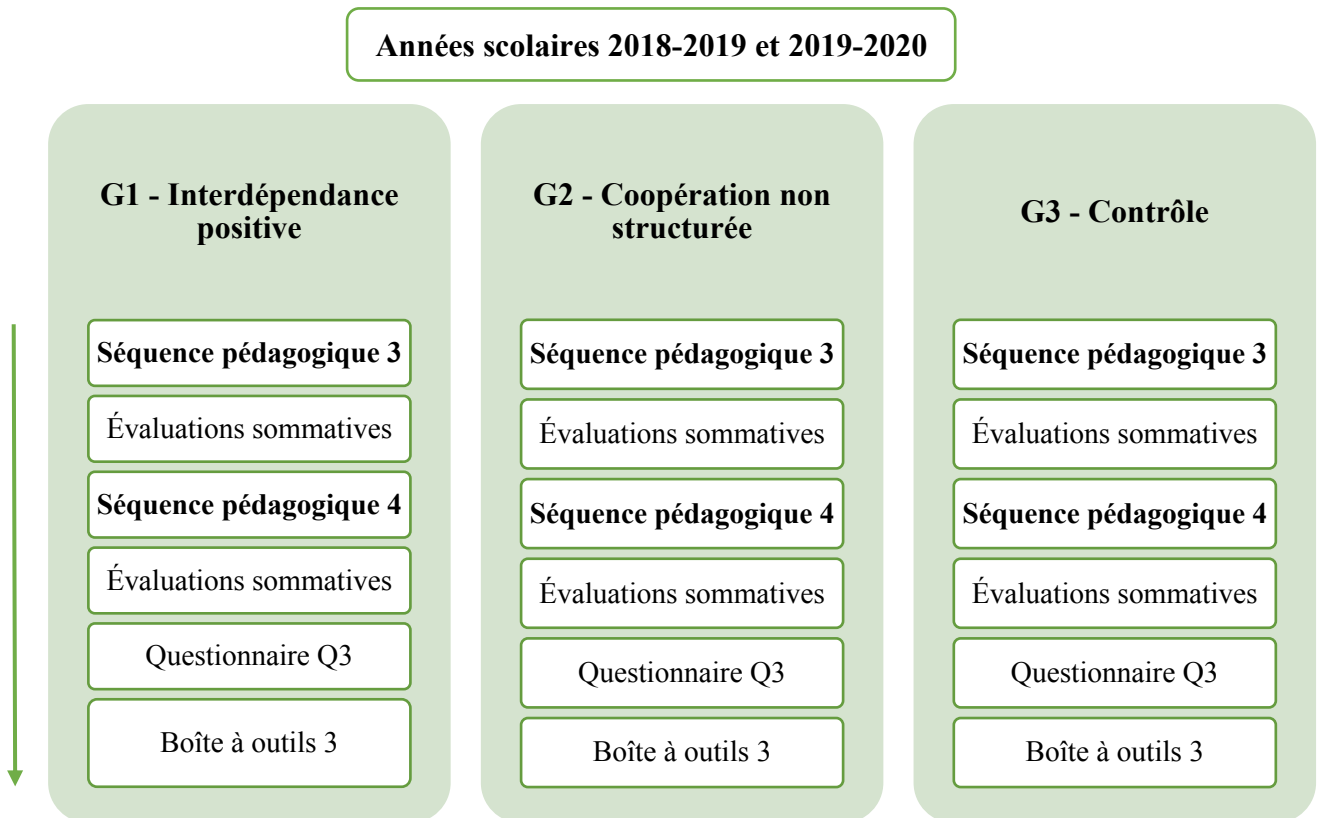


Figure 11

Vue d'ensemble du déroulement de la procédure pour l'année de Terminale (i.e., Temps 2- Année scolaire 2018-2019 et Temps 3-Année scolaire 2019-2020)



Les décisions imposées par la situation sanitaire du pays en Mars 2020 (pandémie de Covid-19) se sont traduites par l'interruption des dispositifs expérimentaux d'enseignement. Ainsi, près de 80% des classes impliquées dans le dispositif ProFan avaient mené et évalué la première séquence pédagogique de l'année, autrement dit la Séquence 3. En revanche, la seconde séquence pédagogique de l'année, la Séquence 4, avait été conduite dans moins de 20% des classes. Par conséquent, le questionnaire Q3 a été administré aux élèves de Terminale à distance, au moyen de la plateforme ProFan sur laquelle chaque élève disposait d'un compte individuel. De plus et au regard des circonstances, une extension a été intégrée au questionnaire Q3 afin notamment d'examiner l'environnement dont disposait les élèves durant le confinement

ainsi que leur ressenti en situation de télétravail. Pour plus de détails concernant le questionnaire Q3 extension, voir Matériel.

3.3 Détails de l'induction expérimentale

Les élèves des établissements G1 (condition Collectif structuré) réalisaient toutes les séquences pédagogiques (enseignements généraux de français, de mathématiques et enseignements professionnels) selon le déroulement suivant :

- Étape 1 : Phase d'introduction des objectifs pédagogiques propres à chaque contenu.
- Étape 2 : Formation de sous-groupes de travail (typiquement de 3 à 5 élèves) avec l'assistance de la plateforme numérique pour en garantir la composition aléatoire. Cette composition demeurait cependant sous le contrôle des enseignant·e·s pour la corriger dans le cas par exemple d'une forte incompatibilité (jugée comme telle à tort ou à raison par l'enseignant·e) entre tels ou tels élèves placé·e·s dans un même sous-groupe de travail. Ils leur suffisaient alors de procéder à un nouveau tirage aléatoire.
- Étape 3 : Chaque élève travaillait individuellement sur une sous-partie de la séquence (Phase « individuelle »).
- Étape 4 : Regroupement des élèves ayant travaillé sur la même sous-partie (Phase des groupes dits « d'experts »)
- Étape 5 : Chaque élève retournait dans son groupe pour présenter la sous-partie sur laquelle il avait travaillé aux membres de son groupe (Phase des groupes dits « puzzles »).
- Étape 6 : Chaque groupe produisait un travail collectif impliquant l'articulation des différentes sous parties.
- Étape 7 (Facultative) : Synthèse en classe entière.

Les élèves des établissements G2 (condition Collectif non structuré) réalisaient toutes les séquences pédagogiques (enseignements généraux de français, de mathématiques et enseignements professionnels) selon le déroulement suivant, identique à G1 seulement sur les deux premières étapes

- Étape 1 : Identique à G1.
- Étape 2 : Identique à G1.
- Étape 3 : Contrairement à G1, chaque élève disposait de tout le contenu de la séquence considérée (toutes les sous-parties donc l'intégralité des éléments disponibles pour le cours). Les enseignant·e·s n'étant informé·e·s préalablement que de la nécessité de faire travailler leurs élèves en groupe, sans aucun guidage sur la manière de structurer le travail des groupes en question, cette organisation était donc laissée à leur initiative.²
- Étape 4 (Facultative) : Synthèse en classe entière.

Enfin, les élèves des établissements G3 (condition de contrôle de type pédagogie Libre) réalisaient eux aussi toutes les séquences pédagogiques (enseignements généraux de français, de mathématiques et enseignements professionnels) dans leur contexte pédagogique habituel, c'est-à-dire sans aucune consigne fournie aux enseignant·e·s s'agissant de la manière d'organiser leurs classes. Ils étaient néanmoins informés de la nécessité : 1) d'organiser une séquence pédagogique dont le contenu était mis à leur disposition, 2) de respecter un certain volume horaire et 3), de prendre connaissance de l'évaluation sommative individuelle proposée dans le cadre de l'enseignement de français et de mathématiques. En effet, cette évaluation devait les guider dans la construction des séquences pédagogiques. Ces trois points ont

² Il est à noter que les élèves des établissements G2 réalisaient les séquences pédagogiques selon les deux options suivantes : *Option A* : Chaque groupe travaillait sur l'ensemble des activités/méthodes/chapitres, les un·e·s après les autres, ou *Option B* : Chaque groupe travaillait une seule activité/méthode/chapitre et la synthèse permettait aux élèves d'aborder toutes les activités. L'option A concernait les filières ASSP et MELEC pour les séquences 1, 3 et 4 de Mathématiques et l'option B les filières COMMERCE et MELEC pour la séquence 2 de Français.

évidemment une grande importance au niveau méthodologique pour rendre comparables les performances des élèves issus des trois ensembles d'établissements.

4 Mesures

4.1 Questionnaire Q1

Le questionnaire Q1 était composé de 37 échelles pour la cohorte 2017 et de 38 échelles pour la cohorte 2018. Dans la mesure où nous n'allons pas traiter l'ensemble des variables mesurées dans ce questionnaire, nous ne détaillerons que les variables d'intérêt pour nos analyses, à savoir la perception par les élèves de leurs compétences en mathématiques et en français. Afin d'évaluer cette perception, nous avons utilisé les questions issues de la « Trousse d'évaluation des décrocheurs potentiels » (TEDP, Janosz et al., 2007). Ces deux échelles étaient composées chacune de 1 item (i.e., « *De mémoire, quelle était ta moyenne en français à la fin de la dernière année scolaire ?* » ; « *De mémoire, quelle était ta moyenne en mathématiques à la fin de la dernière année scolaire ?* », voir Figure 12 ci-dessous). Les réponses étaient données sur une échelle type Likert en 5 points allant de 1 (*0 à 4,9*) à 5 (*16 à 20*). Dans la mesure où ces 2 échelles sont des mesures auto-rapportées, nous avons réalisé des analyses de corrélations avec les variables du questionnaire Q1 (i.e., Sentiment de compétences en français, Sentiment de compétences en mathématiques et Comparaison de ses notes à celles de sa classe) afin de nous assurer de leur fiabilité (cf. Tableau 8 ci-dessous).

Tableau 8

Corrélations entre la perception par les élèves de leurs compétences en Français et en Mathématiques et les variables du questionnaire Q1

Variable	1	2	3	4	5
1. Perception de compétences en Français	–				
2. Perception de compétences en Mathématiques	.345**	–			
3. Sentiment de compétences en Français	.395**	.056**	–		
4. Sentiment de compétences en Mathématiques	.062**	.551**	.105**	–	
5. Comparaison de ses notes à celle de la classe	.443**	.466**	.316**	.350**	–

Note. * $p < .05$; ** $p < .01$; *** $p < .001$

Figure 12

Extrait du questionnaire élève

Voici maintenant quelques questions te concernant:

De mémoire, quelle était ta moyenne en français l'année scolaire dernière ? (réponses obligatoires)

0 à 4,9 5 à 8,9 9 à 12,9 13 à 15,9 16 à 20

De mémoire, quelle était ta moyenne en mathématiques l'année scolaire dernière ? (réponses obligatoires)

0 à 4,9 5 à 8,9 9 à 12,9 13 à 15,9 16 à 20

En pensant à tes notes scolaires, comment te classes-tu par rapport aux autres élèves de ton lycée qui ont le même âge ? (réponses obligatoires)

je suis parmi les moins bons je suis plus faible que la moyenne je suis dans la moyenne je suis plus fort que la moyenne je suis parmi les meilleurs

Suivant →

4.2 Questionnaire Q2

Les conditions de passation du questionnaire Q2 étaient identiques à celles de Q1. Ce dernier était composé de 31 échelles pour la cohorte 2017 et de 32 échelles pour la cohorte 2018. De la même façon que pour le questionnaire présenté précédemment, nous ne détaillerons que les variables d'intérêt pour nos analyses, à savoir la perception des compétences en mathématiques et en français. Ainsi, les deux échelles utilisées pour mesurer la perception de

compétences des élèves en mathématiques et en français étaient similaires en tout point à celles utilisées dans le questionnaire Q1.

4.3 Questionnaire Q3

Les conditions de passation du questionnaire Q3 étaient identiques à celles de Q1 et de Q2.

Ce dernier était composé de 32 échelles pour la cohorte 2017 et de 30 échelles pour la cohorte 2018. De la même façon que pour le questionnaire Q1 et Q2, nous ne détaillerons que les variables d'intérêt pour nos analyses, à savoir la perception des compétences en mathématiques et en français. Ainsi, les deux échelles utilisées pour mesurer la perception de compétences des élèves en mathématiques et en français étaient similaires en tout point à celles utilisées dans le questionnaire Q1 et Q2.

4.4 Mesure des performances académiques

Des évaluations sommatives individuelles conçues par nos collaborateurs inspecteurs et inspectrices avaient pour objectif d'évaluer les acquis des élèves sur l'ensemble des notions abordées dans les séquences pédagogiques. Ces évaluations avaient lieu à l'issue des enseignements généraux (français et mathématiques) et spécifiques (enseignements professionnels propres à chaque filière) après chaque séquence. En français, l'évaluation individuelle était une rédaction écrite en classe par chaque élève au cours de la séquence. En mathématiques, c'était un devoir fait en classe par chaque élève une fois la séquence terminée. Enfin, sur les contenus spécifiques à chaque filière les évaluations individuelles étaient fondées sur des questionnaires à choix multiples que les élèves réalisaient en ligne sur la plateforme une fois que la séquence prenait fin. Les enseignements généraux ou spécifiques ciblant des éléments de programme différents à chaque séquence, les évaluations sommatives différaient nécessairement dans leurs contenus à l'issue de chacune. En résumé, à chaque séquence, un

élève devait répondre à trois évaluations sommatives individuelles, une en français, une en mathématiques et une autre encore sur les contenus relevant de sa filière professionnelle de référence.

Point très important, en français et en mathématiques, donc s'agissant du tronc commun aux trois filières considérées, les productions des élèves étaient également soumises à une évaluation externe. Cette évaluation externe était conduite à l'aveugle de l'induction expérimentale donc sans aucune information sur les établissements de référence des copies (le format papier des copies permettait d'interroger les élèves sous une modalité qui était pour eux familière). Il était donc impossible pour les évaluateurs externes concernés de relier les copies dont ils avaient la charge à une académie ou un établissement particulier. De même, l'appartenance de sexe des auteur-e-s des copies à évaluer était systématiquement masquée (à de rares exceptions près *de facto* écartées des analyses). Ces deux précautions avaient pour objectif de garantir 1- l'absence de biais relatifs à la « demande expérimentale », consistant dans le cas présent pour les évaluateur-ices à surévaluer ou sous-évaluer plus ou moins consciemment les copies en provenance de tel ou tel groupe d'établissements (G1, G2 ou G3) et 2- l'absence de biais de genre consistant, là encore sans nécessairement en avoir l'intention, à sous-évaluer les copies des filles en mathématiques et des garçons en français (évaluations stéréotypiques). Concrètement, ces deux précautions méthodologiques impliquaient de récupérer et centraliser toutes les copies papiers des élèves (au voisinage de 37 000 au total en mathématiques et en français), à les numériser et à les déposer sur la plateforme support de l'étude pour les rendre accessibles en ligne à l'usage des évaluateurs externes. Ce gros travail est l'un des points forts du dispositif ProFan pour la mise en œuvre de comparaisons fiables entre les performances issues des trois groupes d'établissements et éviter de fausser les conclusions de l'étude. En effet, dans la plupart des études décrites dans les chapitres précédents, les évaluateur-ices sont les enseignant-e-s et/ou les expérimentateur-ices avec un

risque élevé de biais d'évaluations dans le sens des hypothèses. Les limites à la fois financières et temporelles afférentes à ces contraintes font que, dans le cas de l'évaluation des apprentissages professionnels, des QCM corrigés automatiquement par la plateforme ont été privilégiés. Enfin, ces différents modes d'évaluation ne se substituaient pas aux évaluations par les enseignant.e.s eux-mêmes qui avaient donc toute liberté quelle que soit leur établissement ou filière de référence de noter comme à leur habitude les productions de leurs élèves.

4.5 Grille d'observation post-séquence pédagogique

À chaque fois que l'enseignement d'une séquence pédagogique prenait fin, les enseignant.e.s étaient invité.e.s à renseigner une grille dite « d'observation » sur la plateforme numérique support de l'étude, soit par classe (si la séquence avait été conduite en classe entière) soit par sous-groupe (si la séquence avait été conduite en sous-groupe). Cette grille avait deux objectifs. Premièrement, garder le contact avec les enseignant.e.s impliqué.e.s dans l'étude sur le terrain pour apprécier le déroulement des séquences pédagogiques (e.g., problèmes techniques, organisationnels, implémentation, etc.). De cette façon, nous avons interrogé les enseignant.e.s quant à leur satisfaction à l'égard des consignes fournies au démarrage du dispositif par le biais des 2 items présentées ci-dessous :

- 1) Pensez-vous que le mode d'organisation pédagogique mis en place pour l'ensemble de la séquence est efficace pour l'apprentissage des élèves ?
- 2) Êtes-vous satisfait.e vous-même de ce mode d'organisation ?

Les réponses étaient données sur une échelle type Likert en 7 points allant de 1 (*Pas du tout*) à 7 (*Énormément*).

Deuxièmement, la grille d'observation permettait à l'ensemble du consortium, en particulier les équipes de recherche, d'apprécier les écarts entre la manière dont les séquences pédagogiques étaient conduites sur le terrain et les consignes fournies aux enseignants des




établissements G1 et G2, un point extrêmement important pour nos propres travaux de thèse et sur lequel nous revenons plus en détail ci-dessous.



En effet, le nombre d'établissements et *a fortiori* de classes étant très élevé dans l'étude ProFan (relativement à la plupart des études quasi-expérimentales publiées dans le champ de l'éducation) qu'il était impossible aux équipes de recherche d'installer des observateurs dans chacune des classes en question pour estimer la qualité de l'opérationnalisation des consignes fournies aux enseignant·e·s des établissements G1 et G2. Par ailleurs, nous ne pouvons pas exclure qu'une telle pratique consistant à installer des observateurs *in situ* fausse le déroulement des pratiques pédagogiques dites « libres » dans G3. Aussi avons-nous opté pour une solution intermédiaire consistant à interroger les enseignant·e·s *via* la grille d'observation évoquée plus haut en ligne sur la plateforme, pour à la fin de chacune des séquences pédagogiques estimer à la fois la nature et l'ampleur des écarts entre les consignes fournies au terrain, et la manière dont les séquences avaient été conduites de l'avis même des enseignant·e·s concerné·e·s par ProFan. Plutôt que de ne rien savoir sur la qualité effective de l'opérationnalisation des consignes sur le terrain, cette procédure intermédiaire permet de construire *post-hoc* la variance liée à l'opérationnalisation des consignes sur chacune des 5 modalités mentionnées dans la grille d'observation (cf. Tableau 2). Ce point est pour nos travaux extrêmement important puisque précisément les effets supposés de l'interdépendance positive suppose qu'elle ait été mise en œuvre, encore faut-il en avoir une estimation minimale. De même, l'avantage de cette procédure intermédiaire était d'accéder à une compréhension minimale des pratiques dites libres des enseignants des établissements G3, dont on ne peut exclure *a priori* qu'ils mettent eux aussi en œuvre spontanément (puisque sans consigne de notre part) des modalités éventuellement proches de celles impulsées par les consignes fournies aux établissements G1 et G2.

L'accès à la grille d'observation *via* la plateforme n'était donné qu'à l'enseignant-e déclaré-e comme « enseignant-e principal-e » de l'enseignement considéré (général ou spécifique). En cas de partage de cet enseignement entre plusieurs enseignant-e-s (un enseignant-e « principal-e » et des enseignant-e-s « associé-e-s »), ces derniers devaient se concerter entre eux pour renseigner la grille. Le recueil de ces informations était réalisé également par les enseignants des établissements G3, pour connaître leurs modalités de réalisation des séquences pédagogiques (libres puisque sans aucune consigne de notre part) et à des fins de comparaison avec les réponses des enseignants des établissements G1 et G2. Concrètement, les enseignant-e-s de G1, G2, et G3 étaient invité-e-s à indiquer le pourcentage de temps alloué durant leurs séquences pédagogique à chacune des modalités d'enseignements proposées dans la grille d'observation ci-dessous avec la contrainte de ne pas dépasser 100% au total sur les 5 modalités en question. Pour une vision plus complète des questions posées aux enseignant-e-s dans cette grille d'observation, voir Annexe E.

Tableau 9

Extrait de la grille d'observation fournies aux enseignant-e-s des établissements G1, G2 et G3

Les élèves ont-ils travaillé... (donnez une répartition en pourcentage du temps en veillant à ne pas dépasser 100% sur le total des réponses)	
En groupe, avec chaque membre du groupe sur la totalité des chapitres	<input type="text" value="0"/> 
En groupe, avec l'ensemble des membres du groupe sur un seul et même chapitre	<input type="text" value="0"/> 
En groupe, avec chaque membre du groupe sur un chapitre différent	<input type="text" value="0"/> 

Individuellement	0	
En classe entière	0	



Note. Chaque item correspondait à une configuration rendant compte des consignes fournies au départ. L'item 1 correspondait à la configuration « Collectif Classique », l'item 2 à la configuration « Expert », l'item 3 à la configuration « Puzzle », l'item 4 à la configuration « Individuel » et enfin, l'item 4 à la configuration « Classe Entière ».

4.6 Mesure des compétences sociales

Enfin comme nous l'avons mentionné précédemment, nous mesurons également les compétences sociales des élèves à partir de quatre tâches *via* un dispositif comportemental dédié, la BAO. Cette dernière est présentée en Annexe car nous ne l'avons pas exploité dans le cadre de cette thèse (pour plus de détails sur les tâches de la BAO, cf. Annexe K).

Chapitre 4 : Problématique générale, hypothèses et premières analyses

1 Problématique générale et hypothèses

La thèse défendue dans nos travaux en référence à l'étude ProFan implique d'étudier les bénéfices attendus de l'interdépendance positive sur les performances des élèves en tenant compte de plusieurs facteurs : 1) la qualité de l'opérationnalisation des consignes fournies aux enseignant·es pour la réalisation de leurs séquences pédagogiques, 2) la satisfaction/insatisfaction de ces dernier·es à l'égard de la méthode pédagogique proposée (définie par les consignes en G1 et G2), et 3- l'estimation, par les élèves eux-mêmes, de leurs performances scolaires antérieures. Ces trois points sont fondamentaux. En effet, comme le montre notre synthèse de la littérature consacrée à l'interdépendance positive de type « Classe Puzzle » (cf. Chapitres 1 et 2), les bénéfices attendus dans ce cadre sont étroitement liés à la qualité de l'implémentation des dispositifs. Or, l'opérationnalisation des consignes supports de l'étude ProFan ne relevait pas d'un tiers formé aux approches quasi-expérimentales, comme cela peut-être le cas dans d'autres études en contexte scolaire, mais des enseignants eux-mêmes (ceux des groupes G1 et G2). Si cette réalité contribue à la validité écologique de l'étude Profan, elle rend néanmoins assez probable certains biais dans l'interprétation et l'application des consignes sur le terrain. Aussi est-il indispensable de chercher à savoir si, et dans quelle mesure, ces consignes ont été suivies. De même, rares sont les études de terrain—nous n'en connaissons aucune—dans lesquelles les effets des dispositifs testés sont évalués auprès des élèves en tenant compte du niveau de satisfaction/insatisfaction des enseignant·es à l'égard de la méthode implémentée. Or, certains des travaux évoqués plus loin suggèrent un sentiment d'insatisfaction chez les enseignant·es confronté·es à une réduction de leur liberté pédagogique, comme cela

peut-être le cas précisément lorsque les consignes de travail ne recouvrent pas strictement leurs pratiques habituelles, *a fortiori* si elles s'en écartent nettement. Enfin, les effets de l'interdépendance positive sont le plus souvent évalués dans la littérature internationale sans tenir compte des performances scolaires antérieures des élèves (i.e., élèves plus ou moins forts ou faibles). Or, par sa capacité supposée à revaloriser les élèves les plus en difficulté, l'interdépendance positive de type « Classe Puzzle » devrait générer un bénéfice plus important chez ces dernier·e·s, relativement à ceux déjà en réussite. Encore faut-il être en mesure d'intégrer aux analyses conduites des éléments liés au statut scolaire et/ou à sa perception par les élèves. C'est donc le cas dans nos travaux. Ces considérations conduisent aux quatre étapes de travail et hypothèses afférentes décrites ci-dessous.

1) *Estimation de la qualité de l'opérationnalisation des consignes.* Dans une première étape, nous estimerons la qualité de l'opérationnalisation du dispositif ProFan à travers les réponses des enseignant·e·s aux items de la grille d'observation (présentée dans le Chapitre 3, point 5). Ces items permettent en effet d'estimer à la fois la nature et l'ampleur des écarts entre les consignes fournies aux enseignant·e·s et la conduite de leurs séquences pédagogiques sur le terrain. En effet, et comme suggéré plus haut, la bonne opérationnalisation des consignes n'est pas garantie dans les études de terrain, en particulier lorsqu'elle dépend exclusivement des enseignant·e·s eux-mêmes, généralement non formé·e·s à la rigueur d'une démarche scientifique de nature expérimentale. Les enseignant·e·s impliqué·e·s dans l'étude ProFan n'avaient pas eux-mêmes suivi de formation particulière à ce sujet. Celles et ceux des groupes G1 et G2 étaient néanmoins fortement sensibilisé·e·s à l'importance d'un respect strict des consignes fournies. Il reste que, sans formation spécifique quant à la mise en œuvre d'une interdépendance positive de type « Puzzle », les enseignant·e·s de G1 n'étaient pas nécessairement en mesure d'intégrer les consignes dans toute leurs subtilités (ce point sera approfondi en discussion général). D'où la nécessité d'accorder une attention particulière à toutes les informations susceptibles de nous

donner une indication de la réalité des pratiques des enseignant·e·s sur le terrain et de l'ampleur des écarts éventuels entre ces pratiques et les consignes fournies pour l'étude. Les réponses fournies par les enseignant·e·s aux 5 items de la « grille d'observation » ne traduisent pas nécessairement la réalité de leurs pratiques dans le cadre de l'étude ProFan, ne serait-ce qu'en raison d'incompréhensions éventuelles s'agissant des consignes, de biais de mémoire (i.e., erreurs de rappel et/ou reconstruction des faits) ou de désirabilité sociale (i.e., affirmer un respect des consignes en dépit de pratiques qui en étaient en réalité éloignées). Ces informations sont néanmoins les seules auxquelles il est possible d'accéder pour estimer les écarts aux consignes fournies pour l'ensemble des 72 séquences pédagogiques soumises à l'épreuve des faits. Il serait par ailleurs sans doute plus critiquable encore de s'affranchir de ce que nous disent les enseignant·e·s à propos de la manière dont ils/elles ont conduit leurs séquences pédagogiques, ce sont tout de même les mieux placé·e·s à ce sujet.

Aussi calculerons-nous pour chacune des 72 séquences pédagogiques ProFan la tendance centrale de réponse des enseignant·e·s sur chacun des 5 items de la grille d'observation à l'issue de chacune de leur séquence. Cela permettra d'estimer la plus ou moins grande conformité des profils de réponse ainsi obtenus avec les profils attendus en relation avec les consignes fournies pour l'étude. Sur la base de ces profils de réponse, nous catégoriserons les séquences pédagogiques selon trois catégories : les séquences 1) plutôt réussies (consignes plutôt respectées), 2) dégradées (respect plutôt faible des consignes) et 3), échouées (aucun respect des consignes, voire même tendances inverses à celles attendues).

En bref, la stratégie retenue est de ne pas analyser les effets de l'interdépendance positive à l'aveugle du respect des consignes sur le terrain. Il s'agit au contraire de se focaliser en priorité sur les séquences jugées plutôt réussies : celles pour lesquelles nos tests d'hypothèse font davantage sens (les analyses relatives aux séquences dégradées et échouées sont toutefois disponibles en Annexe W et X).

Encore une fois, ce n'est qu'en procédant à une évaluation fine et appropriée de la fidélité avec laquelle une intervention a été mise en œuvre qu'une estimation de son efficacité peut-être raisonnablement conduite (Carroll, 2007).

2) *Estimation de la satisfaction/insatisfaction des enseignant·e·s prenant part au dispositif ProFan.* Dans une deuxième étape, nous nous intéresserons aux réponses des enseignant·e·s aux items de la grille d'observation concernant cette fois leur satisfaction/insatisfaction à l'égard des consignes fournies pour l'étude. Ces consignes, de par leur nature, réduisent presque nécessairement la liberté pédagogique des enseignant·e·s. En effet, les consignes invitant à un travail de nature collectif en G1 ou G2 ne sont pas des modalités pédagogiques dominantes dans l'institution scolaire traditionnelle. À ce titre, elles sont susceptibles de réduire la liberté pédagogique des enseignant·e·s. Nous n'affirmons pas que les consignes fournies pour G1 et G2 réduisent *de facto* cette liberté pour tous les enseignant·e·s, mais plutôt qu'elles la réduisent pour celles et ceux qui ne pratiquent pas ou peu une pédagogie fondée sur l'apprentissage coopératif quelle que soit d'ailleurs sa nature.

Cet élément lié à la liberté pédagogique est particulièrement important dans le cadre de nos travaux. Plusieurs études ont en effet mis en évidence l'existence d'un lien étroit entre l'autonomie dont peuvent bénéficier les enseignant·e·s dans leurs pratiques pédagogiques et leur satisfaction professionnelle. Ainsi, plus la marge d'autonomie octroyée aux enseignant·e·s pour réaliser leur travail est importante, en termes par exemple de choix de contenus à enseigner et/ou de méthodes pédagogiques à exploiter, plus ils sont satisfaits professionnellement (e.g., Avanzi et al., 2013 ; Humphrey, 2007 ; Kengatharan, 2020 ; Koustelios et al., 2004 ; Skaalvik & Skaalvik, 2009 ; Skaalvik & Skaalvik, 2010). En gagnant en autonomie, les enseignant·e·s ont le sentiment d'avoir mené un travail de qualité pour atteindre les objectifs pédagogiques qu'ils/elles se fixent, nourrissant par conséquent leur niveau de satisfaction (Amathieu & Chaliès, 2014). C'est pourquoi dans nos travaux nous évaluerons aussi l'impact de l'induction

expérimentale sur la satisfaction exprimée par les enseignant·e·s à l'égard du dispositif pour chaque catégorie de séquence pédagogique (i.e., « plutôt réussies », « dégradées », « échouées »). Considérant la littérature évoquée plus haut, nous devrions observer un effet négatif des consignes de travail collectif (G1 et G2) sur le sentiment de satisfaction en question, relativement à la condition sans aucune consigne de travail (G3), *de facto* la plus compatible avec la liberté pédagogique des enseignant·e·s qui ne pratiquent pas ou peu cette « forme pédagogique » fondée sur le travail en groupe.

Enfin, la troisième et dernière étape de nos analyses sera traitée dans le chapitre 5 afin de faciliter la lecture des résultats. Celle-ci consistera à évaluer si le bénéfice attendu du mécanisme d'interdépendance positive s'exprime 1) quelles que soient les performances antérieures (auto-rapportées) des élèves ou 2) davantage chez ceux rapportant les performances les plus faibles. Nous optons pour la seconde hypothèse, précisément parce que la (re)valorisation permise en principe par le mécanisme d'interdépendance positive au centre de nos travaux fait davantage sens pour les élèves les plus en difficulté. Si l'on ne peut exclure *a priori* un bénéfice pour tous les élèves, l'interdépendance positive et donc la possibilité donnée à chacun de se familiariser puis d'échanger avec les autres membres du groupe des informations que ces derniers ne maîtrisent pas encore constitue en principe une expérience plus nouvelle, marquante et valorisante pour les élèves habituellement en échec que pour ceux en réussite.

Dans leurs travaux sur la régulation sociale des performances scolaires, Monteil et Huguet (2013 ; voir aussi Huguet, 2006 ; Huguet & Kuyper, 2017 ; Monteil & Huguet, 1999) montrent que la valorisation des élèves en difficulté ne produit pas nécessairement les effets escomptés sur leurs performances. Non familière, cette valorisation entrave au contraire leur capacité d'apprentissage, au moins dans un premier temps en particulier lorsqu'elle est rendue publique (perceptible par les autres élèves de la classe). Ce n'est qu'avec la répétition de cette valorisation au fil du temps que les élèves présentant des difficultés montrent de meilleures

performances même en situation de forte visibilité dans l'espace de la classe. Dans ces mêmes travaux, cependant, la valorisation en question était opérationnalisée par une attribution de succès personnelle dans une activité préalable à la tâche principale. La notion d'interdépendance positive impliquée dans nos travaux n'implique donc pas d'exposer les élèves à un tel succès qui, au moins dans un premier temps, pourrait leur paraître incongru ou *a minima* non familier. Elle réfère à la possibilité de devenir expert d'informations partielles néanmoins indispensables à la compréhension d'une leçon et dont les autres membres du groupe ignorent la teneur avant qu'elles ne soient échangées avec eux. Plutôt que d'avoir à assumer publiquement un succès personnel, il s'agit donc pour chaque élève de partager avec les autres membres du groupe des informations qui, une fois combinées avec les leurs, permettent en principe d'accéder à la maîtrise d'un contenu pédagogique dans sa totalité. Cette forme de valorisation référant à la possibilité de devenir expert d'informations indispensables à la réussite de tous les membres du groupe, tout en restant ignorants dans un premier temps des informations détenues par les autres, nous semble moins incongrue—pour les élèves les plus faibles—que l'attribution publique d'un succès personnel que pour la plupart ils n'ont jamais connu. Par ailleurs, et comme suggéré antérieurement, ce sont bien les élèves plus faibles qui sont les plus à même de percevoir cette interdépendance positive comme un élément de valorisation. En effet, la maîtrise des contenus proposés par l'enseignant constitue, par définition, une expérience plus familière voire même ordinaire pour leurs homologues plus en réussite. D'où notre hypothèse d'un bénéfice plus grand de l'interdépendance positive chez les élèves rapportant les performances scolaires antérieures les plus faibles.

Cette hypothèse est également soutenue par les travaux sur la « paresse sociale » (social loafing ; cf. Karau & Williams, 1993 pour une revue en langue anglaise ; Huguet & Monteil, 2001 ; Huguet 1995 pour des revues en langue française). En effet, rappelons que le travail en groupe non structuré encourage le plus souvent une « paresse sociale », définie au sens d'une

réduction de l'effort personnel en situation de travail collectif relativement à un travail de nature strictement individuel. Cette paresse est hautement probable lorsque la tâche est difficile chez les élèves les plus faibles qui alors se reposent sur les efforts des autres membres du groupe et n'apprennent de ce fait rien ou très peu. Les plus forts peuvent dans le même temps vouloir "compenser" cette faiblesse des élèves en difficulté, comme le prévoit aussi Karau et Williams (1993) dans leur modèle de l'effort collectif (« collective effort model »), ne serait-ce que pour afficher leur supériorité, une hypothèse renforcée par d'autres travaux sur les liens entre sentiment de supériorité et paresse sociale (cf. Huguet et al., 1999). Il reste que cette paresse est le plus souvent neutralisée lorsque la condition de travail collectif est structurée de manière à rendre à la fois possible et identifiable une contribution de l'ensemble des membres du groupe (cf. Huguet, 1995), à la manière de G1 dans ProFan. En condition de travail collectif plus classique—non structurée et proche de G2, la combinaison de la paresse sociale des plus faibles et de l'effet éventuel de compensation chez leurs homologues plus forts conduit à attendre un écart maximal entre les élèves les plus faibles et ceux plus forts ; un écart peut-être encore plus fort que celui constaté en condition classe entière dans laquelle les élèves les plus forts n'ont rien à compenser. Cet écart devrait néanmoins s'estomper en condition de travail collectif structuré (d'où une interaction) en raison du mécanisme d'interdépendance positive permettant en principe aux plus faibles une valorisation et un investissement plus grand dans la séquence proposée et donc une meilleure performance.

Cette hypothèse d'interaction entre l'induction expérimentale et les performances auto-rapportées des élèves est au centre de nos travaux. Considérant qu'une telle interaction dans le cadre des séquences dites « dégradées » ou « échouées » serait par nature extrêmement difficile à interpréter du fait de reports verbaux des enseignants peu compatibles avec —si ce n'est même contraire aux—consignes qui leur étaient fournies, nous focaliserons prioritairement sur sa

présence éventuelle dans les séquences jugées « plutôt réussies ». Autrement dit, les séquences pour lesquelles les consignes semblent avoir été davantage respectées.

Une autre facette de nos travaux consiste à déterminer si cette interaction dépend ou non du niveau de satisfaction des enseignant·e·s à l'égard des conditions de réalisation de leurs séquences pédagogiques, séquences que nous savons en effet contraintes en G1 et G2 par nos consignes de travail (i.e., consignes susceptibles d'altérer une liberté pédagogique à laquelle les enseignants sont généralement attachés). Le bénéfice de l'interdépendance positive (G1), que nous envisageons donc plus élevé pour les élèves les plus faibles, devrait s'exprimer nous l'avons vu au moins dans les séquences jugées plutôt conformes aux consignes (séquences jugées plutôt réussies) dans la mesure où cette conformité est une condition à la mise en œuvre effective de l'interdépendance positive en G1. Il reste que même dans le cadre des séquences jugées plutôt réussies, l'impact de l'interdépendance positive sur les performances des élèves peut aussi dépendre du niveau de satisfaction exprimée par les enseignant·e·s à l'égard des conditions qui leur étaient proposées pour ces séquences. L'absence de toute supériorité de G1 sur les deux autres conditions de l'étude est attendue dans les séquences jugées dégradées ou échouées quel que soit ce sentiment de satisfaction des enseignants. Dans les séquences jugées réussies, cette supériorité de G1 est en revanche attendue, au moins dans les séquences conduites par les enseignant·e·s se déclarant satisfait·e·s et donc par hypothèse plutôt en accord avec la méthode proposée (travail en groupe structuré par une interdépendance positive). La question de savoir si un bénéfice de l'interdépendance positive (G1) est possible dans les séquences jugées réussies, mais conduites par des enseignant·e·s se déclarant au contraire insatisfait·e·s reste ouverte. Même si la probabilité de ce bénéfice est faible dans ces conditions *a priori* défavorables à son expression, on ne peut totalement exclure que l'interdépendance positive agisse favorablement même à l'insu des enseignant·e·s qui la mettent en œuvre sans conviction voire même avec un sentiment d'insatisfaction. Aussi rechercherons-nous

l'interaction évoquée plus haut entre l'induction expérimentale et les performances scolaires antérieures (auto-rapportées) des élèves, non seulement de manière prioritaire dans les séquences jugées plutôt réussies mais aussi en tenant compte du sentiment de satisfaction/insatisfaction des enseignant·e·s à l'égard des conditions d'enseignement qui leur étaient proposées pour l'étude (voir Chapitre 5).

2 Qualité de l'opérationnalisation du dispositif ProFan

2.1 Tendances centrales

Dans un premier temps, nous nous sommes intéressés aux réponses des enseignant·e·s à la grille d'observation qui concernaient la conduite de la classe et les consignes qui ont été fournies pour chacune des séquences pédagogiques. L'objectif était d'estimer la nature et l'ampleur des écarts entre les consignes fournies aux enseignant·e·s et la conduite de leurs séquences pédagogiques sur le terrain. Ainsi, à chaque fois que l'enseignement d'une séquence pédagogique prenait fin, les enseignant·e·s étaient invité·e·s à renseigner une grille d'observation par classe (si la séquence avait été conduite en classe entière) ou bien une grille d'observation par sous-groupe (si la séquence avait été conduite en sous-groupe). Parmi les différents items de la grille, les 5 présentés ci-dessous avaient pour objectif d'apprécier le déroulement des séquences pédagogiques au regard des consignes fournies (Pour plus de détails sur les items de la grille d'observation, voir Chapitre 3 p. 87 et Annexe L) :

Les élèves ont-ils travaillé :

- En groupe, avec chaque membre du groupe sur la totalité des chapitres
- En groupe, avec l'ensemble des membres du groupe sur un seul et même chapitre
- En groupe, avec chaque membre du groupe sur un chapitre différent
- Individuellement
- En classe entière

Chaque item correspondait à une configuration rendant compte des consignes fournies au départ. L'item 1 correspondait à la configuration « Collectif Classique », l'item 2 à la configuration « Expert », l'item 3 à la configuration « Puzzle », l'item 4 à la configuration « Individuel » et enfin, l'item 4 à la configuration « Classe Entière ». Les enseignant·e·s étaient invité·e·s à se positionner sur chaque item en donnant une répartition du pourcentage de temps (allant de 0 à 100 %) passé dans chacune de ces configurations sur l'ensemble de la séquence en question. Il leur était demandé de veiller à ne pas dépasser 100% sur le total des réponses. Ainsi, nous avons des attentes concernant les profils de réponse que nous supposons être influencés par les consignes.

2.2 Résultats

Les analyses ont été réalisées avec le logiciel SPSS (Version 26, IBM Corp., 2019). Dans le but d'obtenir les profils de réponses moyen aux 5 items de la grille d'observation et de dégager les tendances centrales, nous avons conduit des ANOVA à mesures répétées avec en variable indépendante la condition, et en variable dépendante chacun des 5 items de la grille d'observation. Pour chacune des ANOVA, nous avons également réalisé trois comparaisons (G1 vs. G2, G1 vs. G3, G2 vs. G3) post-hoc avec correction de Bonferroni afin d'examiner les différences entre chaque condition sur chaque item. Enfin, nous avons généré des graphiques qui rendent compte du pourcentage de temps passé dans les différentes configurations : Classe Entière, Individuel, Collectif Classique, Expert et Puzzle dans chaque condition. Nous avons étudié ces profils selon l'approche suivante³ : cohorte par cohorte, filière par filière et séquence

³ Nous avons également étudié les profils de réponses aux items de la grille selon les deux approches suivantes : 1) Cohorte par cohorte, toutes filières et toutes séquences confondues en Français et en Mathématiques, et filière par filière toutes séquences confondues pour les enseignements professionnels et 2) Cohorte par cohorte, toutes filières confondues et séquence par séquence en Français et en Mathématiques. Pour plus de détails concernant ces deux approches, voir Annexe M et Annexe N.

par séquence en mathématiques, en français et pour les enseignements professionnels. De cette façon, nous avons obtenu 72 graphiques.

2.3 Catégorisation de l'implémentation des séquences ProFan

Dans la mesure où nous avons observé une grande diversité parmi les réponses des enseignant·e·s aux 5 items de la grille d'observation qui se traduit sur le terrain par une grande hétérogénéité en terme de qualité d'implémentation du dispositif, nous avons décidé de catégoriser les séquences sur la base des pourcentages de temps déclarés par les enseignant·e·s.

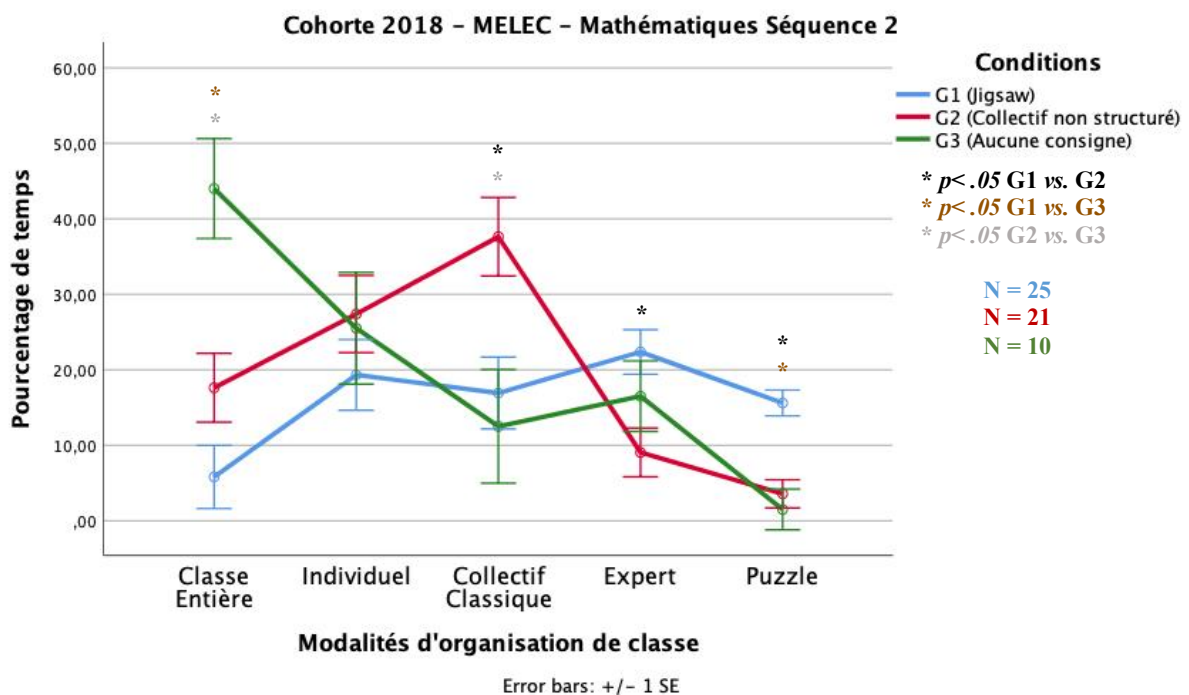
Après avoir examiné les réponses moyennes d'une séquence à l'autre, nous avons remarqué que de façon récurrente, les items les plus discriminants vis à vis des groupes G1, G2 et G3 étaient l'item « Puzzle » pour le groupe G1, l'item « Collectif Classique » pour le groupe G2 et l'item « Classe Entière » pour le groupe G3. Par conséquent, nous avons comparé les pourcentages de temps accordés à chacun de ces 3 items selon le groupe expérimental et nous avons dégagé les 3 profils de réponses attendus suivants. Pour les séquences pour lesquelles les consignes fournies au départ avaient été le mieux respectées, nous attendions que le pourcentage de temps déclaré en configuration Puzzle soit supérieur dans la condition G1 comparativement aux conditions G2 et G3. Nous attendions également que le pourcentage de temps déclaré en configuration Collectif Classique soit supérieur dans la condition G2 comparativement aux conditions G1 et G3. Enfin, nous attendions que le pourcentage de temps déclaré en configuration Classe Entière soit supérieur dans la condition G3 comparativement aux conditions G1 et G2.

De cette façon, nous avons catégorisé l'opérationnalisation des 72 séquences en 3 catégories selon les critères d'inclusions suivants :

- 1) *Opérationnalisation plutôt réussie* : le pourcentage de temps déclaré dans chaque configuration doit correspondre aux attendus dans au moins 2 des 3 items critiques (Pour un exemple, voir Figure 13 ci-dessous).

Figure 13

Pourcentage de temps déclaré dans les modalités d'organisation de classe pour les conditions G1, G2 et G3 en MELEC pour la Séquence 2 de Mathématiques pour la cohorte 2018



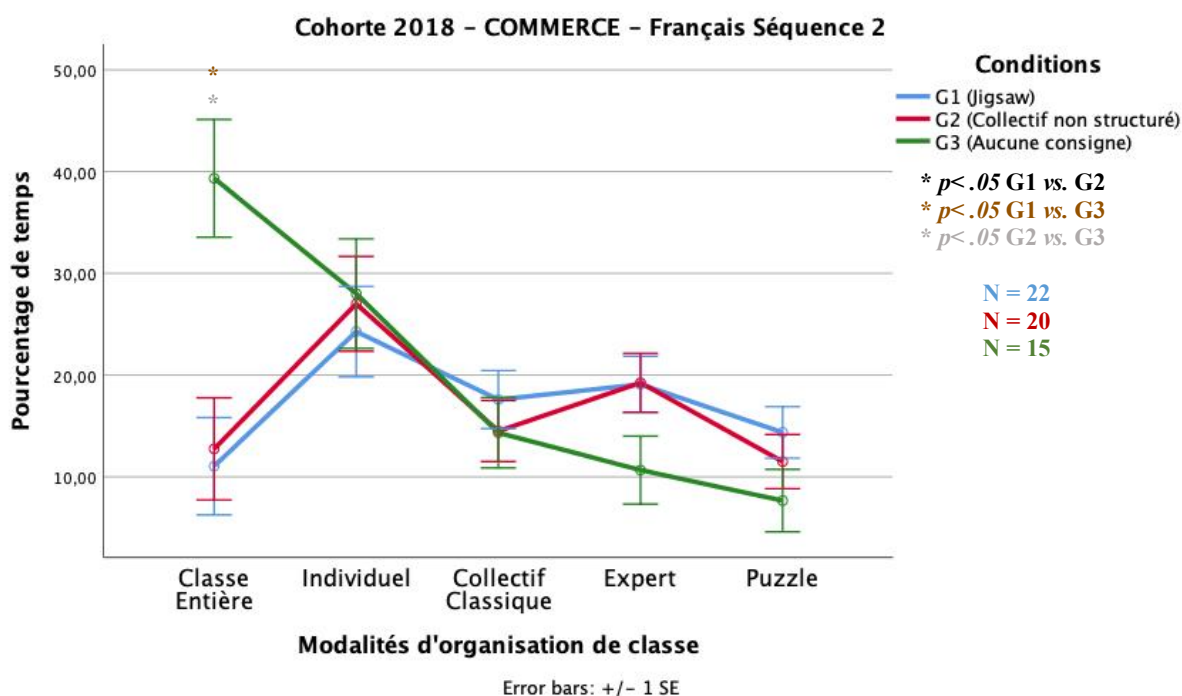
Note. G1 désigné par « Jigsaw » correspond à la condition « Collectif Structuré » et G3 désigné par « Aucune consigne » correspond à la condition « Libre ».

Le pourcentage de temps déclaré en configuration Puzzle est supérieur dans la condition G1 comparativement aux conditions G2 et G3. Le pourcentage de temps déclaré en configuration Collectif Classique est supérieur dans la condition G2 comparativement aux conditions G1 et G3. Le pourcentage de temps déclaré en configuration Classe Entière est supérieur dans la condition G3 comparativement aux conditions G1 et G2.

- 2) *Opérationnalisation dégradée* : le pourcentage de temps déclaré doit correspondre aux attendus dans au moins 1 des 3 items critiques (Pour un exemple, voir Figure 14 ci-dessous).

Figure 14

Pourcentage de temps déclaré dans les modalités d'organisation de classe pour les conditions G1, G2 et G3 en COMMERCE pour la Séquence 2 de Français pour la cohorte 2018



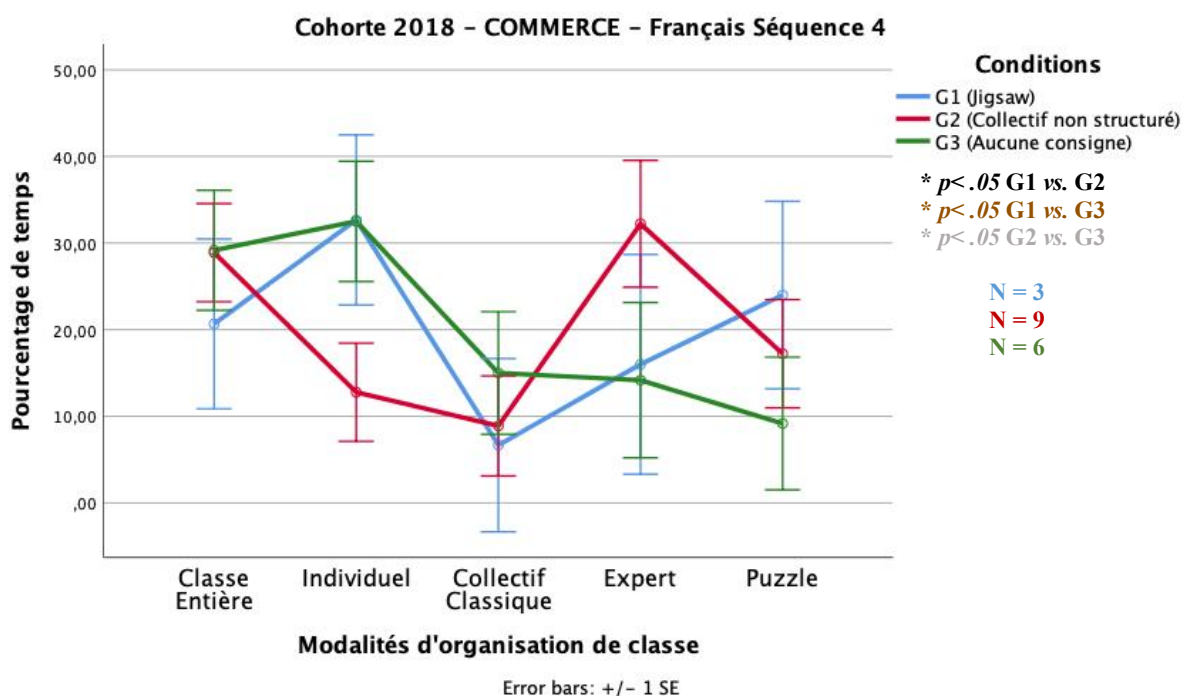
Note. G1 désigné par « Jigsaw » correspond à la condition « Collectif Structuré » et G3 désigne par « Aucune consigne » correspond à la condition « Libre ».

Le pourcentage de temps déclaré en configuration Puzzle n'est pas supérieur dans la condition G1 comparativement aux conditions G2 et G3. Le pourcentage de temps déclaré en configuration Collectif Classique n'est pas supérieur dans la condition G2 comparativement aux conditions G1 et G3. Le pourcentage de temps déclaré en configuration Classe Entière est supérieur dans la condition G3 comparativement aux conditions G1 et G2.

3) *Opérationnalisation échouée* : le pourcentage de temps déclaré dans chaque configuration ne correspond à aucun des attendus sur les 3 items critiques (Pour un exemple, voir Figure 15 ci-dessous).

Figure 15

Pourcentage de temps déclaré dans les modalités d'organisation de classe pour les conditions G1, G2 et G3 en COMMERCE pour la Séquence 4 de Français pour la cohorte 2018



Note. G1 désigné par « Jigsaw » correspond à la condition « Collectif Structuré » et G3 désigné par « Aucune consigne » correspond à la condition « Libre ».

Le pourcentage de temps déclaré en configuration Puzzle n'est pas supérieur dans la condition G1 comparativement aux conditions G2 et G3. Le pourcentage de temps déclaré en configuration Collectif Classique n'est pas supérieur dans la condition G2 comparativement aux conditions G1 et G3. Le pourcentage de temps déclaré en configuration Classe Entière est supérieur dans la condition G3 comparativement aux conditions G1 et G2.

De cette façon et parmi les 72 graphiques générés, 21 séquences ont été jugées plutôt réussies, 32 ont été jugées dégradées et enfin, 19 ont été jugées échouées (Pour plus de détails concernant la catégorisation des séquences, voir Tableau 9 ci-dessous et leurs représentations graphiques, voir Annexe G).

Tableau 9

Catégorisation de l'implémentation des séquences ProFan, cohorte par cohorte, filière par filière pour le Français, les Mathématiques et les Enseignements professionnels

Séquences	
$N = 72$	
Opérationnalisation	
Opérationnalisation plutôt réussie $N = 21$	Cohorte 2018 – ASSP – Français Séquence 1
	Cohorte 2017 – ASSP – Français Séquence 4
	Cohorte 2018 – ASSP – Mathématiques Séquence 1
	Cohorte 2017 – ASSP – Mathématiques Séquence 2
	Cohorte 2018 – ASSP – Mathématiques Séquence 2
	Cohorte 2017 – ASSP – Mathématiques Séquence 3
	Cohorte 2018 – ASSP – Mathématiques Séquence 3
	Cohorte 2017 – ASSP – Mathématiques Séquence 4
	Cohorte 2017 – COMMERCE – Mathématiques Séquence 1
	Cohorte 2017 – COMMERCE – Mathématiques Séquence 2
	Cohorte 2018 – COMMERCE – Mathématiques Séquence 2
	Cohorte 2017 – COMMERCE – Mathématiques Séquence 3
	Cohorte 2017 – COMMERCE – Mathématiques Séquence 4
	Cohorte 2018 – COMMERCE – Enseignement professionnel Séquence 2

	Cohorte 2017 – MELEC – Français Séquence 3
	Cohorte 2017 – MELEC – Français Séquence 4
	Cohorte 2017 – MELEC – Mathématiques Séquence 1
	Cohorte 2018 – MELEC – Mathématiques Séquence 2
	Cohorte 2017 – MELEC – Mathématiques Séquence 3
	Cohorte 2018 – MELEC – Mathématiques Séquence 3
	Cohorte 2017 – MELEC – Mathématiques Séquence 4
	<hr/>
	Cohorte 2017 – ASSP – Français Séquence 1
	Cohorte 2017 – ASSP – Français Séquence 2
	Cohorte 2018 – ASSP – Français Séquence 2
	Cohorte 2017 – ASSP – Français Séquence 3
	Cohorte 2018 – ASSP – Français Séquence 4
	Cohorte 2017 – ASSP – Mathématiques Séquence 1
	Cohorte 2018 – ASSP – Mathématiques Séquence 4
Opérationnalisation	Cohorte 2018 – ASSP – Enseignement professionnel Séquence 2
dégradée	Cohorte 2017 – ASSP – Enseignement professionnel Séquence 3
N = 32	Cohorte 2018 – ASSP – Enseignement professionnel Séquence 3
	Cohorte 2017 – ASSP – Enseignement professionnel Séquence 4
	Cohorte 2018 – ASSP – Enseignement professionnel Séquence 4
	Cohorte 2017 – COMMERCE – Français Séquence 1
	Cohorte 2018 – COMMERCE – Français Séquence 1
	Cohorte 2017 – COMMERCE – Français Séquence 2
	Cohorte 2018 – COMMERCE – Français Séquence 2
	Cohorte 2017 – COMMERCE – Français Séquence 3
	Cohorte 2018 – COMMERCE – Français Séquence 3

Opérationnalisation
échouée
N = 19

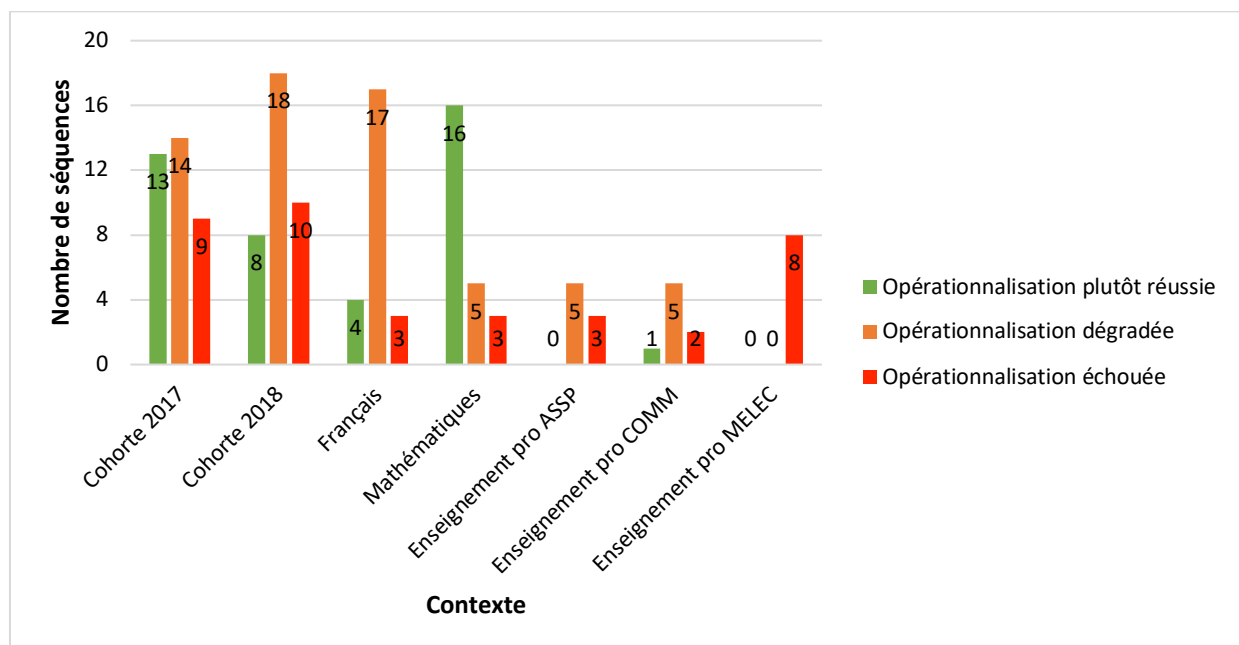
Cohorte 2017 – COMMERCE – Français Séquence 4
 Cohorte 2018 – COMMERCE – Mathématiques Séquence 1
 Cohorte 2018 – COMMERCE – Mathématiques Séquence 3
 Cohorte 2018 – COMMERCE – Enseignement professionnel Séquence 1
 Cohorte 2017 – COMMERCE – Enseignement professionnel Séquence 3
 Cohorte 2018 – COMMERCE – Enseignement professionnel Séquence 3
 Cohorte 2017 – COMMERCE – Enseignement professionnel Séquence 4
 Cohorte 2018 – COMMERCE – Enseignement professionnel Séquence 4
 Cohorte 2017 – MELEC – Français Séquence 1
 Cohorte 2018 – MELEC – Français Séquence 1
 Cohorte 2017 – MELEC – Français Séquence 2
 Cohorte 2018 – MELEC – Français Séquence 2
 Cohorte 2018 – MELEC – Français Séquence 3
 Cohorte 2018 – MELEC – Mathématiques Séquence 1

Cohorte 2018 – ASSP – Français Séquence 3
 Cohorte 2017 – ASSP – Enseignement professionnel Séquence 1
 Cohorte 2018 – ASSP – Enseignement professionnel Séquence 1
 Cohorte 2017 – ASSP – Enseignement professionnel Séquence 2
 Cohorte 2018 – COMMERCE – Français Séquence 4
 Cohorte 2018 – COMMERCE – Mathématiques Séquence 4
 Cohorte 2017 – COMMERCE – Enseignement professionnel Séquence 1
 Cohorte 2017 – COMMERCE – Enseignement professionnel Séquence 2
 Cohorte 2018 – MELEC – Français Séquence 4
 Cohorte 2017 – MELEC – Mathématiques Séquence 2
 Cohorte 2018 – MELEC – Mathématiques Séquence 4

Cohorte 2017 – MELEC – Enseignement professionnel Séquence 1
 Cohorte 2018 – MELEC – Enseignement professionnel Séquence 1
 Cohorte 2017 – MELEC – Enseignement professionnel Séquence 2
 Cohorte 2018 – MELEC – Enseignement professionnel Séquence 2
 Cohorte 2017 – MELEC – Enseignement professionnel Séquence 3
 Cohorte 2018 – MELEC – Enseignement professionnel Séquence 3
 Cohorte 2017 – MELEC – Enseignement professionnel Séquence 4
 Cohorte 2018 – MELEC – Enseignement professionnel Séquence 4

Figure 16

Nombre de séquences « plutôt réussie » vs. « dégradée » vs. « échouée » par cohorte (tout enseignement confondue), en Français, en Mathématiques et pour les Enseignements professionnels par filière



Comme nous pouvons l'observer sur la Figure 16, la catégorie « opérationnalisation plutôt réussie » n'est pas dominante. En effet, nous constatons que la catégorie dominante pour

la cohorte 2017 et pour la cohorte 2018, toutes filières et toutes matières confondues, est celle des séquences « dégradées » ($N_{Cohorte\ 2017} = 14$ séquences, $N_{Cohorte\ 2018} = 18$ séquences). Ce constat est similaire pour le français toutes cohortes et toutes filières confondues ($N = 17$ séquences), les enseignements professionnels en ASSP ($N = 5$ séquences) ainsi qu'en COMMERCE ($N = 5$ séquences), à l'exception des mathématiques où les séquences jugées « plutôt réussies » sont majoritaires ($N = 16$ séquences). Enfin, nous constatons un nombre élevé de séquences échouées pour les enseignements professionnels ($N = 13$ séquences sur 24 au total), et cela tout particulièrement en MELEC où l'intégralité des séquences ont été jugées échouées. Par conséquent, nous ne traiterons pas les enseignements professionnels dans la suite de nos analyses (cf. Partie 2 et 3 de ce chapitre). Nous pouvons évidemment nous interroger sur les raisons de l'échec massif observé dans les enseignements professionnels s'agissant du respect des consignes fournies, tel que mesuré par les réponses des enseignant·e·s aux items de la grille d'observation. Cet échec pourrait tenir au fait que travailler collectivement est une pratique si familière dans ces enseignements qu'elle a conduit les enseignant·e·s des trois groupes à ne pas tenir suffisamment compte des subtilités des consignes pour la conduite de leur classe, et/ou à répondre de manière trop imprécise aux items de la grille d'observation.

Afin d'affiner l'information statistique s'agissant des types de séquences observées en français, en mathématiques et pour les enseignements professionnels, nous proposons l'Annexe P consacrée à la dispersion des réponses attachées aux 3 items critiques de la grille d'observation (i.e., item « Classe Entière », item « Collectif Classique », et item « Puzzle »). Cette analyse permet de repérer les classes qui présentent ou ne présentent pas les profils de réponses attendus.

3 Effet de l'induction expérimentale sur la satisfaction exprimée par les enseignant·es à l'égard du dispositif

En plus de caractériser le respect des consignes qui ont été fournies au démarrage du projet par le biais des réponses aux items de la grille d'observation que nous avons détaillé précédemment, nous interrogeons également les enseignant·es sur leur satisfaction/insatisfaction. Plus précisément, nous avons interrogé ces dernier·es sur la façon dont ils/elles se percevaient par rapport aux consignes qui ont été fournies. En effet, de par leurs natures, ces consignes réduisaient la liberté pédagogique de celles et ceux qui ne pratiquent pas ou peu une pédagogie fondée sur l'apprentissage coopératif. Cet élément est particulièrement important dans le cadre de nos travaux car, comme nous l'avons développé précédemment, la littérature scientifique fait état de plusieurs études suggérant l'existence d'un lien étroit entre l'autonomie dont peuvent bénéficier les enseignant·es dans la réalisation de leurs activités pédagogiques, et leur satisfaction professionnelle. Considérant cette littérature, nous devrions observer un effet négatif des consignes de travail collectif (G1 et G2) sur le sentiment de satisfaction, relativement à la condition sans aucune consigne de travail (G3) *de facto* la plus compatible avec la liberté pédagogique des enseignant·es qui ne pratiquent pas ou peu cette « forme pédagogique » fondée sur le travail en groupe.

Dans ce but, nous nous sommes intéressés aux réponses des enseignant·es à la grille d'observation qui concernaient leur niveau d'efficacité perçue et de satisfaction à l'égard du dispositif par le biais des 2 items suivants (pour plus de détails sur les items de la grille d'observation, voir Annexe D):

- 3) Pensez-vous que le mode d'organisation pédagogique mis en place pour l'ensemble de la séquence est efficace pour l'apprentissage des élèves ?
- 4) Êtes-vous satisfait·e vous-même de ce mode d'organisation ?

Ainsi, à chaque fois que l'enseignement d'une séquence pédagogique prenait fin pour une classe donnée, les enseignant·e·s étaient invité·e·s à exprimer leur niveau d'efficacité perçue et de satisfaction à l'égard du dispositif sur une échelle de type Likert allant de 1 (*Pas du tout*) à 7 (*Énormément*).

3.1 Résultats

Les analyses ont été réalisées avec le logiciel SPSS (Version 26 ; IBM Corp., 2019). Dans un premier temps, nous avons réalisé une analyse de corrélation toutes séquences confondues entre les 2 items mentionnés précédemment qui indiquent que ces derniers sont fortement corrélés ($r > .70, p < .001$). Nous les avons donc agrégés et ainsi obtenu un indicateur de satisfaction plus riche traduisant à la fois l'efficacité perçue par les enseignant·e·s d'un mode donné d'organisation pédagogique, et leur satisfaction à ce sujet. Dans un second temps, nous avons étudié l'effet de l'induction expérimentale, autrement dit des conditions G1, G2 et G3, sur cet indicateur de satisfaction *via* des analyses de variance pour chaque type de séquence (moyenne des séquences réussies, dégradées et échouées) en français et en mathématiques. Ensuite, cet effet a été calculé, toujours dans ces deux disciplines, à l'échelle de chaque séquence (chacune des 20 séquences réussies, des 22 séquences dégradées, et des 6 séquences échouées). Enfin, nous avons réalisé trois comparaisons (G1 vs. G2, G1 vs. G3, G2 vs. G3) *post-hoc* avec correction de Bonferroni afin d'examiner les différences entre chaque condition sur l'item de satisfaction.

3.2 Séquences plutôt réussies

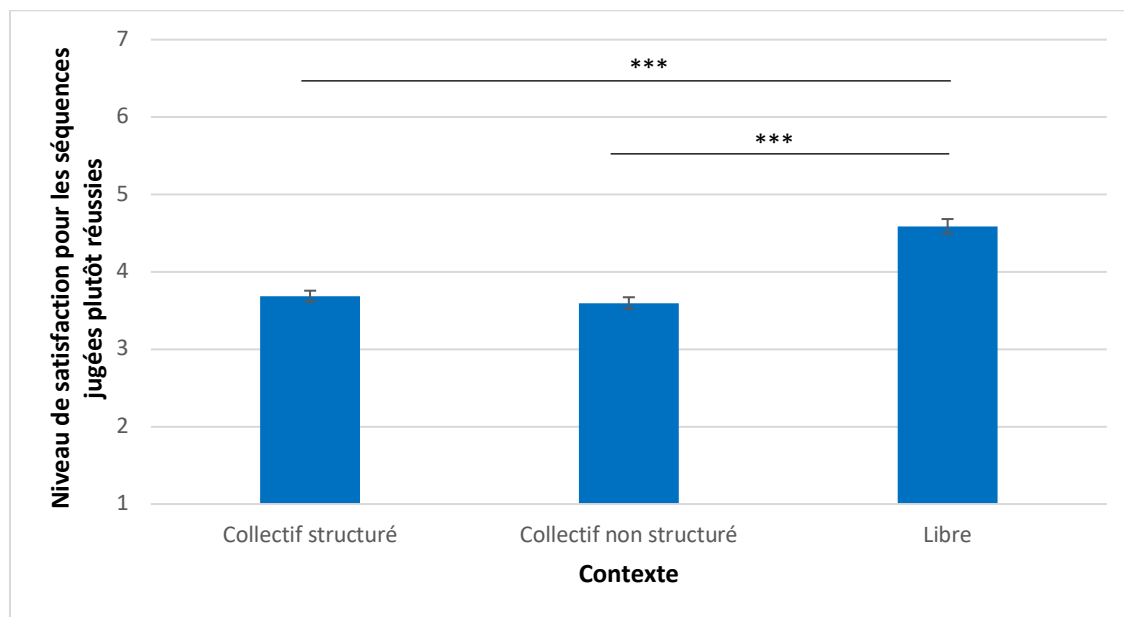
3.2.1 En moyenne

En dépit du caractère plutôt réussies des séquences pédagogiques, les enseignant·e·s expriment un niveau de satisfaction inférieur dans les conditions Collectif Structuré et Collectif

non Structuré que dans la condition Libre ($F(2, 1079) = 38.26, p < .000, \eta_p^2 = .07, G1 : M = 3.69, ET = 1.49, G2 : M = 3.60, ET = 1.53, G3 : M = 4.59, ET = 1.38$, cf. Figure 17 ci-dessous). Le fait de se conformer aux consignes ne semble pas incompatible avec le fait d'exprimer de l'insatisfaction à leurs égards. Ainsi, ces résultats sont en accord avec notre hypothèse d'un effet négatif des consignes de travail collectif (G1 et G2) sur le sentiment de satisfaction, relativement à la condition sans aucune consigne de travail (G3).

Figure 17

Niveau de satisfaction exprimé par les enseignant·es pour les conditions Collectif structuré (G1), Collectif non structuré (G2) et Libre (G3) toutes cohortes et toutes filières confondues en Français et en Mathématiques pour les séquences jugées plutôt réussies



Note. * $p < .05$; ** $p < .01$; *** $p < .001$; Nombre de classes et de demi-classes inclus dans cette analyse : $N_{\text{Total}} = 1079, N_{\text{Collectif Structuré}} = 441, N_{\text{Collectif non structuré}} = 396, N_{\text{Libre}} = 242$.

3.2.2 Séquence par séquence

Séquence par séquence, nous constatons que les enseignant·e·s rapportent un niveau de satisfaction significativement supérieur lorsque la liberté pédagogique est assurée, autrement dit dans la condition G3 (Libre), dans 40 % des cas soit 8 séquences sur 20 au total. Nous observons aussi une tendance, dans le même sens, mais non significative dans 7 autres séquences. Les enseignant·e·s rapportent un niveau de satisfaction plus élevé dans la condition G2 (Collectif non structuré) que dans la condition G1 (Collectif structuré) dans 1 seule séquence (i.e., Cohorte 2017 COMMERCE Séquence 2 de Mathématiques.) Aucun effet ni tendance ne sont observés pour les 4 séquences restantes (pour plus de détails, voir Annexe Q).

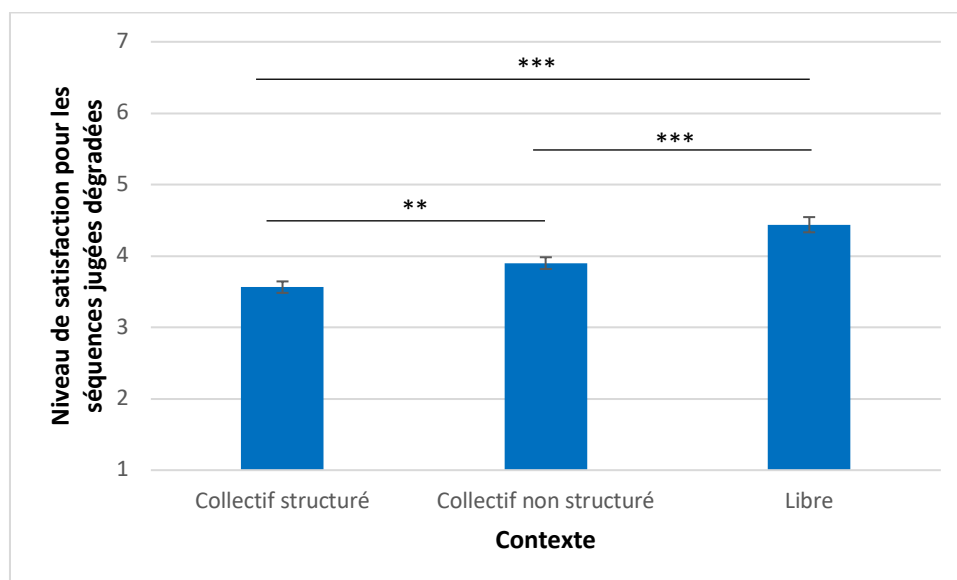
3.3 Séquences dégradées

3.3.1 En moyenne

Le pattern observé précédemment pour les séquences jugées plutôt réussies est sensiblement le même pour les séquences jugées dégradées. Ainsi et conformément à nos hypothèses, les enseignant·e·s expriment un niveau de satisfaction inférieur dans les conditions Collectif Structuré et Collectif non Structuré que dans la condition Libre ($F(2, 858) = 21,53, p < .000, \eta_p^2 = .05$; G1 : $M = 3.56, ET = 1.46$; G2 : $M = 3.90, ET = 1.57$; G3 : $M = 4.44, ET = 1.36$, voir Figure 18 ci-dessous). Nous observons également un niveau de satisfaction inférieur dans la condition Collectif Structuré (G1) comparativement à la condition Collectif non structuré (G2).

Figure 18

Niveau de satisfaction exprimé par les enseignant.e.s pour les conditions Collectif structuré (G1), Collectif non structuré (G2) et Libre (G3) toutes cohortes et toutes filières confondues en Français et en Mathématiques pour les séquences jugées dégradées



Note. * $p < .05$; ** $p < .01$; *** $p < .001$; Nombre de classes et de demi-classes inclus dans cette analyse : $N_{\text{Total}} = 858$, $N_{\text{Collectif Structuré}} = 339$, $N_{\text{Collectif non structuré}} = 325$, $N_{\text{Libre}} = 194$

3.3.2 Séquence par séquence

Séquence par séquence, nous constatons que les enseignant.e.s rapportent un niveau de satisfaction plus élevé dans la condition G3 (Libre) dans 27 % des cas soit 6 séquences sur 22 au total. Nous observons aussi une tendance, dans le même sens mais non significative, dans 3 des séquences examinées. Il est également à noter que les enseignant.e.s rapportent un niveau de satisfaction plus élevé dans la condition G1 (Collectif non structuré) dans 1 séquence (i.e., Cohorte 2018 ASSP Séquence 4 de Mathématiques). Néanmoins pour cette séquence, nous ne pouvons pas affirmer avec certitude que ce sont les consignes fournies, à savoir un travail collectif structuré de type Puzzle, qui déterminent le niveau de satisfaction exprimé par les enseignant.e.s. En effet, cette dernière appartient à la catégorie des séquences jugées dégradées

pour lesquelles le respect des consignes est plutôt faible. Aucun effet ni tendance ne sont observés pour les 12 séquences restantes (pour plus de détails, voir Annexe R).

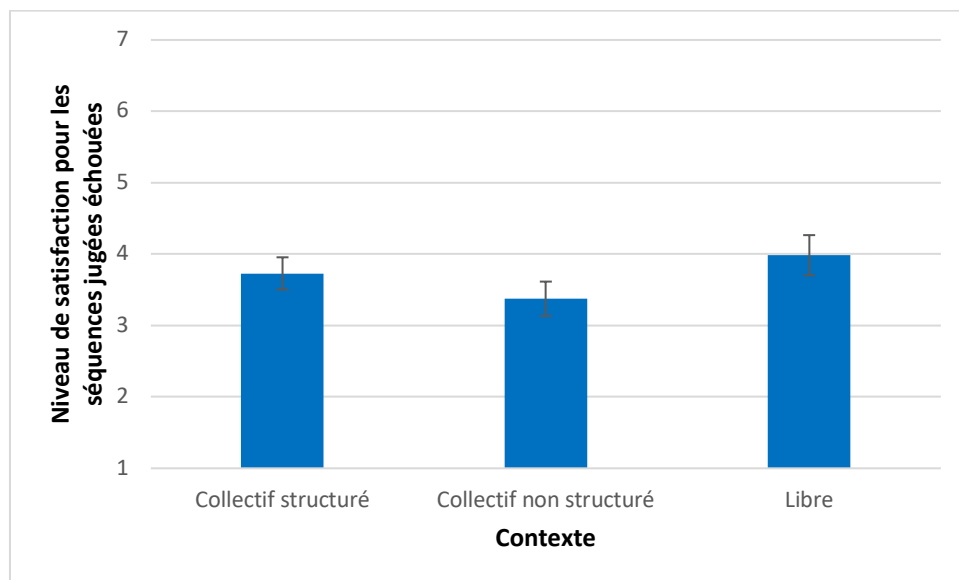
3.4 Séquences échouées

3.4.1 En moyenne

Nous n'observons pas de différence significative de la condition concernant le niveau de satisfaction exprimé par les enseignant·e·s pour les séquences jugées échouées ($F(2, 126) = 1.41, p = .248, \eta_p^2 = .02$; cf. Figure 19 ci-dessous).

Figure 19

Niveau de satisfaction exprimé par les enseignant·e·s pour les conditions Collectif structuré (G1), Collectif non structuré (G2) et Libre (G3) toutes cohortes et toutes filières confondues en Français et en Mathématiques pour les séquences jugées échouées



Note. Nombre de classes et de demi-classes inclus dans cette analyse : $N_{\text{Total}} = 126$, $N_{\text{Collectif Structuré}} = 50$, $N_{\text{Collectif non structuré}} = 44$, $N_{\text{Libre}} = 32$

3.4.2 Séquence par séquence

L'analyse fait apparaître que ce qui est vrai en moyenne pour les séquences jugées échouées l'est aussi séquence par séquence (absence d'effet significatif de la condition ; pour plus de détails, voir Annexe S).

4 En conclusion

Ainsi et conformément à nos hypothèses, nous constatons un niveau de satisfaction significativement réduit dans les séquences jugées plutôt réussies donc lorsque la liberté pédagogique (à laquelle les enseignants sont généralement attachés) est contrainte par les consignes qui ont été fournies dans les conditions G1 (Collectif Structuré) et G2 (Collectif non structuré). L'idée que cette contrainte puisse abaisser le niveau de satisfaction est compatible avec l'absence d'un effet des conditions G1 et G2 dans les séquences jugées échouées pour lesquelles les consignes n'ont pas été respectées et où la liberté pédagogique a donc été « reprise ». Dans les séquences jugées dégradées, le pattern de résultats est assez similaire à celui observé pour les séquences réussies, néanmoins il est impossible de déterminer avec certitude si l'effet de réduction du niveau de satisfaction (G1 et G2 relativement à G3, ou même G1 relativement à G2) correspond à une volonté délibérée des enseignants concernés de se soustraire au moins partiellement aux consignes ou à un éventuel sentiment d'échec face à la difficulté de les respecter.

Dans la suite de nos analyses, nous souhaitons donc non seulement apprécier dans quelle mesure les effets Puzzle et leur taille dépendent de la manière dont les consignes ont été implémentées dans chacun des groupes, mais aussi examiner le rapport entre la nature de ces effets et le niveau de satisfaction exprimé par les enseignants.

Chapitre 5 : Interdépendance positive, niveau de satisfaction des enseignant·e·s et performance des élèves

Notre principal objectif consiste à évaluer si le bénéfice attendu du mécanisme d'interdépendance positive s'exprime *a minima* dans la première catégorie de séquences, c'est-à-dire celles jugées plutôt réussies (les autres étant moins interprétables voire même ininterprétables) et cela quelles que soient les performances antérieures (auto-rapportées) des élèves, ou au contraire davantage chez ceux qui rapportent les performances les plus faibles. Comme nous l'avons développé dans la première partie du chapitre précédent, nous nous attendons à observer un effet d'interaction entre les conditions de travail et le niveau de performance antérieure rapporté par les élèves. Plus précisément, nous attendons dans la condition de travail « Collectif structuré » (G1) un écart réduit entre les élèves les plus faibles et ceux les plus forts, relativement à G2 et G3, en raison du mécanisme d'interdépendance positive. Ce mécanisme devrait, en principe, permettre aux élèves les plus en difficulté une valorisation de soi et un investissement plus élevé dans la séquence proposée, et par conséquent une meilleure performance. Dans la condition de travail « Collectif non structuré » (G2), nous attendons un écart maximal plus large (relativement à G1 et G3) en termes de performance entre les élèves les plus faibles et ceux les plus forts en raison de la combinaison de la paresse sociale des plus faibles (réduction des efforts personnels) et de l'effet éventuel de compensation chez leurs homologues les plus forts (augmentation de leurs efforts personnels).

Nous testerons également si l'hypothèse d'interaction détaillée précédemment dépend ou non du niveau de satisfaction des enseignant·e·s à l'égard des conditions de réalisation de leurs séquences pédagogiques, séquences que nous avons contraintes en G1 et G2 par nos consignes de travail (consignes susceptibles d'altérer une liberté pédagogique à laquelle les

enseignant·e·s sont généralement attaché·e·s, cf. chapitre 4). En effet, les résultats présentés au chapitre 4 de ce manuscrit suggèrent que, pour les séquences jugées réussies, les conditions de travail collectif (G1 et G2) réduisent de manière significative le niveau de satisfaction des enseignant·e·s. La question est de savoir si notre hypothèse d'interaction s'exprime quel que soit le niveau de satisfaction des enseignant·e·s ou exclusivement chez les élèves exposés à des enseignant·e·s satisfait·e·s d'appliquer les méthodes proposées.

Comme constaté précédemment (cf. Point 3 du chapitre précédent), le niveau de satisfaction exprimé par les enseignant·e·s étant influencé dans la majorité des séquences par les conditions de l'étude, nous avons choisi de ne pas réaliser le test d'hypothèse d'une interaction d'ordre 2 (conditions \times niveau initial auto-rapporté des élèves \times niveau de satisfaction de leurs enseignant·e·s) car cela poserait un problème d'endogénéité. Nous proposons en revanche deux estimations multiniveaux pour chaque test d'interaction d'ordre 1 (conditions \times niveau initial auto-rapporté des élèves), une estimation pour les élèves exposés à des enseignant·e·s caractérisé·e·s par un bas niveau de satisfaction (inférieur à la médiane de l'échantillon sur le niveau de satisfaction), et une estimation pour les élèves exposés à des enseignant·e·s caractérisé·e·s par un haut niveau de satisfaction (supérieur à la médiane de l'échantillon), et cela pour chaque séquence réussie considérée.

1 Précisions sur les analyses statistiques

Afin de tenir compte du caractère emboîté des données (i.e., des élèves dans des classes dans des établissements), nous avons réalisé des analyses multiniveaux à l'aide du logiciel SPSS (Version 26 ; IBM Corp., 2019) et en utilisant la Commande Mixed (IBM Corp., 2019). Nous avons respectivement analysé la performance des élèves en mathématiques et en français avec un modèle linéaire mixte incluant les classes et les lycées comme facteurs aléatoires. Ce modèle comprenait le niveau initial auto-rapporté, la condition ainsi que leur interaction en tant

que facteurs fixes. Outre les facteurs fixes, le modèle retenu incluait également des intercepts aléatoires au niveau des établissements et des classes, permettant ainsi de rendre compte de la variabilité entre établissements et classes. Pour des raisons de standardisation, nous avons choisi de conserver ces deux niveaux dans nos analyses quel que soit le pourcentage de variance associée (les résultats étant la plupart du temps les mêmes que l'on retire ou pas le niveau expliquant très peu de variance, généralement le niveau inter-établissement). Sauf dans le cas où le modèle rencontrait un problème de matrice de Hess en raison d'un nombre d'établissements considérés dans l'analyse trop faible. Dans ce cas, seul le niveau classe était retenu comme facteur aléatoire.

Ces analyses ont été réalisées sur la base de la catégorisation précédemment établie pour les séquences dont l'opérationnalisation a été jugée plutôt réussie, dégradée et échouée, pour chaque cohorte, chaque séquence, filière par filière en français et en mathématiques pour davantage de clarté. Néanmoins, et considérant que l'interaction entre l'induction expérimentale et les performances auto-rapportées des élèves serait extrêmement difficile à interpréter dans le cadre des séquences jugées « dégradées » et « échouées » (puisque hors consignes et peu interprétables), nous nous focaliserons sur sa présence éventuelle uniquement dans les séquences jugées « plutôt réussies ». Ces dernières seront présentées de la façon suivante : par cohorte, par filière et par séquence pour les mathématiques puis pour le français. Nous présenterons dans un premier temps les résultats rendant compte du pourcentage de temps déclaré dans les modalités d'organisation de classe pour la séquence considérée. Puis dans un second temps, nous présenterons les résultats qui correspondent à l'effet d'intérêt, soit l'effet principal de la condition, soit l'effet d'interaction entre la condition et le niveau initial auto-rapporté des élèves. Ces résultats seront présentés séparément pour les élèves exposés à des enseignant·e·s caractérisé·e·s par un bas niveau de satisfaction (inférieur à la médiane de l'échantillon sur le niveau de satisfaction) et pour ceux exposés à des enseignant·e·s

caractérisé-e-s par un haut niveau de satisfaction (supérieur à la médiane de l'échantillon sur le niveau de satisfaction). Pour plus de détails concernant les résultats des analyses multiniveaux pour les séquences jugées dégradées et échouées, voir Annexe W et X.

Nos hypothèses impliquant une variable censée nous renseigner sur le statut scolaire des élèves (i.e., niveau initial auto-rapporté), nous avons testé sa validité avant intégration à nos analyses multi-niveaux. En effet, si le niveau initial auto-rapporté par les élèves en français et en mathématiques renseigne correctement sur leur statut scolaire, alors nous devrions observer que ce niveau prédit significativement la performance aux tests standardisés administrés en fin de séquence et corrigés à l'aveugle (pour rappel : par un groupe d'enseignant-e-s indépendant-e-s de l'étude ProFan). Ce point très important a été vérifié *via* le modèle statistique multiniveaux décrit précédemment mais sans tenir compte des conditions de l'étude (« Collectif structuré », « Collectif non structuré » et « Libre ») à ce stade préalable à l'examen de notre hypothèse d'interaction. Ces calculs préliminaires montrent que le niveau initial prédit significativement la performance aux tests standardisés dans 68 séquences sur les 72 disponibles (cf. Annexe U). Cette observation atteste la validité du niveau auto-rapporté par les élèves que nous appellerons ci-après « niveau initial des élèves ».

2 Résultats principaux

Sur les 20 séquences jugées plutôt réussies, nous obtenons soit un effet principal soit un effet d'interaction dans 6 séquences portant sur le français ou les mathématiques. L'effet principal (2 séquences sur les 6) est observé pour la Cohorte 2018 ASSP Français Séquence 1 et la Cohorte 2017 MELEC Français Séquence 3. L'effet d'interaction (4 séquences sur les 6 et toutes concentrées sur les mathématiques Séquence 2 et 3) est observé pour la Cohorte 2018 ASSP Mathématiques Séquence 3, la Cohorte 2018 COMMERCE Mathématiques Séquence 2,

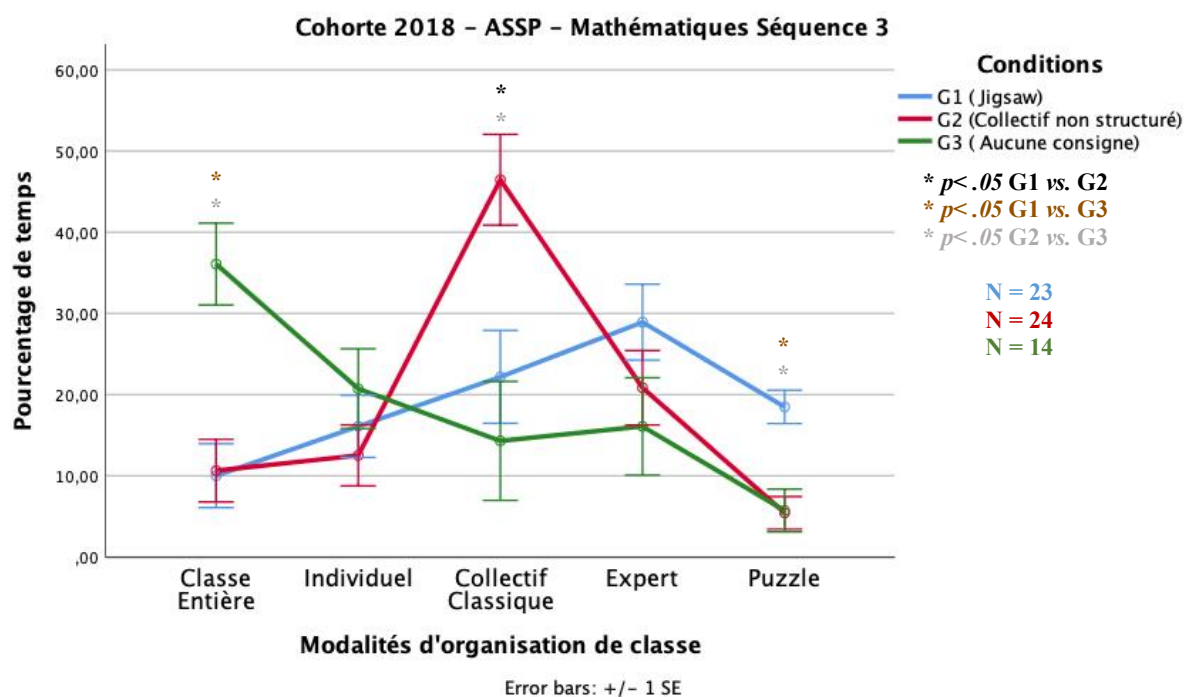
la Cohorte 2017 COMMERCE Mathématiques Séquence 3 et la Cohorte 2018 MELEC Mathématiques Séquence 3. Le détail de ces résultats est donné ci-dessous.

2.1 Cohorte 2018 ASSP Mathématiques Séquence 3

La Figure 20 ci-dessous rappelle le profil de réponses fourni par les enseignant·es à la grille d'observation.

Figure 20

Pourcentage de temps déclaré dans les modalités d'organisation de classe pour les conditions G1, G2 et G3 en ASSP pour la Séquence 3 de Mathématiques pour la cohorte 2018



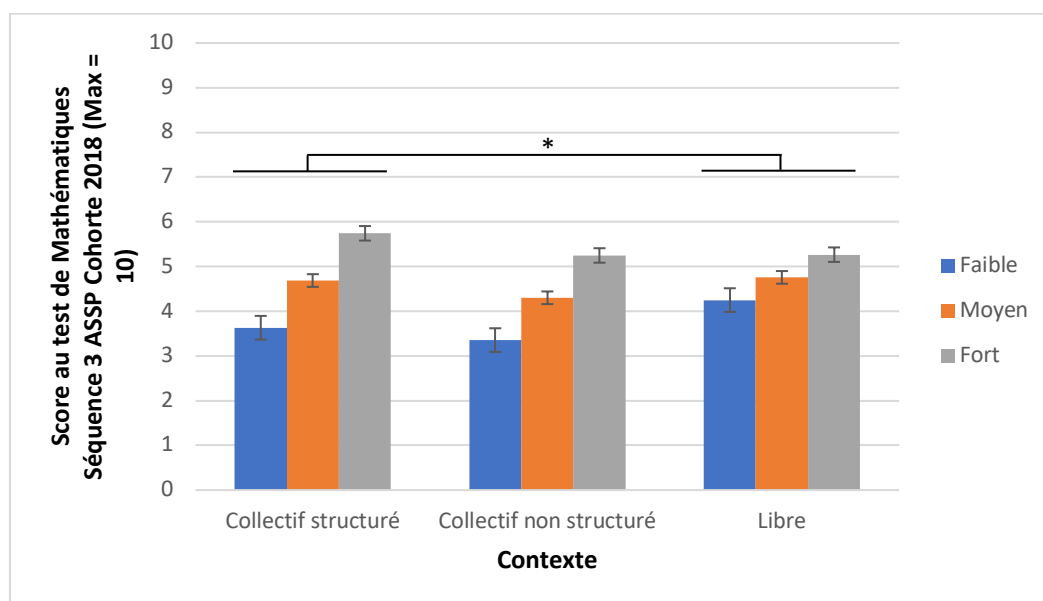
Note. G1 désigné par « Jigsaw » correspond à la condition « Collectif Structuré » et G3 désigné par « Aucune consigne » correspond à la condition « Libre ».

Chez les élèves exposés à des enseignant·es caractérisé·es par un haut niveau de satisfaction, l'analyse des résultats indique que l'effet principal du niveau initial est significatif, $F(1,368.700) = 66.100, p < .001$. Les élèves qui rapportent les notes les plus basses sont sans surprise ceux qui performent le moins sur le test standardisé administré en fin de séquence.

L'effet principal de la condition s'est révélé, quant à lui, non significatif, $F(2,23.250) = 0.342$, $p = .714$. En revanche, l'interaction entre la condition et le niveau initial des élèves est significative, $F(2,369.200) = 3.645$, $p = .027$. Les élèves dans la condition « Collectif structuré », particulièrement ceux au niveau initial élevé, produisent une performance plus élevée ($B = 0.549$, 95% IC = [0.097, 1.002], $p = .017$) que leurs homologues de même niveau dans la condition « Libre ». Il en résulte une hiérarchie entre les élèves plus marquée dans la condition « Collectif structuré » que dans la condition « Libre » (cf. Figure 21). Enfin, les élèves dans la condition « Collectif non structuré » produisent une performance non significativement différente de celles produites par leurs homologues des deux autres conditions, la condition "Collectif Structuré" ($B = -0.108$, 95% IC = [-1.001, -0.097], $p = .702$) et la condition "Libre" ($B = 0.441$, 95% IC = [-0.028, 0.909], $p = .065$) et cela quel que soit leur niveau initial.

Figure 21

Score au test de Mathématiques Séquence 3 en ASSP pour la cohorte 2018 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) et du niveau initial (Faible vs. Moyen vs. Fort) chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un haut niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

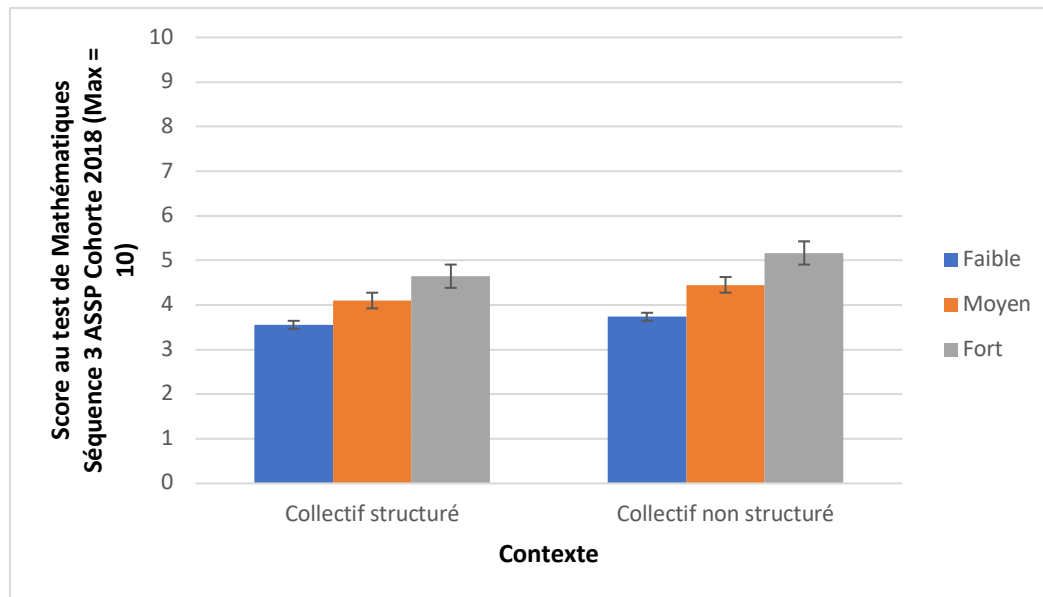


Note. * $p < .05$; ** $p < .01$; *** $p < .001$

Chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un bas niveau de satisfaction (cf. Figure 22), l'effet principal du niveau initial est significatif $F(1,213.300) = 25.310, p < .001$. Les élèves qui rapportent les notes les plus basses sont sans surprise ceux qui performent le moins sur le test standardisé administré en fin de séquence. En revanche, l'effet principal de la condition et l'effet d'interaction entre la condition et le niveau initial sont tous deux non significatifs (respectivement, $F(1,10.970) = 0.278, p = .608$ et $F(1,213.300) = 0.460, p = .498$).

Figure 22

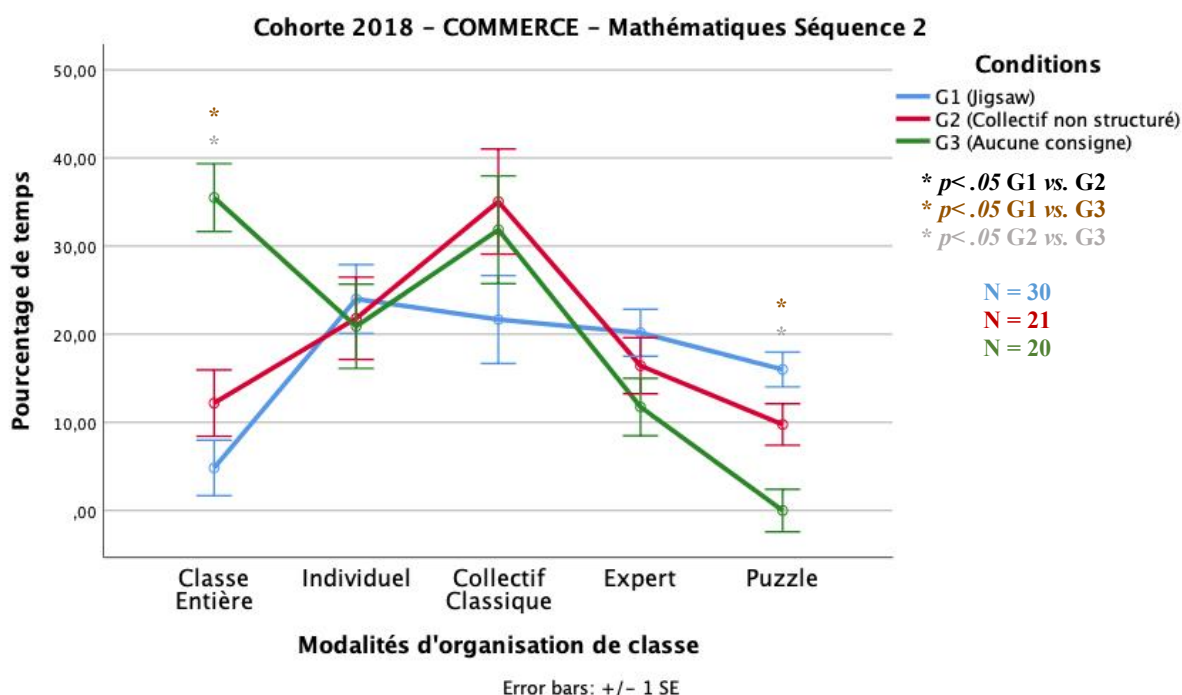
Score au test de Mathématiques Séquence 3 en ASSP pour la cohorte 2018 en fonction du contexte (G1, Collectif Structuré vs. G2 vs. Collectif non structuré) et du niveau initial (Faible vs. Moyen vs. Fort) chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un bas niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

**2.2 Cohorte 2018 COMMERCE Mathématiques Séquence 2**

La Figure 23 ci-dessous rappelle le profil de réponses fourni par les enseignant·e·s à la grille d'observation.

Figure 23

Pourcentage de temps déclaré dans les modalités d'organisation de classe pour les conditions G1, G2 et G3 en COMMERCE pour la Séquence 2 de Mathématiques pour la cohorte 2018

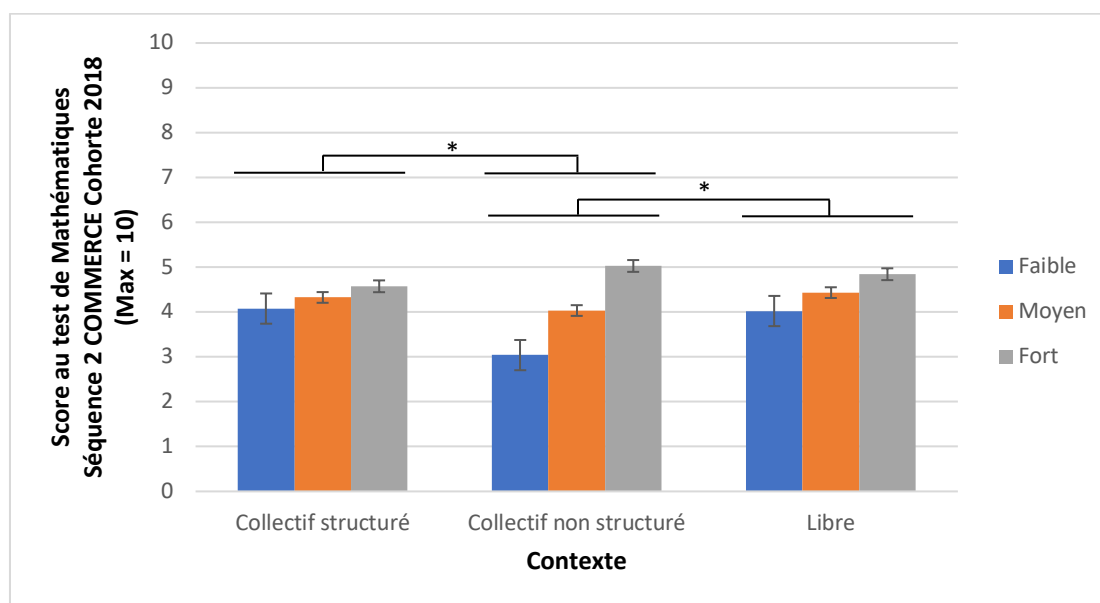


Note. G1 désigné par « Jigsaw » correspond à la condition « Collectif Structuré » et G3 désigné par « Aucune consigne » correspond à la condition « Libre ».

Chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un haut niveau de satisfaction, l'effet principal du niveau initial est significatif $F(1,519.709) = 19.700, p < .001$. Les élèves qui rapportent les notes les plus basses sont sans surprise ceux qui performant le moins sur le test standardisé administré en fin de séquence. L'effet principal de la condition est quant à lui non significatif $F(2,26.143) = 0.081, p = .922$. En revanche, l'interaction entre la condition et le niveau initial des élèves est significative $F(2,518.924) = 3.128, p = .045$. Elle montre que la hiérarchie entre les trois groupes d'élèves est amplifiée en condition « Collectif non structuré » ($B = 0.543, 95\% \text{ IC} = [0.071, 1.014], p = .024$) relativement aux deux autres conditions, « Collectif structuré » ($B = 0.697, 95\% \text{ IC} = [0.056, 1.337], p = .033$) et « Libre » ($B = -0.015, 95\% \text{ IC} = [-0.700, 0.392], p = .580$), voir Figure 24.

Figure 24

Score au test de Mathématiques Séquence 2 en COMMERCE pour la cohorte 2018 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) et du niveau initial (Faible vs. Moyen vs. Fort) chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un haut niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

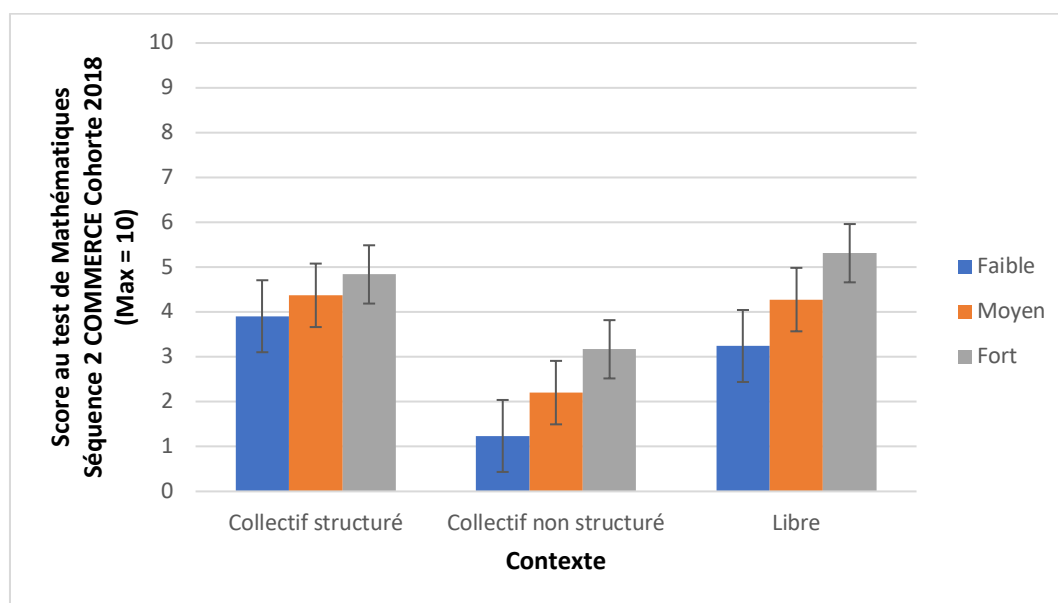


Note. * $p < .05$; ** $p < .01$; *** $p < .001$

Chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un bas niveau de satisfaction (cf. Figure 25), l'effet principal du niveau initial est significatif $F(1,288.804) = 23.657, p < .001$. Les élèves qui rapportent les notes les plus basses sont sans surprise ceux qui performent le moins sur le test standardisé administré en fin de séquence. En revanche, l'effet principal de la condition et l'effet d'interaction entre la condition et le niveau initial sont tous deux non significatifs (respectivement, $F(2,14.717) = 5.459, p = .117$ et $F(2,290.754) = 1.809, p = .166$).

Figure 25

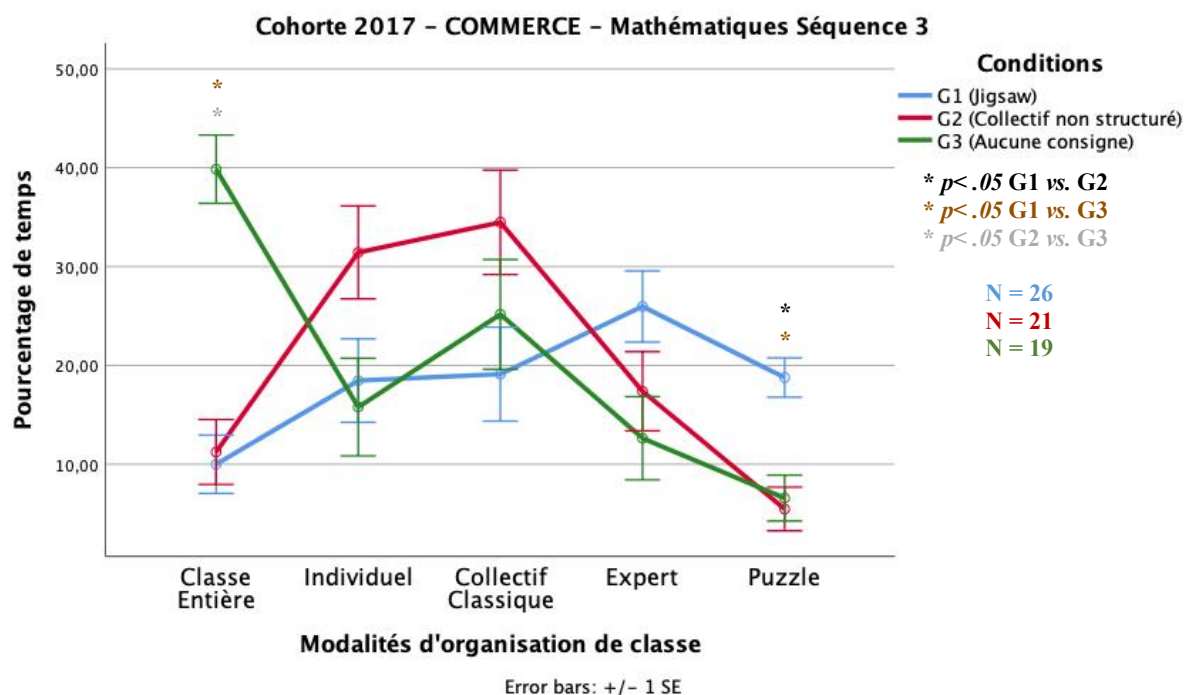
Score au test de Mathématiques Séquence 2 en COMMERCE pour la cohorte 2018 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) et du niveau initial (Faible vs. Moyen vs. Fort) chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un bas niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

**2.3 Cohorte 2017 COMMERCE Mathématiques Séquence 3**

La Figure 26 ci-dessous rappelle le profil de réponses fourni par les enseignant·e·s à la grille d'observation.

Figure 26

Pourcentage de temps déclaré dans les modalités d'organisation de classe pour les conditions G1, G2 et G3 en COMMERCE pour la Séquence 3 de Mathématiques pour la cohorte 2017



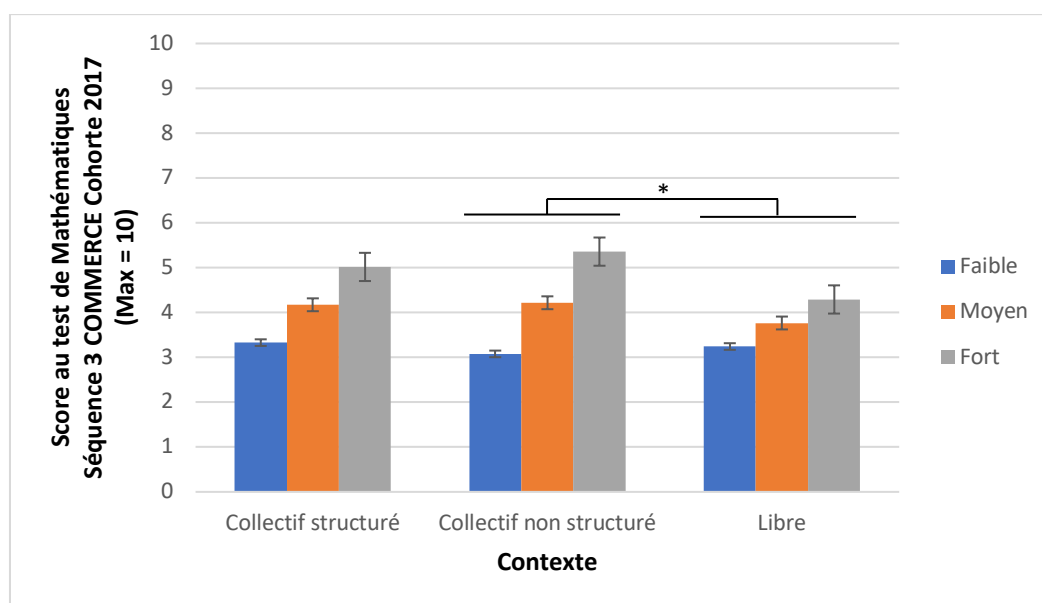
Note. G1 désigné par « Jigsaw » correspond à la condition « Collectif Structuré » et G3 désigné par « Aucune consigne » correspond à la condition « Libre ».

Chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un haut niveau de satisfaction, l'effet principal du niveau initial est significatif, $F(1,431.768) = 70.019, p < .001$. Les élèves qui rapportent les notes les plus basses sont sans surprise ceux qui performant le moins sur le test standardisé administré en fin de séquence. L'effet principal de la condition est quant à lui non significatif $F(2,30.359) = 0.607, p = .552$. En revanche, l'interaction entre la condition de travail et le niveau initial des élèves est significative $F(2,430.858) = 3.513, p = .031$. Les élèves dans la condition « Collectif non structuré », particulièrement les plus forts, produisent une performance plus élevée ($B = 0.543, 95\% \text{ IC} = [0.092, 0.994], p = .018$) que ceux dans la condition « Libre ». Cet effet différenciateur au profit des élèves les plus forts dans la condition « Collectif non structuré » s'estompe dans la condition « Libre » où nous constatons moins de différence entre les élèves selon leur niveau initial (cf. Figure 27). Les

élèves dans la condition « Collectif structuré » produisent quant à eux une performance non significativement différente de celles produites par leurs homologues dans la condition "Collectif non structuré " ($B = 0.258$, 95% IC = $[-0.221, 0.738]$, $p = .291$) et dans la condition "Libre" ($B = 0.284$, 95% IC = $[-0.092, 0.994]$, $p = .079$) et cela, quel que soit leur niveau initial.

Figure 27

Score au test de Mathématiques Séquence 3 en COMMERCE pour la cohorte 2017 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) et du niveau initial (Faible vs. Moyen vs. Fort) chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un haut niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne



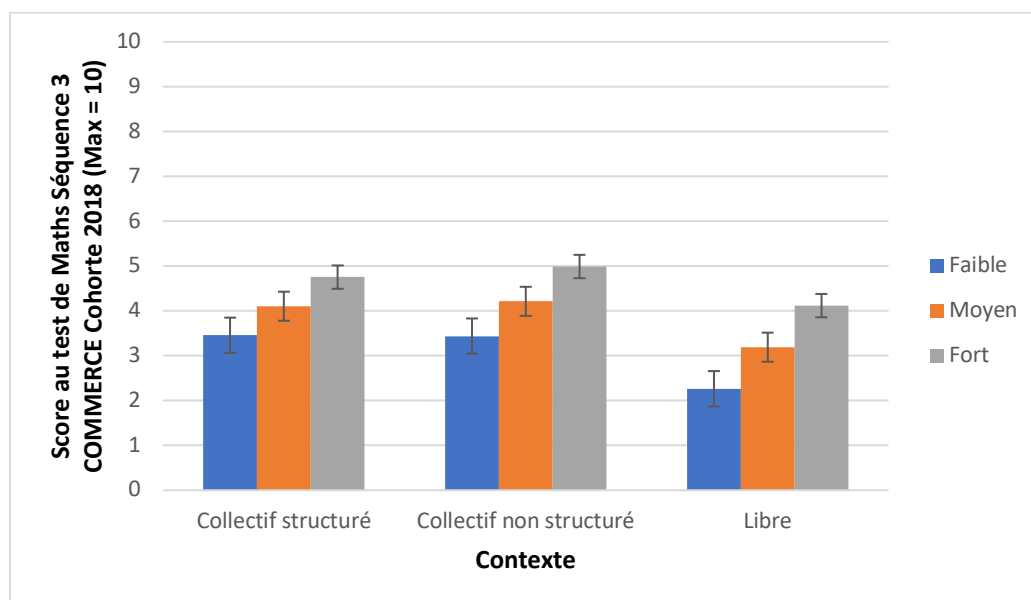
Note. * $p < .05$; ** $p < .01$; *** $p < .001$

Chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un bas niveau de satisfaction (cf. Figure 28), l'effet principal du niveau initial est significatif, $F(1,379.084) = 67.209$, $p < .001$. Les élèves qui rapportent les notes les plus basses sont sans surprise ceux qui performent le moins sur le test standardisé administré en fin de séquence. En revanche, l'effet principal de la condition et l'effet d'interaction entre la condition et le niveau initial sont

tous deux non significatifs (respectivement, $F(2,20.838) = 1.567, p = .232$ et $F(2,378.782) = 0.640, p = .528$).

Figure 28

Score au test de Mathématiques Séquence 3 en COMMERCE pour la cohorte 2017 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) et du niveau initial (Faible vs. Moyen vs. Fort) chez les élèves exposés à des enseignant·es caractérisé·es par un bas niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

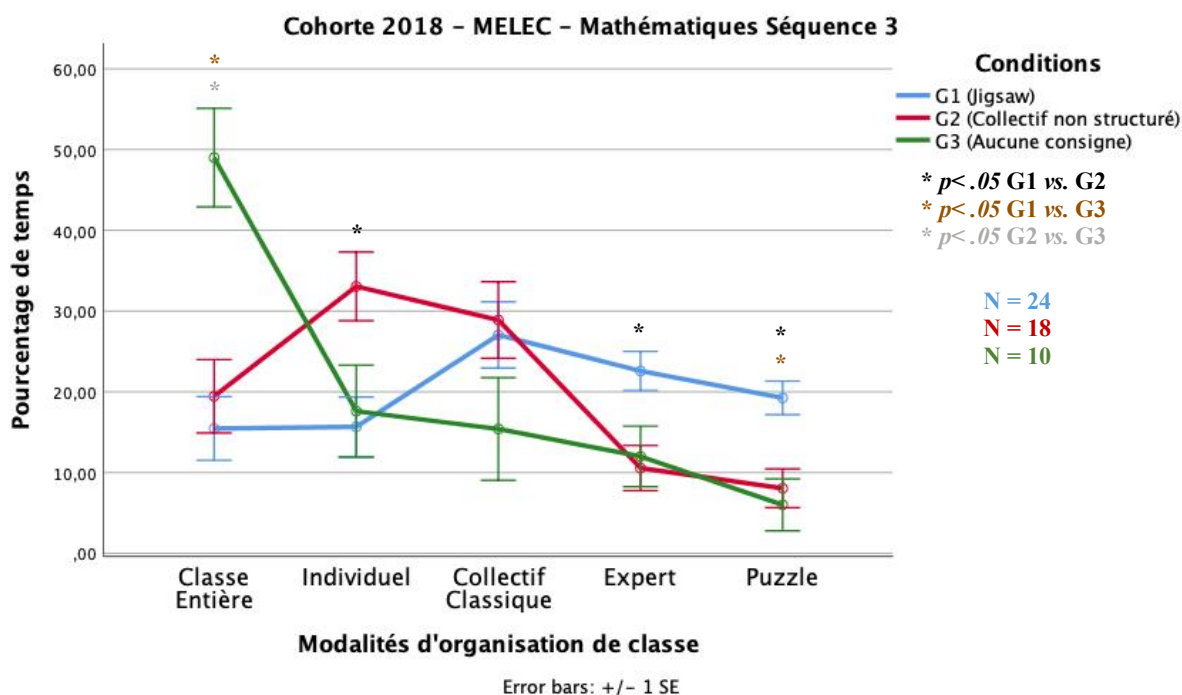


2.4 Cohorte 2018 MELEC Mathématiques Séquence 3

La Figure 29 ci-dessous rappelle le profil de réponses fourni par les enseignant·es à la grille d'observation.

Figure 29

Pourcentage de temps déclaré dans les modalités d'organisation de classe pour les conditions G1, G2 et G3 en MELEC pour la Séquence 3 de Mathématiques pour la cohorte 2018



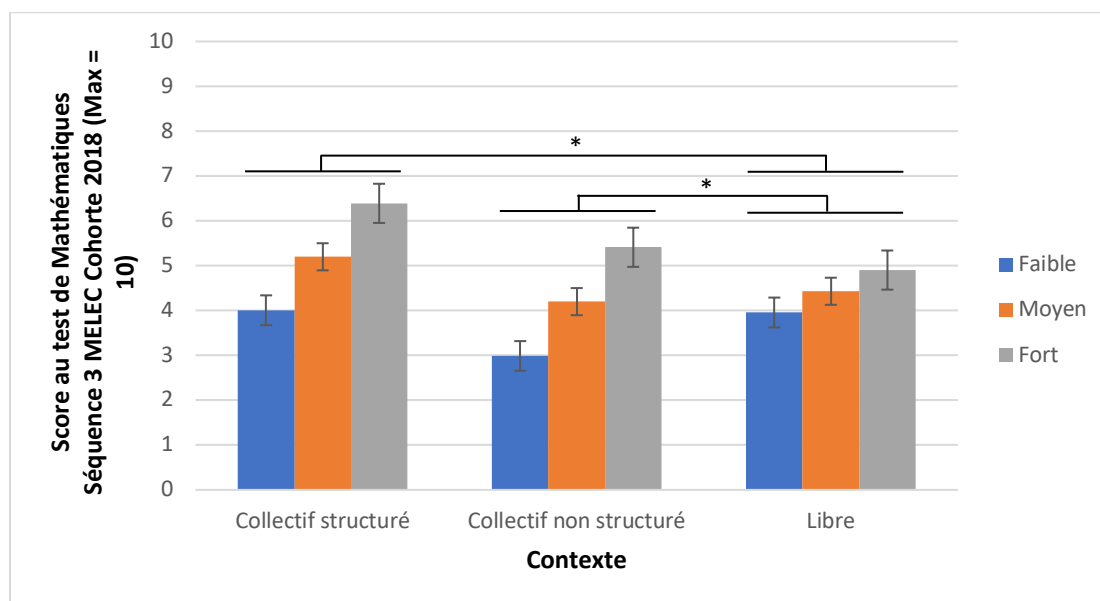
Note. G1 désigné par « Jigsaw » correspond à la condition « Collectif Structuré » et G3 désigné par « Aucune consigne » correspond à la condition « Libre ».

Chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un haut niveau de satisfaction, l'effet principal du niveau initial est significatif, $F(1,133.084) = 44.247, p < .001$. Les élèves qui rapportent les notes les plus basses sont sans surprise ceux qui performant le moins sur le test standardisé administré en fin de séquence. L'effet principal de la condition est quant à lui non significatif $F(2,5.306) = 1.270, p = .354$. En revanche, l'interaction entre la condition et le niveau initial des élèves est significative $F(2,132.896) = 3.119, p = .047$. Les élèves dans la condition « Collectif structuré » ($B = 0.699, 95\% \text{ IC} = [-0.001, 1.400], p = .050$) et « Collectif non structuré » ($B = 0.724, 95\% \text{ IC} = [0.090, 1.359], p = .026$), particulièrement ceux les plus forts, produisent une performance plus élevée que ceux dans la condition « Libre », voir Figure 30. L'effet différenciateur au profit des élèves les plus forts dans la condition « Collectif structuré » et « Collectif non structuré », s'estompe dans la

condition « Libre » où nous constatons moins de différence entre les élèves selon leur niveau initial. Enfin, les élèves dans la condition « Collectif structuré » produisent une performance non significativement différente de celle produite par leurs homologues dans la condition « Collectif non structuré » ($B = 0.025$, 95% IC = [-0.674, 0.725], $p = .943$)

Figure 30

Score au test de Mathématiques Séquence 3 en MELEC pour la cohorte 2018 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) et du niveau initial (Faible vs. Moyen vs. Fort) chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un haut niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne



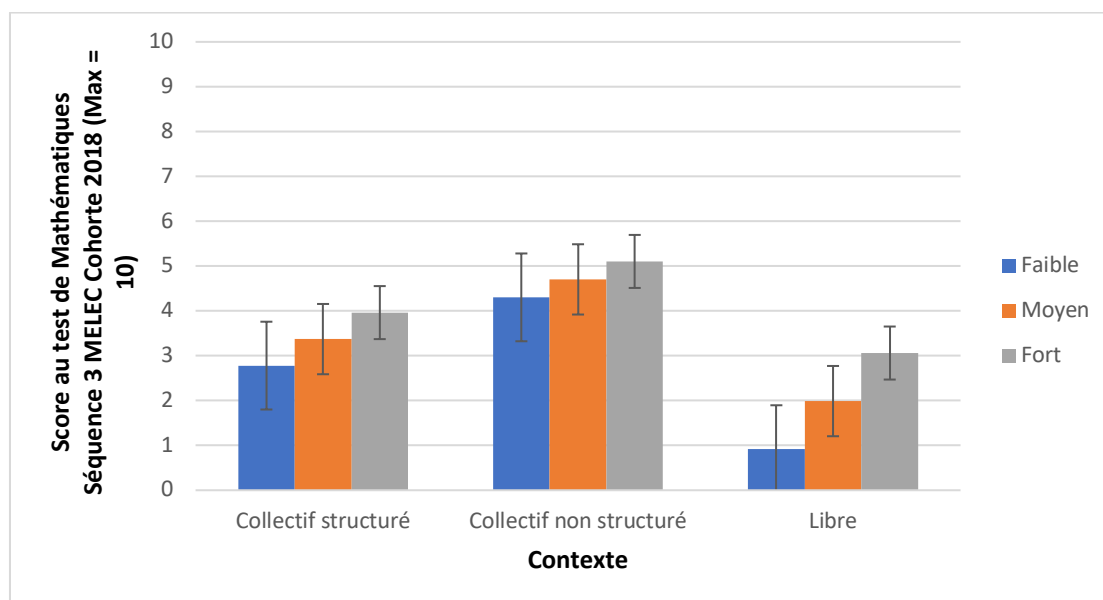
Note. * $p < .05$; ** $p < .01$; *** $p < .001$

Chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un bas niveau de satisfaction (cf. Figure 31), l'effet principal du niveau initial est significatif, $F(1,141.151) = 9.865$, $p < .001$. Les élèves qui rapportent les notes les plus basses sont sans surprise ceux qui performent le moins sur le test standardisé administré en fin de séquence. En revanche, l'effet principal de la condition et l'effet d'interaction entre la condition et le niveau initial sont

tous deux non significatifs (respectivement, $F(2,9.184) = 2.449, p = .140$ et $F(2,141.845) = 0.636, p = .531$).

Figure 31

Score au test de Mathématiques Séquence 3 en MELEC pour la cohorte 2018 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) et du niveau initial (Faible vs. Moyen vs. Fort) chez les élèves exposés à des enseignant·es caractérisé·es par un bas niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne.

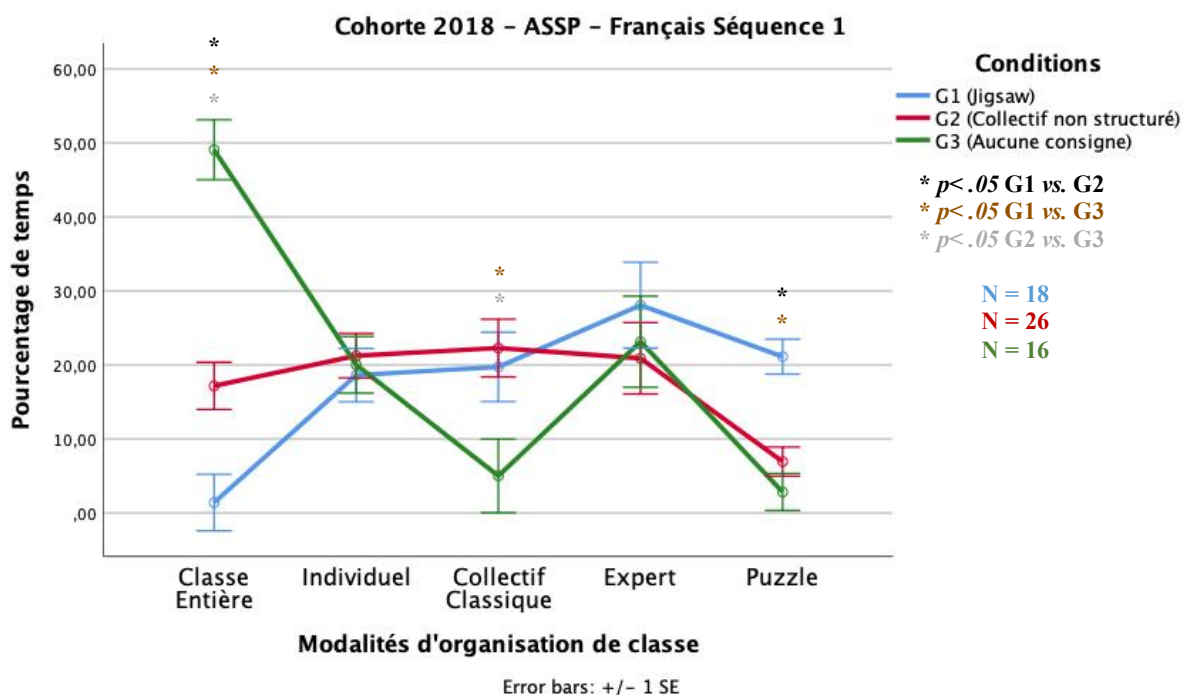


2.5 Cohorte 2018 ASSP Français Séquence 1

La Figure 32 ci-dessous rappelle le profil de réponses fourni par les enseignant·es à la grille d'observation.

Figure 32

Pourcentage de temps déclaré dans les modalités d'organisation de classe pour les conditions G1, G2 et G3 en ASSP pour la Séquence 1 de Français pour la cohorte 2018



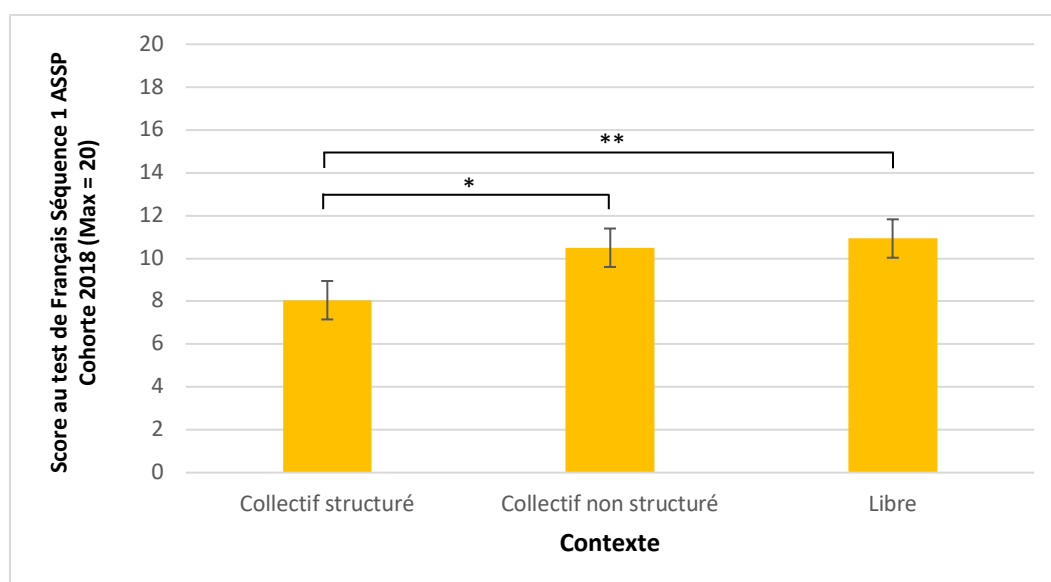
Note. G1 désigné par « Jigsaw » correspond à la condition « Collectif Structuré » et G3 désigné par « Aucune consigne » correspond à la condition « Libre ».

Chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un haut niveau de satisfaction, l'effet principal du niveau initial est non significatif, $F(1,523.212) = 0.044$, $p = .833$ (la seule observation de ce type sur les 6 séquences considérées dans ce chapitre rassemblant les séquences jugées réussies pour lesquelles nous obtenons des effets). En revanche, l'effet principal de la condition est significatif $F(2,21.012) = 4.463$, $p = .024$. Les élèves dans la condition « Collectif Structuré » produisent une performance moins élevée que ceux dans la condition « Collectif non structuré » ($B = 2.450$, 95% IC = [0.331, 4.569], $p = .026$) et « Libre » ($B = -2.876$, 95% IC = [-4.944, -0.808], $p = .009$). Les élèves dans la condition « Collectif non structuré » produisent quant à eux une performance non significativement différente de celle produite par leurs homologues dans la condition "Libre" (B

= -0.426, 95% IC = [-2.157, 1.303], $p = .615$) (cf. Figure 33). Enfin, l'effet d'interaction entre la condition et le niveau initial des élèves est non significatif, $F(2,521.919) = 1.632, p = .196$.

Figure 33

Score au test de Français Séquence 1 en ASSP pour la cohorte 2018 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) chez les élèves exposés à des enseignantes caractérisées par un haut niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

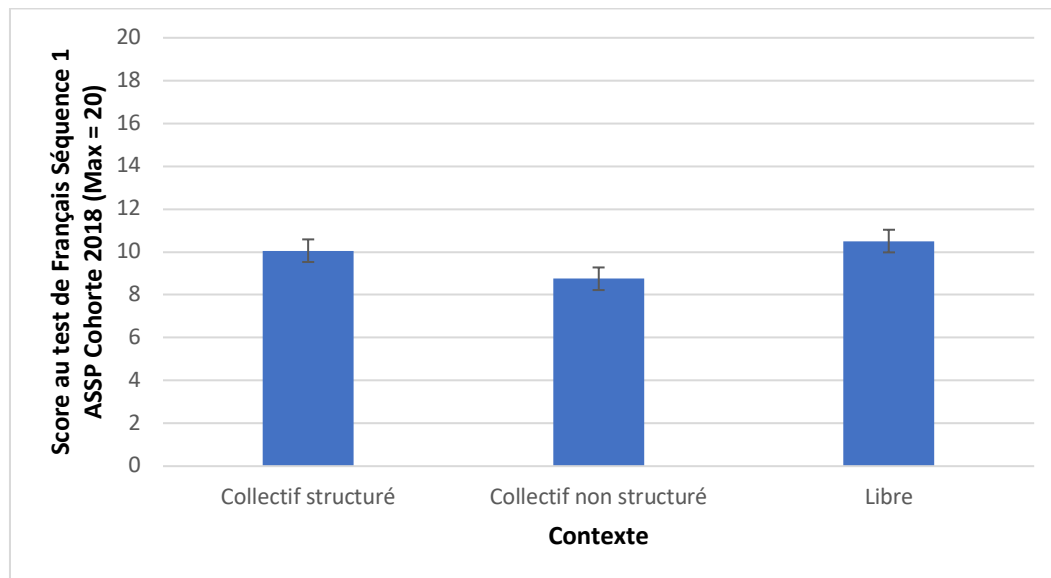


Note. * $p < .05$; ** $p < .01$; *** $p < .001$

Chez les élèves exposés à des enseignantes caractérisées par un bas niveau de satisfaction (cf. Figure 34), l'effet principal du niveau initial, de la condition et l'effet d'interaction entre la condition et le niveau initial sont tous trois non significatifs (respectivement, $F(1,385.142) = 2.010, p = .157$, $F(2,23.106) = 1.587, p = .227$ et $F(2, 384.654) = 0.248, p = .781$).

Figure 34

Score au test de Français Séquence 1 en ASSP pour la cohorte 2018 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) chez les élèves exposés à des enseignant·es caractérisés·es par un bas niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

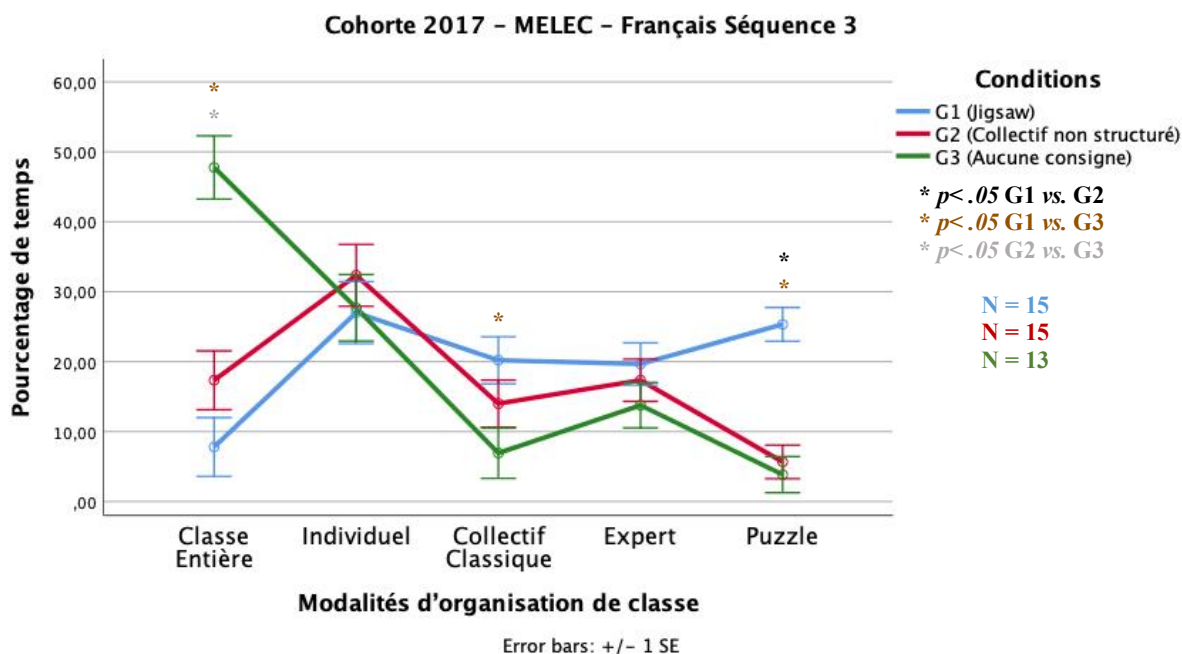


2.6 Cohorte 2017 MELEC Français Séquence 3

La Figure 35 ci-dessous rappelle le profil de réponses fourni par les enseignant·es à la grille d'observation.

Figure 35

Pourcentage de temps déclaré dans les modalités d'organisation de classe pour les conditions G1, G2 et G3 en MELEC pour la Séquence 3 de Français pour la cohorte 2017



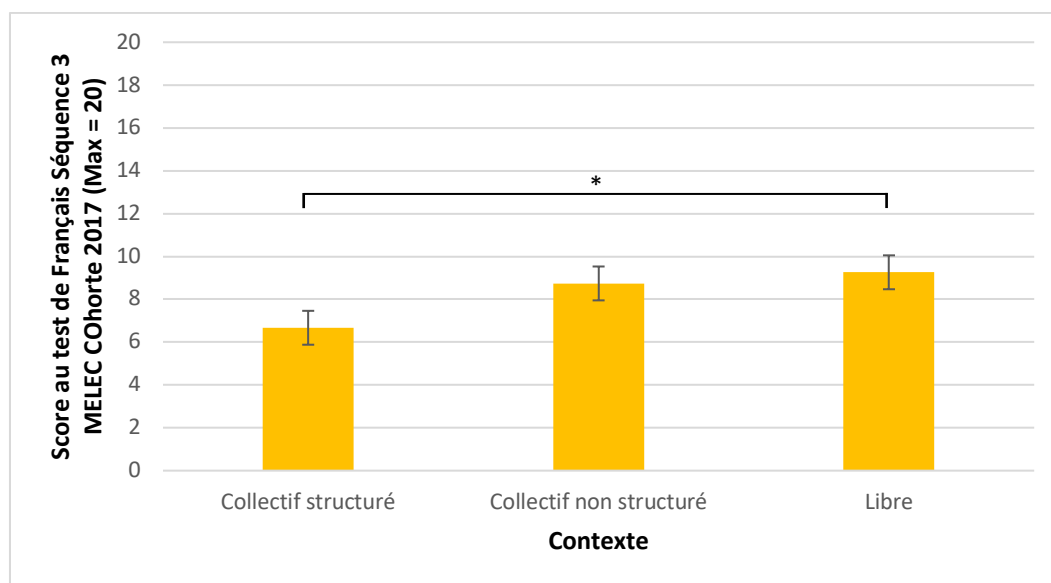
Note. G1 désigné par « Jigsaw » correspond à la condition « Collectif Structuré » et G3 désigné par « Aucune consigne » correspond à la condition « Libre ».

Chez les élèves exposés à des enseignant·es caractérisé·es par un haut niveau de satisfaction, l'effet principal du niveau initial est significatif, $F(1,280.817) = 5.536, p = .019$. Les élèves qui rapportent les notes les plus basses sont sans surprise ceux qui performant le moins sur le test standardisé administré en fin de séquence. L'effet principal de la condition est lui aussi significatif $F(2,18.991) = 4.214, p = .031$. Les élèves de la condition « Collectif Structuré » produisent une performance moins élevée ($B = -2.587, 95\% \text{ IC} = [-4.534, -0.639], p = .012$) que ceux dans la condition « Libre ». Les élèves de la condition « Collectif non structuré » produisent quant à eux une performance non significativement différente de celles produites par leurs homologues dans la condition "Collectif Structuré" ($B = 2.069, 95\% \text{ IC} = [-0.375, 4.513], p = .093$) et dans la condition "Libre" ($B = -0.517, 95\% \text{ IC} = [-3.014, 1.978], p$

= .670) (cf. Figure 36). En revanche, l'interaction entre la condition et le niveau initial des élèves est non significative, $F(2,280.869) = 1.809, p = .166$.

Figure 36

Score au test de Français Séquence 3 en MELEC pour la cohorte 2017 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) chez les élèves exposés à des enseignantes caractérisées par un haut niveau de satisfaction. Les barres d'erreurs représentent les erreurs-standards de la moyenne

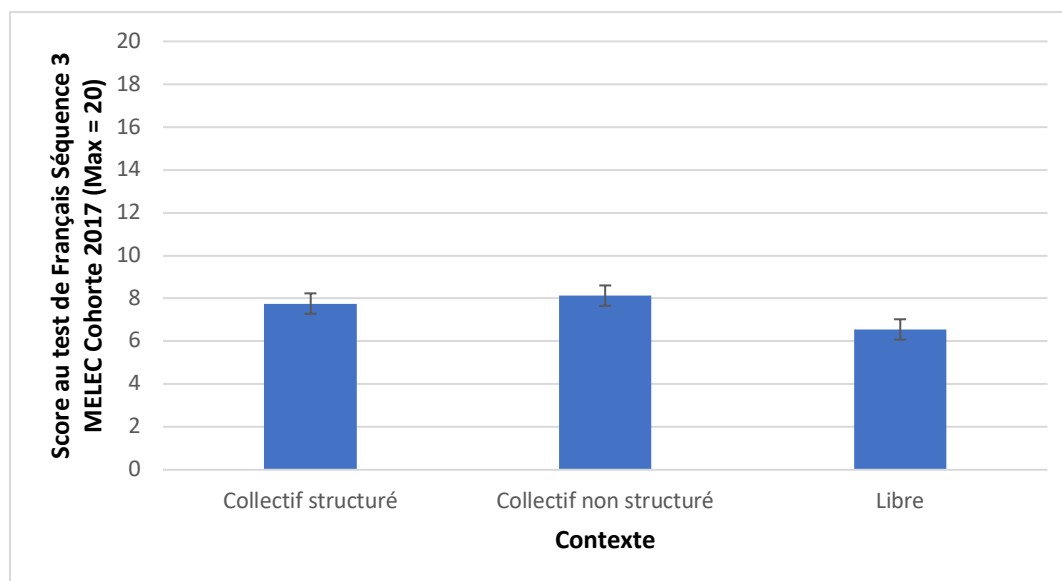


Note. * $p < .05$; ** $p < .01$; *** $p < .001$

Chez les élèves exposés à des enseignantes caractérisées par un bas niveau de satisfaction (cf. Figure 37), l'effet principal du niveau initial, de la condition et l'effet d'interaction entre la condition et le niveau initial sont tous trois non significatifs (respectivement, $F(1,190.476) = 0.759, p = .385$, $F(2,14.951) = 1.027, p = .382$ et $F(2, 190.239) = 1.807, p = .167$).

Figure 37

Score au test de Français Séquence 3 en MELEC pour la cohorte 2017 en fonction du contexte (G1, Collectif Structuré vs. G2, Collectif non structuré vs. G3, Libre) chez les élèves exposés à des enseignante:s caractérisé:s par un bas niveau de satisfaction.. Les barres d'erreurs représentent les erreurs-standards de la moyenne

**3 Résumé et discussion**

L'effet principal du niveau initial auto-rapporté sur le score aux tests standardisés est significatif dans 5 des 6 séquences pour lesquelles nous rapportons un effet principal de la condition ou d'interaction entre la condition et le niveau initial des élèves.

L'effet principal de la condition s'exprime quant à lui dans 2 séquences (Cohorte 2018 ASSP Français Séquence 1 et Cohorte 2017 MELEC Français Séquence 3) et exclusivement en français chez les élèves exposés à des enseignante:s caractérisé:s par un haut niveau de satisfaction. Dans ces 2 séquences et contrairement à notre hypothèse, les élèves produisent une performance plus élevée lorsque la liberté pédagogique de leurs enseignante:s n'est pas

contrainte par l'interdépendance positive de type Puzzle, autrement dit dans la condition « Collectif non structuré » et dans la condition « Libre » qui ne diffèrent pas l'une de l'autre.

Enfin, l'effet d'interaction entre la condition et le niveau initial des élèves s'exprime dans 4 séquences et exclusivement en mathématiques chez les élèves exposés à des enseignant·e·s caractérisé·e·s par un haut niveau de satisfaction. Contrairement à notre hypothèse suggérant un bénéfice de la condition "Collectif Structuré" (par rapport aux deux autres conditions) plus prononcé pour les élèves les plus faibles relativement à leurs homologues moins en difficulté qu'eux, les résultats observés dans 3 des 4 séquences soutiennent plutôt l'idée d'une amplification de la hiérarchie entre les trois groupes d'élèves lorsqu'ils sont placés dans une situation de travail collectif, qu'elle implique ou non de l'interdépendance positive. Cet effet, au profit des élèves les plus forts qui performant donc davantage en comparaison de leurs homologues plus en difficulté, s'estompe dans la condition « Libre », condition dans laquelle la liberté pédagogique des enseignant·e·s a été préservée. Nous observons également dans la séquence restante (i.e., Cohorte 2018 COMMERCE Mathématiques Séquence 2), et conformément à notre hypothèse, un effet de hiérarchie plus marqué entre les élèves dans la condition « Collectif non structuré » qui s'estompe dans la condition « Collectif Structuré », condition dans laquelle une relation d'interdépendance positive entre les élèves les plus en difficulté et leurs homologues les plus forts était plus probable. Bien que ce résultat semble indiquer un bénéfice de l'interdépendance positive de type Puzzle chez les élèves les plus en difficulté, il convient de rester prudent quant à son interprétation dans la mesure où nous ne l'observons que dans une seule séquence. Enfin, et conformément à notre hypothèse, nous constatons que l'effet d'interaction entre la condition et le niveau initial auto-rapporté par les élèves s'exprime exclusivement dans les séquences conduites par les enseignant·e·s se déclarant satisfait·e·s et donc par hypothèse plutôt en accord avec la méthode proposée.

Les effets significatifs décrit précédemment sont peu nombreux (pour rappel 6 effets au total sur les 20 séquences jugées plutôt réussies de français et de mathématiques). D'un point de vue méthodologique, ce nombre restreint de résultats ne peut pas être expliqué par des biais relatifs à la correction des copies puisque nous avons pris la précaution de soumettre les productions des élèves en français et en mathématiques à une évaluation externe qui était conduite à l'aveugle de l'induction expérimentale. De cette façon, nous avons pu garantir 1) l'absence de biais relatifs à la « demande expérimentale », consistant dans le cas présent pour les évaluateurs à surévaluer ou sous-évaluer plus ou moins consciemment les copies en provenance de tel ou tel groupe d'établissements (G1, G2 ou G3) et 2) l'absence de biais de genre consistant, là encore sans nécessairement en avoir l'intention, à sous-évaluer les copies des filles en mathématiques et des garçons en français (évaluations stéréotypiques). De la même façon, nous avons pris le soin de tester la crédibilité du niveau initial auto-rapporté par les élèves en français et en mathématiques avant de l'intégrer à nos analyses multi-niveaux afin de nous assurer de la fiabilité de cette variable en tant que « proxy » du statut scolaire des élèves.

En revanche, les visites réalisées par le groupement de chercheur·es début 2019 dans les établissements scolaires concernés par le dispositif ont fait émerger un certain nombre de difficultés qui pourraient expliquer le peu de résultats significatifs. En français, où seuls deux effets principaux sont significatifs, les enseignant·es ont évoqué des difficultés à piloter certaines séquences (particulièrement la Séquence 1 et la Séquence 3 qui sont celles pour lesquelles les deux effets ont été observés) et à faire travailler les élèves sur des contenus parfois jugés trop denses voire même inadaptés pour un public de lycée professionnel qui, selon les enseignant·es interrogé·es, présente de grandes difficultés en lecture et des lacunes en vocabulaire. Les enseignant·es ont par conséquent rapporté un manque de motivation de la part de leurs élèves et des effets négatifs en termes de compréhension et d'apprentissage. Cela les a par exemple poussé·es à opérer des remédiations pédagogiques lorsque les contenus étaient

jugés trop complexes, à donner plus de temps aux élèves pour lire individuellement les textes ou même dans certains cas (qui semblent cependant minoritaires) à prendre la décision d'attribuer une activité par groupe et non plus par élève.

En mathématiques où nous avons observé 4 effets d'interaction significatifs, les enseignant·e·s ont rapporté que les élèves étaient dans l'ensemble davantage impliqués malgré des difficultés relatives à la complexité des contenus sur la Séquence 3 (séquence sur laquelle se concentrent pourtant ces effets d'interaction) et sur la Séquence 4. Selon Aronson et Patnoe (2011), l'interdépendance positive de type classe Puzzle est un changement de paradigme qui nécessite du temps pour que les élèves apprennent à coopérer entre eux, ce qui semble bien être le cas puisque nos effets sont concentrés sur la cohorte 2018, c'est à dire la deuxième cohorte impliquée dans le dispositif et sur les séquences de fin de Première (Séquence 2) et de début de Terminale (Séquence 3). Néanmoins, la majorité de ces derniers (3 effets sur 4) ne vont pas dans le sens d'un bénéfice de la méthode Puzzle. Nous avons en effet au contraire observé une amplification de la hiérarchie entre les trois groupes d'élèves (Faible vs. Moyen vs. Fort) lorsque ces derniers étaient placés dans une situation de travail collectif, que cette dernière soit structurée ou non structurée. Ces observations majoritairement à l'encontre de nos hypothèses pourraient être expliquées par la difficulté que les enseignant·e·s ont rapporté s'agissant de la posture à adopter face aux élèves lorsqu'ils travaillaient en groupe. Un tel dispositif demande en effet une réorganisation des interventions en classe et contraint l'enseignant·e à une posture particulière dans laquelle il n'est plus le seul·e détenteur·ice du savoir à construire (Reverdy, 2020 ; Topping et al., 2017). En G1, certain·e·s enseignant·e·s ont même rapporté s'être senti·e·s bridé·e·s s'agissant des interventions qu'ils/elles pouvaient engager avec les élèves durant la séquence. Ces dernier·e·s avaient compris qu'ils n'avaient pas le droit d'intervenir et que les élèves devaient « *se débrouiller tout·e·s seul·e·s* » lors de l'étape des groupes experts. Or dans l'apprentissage coopératif, il n'est pas attendu de l'enseignant·e

qu'il/elle soit passive. Au contraire, son rôle est primordial puisqu'il/elle a pour mission d'accompagner les élèves pendant le travail coopératif, d'animer et de réguler les groupes, mais aussi de veiller à ce que toutes les élèves participent de manière équitable au travail coopératif (Plante, 2012 ; Reverdy, 2016).

La faible proportion de résultats significatifs pourrait aussi tenir compte de la méthode employée pour catégoriser les séquences pédagogiques. Pour rappel, les réponses fournies par les enseignant·es aux 5 items de la « grille d'observation » nous ont permis d'estimer la conduite des séquences pédagogiques sur le terrain mais également de catégoriser ces dernières selon trois catégories : les séquences 1) plutôt réussies (consignes plutôt respectées), 2) dégradées (respect plutôt faible des consignes) et 3), échouées (aucun respect des consignes voire même tendances inverses à celles attendues). Cette méthode pour laquelle nous avons opté ne traduit peut-être pas la réalité de leurs pratiques dans le cadre de l'étude ProFan. Outre des biais de mémoire (i.e., erreurs de rappel et/ou reconstruction des faits) ou de désirabilité sociale (i.e., affirmer un respect des consignes en dépit de pratiques qui en étaient éloignées), les incompréhensions évoquées plus haut s'agissant des consignes ont pu altérer les réponses fournies par les enseignant·es aux 5 items de la « grille d'observation » et par conséquent la catégorisation des séquences qui en découle. En effet, au-delà d'avoir été fortement sensibilisés à l'importance de respecter strictement les consignes fournies dans les groupes G1 et G2, les enseignant·es de G1 n'ont pas reçu de formation spécifique à la mise en œuvre d'une interdépendance positive de type classe Puzzle. Lors des entretiens certain·es enseignant·es ont d'ailleurs formulé le regret de ne pas avoir été formé à ce propos.

Ainsi, si l'importance accordée au climat positif n'est pas spécifique à la mise en place de l'interdépendance positive, il est tout de même essentiel afin que les élèves se sentent à l'aise et osent coopérer (Topping et al., 2017). C'est pourquoi il n'est pas étonnant d'avoir constaté que l'effet principal de la condition et l'effet d'interaction entre la condition et le niveau initial

auto-rapporté par les élèves se sont exclusivement exprimés dans les séquences conduites par des enseignant·e·s se déclarant satisfait·e·s et donc par hypothèse plutôt en accord avec la méthode proposée. Certain·e·s enseignant·e·s ont en effet rapporté avoir des regrets quant au fait de devoir faire travailler les élèves sur des contenus qu'ils/elles « *n'auraient jamais mis en place eux/elles-mêmes* » et selon un format qu'ils/elles « *n'auraient pas choisi* ». Certain·e·s rapportent même avoir eu le sentiment de perdre leur liberté pédagogique. En effet, et même si ils/elles étaient soucieux de bien appliquer les consignes fournies, cela les a parfois amené à se sentir comme de simples « exécutant·e·s », et par conséquent à réaliser les activités dans le cadre de l'étude ProFan avec moins d'enthousiasme que leurs cours habituels. Comme nous l'avons noté dans le Chapitre 4, nous avons aussi observé une diminution significative du niveau de satisfaction des enseignant·e·s dans les séquences jugées plutôt réussies lorsque leur liberté pédagogique était contrainte par les consignes fournies dans les conditions G1 (Collectif Structuré) et G2 (Collectif non structuré). Par conséquent, il est possible que le sentiment d'autonomie, en affectant le niveau de satisfaction des enseignant·e·s à l'égard du dispositif, ait déterminé au moins en partie leur efficacité pédagogique et donc les performances de leurs élèves. Ce raisonnement conduit cependant à attendre des performances en moyenne plus élevée en français et en mathématiques en G3, relativement aux deux autres conditions en particulier G1 qui est la plus contraignante pour les enseignant·e·s. Or comme nous l'avons noté antérieurement, cette supériorité n'est observée que dans 2 séquences concentrée sur l'enseignement de français. On ne peut donc pas dire que la restriction de la liberté pédagogique des enseignant·e·s détériore automatiquement la performance de leurs élèves.

Globalement et d'un point de vue organisationnel, les enseignant·e·s ont également rapporté un fort taux d'absentéisme des élèves (pouvant aller jusqu'à 50 % dans certains établissements selon les référent·e·s de l'étude sur le terrain) qui s'est inévitablement répercuté sur le dispositif, les taux de réponses et de participation, ainsi que sur la gestion des groupes et leur capacité à

collaborer. Ainsi lorsqu'il y avait des absences, les synthèses en groupe puzzle étaient incomplètes puisqu'il manquait des expertises. Afin de contourner ce problème, soit un élève experte était désigné-e afin de se rendre dans les différents groupes puzzle ce qui engendrait éventuellement un sentiment d'injustice quant au fait d'avoir une charge de travail supplémentaire, soit l'enseignante prenait la place de l'élève experte manquante, déséquilibrant nécessairement les dynamiques à l'intérieur des groupes. Le dispositif semble ainsi avoir été parfois vécu par les lycéen·n·e·s comme un évènement contraignant, imposé et secondaire à leur scolarité, expliquant probablement le peu de résultats observés en français et en mathématiques.

Enfin, les enseignant·e·s ont évoqué des difficultés relatives au fait de terminer leurs séquences dans les temps, notamment à cause de la longueur et de la difficulté de certaines activités, mais aussi à cause d'autres problèmes tels que l'agencement des calendriers, la difficulté à lancer le travail de groupe et/ou à récupérer certains documents sur la plateforme support de l'expérimentation (liens morts, problèmes de réseau, mots de passe perdus, etc.). Ces difficultés qui ont pu subsister dans certains établissements semblent toutefois s'être lissées à partir de la deuxième année. Les enseignant·e·s se sont en effet accordé·e·s à dire que par rapport à la première année d'expérimentation, certaines améliorations avaient été apportées facilitant l'organisation et la compréhension des consignes, ce qui explique probablement pourquoi la majorité de nos effets sont concentrés sur la cohorte 2018 (4 séquences sur les 6).

Chapitre 6 : Discussion générale

Comme nous l'avons évoqué dans l'introduction générale, la transition numérique entraîne des modifications comportementales profondes dans le monde du travail. Le changement dans les relations hiérarchiques traditionnelles au sein des organisations remet en question les modes classiques de communication et de collaboration. En favorisant les échanges en réseau, le numérique invite à davantage de polyvalence et des compétences à interagir. Pourvoir les futures générations d'un répertoire de compétences qui leur permettront de faire face aux transformations du travail dans l'économie du futur apparaît alors comme un enjeu éducatif majeur. Aussi, et même si les salles de classe ont traditionnellement été organisées autour d'un apprentissage individuel, il apparaît aujourd'hui nécessaire d'offrir aux élèves la possibilité de coopérer et de travailler en collaboration.

Cependant, et même si le recours à l'apprentissage par petits groupes d'élèves est une modalité pédagogique fortement encouragée par le Ministère de l'Éducation Nationale (Article L111-1 du Code de l'éducation datant du 24 Août 2021), elle semble peu répandue dans les écoles françaises (OCDE, 2018). Son implémentation représente en effet un défi aussi bien pour les enseignant·e·s que pour les élèves (e.g., Baloche & Brody, 2017 ; Buchs et al., 2017 ; Volpé & Buchs, 2019). Les méthodes pédagogiques par le biais desquelles les enseignant·e·s réunissent les élèves en petits groupes afin qu'ils travaillent de façon conjointe en vue d'atteindre un objectif commun, sont désignées sous le terme « d'apprentissage coopératif » (Slavin, 2011). Issus de la Psychologie et des Sciences de l'éducation, les travaux sur l'apprentissage coopératif sont consignés dans plusieurs centaines d'études réalisées de l'école primaire à l'université, qui attestent des bénéfices sociaux, motivationnels et cognitifs des

travaux de groupes coopératifs (Hattie, 2009 ; Johnson & Johnson, 2009 ; Johnson et al., 2000 ; Kyndt et al., 2013).

Une méthode fréquemment évoquée par la littérature scientifique concernant l'apprentissage coopératif est la méthode de la « classe Puzzle » (*Jigsaw classroom*, Aronson et al., 1976 ; 1978). Développée à la suite de la désagrégation des écoles aux États-Unis dans le but de favoriser l'intégration des minorités ethniques, cette méthode repose en principe sur la mise en place d'un mécanisme d'interdépendance positive au sens des ressources. Ce mécanisme, qui a fait l'objet de notre rapport de thèse, suppose 1) que chaque membre du groupe dispose d'une partie du matériel à apprendre et 2) que l'accès à son intégralité exige l'articulation des différentes parties. Ainsi, c'est la coordination du travail de chacun des membres du groupe qui devrait les amener à mettre en œuvre des interactions facilitatrices (e.g., entraide, explication, coopération, etc.), afin de parvenir à un meilleur apprentissage (Johnson & Johnson, 2002).

1 La littérature Puzzle et ses faiblesses

Malgré sa popularité (Topping et al., 2017), les fondements théoriques et même empiriques de la classe Puzzle suscitent aujourd'hui de nombreuses interrogations (e.g., Roseth et al., 2018, Stanczack et al., 2022). C'est pourquoi dans le premier chapitre de cette thèse, nous avons présenté une synthèse narrative de la littérature consacrée à la classe Puzzle dans le but d'identifier précisément ses forces et faiblesses. Cette synthèse, adossée à la méthode PRISMA (Liberati et al., 2009), a fait apparaître une incohérence importante entre l'efficacité supposée de la méthode Puzzle et la qualité des preuves empiriques à l'appui de cette idée. Notre synthèse (cf. chapitre 1) a révélé que les travaux au sujet de cette méthode étaient en réalité peu nombreux, majoritairement focalisés sur la question des performances académiques plutôt que sur le climat de classe ou d'autres variables fréquemment invoquées dans la littérature

spécialisée, comme par exemple l'estime de soi, et ce, chez les étudiant-e-s de l'enseignement supérieur plutôt que chez les élèves plus jeunes comme ceux des premières études réalisées par Aronson. Mais surtout, cette synthèse suggère que ces travaux présentent de nombreuses limites méthodologiques, rendant difficile, voire impossible, toute conclusion ferme s'agissant de l'efficacité de la méthode Puzzle.

Dans le but d'évaluer précisément les soutiens empiriques à cette méthode, nous avons dans le chapitre 2 réalisé une synthèse quantitative des effets de la classe Puzzle sur les performances académiques, la seule variable dépendante véritablement exploitable car très majoritaire. Notre approche n'a pas consisté en une méta-analyse quantitative au sens strict du terme, mais plutôt à requalifier l'intérêt méthodologique des études publiées, pour conclure *in fine* à l'efficacité de la méthode Puzzle. Or, ce travail plus qualitatif est généralement absent des méta-analyses quantitatives, ces dernières pouvant donc aboutir à des conclusions à l'aveugle de la qualité méthodologique des travaux publiés et sélectionnés dans ce cadre. Cependant, le fait que ces derniers soient publiés ne garantit absolument pas qu'ils réunissent des conditions méthodologiques et statistiques satisfaisantes, et cela est bien illustré par les travaux réalisés sur la méthode Puzzle. Dans ce but, nous avons dans un premier temps calculé sur cette variable la taille de l'effet Puzzle à partir des paramètres disponibles dans les études sélectionnées dans notre synthèse. Dans un second temps, nous avons construit un indice de « qualité méthodologique » sur la base de 6 critères qui nous a permis de classer les études sélectionnées selon 5 degrés de qualité méthodologique : « Médiocre », « Plutôt faible », « Intermédiaire », « Plutôt Forte » et « Excellente ». A l'issue de cet examen, nous avons constaté que ces études présentaient de nombreuses faiblesses méthodologiques telles que des échantillons de petite taille, une faiblesse au niveau des groupes contrôles ou encore une absence d'informations au sujet du statut scolaire des élèves permettant ensuite de voir si l'efficacité de la méthode Puzzle s'exprime plus pour certain-e-s élèves que pour d'autres. Aussi,

nous avons constaté que très peu d'études étaient satisfaisantes et que celles jugées les plus faibles présentent des tailles d'effets anormalement élevées ($g > .80$). Les deux études jugées les plus solides d'un point de vue méthodologique présentent tout de même une taille d'effet moyenne ($M_g \text{ de Hedge} = 0.35$), une valeur plus proche des estimations habituelles des effets attribués à l'apprentissage coopératif, qu'il s'agisse de la méthode Puzzle ou d'autres méthodes (Hattie, 2009). Ces résultats, tout en invitant à la prudence s'agissant des effets de la méthode Puzzle sur les performances académiques, suggéraient tout de même de poursuivre les travaux à l'aide de dispositifs plus ambitieux que les précédents, s'agissant en particulier de la puissance statistique de test ou encore des standards de qualité méthodologique à réunir.

2 ProFan : une étude plus ambitieuse

L'étude ProFan, dont la méthodologie a été décrite dans le troisième chapitre, permettait précisément de tester l'efficacité de la méthode Puzzle dans des conditions méthodologiques et statistiques plus satisfaisantes que celles employées dans les études réalisées jusqu'ici. Déployé de Septembre 2017 à Juin 2020 dans 109 lycées professionnels de dix académies couvrant 5 régions de France métropolitaine, le dispositif ProFan a impliqué 10 163 élèves issus de trois filières de formation (ASSP, COMMERCE et MELEC), 1263 enseignant·e·s et un groupement de chercheur·e·s (15 statutaires, 3 post-doctorant·e·s et 6 doctorant·e·s) en interaction permanente avec des inspecteurs et inspectrices de l'éducation nationale spécialistes de la voie professionnelle. Cette étude a été réalisée sur un ensemble de 3 groupes d'établissements distincts répondant à des modes d'organisation pédagogiques spécifiques dans les enseignements de français, de mathématiques et professionnels selon un calendrier commun (72 séquences pédagogiques au total), en classe de première et en classe de terminale selon un suivi longitudinal sur 2 promotions.

Les variables dépendantes testées dans ce cadre étaient de trois types. Premièrement, les performances individuelles des élèves (tests standardisés construits pour l'étude) sur des

contenus développés dans les séquences pédagogiques et impliquant ou non un travail collectif (l'observable central dans nos travaux). Nous avons d'ailleurs pris la précaution de soumettre les productions des élèves en français et en mathématiques à une évaluation externe qui était conduite à l'aveugle de l'induction expérimentale. De cette façon, nous avons pu garantir 1) l'absence de biais relatifs à la « demande expérimentale », consistant dans le cas présent pour les évaluateurs à surévaluer ou sous-évaluer plus ou moins consciemment les copies en provenance de tel ou tel groupe d'établissements (G1, G2 ou G3) et 2) l'absence de biais de genre consistant, là encore sans nécessairement en avoir l'intention, à sous-évaluer les copies des filles en mathématiques et des garçons en français (évaluations stéréotypiques). Deuxièmement, les performances et autres comportements des élèves en situation de résolution collectives de problèmes non issus des séquences pédagogiques (en référence à ce que nous avons nommé « la boîte à outils » ou BAO). Et enfin troisièmement, la réponse des élèves à plusieurs questionnaires permettant de sonder leurs représentations et auto-évaluations dans différents domaines (e.g., perception de compétences, estime de soi, etc. ; cf. Annexes G, H, I et J) et à différents moments de l'étude (pour en saisir l'évolution dans le temps et en relation avec les modalités de travail considérées). Dans notre rapport de thèse, nous avons décidé de concentrer nos efforts sur les variables suivantes. Premièrement, les performances individuelles des élèves en français et en mathématiques puisque cette variable est très majoritaire dans les travaux consacrés à la classe Puzzle (cf. Chapitre 1). Deuxièmement, le niveau initial auto-rapporté par les élèves en français et en mathématiques, dont nous avons testé la fiabilité en tant que « proxy » du statut scolaire avant de l'intégrer à nos analyses (cf. Chapitre 5). Troisièmement, les réponses des enseignant·e·s à différents items de la grille d'observation post-séquence pédagogique (cf. Chapitre 3) relatives à la conduite des séquences et à leur satisfaction/insatisfaction à l'égard des consignes fournies pour l'étude.

Ce dispositif, d'une ampleur inédite dans le champ des travaux consacrés à l'évaluation de l'interdépendance positive au sens de la classe Puzzle, a également pu être apprécié à la lumière de la quantité de données produites et centralisées sur la plateforme numérique support de l'étude, soit plusieurs millions au total. Cette action, ainsi que les précautions méthodologiques qui ont été prises, devaient en principe créer les conditions favorables à l'expression de l'effet de l'interdépendance positive au sens de la classe Puzzle, si tenté qu'il existe.

3 Les résultats du dispositif ProFan

Les quatrième et cinquième chapitres étaient quant eux consacrés à la présentation des résultats du dispositif ProFan. Dans le quatrième chapitre, nous avons d'abord estimé la qualité de l'opérationnalisation des consignes fournies aux enseignant·e·s pour la réalisation de leurs séquences pédagogiques. Ce point consacré à la qualité méthodologique des travaux déployés sur le terrain, qui constitue le fil rouge de notre rapport de thèse, est fondé au moins en partie sur notre synthèse de la littérature spécialisée qui a en effet montré que les bénéfices attendus de l'interdépendance positive au sens de la classe Puzzle sont étroitement liés à la qualité de l'implémentation des dispositifs (cf. Chapitres 1 et 2). Encore une fois, nous ne pouvons pas conclure sur l'efficacité d'une méthode sans évaluer précisément les qualités méthodologiques des travaux censés la valider. C'était le sens des travaux examinés au cours du chapitre 2, et c'est aussi ce que nous avons cherché à faire dans le cadre du dispositif ProFan.

Ainsi avons-nous calculé, pour les 72 séquences pédagogiques ProFan, la tendance centrale de réponse des enseignant·e·s sur chacun des 5 items de la grille d'observation à l'issue de chaque séquence, dans le but d'estimer la plus ou moins grande conformité des profils de réponse obtenus avec les profils attendus en relation avec les consignes fournies pour l'étude. Sur la base de ces profils de réponse, nous avons catégorisé les séquences pédagogiques selon trois catégories les séquences : 1) plutôt réussies (consignes plutôt respectées), 2) dégradées

(respect plutôt faible des consignes) et 3), échouées (aucun respect des consignes voire même tendances inverses à celles attendues). Néanmoins, nous avons constaté que la catégorie dominante toutes cohortes, toutes filières et toutes matières prises dans leur ensemble était celle des séquences « dégradées » ($N_{\text{Cohorte 2017}} = 14$ séquences, $N_{\text{Cohorte 2018}} = 18$ séquences ; Séquences « plutôt réussies » : $N_{\text{Cohorte 2017}} = 13$, $N_{\text{Cohorte 2018}} = 8$; Séquences « échouées » : $N_{\text{Cohorte 2017}} = 9$, $N_{\text{Cohorte 2018}} = 10$).

Dans un deuxième temps, nous nous sommes intéressés aux réponses des enseignant·e·s aux items de la grille d'observation qui concernaient leur satisfaction/insatisfaction à l'égard des consignes fournies pour l'étude. Ces consignes, de par leur nature, réduisaient presque nécessairement la liberté pédagogique des enseignant·e·s. En effet, les consignes invitant à un travail de nature collectif ne sont pas des modalités pédagogiques dominantes dans l'institution scolaire traditionnelle. À ce titre, elles étaient susceptibles de réduire la liberté pédagogique des enseignant·e·s qui ne pratiquent pas ou peu une pédagogie fondée sur l'apprentissage coopératif quelle que soit d'ailleurs sa nature. Cet élément est particulièrement important dans le cadre de nos travaux car, comme nous l'avons développé dans le chapitre 4, la littérature scientifique fait état de plusieurs études démontrant l'existence d'un lien étroit entre l'autonomie dont peuvent bénéficier les enseignant·e·s dans la réalisation de leurs activités pédagogiques, et leur satisfaction professionnelle (e.g., Avanzi et al., 2013 ; Humphrey, 2007 ; Kengatharan, 2020 ; Koustelios et al., 2004 ; Skaalvik & Skaalvik, 2009 ; Skaalvik & Skaalvik, 2010). Considérant cette littérature, nous nous attendions à observer un effet négatif des consignes de travail collectif (G1 et G2) sur le sentiment de satisfaction, relativement à la condition sans aucune consigne de travail (G3), *de facto* la plus compatible avec la liberté pédagogique des enseignant·e·s qui ne pratiquent pas ou peu cette « forme pédagogique » fondée sur le travail en groupe. Conformément à nos hypothèses, nous avons observé un niveau de satisfaction significativement réduit dans les séquences jugées plutôt réussies autrement dit lorsque la

liberté pédagogique (à laquelle les enseignant·e·s sont généralement attachés) était contrainte par les consignes qui avaient été fournies dans les conditions G1 (Collectif Structuré) et G2 (Collectif non structuré). Alors même que les enseignant·e·s ont plutôt bien suivi les consignes (puisque'il s'agit des séquences jugées plutôt réussies), ils sont significativement moins satisfait·e·s dans la condition G3 (Libre). C'est donc bel et bien le fait d'avoir réduit leur liberté pédagogique qui semble expliquer la différence observée en terme de niveau de satisfaction entre les enseignant·e·s des conditions G1 (Collectif Structuré) et G2 (Collectif non structuré), comparativement à ceux de la condition G3 (Libre).

Enfin, la dernière étape de nos analyses, présentée dans le chapitre 5, consistait à examiner si le bénéfice attendu du mécanisme d'interdépendance positive s'exprimait *a minima* dans les séquences jugées plutôt réussies (les autres étant *de facto* difficilement interprétables) et cela, indépendamment des performances antérieures (auto-rapportées) des élèves, ou au contraire davantage chez ceux qui rapportaient les performances les plus faibles. Nous nous attendions à observer un effet d'interaction entre les conditions de travail et le niveau de performance antérieure rapporté par les élèves. Plus précisément, nous attendions dans la condition de travail « Collectif structuré » (G1) un écart réduit entre les élèves les plus faibles et ceux les plus forts, relativement à G2 et G3, en raison du mécanisme d'interdépendance positive. Ce mécanisme devait, en principe, permettre aux élèves les plus en difficulté une valorisation de soi et un investissement plus élevé dans la séquence proposée, et par conséquent une meilleure performance. Dans la condition de travail « Collectif non structuré » (G2), nous attendions un écart de performance maximal entre les élèves les plus faibles et ceux les plus forts (relativement à G1 et G3) en raison de la combinaison de la paresse sociale des plus faibles (réduction des efforts personnels) et de l'effet éventuel de compensation chez leurs homologues les plus forts (augmentation de leurs efforts personnels).

Nous souhaitions également tester si l'hypothèse d'interaction détaillée précédemment dépendait du niveau de satisfaction des enseignant·e·s à l'égard des conditions de réalisation de leurs séquences pédagogiques, séquences que nous avons pour rappel contraintes en G1 et G2 par nos consignes de travail. Autrement dit, nous souhaitions savoir si notre hypothèse d'interaction s'exprimait quel que soit le niveau de satisfaction des enseignant·e·s ou exclusivement chez les élèves exposé·e·s à des enseignant·e·s satisfaits d'appliquer les méthodes proposées.

Sur les 20 séquences jugées plutôt réussies, nous avons obtenu soit un effet principal soit un effet d'interaction seulement dans 6 séquences portant sur le français ou les mathématiques.

L'effet principal de la condition s'exprimait dans 2 séquences et exclusivement en français chez les élèves exposé·e·s à des enseignant·e·s caractérisé·e·s par un haut niveau de satisfaction. Dans ces 2 séquences, contrairement à notre hypothèse, les élèves produisaient une performance plus élevée lorsque la liberté pédagogique de leurs enseignant·e·s n'était pas contrainte par l'interdépendance positive de type Puzzle, autrement dit dans la condition « Collectif non structuré » et dans la condition « Libre », qui ne différaient pas l'une de l'autre.

L'effet d'interaction entre la condition et le niveau initial des élèves s'est quant à lui exprimé dans 4 séquences et exclusivement en mathématiques chez les élèves exposé·e·s à des enseignant·e·s caractérisé·e·s par un haut niveau de satisfaction. Contrairement à notre hypothèse, une amplification de la hiérarchie entre les trois groupes d'élèves (Faible vs. Moyen vs. Fort) était observée lorsqu'ils étaient placés dans une situation de travail collectif, que cette dernière soit structurée ou non structurée. Cet effet, au profit des élèves les plus forts s'estompait dans la condition « Libre » dans laquelle la liberté pédagogique des enseignant·e·s était préservée. Enfin, l'effet principal de la condition et l'effet d'interaction entre la condition et le niveau initial auto-rapporté par les élèves s'exprimaient exclusivement dans les séquences

conduites par des enseignant·e·s se déclarant satisfait·e·s, et donc par hypothèse plutôt en accord avec la méthode proposée. Certain·e·s enseignant·e·s ont en effet rapporté durant les entretiens réalisés dans les établissements qu'ils/elles avaient des regrets quant au fait de devoir faire travailler les élèves sur des contenus et selon un format qui leur ont été imposé·e·s. Néanmoins, nous ne pouvons pas conclure que la restriction de leur liberté pédagogique ait nécessairement détérioré les performances de leurs élèves. En effet, sur les 20 séquences jugées « plutôt réussies », les performances moyennes en G3 (Libre) relativement aux deux autres conditions, en particulier G1 (Collectif structuré), sont plus élevées dans seulement 2 séquences concentrées sur l'enseignement de français.

4 Discussion et limites des résultats de l'étude ProFan

Nous obtenons peu d'effets significatifs (pour rappel 6 effets au total sur les 20 séquences jugées plutôt réussies). Cette faible proportion de résultats significatifs pourrait être expliquée par le fait que la classe Puzzle, et l'apprentissage coopératif de façon générale, est une méthode pédagogique complexe et exigeante qui peut représenter un défi en terme d'implémentation aussi bien pour les élèves que pour leurs enseignant·e·s (Buchs et al., 2017 ; Topping et al., 2017).

Les objections habituelles consistent à affirmer que les principaux inconvénients de la méthode Puzzle résident dans le fait que les élèves peuvent tout d'abord éprouver des difficultés à 1) comprendre un contenu qui peut être jugé difficile dans un temps limité (comme cela a pu être le cas dans Profan, cf. Chapitre 5) et 2) à trouver le moyen de l'enseigner de façon à ce que les autres élèves puissent le comprendre. De plus, les élèves étant exposé·e·s de manière limitée aux contenus pédagogiques qu'ils/elles n'ont pas étudié eux-mêmes, ils/elles sont par conséquent dépendants de l'apport de leurs pair·e·s pour accéder à l'ensemble du matériel (Buchs, 2020). Cela peut être problématique dans les cas où les élèves ont du mal à comprendre le matériel que leurs pair·e·s leur enseignent, cela d'autant plus si les élèves n'ont pas été

spécifiquement entraîné à le faire. Les élèves impliqués dans le dispositif ProFan n'ont en effet pas reçu de formation spécifique s'agissant des compétences sociales et cognitives nécessaires à la mise en œuvre d'un travail coopératif de qualité. Pendant les entretiens réalisés par le groupement de chercheur·es dans les établissements impliqués dans le dispositif, les enseignant·es ont par exemple rapporté que les élèves éprouvaient des difficultés pendant la phase Puzzle lorsque qu'ils/elles devaient élaborer puis transmettre les connaissances sur la partie de la leçon sur laquelle ils avaient travaillé. Or, c'est bien le fait de résumer des informations, d'apporter des explications ou encore de co-construire les connaissances, qui soutient l'apprentissage des élèves dans un contexte coopératif (Buchs, 2017). Plusieurs auteur·rices soutiennent ainsi l'idée que les élèves ne coopèrent pas de façon spontanée et qu'il est indispensable de les préparer à la mise en œuvre d'interactions constructives (e.g., Buchs, 2017 ; Plante, 2012 ; Topping et al, 2017). En effet, le contexte scolaire ne permettrait pas aux élèves de développer efficacement les compétences nécessaires à la coopération puisqu'ils sont socialisés dans un environnement compétitif qui valorise la réussite individuelle de chacune (Buchs et Butera, 2019). Préparer les élèves à travailler ensemble afin de promouvoir des interactions constructives apparaît alors indispensable. Topping et al. (2017) proposent premièrement que l'enseignant·e, à travers des activités de différentes natures, peut encourager un climat positif favorable à l'esprit d'équipe et à l'apprentissage entre pair·es, dans le but d'orienter la motivation des élèves vers des buts de maîtrise et non pas des buts de performance (Buchs, 2017). Deuxièmement, l'enseignant·e peut réaliser un travail spécifique sur les compétences coopératives durant lequel il/elle explicite les compétences nécessaires et la manière dont les élèves peuvent les déployer lors du travail en petits groupes afin de favoriser son efficacité. Troisièmement, l'enseignant·e peut proposer aux élèves de réfléchir sur la manière d'améliorer le fonctionnement des groupes de travail.

Ainsi, l'enseignement des compétences sociales et cognitives nécessaires à la mise en œuvre d'un apprentissage coopératif de qualité semble demander de la réflexion et du temps, dans un contexte que l'on sait contraint par un calendrier pédagogique strict. Il n'est donc pas étonnant qu'il puisse être souvent négligé, comme cela a été le cas dans le dispositif Profan. Par conséquent, il est probable que les groupes d'élèves n'aient pas été en mesure de fonctionner de manière optimale.

Plus récemment, Roseth et al., (2019) ont fait valoir de nouveaux éléments susceptibles d'expliquer les effets contrastés de la méthode Puzzle s'agissant des performances académiques, variable dépendante que nous testons dans le cadre de nos travaux. Ces dernières suggèrent que la procédure employée dans la méthode Puzzle (i.e., Phase Expert et Phase Puzzle) pourrait en réalité engager différents types de relations entre les élèves qui provoqueraient des effets opposés sur l'apprentissage. En effet dans la phase Expert, le fait de travailler avec d'autres élèves sur le même matériel engendrerait une indépendance des ressources, qui aurait pour conséquence d'orienter les élèves vers la compétition. Dans la phase Puzzle, le fait de travailler avec des pairs sur un corpus d'informations complémentaire impliquerait non seulement des relations d'interdépendance positive, mais également des relations compétitives dès que les membres du groupe ont accès aux ressources des un·es et des autres. Toutefois, Roseth et al. (2019) notent tout de même un effet de la méthode Puzzle sur les performances des étudiant·es à la fin du semestre d'expérimentation, en comparaison à un groupe exposé à une méthode d'enseignement habituelle. Cet effet, bien qu'il soit faible, suggère là encore que les compétences sociales et cognitives nécessaires à la mise en œuvre d'une interdépendance positive de type classe Puzzle nécessitent du temps et de l'entraînement.

Par conséquent, un tel dispositif demande une réorganisation des interventions en classe et contraint l'enseignant·e à une posture particulière dans laquelle son rôle est reconfiguré (Buchs, 2017 ; Topping et al., 2017). Ce changement de perspective semble en effet avoir

déstabilisé les enseignant·e·s impliqué·e·s dans le dispositif Profan puisque certain·e·s ont rapporté durant les entretiens avoir eu des difficultés s'agissant de la posture qu'ils/elles devaient adopter face aux élèves lorsque ces dernier·e·s travaillaient en groupe. Comme les élèves, les enseignant·e·s n'ont pas reçu de formation spécifique à la mise en œuvre d'une interdépendance positive de type classe Puzzle. Or, certains auteurs·rices affirment que les connaissances des enseignant·e·s sur l'apprentissage coopératif affectent leur capacité à la mettre en œuvre avec succès (Hennessey & Dionigi, 2013). Ainsi, la classe Puzzle exige de nombreux aménagements de la part des enseignant·e·s concernant leurs pratiques pédagogiques puisqu'elle implique notamment le fait de transférer une grande partie de la responsabilité de l'enseignant·e à l'élève. Cela implique également que les enseignant·e·s aient confiance en la capacité des élèves à apprendre ensemble, ce qui n'est pas toujours le cas (e.g. Blatchford et al., 2003). De plus, dans l'apprentissage coopératif, le rôle des enseignant·e·s change. Ils/elles n'incarnent plus la posture traditionnelle de transmission des savoirs mais deviennent des facilitateurs qui observent, assistent et guident les groupes en action (Buchs et al., 2017 ; Topping et al. 2017). Même si l'enseignant·e devient observateur·trice, son rôle reste toutefois primordial puisqu'il/elle a pour mission d'accompagner les élèves pendant le travail coopératif en intervenant par exemple pour réguler les comportements et les activités cognitives (Volpé et Buchs, 2019). Un rôle parfois mal interprété dans Profan, puisque certain·e·s enseignant·e·s ont rapporté avoir compris qu'ils/elles n'avaient pas le droit d'intervenir et que les élèves devaient « *se débrouiller tout·e·s seul·e·s* » lors de l'étape des groupes experts par exemple, d'où la nécessité de proposer des formations spécifiques aux enseignant·e·s à ce sujet. Une demande qui a d'ailleurs été formulée par certain·e·s lors des entretiens réalisés dans les établissements concernés par le dispositif ProFan. De la même façon, le temps nécessaire à l'implémentation de l'apprentissage coopératif est considéré par les enseignant·e·s comme coûteux et peut représenter un frein à la mise en place de cet outil pédagogique (Buchs et al., 2017). Les

enseignant·e·s du dispositif ProFan ont en effet rapporté des difficultés relatives au fait de terminer leurs séquences dans les temps, notamment à cause de la longueur et de la difficulté de certaines activités, mais aussi à cause d'autres problèmes tels que l'agencement des calendriers. Outre ces difficultés, Buchs et al. (2017) suggèrent que l'application des principes relatifs à l'apprentissage coopératif (i.e., préparation des élèves à la coopération, interdépendance positive et responsabilité individuelle), l'harmonisation avec le programme pédagogique ou encore l'évaluation dans le cadre de l'apprentissage coopératif peuvent représenter des freins quant à son implémentation.

Les éléments décrits plus haut représentent des défis importants pour la mise en œuvre de l'interdépendance positive au sens de la classe Puzzle. Même si les enseignant·e·s étaient tous/toutes disposé·e·s à implémenter cette méthode pédagogique du mieux possible, ils/elles ont pu rencontrer un certain nombre de difficultés susceptibles d'expliquer 1) le nombre réduit de séquences pédagogiques jugées plutôt réussies du point de vue du respect des consignes (pour rappel 21 sur 72 séquences au total) et 2) le peu d'effets significatifs observés dans nos travaux.

La faible proportion de résultats significatifs pourrait aussi tenir à la méthode employée pour catégoriser les séquences pédagogiques. Nous avons en effet reconstruit post-hoc la réalité des séquences pédagogiques sur la base de ce que nous ont dit les enseignant·e·s sur leurs déroulements et cela à partir de leurs réponses aux items de la grille d'observation (cf. Chapitre 4). Néanmoins, et comme nous l'avons évoqué dans le Chapitre 5, des biais de mémoire (i.e., erreurs de rappel et/ou reconstruction des faits), de désirabilité sociale (i.e., affirmer un respect des consignes en dépit de pratiques qui en étaient éloignées), ou des incompréhensions s'agissant des consignes ont pu altérer les réponses fournies par les enseignant·e·s aux 5 items de la « grille d'observation » et par conséquent la catégorisation des séquences qui en découle. D'autre part, la principale limite de cette catégorisation réside dans le fait qu'elle se base sur

les réponses moyennes des enseignant·e·s aux 5 items de la grille d'observation, sans tenir compte de la dispersion des réponses autour de cette moyenne. Ainsi, les tendances centrales (cf. Chapitre 4) nous ont permis d'estimer si, en moyenne, les réponses des enseignant·e·s aux 5 items de la grille d'observation étaient différentes selon les conditions G1, G2 et G3, cohorte par cohorte, filière par filière et séquence par séquence pour chaque enseignement. Néanmoins avec cette procédure, il subsiste des enseignant·e·s pour qui les réponses s'écartent sensiblement des profils de réponses moyens et par conséquent des profils de réponses attendus dans les séquences considérées. La dispersion des profils de réponses aux items « Collectif Classique », « Classe Entière » et « Puzzle » (cf. Annexe P) nous a ainsi permis d'apercevoir que même si au niveau du calcul des tendances centrales le profil de réponse moyen dans la séquence considérée est correct, certains enseignant·e·s peuvent tout de même avoir produit des réponses qui ne correspondent pas à un profil compatible avec les consignes. Nous aurions pu employer une autre technique pour catégoriser les séquences pédagogiques, notamment celle de l'analyse en « *clustering* » qui a été utilisée pour la synthèse des travaux du consortium ProFan (accessible sur demande).

La logique de l'analyse par « *clusters* » consiste à examiner les réponses des enseignant·e·s aux 3 items critiques de la grille d'observation post-séquence pédagogique non plus en moyenne, mais de façon individuelle pour chaque enseignant. Ainsi, les enseignant·e·s sont regroupé·e·s par « *clusters* » en fonction de leur degré de similitude s'agissant de leurs réponses aux 3 items critiques de la grille d'observation. Par exemple, on aurait pu imaginer un premier cluster qui regrouperait les enseignant·e·s qui ont accordé un haut pourcentage à l'item « Puzzle », et un faible pourcentage à l'item « Collectif Classique » et « Classe Entière ». Puis un autre, qui regrouperait les enseignant·e·s qui ont répondu un haut pourcentage à l'item « Puzzle » et à l'item « Collectif Classique », et un faible pourcentage à l'item « Classe Entière », etc. De cette façon, en ne conservant que les « *clusters* » dont les profils de réponses

correspondent aux consignes fournies dans chacun des groupes G1, G2 et G3, le « bruit » associé aux réponses des enseignant·e·s qui s'écartent des consignes est réduit.

Les analyses pour la synthèse ProFan ont été réalisées sur la base du même modèle multi-niveaux que celui utilisé pour notre rapport de thèse (cf. Chapitre 5). Pour la synthèse, le modèle était cependant appliqué non plus à chaque séquence pédagogique prise isolément, mais à des séquences regroupées par « *clusters* » (les performances des élèves ont donc été centrées et réduites pour permettre ces regroupements).

Les résultats de ces analyses ont montré que le bénéfice du travail collectif à l'échelle de la classe entière était observé pour les trois groupes d'élèves (Faible vs. Moyen vs. Fort) en Français et en Mathématiques, et davantage pour celles et ceux qui se percevaient plutôt en réussite ou en position intermédiaire. Mais cela, uniquement pour les élèves exposés à des enseignant·e·s se déclarant satisfait·e·s des consignes à respecter. Pour les élèves exposés à des enseignant·e·s insatisfait·e·s, le bénéfice était observé uniquement pour les élèves qui se percevaient plutôt en réussite.

Des analyses de même nature conduites sur les performances obtenues dans les matières professionnelles ont montré des résultats différents selon les filières. En ASSP, un bénéfice de l'interdépendance positive (G1) a été observé pour les trois groupes d'élèves (Faible vs. Moyen vs. Fort), relativement aux configurations de travail collectif non structuré (G2) et de pédagogie libre (G3). Le même pattern a été observé pour MELEC. En revanche, aucune différence n'a été observée entre les conditions G1, G2 et G3 pour les élèves de la filière COMMERCE.

Les résultats décrits précédemment suggèrent donc que la technique employée pour catégoriser les séquences pédagogiques explique, au moins en partie, les différences observées de part et d'autre (synthèse vs. rapport de thèse). Autrement dit, l'analyse de « *clustering* » offre des résultats plus favorables à la méthode Puzzle, comparativement à l'analyse utilisée pour ce rapport de thèse (i.e., catégorisation selon le profil de réponse par des

comparaisons de moyennes). Ces analyses, plus favorables, peuvent traduire aussi le fait de l'augmentation de la puissance de test lié au regroupement des séquences pédagogiques. En bref, même s'il est encore difficile de conclure à ce stade sur l'ampleur des bénéfices de la méthode Puzzle en matière de performances scolaires, cette divergence entre les résultats présentés dans la synthèse ProFan et les résultats présentés dans le présent rapport de thèse est intéressante. Elle suggère que la catégorisation des séquences pédagogiques sur la base des comparaisons de moyennes n'est en effet pas nécessairement optimale.

Par ailleurs, même si nous avons observé un faible nombre d'effets significatifs dans ce rapport de thèse, nous ne pouvons pas conclure à l'inutilité de la méthode Puzzle dans la mesure où les conditions favorables à son expression ne semblent pas avoir été toutes réunies. En effet, nous avons vu précédemment que le manque de formation des enseignant·e·s et le manque d'entraînement des élèves pouvaient constituer des freins à l'implémentation de la coopération à l'école. D'où la recommandation de certains auteurs consistant à favoriser cet entraînement (e.g., Buchs et al., 2017 ; Topping et al. 2017), ou à étudier *a minima* les challenges que peuvent rencontrer les enseignant·e·s pour l'implémentation de dispositifs coopératifs à l'école (Baloche & Brody, 2017).

Enfin, une limite générale de l'étude ProFan est qu'il ne s'agit pas d'une étude randomisée. Il était en effet très difficile de choisir au hasard les établissements impliqués dans le groupe G1, par définition soumis à une procédure assez chargée en consignes que tous les établissements ne pouvaient d'emblée accepter. D'où une procédure consistant d'abord à choisir des établissements capables, pour diverses raisons, de soutenir la lourdeur du dispositif à mettre en place dans le groupe G1, pour ensuite choisir des établissements comparables au titre de groupes contrôles (G2 et G3). Cette procédure non randomisée est assez courante sur le terrain en raison même des contraintes souvent insolubles en dehors du laboratoire. Par ailleurs, malgré la popularité croissante des expérimentations longitudinales à essais randomisés et

contrôlés dans le domaine de l'éducation (« *Randomized controlled trials* » ou RCT), la communauté scientifique prend de plus en plus conscience de leurs limites (Sims, 2020). Plusieurs études ont en effet révélé que ce type d'expérimentation disposait d'une puissance statistique bien inférieure à 80 % pour détecter des effets de la taille de ceux que l'on trouve habituellement dans la littérature sur l'éducation (e.g., Cheung & Slavin, 2016 ; Spybrook et al., 2016). Ainsi, ces expérimentations ne permettent pas toujours d'évaluer correctement l'efficacité des interventions académiques supposées améliorer l'apprentissage. Lortie-Forgues & Inglis (2019) estiment que 40% des expérimentations de ce type en éducation sont inefficaces et non-informatives car elles seraient réalisées dans des conditions qui sous-estiment un « bruit expérimental ». Selon eux, ce « bruit » proviendrait entre autres du fait que la majorité des interventions déployées seraient mal implémentées.

En conclusion, nous suggérons de poursuivre l'effort engagé pour tester les méthodes pédagogiques coopératives à grande échelle, en prêtant une attention particulière aux conditions à réunir sur le terrain pour que la coopération puisse exprimer l'entièreté de ses bénéfices. Selon le Comité Consultatif National d'Éthique (CCNE), et dans le but de fournir des recommandations pédagogiques précises et valides, les résultats obtenus dans les « conditions artificielles » du laboratoire ne sont pas suffisants. En effet, il serait nécessaire de « mesurer directement les conséquences d'une pratique pédagogique en conditions réelles » (CCNE, 2019, p. 10). Une collaboration entre chercheur·es et enseignant·es apparaît alors aujourd'hui indispensable afin d'implanter des méthodes pédagogiques sur la base de connaissances scientifiques solides, tout en tenant comptes des spécificités et contraintes du terrain scolaire.

Bibliographie

- Akçay, N. O. (2016). Implementation of Cooperative Learning Model in Preschool. *Journal of Education and Learning*, 5(3), 83. <https://doi.org/10.5539/jel.v5n3p83>
- Alamri, H. R. H. (2018). The Effect of Using the Jigsaw Cooperative Learning Technique on Saudi EFL Students' Speaking Skills. *Journal of Education and Practice*, 9(6), 65-77. <http://www.iiste.org/Journals/index.php/JEP/article/view/41150>
- Allport, G. W. (1954). *The nature of prejudice*. Addison-Wesley.
- Al-Salkhi, M. (2015). The effectiveness of jigsaw strategy on the achievement and learning motivation of the 7th primary grade students in the islamic education. ... *Journal of humanities and social science*, Query date: 2020-05-13 10:45:36. <https://pdfs.semanticscholar.org/db73/372db0d6a6d459dea978f941f29e1c1284df.pdf> *⁴
- Amathieu, J., & Chaliès, S. (2014). Satisfaction professionnelle, formation et santé au travail des enseignants. *Carrefours de l'éducation*, 38(2), 211. <https://doi.org/10.3917/cdle.038.0211>
- Aronson, E. (1978). *The jigsaw classroom*. psycnet.apa.org. <https://psycnet.apa.org/record/1980-51351-000>
- Aronson, E. (2011). Reducing prejudice and building empathy in the classroom. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Éds.), *Psychology and the real world : Essays illustrating fundamental contributions to society*. (2011-19926-028; p. 230-236). Worth Publishers; APA PsycInfo. <http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2011-19926-028&lang=fr&site=ehost-live>

⁴ Les références suivies d'un * sont incluses dans notre synthèse plus quantitative de la littérature Puzzle.

- Aronson, E., & Bridgeman, D. (1979). Jigsaw groups and the desegregated classroom : In pursuit of common goals. *Personality and social psychology ...*, *Query date: 2020-05-13* 10:45:36. <https://journals.sagepub.com/doi/abs/10.1177/014616727900500405>
- Aronson, E., & Patnoe, S. (2011). *Cooperation in the classroom : The jigsaw method*. Pinter & Martin.
- Aronson, E., Stephan, C., Sikes, J., Blaney, N., & Snapp, M. (1978). *The Jigsaw Classroom*. Sage.
- Artut, P. D., & Tarim, K. (2007). The Effectiveness of Jigsaw II on Prospective Elementary School Teachers. *Asia-Pacific Journal of Teacher Education*, 35(2), 129-141. ERIC. <http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ763910&lang=fr&site=ehost-live> *
- Avanzi, L., Miglioretti, M., Velasco, V., Balducci, C., Vecchio, L., Fraccaroli, F., & Skaalvik, E. M. (2013). Cross-validation of the Norwegian Teacher's Self-Efficacy Scale (NTSES). *Teaching and Teacher Education*, 31, 69-78. <https://doi.org/10.1016/j.tate.2013.01.002>
- Aydin, A., & Biyikli, F. (2017). The Effect of Jigsaw Technique on the Students' Laboratory Material Recognition and Usage Skills in General Physics Laboratory-I Course. *Universal Journal of Educational Research*, 5(7), 1073-1082. ERIC. <http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1147794&lang=fr&site=ehost-live> *
- Bakker, A., Cai, J., English, L., Kaiser, G., Mesa, V., & Van Dooren, W. (2019). Beyond small, medium, or large : Points of consideration when interpreting effect sizes. *Educational Studies in Mathematics*, 102(1), 1-8. <https://doi.org/10.1007/s10649-019-09908-4>
- Baloche, L., & Brody, C. M. (2017). Cooperative learning : Exploring challenges, crafting innovations. *Journal of Education for Teaching*, 43(3), 274-283. <https://doi.org/10.1080/02607476.2017.1319513>

- Basyah, N. A., Muslem, A., & Usman, B. (2018). The Effectiveness of Using the Jigsaw Model to Improve Students' Economics Teaching-Learning Achievement. *The New Educational Review*, 51(1), 30-40. https://tner.polsl.pl/dok/volumes/tner_1_2018b.pdf#page=30 *
- Berger, R., & Hänze, M. (2009). Comparison of Two Small-group Learning Methods in 12th-grade Physics Classes Focusing on Intrinsic Motivation and Academic Performance. *International Journal of Science Education*, Query date: 2020-05-13 10:45:36. <https://www.tandfonline.com/doi/abs/10.1080/09500690802116289>
- Blaney, N. T., Stephan, C., Rosenfield, D., Aronson, E., & Sikes, J. (1977). Interdependence in the classroom : A field study. *Journal of Educational Psychology*, 69(2), 121-128. <https://doi.org/10.1037/0022-0663.69.2.121>
- Blatchford, P., Kutnick, P., Baines, E., & Galton, M. (2003). Toward a social pedagogy of classroom group work. *International Journal of Educational Research*, 39(1-2), 153-172. [https://doi.org/10.1016/S0883-0355\(03\)00078-8](https://doi.org/10.1016/S0883-0355(03)00078-8)
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1(2), 97-111. <https://doi.org/10.1002/jrsm.12>
- Bratt, C. (2008). The Jigsaw classroom under test : No effect on intergroup relations evident. *Journal of Community & Applied Social Psychology*, 18(5), 403-419. <https://doi.org/10.1002/casp.946>
- Brickner, M. A., Harkins, S. G., & Ostrom, T. M. (1986). Effects of personal involvement : Thought-provoking implications for social loafing. *Journal of Personality and Social Psychology*, 51(4), 763-769. <https://doi.org/10.1037/0022-3514.51.4.763>
- Bridgeman, D. L. (1981). Enhanced Role Taking Through Cooperative Interdependence : A Field Study. *Child Development*, 52(4), 1231. <https://doi.org/10.2307/1129511>

- Buchs, C. (2017). Comment organiser l'apprentissage des élèves par petits groupes? *Différenciation pédagogique : comment adapter l'enseignement pour la réussite de tous les élèves?* Conseil National d'évaluation du Système Scolaire. <https://archive-ouverte.unige.ch/unige:95551>
- Buchs, C. (2020). Reflection on the Jigsaw method. *IASCE Newsletter*, 39(1), 3. https://orfee.hepl.ch/bitstream/handle/20.500.12162/5615/20_Buchs_IASCE_Jigsaw.pdf?sequence=1
- Buchs, C., Filippou, D., Pulfrey, C., & Volpé, Y. (2017). Challenges for cooperative learning implementation : Reports from elementary school teachers. *Journal of Education for Teaching*, 43(3), 296-306. <https://doi.org/10.1080/02607476.2017.1321673>
- Buchs, C., Gilles, I., Antonietti, J.-P., & Butera, F. (2016). Why students need to be prepared to cooperate : A cooperative nudge in statistics learning at university. *Educational Psychology*, 36(5), 956-974. <https://doi.org/10.1080/01443410.2015.1075963>
- Butera, F., Świątkowski, W., & Dompnier, B. (2021). Competition in Education. In S. M. Garcia, A. Tor, & A. J. Elliot (Éds.), *The Oxford Handbook of the Psychology of Competition*. Oxford University Press. <https://doi.org/10.1093/oxfordhb/9780190060800.013.24>
- Çağatay, G., & Demircioğlu, G. (2013). The effect of Jigsaw-I cooperative learning technique on students' understanding about basic organic chemistry concepts. *International Journal of Educational Researchers*, Query date: 2020-05-13 10:45:36. http://ijer.eab.org.tr/media/volume4/issue2/g_cagatay.pdf *
- Caprara, G. V., Barbaranelli, C., Steca, P., & Malone, P. S. (2006). Teachers' self-efficacy beliefs as determinants of job satisfaction and students' academic achievement : A study at the school level. *Journal of School Psychology*, 44(6), 473-490. <https://doi.org/10.1016/j.jsp.2006.09.001>

- Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science*, 2(1), 40.
<https://doi.org/10.1186/1748-5908-2-40>
- Charbonnier, E., Huguet, P., Brauer, M., & Monteil, J.-M. (1998). Social loafing and self-beliefs : people's collective effort depends on the extent to which they distinguish themselves as better than others. *Social Behavior and Personality: An International Journal*, 26(4), 329-340. <https://doi.org/10.2224/sbp.1998.26.4.329>
- Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, 45(5), 283-292.
<https://files.eric.ed.gov/fulltext/ED567237.pdf>
- Christley, R. M. (2010). Power and Error : Increased Risk of False Positive Results in Underpowered Studies. *The Open Epidemiology Journal*, 3(1), 16-19.
<https://doi.org/10.2174/1874297101003010016>
- Chu, S. (2014). Application of the jigsaw cooperative learning method in economics course. *International Journal of Managerial Studies and Research (IJMSR)*, Query date: 2020-05-13 10:45:36.
<https://pdfs.semanticscholar.org/9c47/c9825dfade109cf75fc13febb0aca6909952.pdf> *
- Clark, K. B., & Clark, M. P. (1950). Emotional Factors in Racial Identification and Preference in Negro Children. *The Journal of Negro Education*, 19(3), 341.
<https://doi.org/10.2307/2966491>
- Cohen, J. (1962). The statistical power of abnormal-social psychological research : A review. *The Journal of Abnormal and Social Psychology*, 65(3), 145-153.
<https://doi.org/10.1037/h0045186>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). NJ: Lawrence Erlbaum Associates.

- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112(1), 155-159.
<https://doi.org/10.1037/0033-2909.112.1.155>
- Comité Consultatif National d'Ethique. (2019). *Cadre éthique de l'expérimentation pédagogique en situation réelle. Avis 131. Du laboratoire à l'école. Communiqué de presse.* <https://www.ccne-ethique.fr/fr/publications/cadre-ethique-delexperimentation-pedagogique-en-situation-reelle>
- Crone, T. S., & Portillo, M. C. (2013). Jigsaw Variations and Attitudes about Learning and the Self in Cognitive Psychology. *Teaching of Psychology*, 40(3), 246-251. ERIC.
<http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1014333&lang=fr&site=ehost-live>
- Darnon, C., Buchs, C., & Desbar, D. (2012). The jigsaw technique and self-efficacy of vocational training students: A practice report. *European Journal of Psychology of Education*, 27(3), 439-449. <https://doi.org/10.1007/s10212-011-0091-4>
- de Boer, H., Donker, A. S., & van der Werf, M. P. C. (2014). Effects of the Attributes of Educational Interventions on Students' Academic Performance: A Meta-Analysis. *Review of Educational Research*, 84(4), 509-545.
<https://doi.org/10.3102/0034654314540006>
- Deci, E. L., & Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality*, 19(2), 109-134.
[https://doi.org/10.1016/0092-6566\(85\)90023-6](https://doi.org/10.1016/0092-6566(85)90023-6)
- Deci, E. L., & Ryan, R. M. (2000). The « What » and « Why » of Goal Pursuits: Human Needs and the Self-Determination of Behavior. *Psychological Inquiry*, 11(4), 227-268.
https://doi.org/10.1207/S15327965PLI1104_01

- Dirlikli, M., Aydın, K., & Akgün, L. (2016). Cooperative Learning in Turkey : A Content Analysis of Theses. *Educational Sciences: Theory & Practice*, 16(4). <https://doi.org/10.12738/estp.2016.4.0142>
- Dori, Y., Yeroslavski, O., & Lazarowitz, R. (1995). *The Effect of Teaching the Cell Topic Using the Jigsaw Method on Students Achievement and Learning Activity*. 68th Annual National Association for Research in Science Teaching Conference, San Francisco, CA. *
- Doymus, K. (2007). Effects of a cooperative learning strategy on teaching and learning phases of matter and one-component phase diagrams. *Journal of Chemical Education*, Query date: 2020-05-13 10:45:36. <https://pubs.acs.org/doi/abs/10.1021/ed084p1857>
- Doymus, K. (2008). Teaching Chemical Equilibrium with the Jigsaw Technique. *Research in Science Education*, 38(2), 249-260. <https://doi.org/10.1007/s11165-007-9047-8> *
- Doymus, K., Karacop, A., & Simsek, U. (2010). Effects of jigsaw and animation techniques on students' understanding of concepts and subjects in electrochemistry. *Educational technology research and development*, Query date: 2020-05-13 10:45:36. <https://link.springer.com/article/10.1007/s11423-010-9157-2>
- Evcim, H., & İpek, Ö. F. (2013). Effects of Jigsaw II on Academic Achievement in English Prep Classes. *Procedia - Social and Behavioral Sciences*, 70, 1651-1659. <https://doi.org/10.1016/j.sbspro.2013.01.236> *
- Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3), 891-904. <https://doi.org/10.1007/s11192-011-0494-7>
- Farahnaz, R. N., Parviz, A., & Nazila, K. (2013). The Effect of Using Jigsaw to Enhance Female Iranian Intermediate EFL Learners' Oral Proficiency. *Australian Journal of Basic and Applied Sciences*, 7(9), 315-326. *

- Ferguson, C. J., & Heene, M. (2012). A Vast Graveyard of Undead Theories : Publication Bias and Psychological Science's Aversion to the Null. *Perspectives on Psychological Science*, 7(6), 555-561. <https://doi.org/10.1177/1745691612459059>
- Franco, A., Malhotra, N., & Simonovits, G. (2016). Underreporting in Psychology Experiments : Evidence From a Study Registry. *Social Psychological and Personality Science*, 7(1), 8-12. <https://doi.org/10.1177/1948550615598377>
- Funder, D. C., & Ozer, D. J. (2019). Evaluating Effect Size in Psychological Research : Sense and Nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2), 156-168. <https://doi.org/10.1177/2515245919847202>
- Gambari, I., & Yusuf, M. (2016). Effects of Computer-Assisted Jigsaw II Cooperative Learning Strategy on Physics Achievement and Retention. *Contemporary Educational Technology*, Query date: 2020-05-13 10:45:36. <https://eric.ed.gov/?id=EJ1117586>
- Geffner, R. A. (1978). *The effects of interdependent learning on self-esteem, inter-ethnic relations, and intra-ethnic attitudes of elementary school children : A field experiment*. University of California.
- Gerber, A. (2008). Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals. *Quarterly Journal of Political Science*, 3(3), 313-326. <https://doi.org/10.1561/100.00008024>
- Gerber, A. S., & Malhotra, N. (2008). Publication Bias in Empirical Sociological Research : Do Arbitrary Significance Levels Distort Published Results? *Sociological Methods & Research*, 37(1), 3-30. <https://doi.org/10.1177/0049124108318973>
- Gersten, R., Chard, D. J., Jayanthi, M., Baker, S. K., Morphy, P., & Flojo, J. (2009). Mathematics Instruction for Students With Learning Disabilities : A Meta-Analysis of Instructional Components. *Review of Educational Research*, 79(3), 1202-1242. <https://doi.org/10.3102/0034654309334431>

- Ghaith, G., & El-Malak, M. (2004). Effect of Jigsaw II on literal and higher order EFL reading comprehension. *Educational Research and Evaluation*, *Query date: 2020-05-13* 10:45:36. <https://www.tandfonline.com/doi/abs/10.1076/edre.10.2.105.27906> *
- Gillies, R. M. (2014). Cooperative Learning : Developments in Research. *International Journal of Educational Psychology*, 3, 125-140. <https://doi.org/10.4471/ijep.2014.08>
- Gocer, A. (2010). A Comparative Research on the Effectivity of Cooperative Learning Method and Jigsaw Technique on Teaching Literary Genres. *Educational Research and Reviews*, 5(8), 439-445. ERIC. <http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ911830&lang=fr&site=ehost-live> *
- Gömleksiz, M. N. (2007). Effectiveness of cooperative learning (jigsaw II) method in teaching English as a foreign language to engineering students (Case of Firat University, Turkey). *European Journal of Engineering Education*, 32(5), 613-625. <https://doi.org/10.1080/03043790701433343> *
- Hänze, M., & Berger, R. (2007). Cooperative learning, motivational effects, and student characteristics : An experimental study comparing cooperative learning and direct instruction in 12th grade physics classes. *Learning and instruction*, *Query date: 2020-05-13* 10:45:36. <https://www.sciencedirect.com/science/article/pii/S0959475206001174>
- Hattie, J. (2009). *Visible learning : A synthesis of over 800 meta-analyses relating to achievement*. Routledge.
- Hedeen, T. (2003). The Reverse Jigsaw : A Process of Cooperative Learning and Discussion. *Teaching Sociology*, 31(3), 325-332. ERIC. <http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ679820&lang=fr&site=ehost-live>

- Hennessey, A., & Dionigi, R. (2013). Implementing cooperative learning in Australian primary schools : Generalist teachers' perspectives. *Issues in Educational Research*, 23(1), 52-68.
<http://www.iier.org.au/iier23/hennessey.html>
- Hilk, C. L. (2013). *Effects of cooperative, competitive, and individualistic learning structures on college student achievement and peer relationships : A series of meta-analyses*. University of Minnesota.
- Holliday, D. (2000). *The Development of Jigsaw IV in a Secondary Social Studies Classroom*. ERIC. <https://eric.ed.gov/?id=ED447045>
- Holliday, D. C. (1995). *The effect of the cooperative learning strategy jigsaw II on academic achiever and cross race relationships in a secondary social studies classroom*. Hattiesburg MS.
- Hornby, G. (2009). The effectiveness of cooperative learning with trainee teachers. *Journal of Education for Teaching*, Query date: 2020-05-13 10:45:36.
<https://www.tandfonline.com/doi/abs/10.1080/02607470902771045> *
- Hosseini, S. M., Maleki, R., & Mehrizi, A. A. H. (2014). On the impact of using Jigsaw II technique on the development of writing performance of Iranian intermediate EFL learners. *International Journal of Language Learning and Applied Linguistics World*, 7(3), 198-215. *
- Huang, Y.-M., Liao, Y.-W., Huang, S.-H., & Chen, H.-C. (2014). A jigsaw-based cooperative learning approach to improve learning outcomes for mobile situated learning. *Journal of Educational Technology & Society*, 17(1), 128-140. APA PsycInfo.
<http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2014-06607-012&lang=fr&site=ehost-live> *
- Huddy, W. P. (2013). *A meta-analytic review of cooperative learning practices in higher education : A human communication perspective* (2013-99151-173; Numéros 2-A(E))

[ProQuest Information & Learning]. APA PsycInfo.

<http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2013-99151-173&lang=fr&site=ehost-live>

Huguet, P. (1995). Travail Collectif et Performance Individuelle. In G. Mugny, D. Oberlé, & J. L. Beauvois (Éds.), *Traité de Psychologie Sociale* (Vol. 1, p. 31-41). Presses universitaires de Grenoble.

Huguet, P. (2006). Apprendre en groupe : La classe dans sa réalité sociale et émotionnelle. In E. Bourgeois & G. Chapelle (Éds.), *Apprendre et faire apprendre* (p. 153-167). PUF.

Huguet, P., Charbonnier, E., & Monteil, J.-M. (1999). Productivity loss in performance groups : People who see themselves as average do not engage in social loafing. *Group Dynamics: Theory, Research, and Practice*, 3(2), 118-131. <https://doi.org/10.1037/1089-2699.3.2.118>

Huguet, P., & Kuyper, H. (2017). Applying social psychology to the classroom. In *Applied Social Psychology: Understanding and Managing Social Problems*. (Cambridge University Press, p. 172-192). <https://hal.archives-ouvertes.fr/hal-03012898>

Huguet, P., & Monteil, J.-M. (2001). Coaction, Interdépendance et Performances Cognitives : Le Cas de la Paresse Sociale ou l'Effet « Social Loafing ». In J.-M. Monteil & J. L. Beauvois (Éds.), *La Psychologie Sociale : Des Compétences pour l'Application*. Presses Universitaires de Grenoble.

Huguet, P., & Monteil, J.-M. (2015). *Social context and cognitive performance : Towards a social psychology of cognition*.

Humphrey, S. E., Nahrgang, J. D., & Morgeson, F. P. (2007). Integrating motivational, social, and contextual work design features : A meta-analytic summary and theoretical extension of the work design literature. *Journal of Applied Psychology*, 92(5), 1332-1356. <https://doi.org/10.1037/0021-9010.92.5.1332>

IBM Corp. (2019). *IBM SPSS Statistics for Macintosh* (Version 26.0) [Computer software].

IBM Corp.

Iweka, F. (2017). Effects of authentic and Jigsaw II learning techniques on students academic achievement in mathematics. *Global Journal of Arts, Humanities and Social Sciences*, 5(5), 18-24. *

Jackson, J. M., & Williams, K. D. (1985). Social loafing on difficult tasks: Working collectively can improve performance. *Journal of Personality and Social Psychology*, 49(4), 937-942. <https://doi.org/10.1037/0022-3514.49.4.937>

Jafariyan, M., Matlabi, M., Esmaili, R., & Kianmehr, M. (2017). Effectiveness of teaching : Jigsaw technique vs lecture for medical students' Physics course. *Bali Medical Journal*, 6(3), 529. <https://doi.org/10.15562/bmj.v6i3.400> *

Janosz, M., Archambault, I., Lacroix, M., & Lévesque, J. (2007). *Trousse d'évaluation des décrocheurs potentiels (TEDP) : Manuel d'utilisation*. Montréal : Groupe de recherche sur les environnements scolaires. https://cpe.ac-noumea.nc/IMG/pdf/trousse_evaluation_decrocheurs_potentiels.pdf

Johnson, D., & Johnson, R. (1989). Cooperative learning : What special education teachers need to know. *The Pointer*, Query date: 2020-05-13 10:45:36. <https://www.tandfonline.com/doi/pdf/10.1080/05544246.1989.9945370>

Johnson, D., & Johnson, R. (2002). Cooperative learning and social interdependence theory. *Theory and research on small groups*, Query date: 2020-05-13 10:45:36. https://link.springer.com/content/pdf/10.1007/0-306-47144-2_2.pdf

Johnson, D., & Johnson, R. (2005). Cooperative learning, values, and culturally plural classrooms. *Classroom Issues*, Query date: 2020-05-13 10:45:36.

Johnson, D., & Johnson, R. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational researcher*, Query date:

2020-05-13

10:45:36.

<https://journals.sagepub.com/doi/abs/10.3102/0013189X09339057>

- Johnson, D. W., Johnson, R. T., & Stanne, M. B. (2000). *Cooperative learning methods : A meta-analysis*.
- Karacop, A., & Diken, E. (2017). The Effects of Jigsaw Technique Based on Cooperative Learning on Prospective Science Teachers' Science Process Skill. *Journal of Education and Practice*, Query date: 2020-05-13 10:45:36. <https://eric.ed.gov/?id=EJ1133003> *
- Karau, S. J., & Williams, K. (1993). Social Loafing : A Meta-Analytic Review and Theoretical Integration. *Journal of Personality and Social Psychology*, 65(4), 681-706.
- Karlsson, P., & Bergmark, A. (2015). Compared with what? An analysis of control-group types in Cochrane and Campbell reviews of psychosocial treatment efficacy with substance use disorders. *Addiction*, 110(3), 420-428. <https://doi.org/10.1111/add.12799>
- Kengatharan, N. (2020). The Effects of Teacher Autonomy, Student Behavior and Student Engagement on Teacher Job Satisfaction. *Educational Sciences: Theory & Practice*, 20(4), 1-15. https://www.researchgate.net/profile/Kengatharan-Navaneethakrishnan/publication/347932619_The_Effects_of_Teacher_Autonomy_Student_Behavior_and_Student_Engagement_on_Teacher_Job_Satisfaction/links/5fea9c5892851e13fecfd9ce/The-Effects-of-Teacher-Autonomy-Student-Behavior-and-Student-Engagement-on-Teacher-Job-Satisfaction.pdf
- Kerr, N. L. (1983). Motivation losses in small groups : A social dilemma analysis. *Journal of Personality and Social Psychology*, 45(4), 819-828. <https://doi.org/10.1037/0022-3514.45.4.819>
- Kerr, N. L., & Bruun, S. E. (1981). Ringelmann Revisited : Alternative Explanations for the Social Loafing Effect. *Personality and Social Psychology Bulletin*, 7(2), 224-231. <https://doi.org/10.1177/014616728172007>

- Koustelios, A. D., Karabatzaki, D., & Kousteliou, I. (2004). Autonomy and Job Satisfaction for a Sample of Greek Teachers. *Psychological Reports*, 95(3), 883-886. <https://doi.org/10.2466/pr0.95.3.883-886>
- Kraft, M. A. (2020). Interpreting Effect Sizes of Education Interventions. *Educational Researcher*, 49(4), 241-253. <https://doi.org/10.3102/0013189X20912798>
- Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication Bias in Psychology : A Diagnosis Based on the Correlation between Effect Size and Sample Size. *PLoS ONE*, 9(9), e105825. <https://doi.org/10.1371/journal.pone.0105825>
- Kumar, C. S. V., Kalasuramath, S., Patil, S., Kumar, K. G. R., Taj, K. R. S., Jayasimha, V. L., Basavarajappa, K. G., Shashikala, P., Kukkamalla, A., & Chacko, T. (2017). Effect of Jigsaw Co-Operative Learning Method in Improving Cognitive Skills among Medical Students. *International Journal of Current Microbiology and Applied Sciences*, 6(3), 164-173. <https://doi.org/10.20546/ijemas.2017.603.018> *
- Lai, C.-H., Liu, M.-C., Huang, S.-H., & Huang, Y.-M. (2015). Effectiveness of Jigsaw-based cooperative report writing in a vocational high school. *2015 International Conference on Interactive Collaborative Learning (ICL)*, 798-802. <https://doi.org/10.1109/ICL.2015.7318130>
- Lakens, D. (2021). *Sample Size Justification* [Preprint]. PsyArXiv. <https://doi.org/10.31234/osf.io/9d3yf>
- Lakens, D., & Etz, A. J. (2017). Too True to be Bad : When Sets of Studies With Significant and Nonsignificant Findings Are Probably True. *Social Psychological and Personality Science*, 8(8), 875-881. <https://doi.org/10.1177/1948550617693058>
- Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work : The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, 37(6), 822-832. <https://doi.org/10.1037/0022-3514.37.6.822>

- Law, Y. (2011). The effects of cooperative learning on enhancing Hong Kong fifth graders' achievement goals, autonomous motivation and reading proficiency. *Journal of Research in Reading*, 34(4), 402-425. <https://doi.org/10.1111/j.1467-9817.2010.01445.x> *
- Lazarowitz, R., Hertz-Lazarowitz, R., & ... (1994). Learning science in a cooperative setting : Academic achievement and affective outcomes. *Journal of research in science teaching, Query date: 2020-05-13 10:45:36.* <https://onlinelibrary.wiley.com/doi/abs/10.1002/tea.3660311006> *
- Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., & Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions : Explanation and elaboration. *Journal of Clinical Epidemiology*, 62(10), e1-e34. <https://doi.org/10.1016/j.jclinepi.2009.06.006>
- Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., Roberts, M., Anthony, K. S., & Busick, M. D. (2012). Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms. *National Center for Special Education Research*.
- Lortie-Forgues, H., & Inglis, M. (2019). Rigorous Large-Scale Educational RCTs Are Often Uninformative : Should We Be Concerned? *Educational Researcher*, 48(3), 158-166. <https://doi.org/10.3102/0013189X19832850>
- Lucker, G. W., Rosenfield, D., Sikes, J., & Aronson, E. (1976). Performance in the Interdependent Classroom : A Field Study. *American Educational Research Journal*, 13(2), 115-123. <https://doi.org/10.3102/00028312013002115>
- Marhamah, M., & Mulyadi, M. (2013). Jigsaw cooperative learning : A viable teaching-learning strategy? *Journal of educational and social research, Query date: 2020-05-13 10:45:36.* <https://www.mcser.org/journal/index.php/jesr/article/view/1027> *

- Mari, J. S., & Gumel, S. A. (2015). Effects of Jigsaw Model of Cooperative Learning on Self-Efficacy and Achievement in Chemistry among Concrete and Formal Reasoners in Colleges of Education in Nigeria. *International Journal of Information and Education Technology*, 5(3), 196-199. <https://doi.org/10.7763/IJiet.2015.V5.501> *
- Mattingly, R. M., & VanSickle, R. L. (1991). Cooperative Learning and Achievement in Social Studies : Jigsaw II. *Social Education*, 55(6), 392-395. ERIC. <http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ443698&lang=fr&site=ehost-live>
- Mons, N., Duru-Bellat, M., & Savina, Y. (2012). Modèles éducatifs et attitudes des jeunes : Une exploration comparative internationale. *Revue française de sociologie*, 53(4), 589. <https://doi.org/10.3917/rfs.534.0589>
- Monteil, J.-M., & Huguet, P. (1999a). *Social context and cognitive performance : Towards a social psychology of cognition*. Psychology press.
- Monteil, J.-M., & Huguet, P. (1999b). *Social Context and Cognitive Performance : Towards a Social Psychology of Cognition*. Hove, East Sussex: Psychology Press.
- Monteil, J.-M., & Huguet, P. (2013). *Réussir ou échouer à l'école : Une question de contexte?* PUG.
- Moskowitz, J., Malvin, J., & ... (1983). Evaluation of a cooperative learning strategy. *American Educational Research Journal*, Query date: 2020-05-13 10:45:36. <https://journals.sagepub.com/doi/abs/10.3102/00028312020004687>
- Moskowitz, J., Malvin, J., Schaeffer, G., & ... (1985). Evaluation of jigsaw, a cooperative learning technique. *Contemporary educational psychology*, Query date: 2020-05-13 10:45:36. <https://www.sciencedirect.com/science/article/pii/0361476X85900116>
- Mutlu, A. (2018). Comparison of Two Different Techniques of Cooperative Learning Approach : Undergraduates' Conceptual Understanding in the Context of Hormone

Biochemistry. *Biochemistry and Molecular Biology Education*, 46(2), 114-120. ERIC. <http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1173861&lang=fr&site=ehost-live> *

Nebel, S., Schneider, S., Beege, M., Kolda, F., Mackiewicz, V., & Rey, G. D. (2017). You cannot do this alone! Increasing task interdependence in cooperative educational videogames to encourage collaboration. *Educational Technology Research and Development*, 65(4), 993-1014. <https://doi.org/10.1007/s11423-017-9511-8> *

Newmann, F., & Thompson, J. (1987). *Effects of Cooperative Learning on Achievement in Secondary Schools : A Summary of Research*. ERIC. <https://eric.ed.gov/?id=ED288853>

Nolan, J. M., Hanley, B. G., DiVietri, T. P., & Harvey, N. A. (2018). She who teaches learns : Performance benefits of a jigsaw activity in a college classroom. *Scholarship of Teaching and Learning in Psychology*, 4(2), 93-104. APA PsycInfo. <https://doi.org/10.1037/stl0000110>

Olson, M. (1965). *The logic of collective action : Public goods and the theory of groups*. Harvard Univ. Press.

Özdemir, E., & Arslan, A. (2016). The Effect of Self-Regulated Jigsaw IV on University Students' Academic Achievements and Attitudes towards English Course. *Journal of Education and Training Studies*, 4(5), 173-182. ERIC. <http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1095789&lang=fr&site=ehost-live> *

Pansu, P., Dubois, N., & Beauvois, J.-L. (2013). *Dis-moi qui te cite, et je saurai ce que tu vauX : Que mesure vraiment la bibliométrie ?* Presses universitaires de Grenoble.

- Perugini, M., Gallucci, M., & Costantini, G. (2018). A Practical Primer To Power Analysis for Simple Experimental Designs. *International Review of Social Psychology*, 31(1), 20. <https://doi.org/10.5334/irsp.181>
- Pettigrew, T. F. (1997). Chapter 17—Personality and Social Structure : Social Psychological Contributions. In R. Hogan, J. Johnson, & S. Briggs (Éds.), *Handbook of Personality Psychology* (p. 417-438). Academic Press. <https://doi.org/10.1016/B978-012134645-4/50018-4>
- Plante, I. (2012). L'apprentissage coopératif : Des effets positifs sur les élèves aux difficultés liées à son implantation en classe. *Canadian Journal of Education/Revue canadienne de l'éducation*, 35(4), 252-283. <https://www.jstor.org/stable/canajeducrevucan.35.4.252>
- Rego, M. S., & Moledo, M. (2005). Promoting interculturality in Spain : Assessing the use of the Jigsaw classroom method. *Intercultural education, Query date: 2020-05-13 10:45:36*. <https://www.tandfonline.com/doi/abs/10.1080/14675980500212020>
- Reverdy, C. (2016). La coopération entre élèves : Des recherches aux pratiques. *Dossier de veille de l'IFÉ*. <http://ife.ens-lyon.fr/vst/DA/detailsDossier.php?parent=accueil&dossier=114&lang=fr>
- Reverdy, C. (2020). Apprendre et coopérer en classe. *Edubref*. <https://edupass.hypotheses.org/1917>
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638-641. <https://doi.org/10.1037/0033-2909.86.3.638>
- Roseth, C. J., Lee, Y., & Saltarelli, W. A. (2018). Reconsidering jigsaw social psychology : Longitudinal effects on social interdependence, sociocognitive conflict regulation, motivation, and achievement. *Journal of Educational Psychology*, 1-21. <https://doi.org/10.1037/edu0000257> *

- Şahin, A. (2010). Effects of Jigsaw II technique on academic achievement and attitudes to written expression course. *Educational Research and Reviews*, 5(12), 777-787. <https://www.scopus.com/inward/record.uri?partnerID=HzOxMe3b&scp=78751633317&origin=inward> *
- Sahin, A. (2011). Effects of Jigsaw III Technique on Achievement in Written Expression. *Asia Pacific Education Review*, 12(3), 427-435. ERIC. <http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ935156&lang=fr&site=ehost-live> *
- Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological Research : Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology*, 10, 813. <https://doi.org/10.3389/fpsyg.2019.00813>
- Sezek, F. (2013). A New Approach in Teaching the Features and Classifications of Invertebrate Animals in Biology Courses. *Mevlana International Journal of Education*, 3(2), 99-111. <https://doi.org/10.13054/mije.13.25.3.2>
- Sherif, M. (1961). Conformity-Deviation, Norms, and Group Relations. In I. A. Berg & B. M. Bass (Éds.), *Conformity and deviation*. (p. 159-198). Harper and Brothers. <https://doi.org/10.1037/11122-006>
- Sims, S., Anders, J., Inglis, M., & Lortie-Forgues, H. (2020). *Quantifying 'promising trials bias' in randomised controlled trials in education*. 16(20), 31. <https://repec-cepeo.ucl.ac.uk/cepeow/old-style/cepeowp20-16.pdf>
- Skaalvik, E. M., & Skaalvik, S. (2009). Does school context matter? Relations with teacher burnout and job satisfaction. *Teaching and Teacher Education*, 25(3), 518-524. <https://doi.org/10.1016/j.tate.2008.12.006>

- Skaalvik, E. M., & Skaalvik, S. (2010). Teacher self-efficacy and teacher burnout : A study of relations. *Teaching and Teacher Education*, 26(4), 1059-1069. <https://doi.org/10.1016/j.tate.2009.11.001>
- Slavin, R. (1980). Cooperative learning. *Review of educational research*, Query date: 2020-05-13 10:45:36. <https://journals.sagepub.com/doi/abs/10.3102/00346543050002315>
- Slavin, R. (1983). When does cooperative learning increase student achievement? *Psychological bulletin*, Query date: 2020-05-13 10:45:36. <https://psycnet.apa.org/record/1984-07975-001>
- Slavin, R. (1996). Research on cooperative learning and achievement : What we know, what we need to know. *Contemporary educational psychology*, Query date: 2020-05-13 10:45:36. http://www.academia.edu/download/32134643/Cooperative_Learning_-_SLAVIN_Robert.pdf
- Slavin, R. E. (1978). Student teams and comparison among equals : Effects on academic performance and student attitudes. *Journal of Educational Psychology*, 70(4), 532-538. <https://doi.org/10.1037/0022-0663.70.4.532>
- Slavin, R., & Smith, D. (2009). The Relationship Between Sample Sizes and Effect Sizes in Systematic Reviews in Education. *Educational Evaluation and Policy Analysis*, 31(4), 500-506. <https://doi.org/10.3102/0162373709352369>
- Souvignier, E., & Kronenberger, J. (2007). Cooperative learning in third graders' jigsaw groups for mathematics and science with and without questioning training. *British Journal of Educational Psychology*, 77(4), 755-771. APA PsycInfo. <https://doi.org/10.1348/000709906X173297> *
- Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade : An examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences.

- International Journal of Research & Method in Education*, 39(3), 255-267.
<https://doi.org/10.1080/1743727X.2016.1150454>
- Stahl, R. (1994). *The Essential Elements of Cooperative Learning in the Classroom*. ERIC Digest. ERIC. <https://eric.ed.gov/?id=ED370881>
- Stanczak, A. (2020). *La méthode de la "classe puzzle" est-elle efficace pour améliorer l'apprentissage ?*. [Psychologie, Université Clermont Auvergne]. <https://tel.archives-ouvertes.fr/tel-03170791>
- Stanczak, A., Darnon, C., Robert, A., Demolliens, M., Sanrey, C., Bressoux, P., Huguet, P., Buchs, C., Butera, F., & PROFAN Consortium. (2022). Do jigsaw classrooms improve learning outcomes? Five experiments and an internal meta-analysis. *Journal of Educational Psychology*. <https://doi.org/10.1037/edu0000730>
- Stephan, W. G., & Rosenfield, D. (1978). Effects of desegregation on racial attitudes. *Journal of Personality and Social Psychology*, 36(8), 795-804. <https://doi.org/10.1037/0022-3514.36.8.795>
- Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rucker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, 343(jul22 1), d4002-d4002. <https://doi.org/10.1136/bmj.d4002>
- Świątkowski, W., & Dompnier, B. (2017). Replicability Crisis in Social Psychology : Looking at the Past to Find New Pathways for the Future. *International Review of Social Psychology*, 30(1), 111. <https://doi.org/10.5334/irsp.66>
- Tarhan, L., Ayyildiz, Y., Ogunc, A., & Sesen, B. (2013). A jigsaw cooperative learning application in elementary science and technology lessons: Physical and chemical

- changes. *Research in Science & Technological Education*, 31(2), 184-203.
<https://doi.org/10.1080/02635143.2013.811404>
- Tarhan, L., & Sesen, B. (2012). Jigsaw cooperative learning : Acid-base theories. *Chemistry Education Research and Practice*, 13(3), 307-313. <https://doi.org/10.1039/c2rp90004a> *
- Topping, K., Buchs, C., Duran, D., & Keer, H. van. (2017). *Effective peer learning : From principles to practical implementation*. Routledge, Taylor & Francis Group.
- Tran, V. D., & Lewis, R. (2012). The Effects of Jigsaw Learning on Students' Attitudes in a Vietnamese Higher Education Classroom. *International Journal of Higher Education*, 1(2), 9-20. ERIC.
<http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1057193&lang=fr&site=ehost-live> *
- Turkmen, H., & Buyukaltay, D. (2015). Which One Is Better ? Jigsaw II versus Jigsaw IV on the Subject of the Building Blocks of Matter and Atom. *Journal of Education in Science, Environment and Health*, 1(2), 88-94. ERIC.
<http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1105340&lang=fr&site=ehost-live>
- Ural, E., Ercan, O., & Gençoglan, D. M. (2017). The Effect of Jigsaw Technique on 6th Graders' Learning of Force and Motion Unit and Their Science Attitudes and Motivation. *Asia-Pacific Forum on Science Learning and Teaching*, 18. ERIC.
<http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1160102&lang=fr&site=ehost-live> *
- Volpe, Y., & Buchs, C. (2019). Pédagogie coopérative : Pratiques déclarées et facteurs d'appropriation. *Swiss Journal of Educational Research*, 41(1), 99-120.
<https://doi.org/10.24452/sjer.41.1.8>

- Walker, I., & Crogan, M. (1998). Academic performance, prejudice, and the Jigsaw classroom : New pieces to the puzzle. *Journal of Community & Applied Social Psychology*, 8(6), 381-393. APA PsycInfo. [https://doi.org/10.1002/\(SICI\)1099-1298\(199811/12\)8:6<381::AID-CASP457>3.0.CO;2-6](https://doi.org/10.1002/(SICI)1099-1298(199811/12)8:6<381::AID-CASP457>3.0.CO;2-6)
- Wolf, R., Morrison, J., Inns, A., Slavin, R., & Risman, K. (2020). Average Effect Sizes in Developer-Commissioned and Independent Evaluations. *Journal of Research on Educational Effectiveness*, 13(2), 428-447. <https://doi.org/10.1080/19345747.2020.1726537>
- Yapici, H. (2016). Use of Jigsaw Technique to Teach the Unit « Science within Time » in Secondary 7th Grade Social Sciences Course and Students' Views on This Technique. *Educational Research and Reviews*, 11(8), 773-780. ERIC. <http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1098250&lang=fr&site=ehost-live> *
- Yoruk, A. (2016). Effect of jigsaw method on students' chemistry laboratory achievement. *International Journal of Educational Sciences*, Query date: 2020-05-13 10:45:36. <https://www.tandfonline.com/doi/abs/10.1080/09751122.2016.11890547>
- Zahra, R. (2014). The Use of Jigsaw Technique in Improving Students' Ability in Writing a Descriptive Text. *Journal of English and Education*, Query date: 2020-05-13 10:45:36. <http://ejournal.upi.edu/index.php/LE/article/view/748> *
- Ziegler, S. (1981). The Effectiveness of Cooperative Learning Teams for Increasing Cross-ethnic Friendship: Additional Evidence. *Human Organization*, 40(3), 264-268. <https://doi.org/10.17730/humo.40.3.0m0q1143l43r4x44> *

