Évaluer l'utilité d'un EIAH : difficultés rencontrées lors d'une expérience randomisée

▶ **Matthieu CISEL** (IDHN, Institut des humanités numériques, CY Cergy Paris Université)

■ **RÉSUMÉ** • Cette rubrique présente, sous la forme d'un retour d'expérience, une tentative d'évaluer l'utilité d'un module du Carnet Numérique de l'Élève-Chercheur (CNEC), un environnement informatique pour l'apprentissage humain (EIAH) à destination de classes de collège. Cette application vise notamment à étayer la rédaction de propositions scientifiques – hypothèses, protocoles, etc. – dans le cadre de démarches d'investigation. Nous avons étudié la possibilité de mettre en place une expérience randomisée. Cette étude de faisabilité, où nous avons été simultanément chercheur et enseignant, a permis une réflexion sur les difficultés de cette approche, que nous exposons ici. L'accumulation des obstacles nous a conduit à préférer des approches qualitatives pour évaluer l'utilité de l'EIAH.

■ **MOTS-CLÉS** • EIAH, évaluation, démarche d'investigation, expérience randomisée.

■ ***ABSTRACT*** • *This paper presents, as an experience feedback, an attempt to assess the usefulness of a module of the Student-Researcher Digital Notebook, a computer-based learning environment intended for middle schools. This artefact aims at scaffolding the writing of scientific propositions –hypotheses, protocols, etc.– in the context of inquiry learning. We investigated the possibility of setting up a randomized experiment. This feasibility study, during which we were simultaneously a biology teacher and a researcher, led to a reflection on the difficulties of this approach, which we set out here. The accumulation of obstacles led us to prefer qualitative approaches for assessing the usefulness of the learning environment.*

■ ***KEYWORDS*** • *Learning environment, assessment, inquiry learning, randomized experiment.*

1. Introduction

Dans cette rubrique, nous menons une réflexion sur la faisabilité d'une expérience randomisée visant à évaluer en classe l'utilité du module « Brouillon de Recherche » du Carnet Numérique de l'Elève-Chercheur (CNEC, https://www.cnec.fr/l) (Baron *et al.,* 2019), un EIAH développé dans le cadre du projet « Les Savanturiers du Numérique » (Cisel *et al.*, 2019). La fonction de ce module est d'étayer la rédaction de productions scientifiques (hypothèses, protocoles, etc.), dans le cadre de démarches d'investigation menées à l'école primaire et au collège. Après avoir rappelé le contexte politique et scientifique dans lequel s'inscrit ce projet, nous présentons plus avant la démarche de cet article : mettre en lumière les obstacles que nous avons rencontrés en tâchant de mettre en œuvre une telle approche quantitative de l'évaluation. Nous décrivons dans la section 2 certaines caractéristiques saillantes du CNEC, et notamment du module Brouillon de Recherche qui est au centre de cette étude, ainsi que le rôle que nous avons joué dans sa conception. Puis nous présentons en section 3 les principaux obstacles que nous avons identifiés. La discussion s'ouvre ensuite, d'une part, sur la question du caractère généralisable des résultats obtenus et, d'autre part, sur les déterminants de notre réorientation vers des méthodes qualitatives.

1.1. Engouement contemporain pour les expériences randomisées

Selon les promoteurs du courant que les anglo-saxons qualifient d'*Evidence-Based Education* (EBE) (Slavin, 2002), que l'on peut traduire par « Éducation fondée sur des données probantes », les pratiques pédagogiques ont vocation à faire la preuve de leur efficacité, sur la base d'études scientifiques. Celles qui mobilisent le numérique ne font pas exception (Chaptal, 2003). L'enjeu réside alors dans l'identification de ce qui fait preuve. Pour les défenseurs de ce courant, les expérimentations randomisées constituent l'approche dont la valeur probatoire est la plus forte. Celles-ci désignent les protocoles fondés sur la constitution randomisée d'un groupe expérimental avec lequel on met en œuvre la pratique à évaluer, et dont on compare les performances avec celles d'un groupe témoin qui n'est pas concerné par la pratique. Cela va généralement de pair avec l'organisation d'un pré-test, en amont de la pratique évaluée, et d'un post-test, en aval, qui permet de comparer les performances avant et après l'intervention.

Relativement marginales jusqu'à peu dans la littérature scientifique (Cook, 2002), ces expérimentations semblent être de plus en plus appréciées dans les instances dirigeantes de l'Éducation nationale en France, en particulier dans un contexte de renforcement des protocoles d'évaluation standardisés des compétences des élèves (Yerly, 2017). Il est vraisemblable que nous rejouions désormais en France ce qui s'est passé outre-Atlantique il y a de cela quelques décennies, lorsque l'administration Bush lançait le programme *No Child Left Behind* (Abedi, 2004). Les pratiques reconnues comme les plus efficaces étaient mutualisées *via* des sites comme *What works – clearing house* (https://ies.ed.gov/ncee/wwc/), qui répertorient un certain nombre de pratiques évaluées à l'échelle fédérale sur la base de ces méthodes expérimentales. Les Britanniques, à travers l'*Education Endowment Fundation* (https://educationendowmentfoundation.org.uk/), ont suivi une démarche relativement similaire.

En France, l'ingénierie didactique (Artigue, 1989) a précédé de plusieurs décennies l'engouement contemporain pour l'expérimentation randomisée, mettant en œuvre des approches analogues. Néanmoins, un changement d'échelle semble être recherché, tant en termes de nombre d'expérimentations que de nombre d'enseignants participant aux expérimentations. L'analyse des sujets présentés au colloque des 15 et 16 octobre 2019 consacré à la présentation des projets eFRAN (espaces de Formation, de Recherche et d'Animation Numérique) suggère notamment que les organismes de financement et les rectorats semblent de plus en plus séduits par les projets qui mettent en avant des expérimentations randomisées. On peut craindre qu'une telle orientation ne favorise la propagation de l'idée selon laquelle la recherche hexagonale doit se concentrer sur ces approches, imitant ainsi des approches promues dans le monde anglo-saxon.

1.2. Présentation de la démarche

Cette rubrique vise à faire le point sur les problèmes rencontrés lors de l'étude de faisabilité d'une expérience randomisée. Si la tendance actuelle doit amener à convaincre un nombre croissant d'enseignants de s'impliquer dans des évaluations de leurs pratiques, il convient de communiquer davantage auprès des praticiens sur les contraintes inhérentes aux protocoles d'évaluation. Les enseignants qui se lancent pour la première fois dans une expérimentation randomisée, tout comme les chercheurs, gagneraient à mieux les appréhender au moment de la mise au point des protocoles.

Ces considérations nous ont amené à proposer cette contribution fondée sur une expérience personnelle, où nous nous sommes placé simultanément dans le rôle du chercheur et dans le rôle de l'enseignant, difficulté méthodologique de taille. Les obstacles que nous avons rencontrés découlent dans une large mesure de contingences spécifiques à notre terrain et à nos problématiques ; ils peuvent néanmoins illustrer des tensions plus générales associées ce type de démarche. Nous nous inscrivons ainsi dans la lignée de travaux critiques de cette approche (Baron et Bruillard, 2007 ; Biesta, 2010).

2. Contexte de la réflexion et éléments de description du protocole

Cette section présente d'abord le contexte dans lequel a été menée notre réflexion. Nous revenons ensuite sur les travaux qui ont inspiré la mise au point du protocole d'expérimentation, sur le module Brouillon de Recherche et sur les étayages qu'il porte, pour nous pencher enfin sur le protocole envisagé.

2.1. Présentation du projet de conception du CNEC

La vocation du CNEC est d'instrumenter les projets du programme « Savanturiers », fondé en 2013 par une ancienne professeure des écoles et à destination du primaire et du secondaire (Barbier, 2019). Le programme vise à développer des mini-projets de recherche, encadrés par des mentors généralement issus du milieu académique, afin d'initier les élèves aux méthodes de l'investigation scientifique. Ces derniers prennent une part active aux différentes étapes de la démarche, de la formulation de la question de recherche à l'interprétation des résultats. Le développement informatique est assuré par l'entreprise *Tralalère*, qui est propriétaire du code et décide en dernière instance des développements effectués. Elle n'est pas en position de prestataire, mais de partenaire. Les académies de Paris et de Créteil facilitent l'accès au terrain et effectuent des retours utilisateurs, et le laboratoire EDA, où j'effectuais mon post-doctorat, est responsable de la conduite d'une recherche, dont les axes n'étaient pas pleinement déterminés au moment du dépôt du projet.

La participation à l'évaluation de l'utilité du CNEC, une fois les prototypes suffisamment robustes pour être mobilisés en classe, faisait explicitement partie des missions attribuées au laboratoire au moment de la constitution du consortium. Nous avons pour cette raison exploré la diversité des approches possibles (Tricot *et al.*, 2003 ; Nogry *et al.*, 2004 ;

Jamet, 2006). Le positionnement des financeurs et décideurs du programme eFRAN en faveur des expérimentations randomisées nous a amené à envisager une approche quantitative de l'évaluation. Une étude de faisabilité s'imposait en amont avant de mettre en place des expérimentations mobilisant près d'une dizaine de classes. Pour cette raison, nous avons pris la responsabilité de deux classes de quatrième en tant qu'enseignant de SVT, avec l'accord des autorités académiques, afin de mener cette étude de faisabilité. L'étude visait à identifier les biais qui pourraient affecter les résultats de l'expérience, pour les prendre en compte dans la mise au point de la version finale du protocole. L'objectif de la démarche consistait à évaluer l'utilité d'un module en particulier, le Brouillon de Recherche, dont la vocation est de structurer la rédaction de productions scientifiques comme des questions de recherche, des hypothèses ou des protocoles. La démarche était assez proche de celle qui a été suivie par Bonnat (2017) ou Saavedra (2015) dans leurs thèses respectives. L'échantillon d'une soixantaine d'élèves avec lequel nous pouvions expérimenter était *a priori* d'une taille suffisante pour tester la faisabilité des protocoles expérimentaux envisagés.

Dans la mesure où nous avons enseigné dans deux classes pour mettre en œuvre des expériences en éducation, nous parlerons ici d'un double point de vue : celui du chercheur, co-concepteur d'une technologie éducative qu'il souhaite évaluer, et celui du professeur de sciences de la vie et de la terre, prenant ses premières classes spécialement pour l'occasion.

2.2. Fonctionnement des étayages du Brouillon de Recherche

Le CNEC est composé de six modules interconnectés (Cisel, Barbier et Baron, 2019). Le module dit du Brouillon de Recherche a été choisi pour le protocole expérimental car il permet de mesurer des performances individuelles en matière de rédaction de propositions scientifiques. Selon la logique de co-conception qui caractérisait ce projet eFRAN, nous nous sommes considérablement impliqués dans le développement de son cahier des charges, avant même que ne soit envisagée l'étude de faisabilité présentée ici.

Sur la base de travaux de synthèse consacrés à la conception d'EIAH (Quintana *et al.,* 2004) et en amont de l'étude de faisabilité, nous avons effectué un certain nombre de recommandations aux développeurs, et notamment celle de développer des étayages. Nous nous sommes inspirés d'environnements comme *Knowledge Forum* produit au Canada

(Scardamalia et Bereiter, 2003), *WISE* de Stanford (Slotta et Linn, 2009), *LabNbook* (Girault et d'Ham, 2014 ; Wajeman *et al.,* 2015), ou *Hypothesis Scratchpad* (Joolingen et Jong, 1991) pour produire des étayages visant à aider les élèves à formuler des hypothèses, des questions, etc.

Le principe, illustré en figure 1, est le suivant : au lieu de partir d'une page blanche pour commencer une recherche, les élèves choisissent, dans une liste, une section correspondant à une étape donnée de la recherche (Question, Hypothèse, Protocole, Collecte, Analyse et Interprétation, Conclusion). Un brouillon de recherche apparaît alors et, en fonction de la section choisie, des ouvreurs de phrase spécifiques apparaissent, afin de mieux faire comprendre ce qui est attendu.

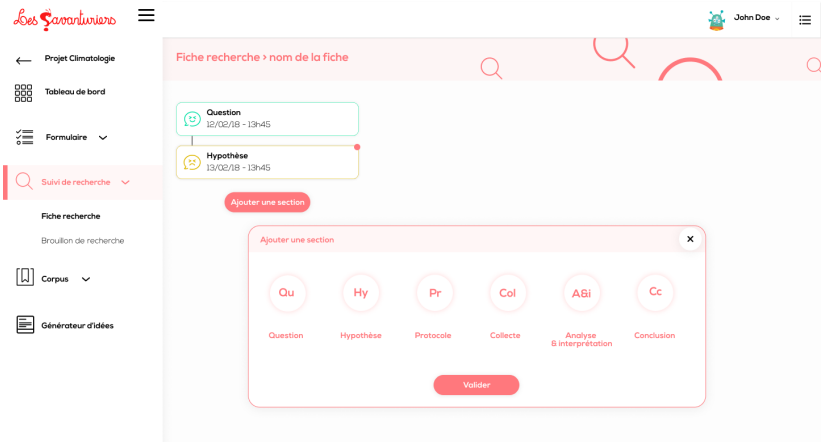

Figure 1 • Interface de la Fiche-Recherche permettant de choisir un Brouillon de Recherche spécifique à une étape de la démarche

Par exemple, pour la rédaction d'une question, nous pouvons proposer un ouvreur de phrase comme : *Nous cherchons à savoir pourquoi,* ou *Nous cherchons à savoir comment,* selon le type de question que l'on se pose. Pour un protocole, on pourra proposer un étayage comme *La première étape du protocole consiste à,* pour indiquer qu'il s'agit de dresser une liste d'actions, à la manière d'une recette de cuisine. Lorsque l'élève clique sur l'étayage, celui-ci s'affiche alors dans le champ texte situé en dessous, champ qu'il complète en fonction de ce qu'il désire écrire. Des questions d'autoévaluation sont ensuite proposées pour amener l'élève à une certaine réflexivité sur sa production. Une fois ces différentes étapes passées, l'utilisateur peut envoyer la phrase ainsi construite à l'enseignant pour évaluation.

2.3. Protocole envisagé pour l'évaluation

Le protocole retenu est inspiré des travaux portant sur l'évaluation empirique de tels étayages, que cela soit dans le monde anglo-saxon (Azevedo *et al.,* 2004) ou dans des recherches francophones portant sur le *LabNBook* (Bonnat, 2017 ; Saavedra, 2015). Ce dernier EIAH instrumente la conception expérimentale dans les sciences naturelles. Les auteurs définissent des critères de qualité pour les productions des élèves et s'attachent à montrer que les étayages proposés améliorent la qualité des productions écrites.

Dans le cadre de notre étude de faisabilité, les élèves du groupe témoin comme ceux du groupe expérimental se voient proposer un phénomène à expliquer – blanchiment de récifs coraliens, disparition des amphibiens en Amazonie, etc. – et doivent proposer des explications. Un groupe travaille à partir des étayages numériques du Brouillon de Recherche, l'autre reçoit des questions équivalentes sur papier. À l'issue des tests organisés a posteriori, lors de contrôles, nous comparons les hypothèses produites par les deux groupes sur la base d'un certain nombre de critères : adéquation avec le problème posé, existence d'une conséquence vérifiable, c'est-à-dire une manière de la tester, qualités rédactionnelles. Si l'application a une valeur ajoutée par rapport au papier, on devrait constater que les élèves utilisant le Brouillon de Recherche réalisent *a priori* des productions de meilleure qualité que ceux qui disposent d'instructions sur papier, au regard des différents critères retenus. Le choix de retenir le papier pour l'activité témoin est à mettre au regard de l'objectif de l'expérience : montrer la valeur ajoutée du Brouillon de Recherche sur les activités « papier » que mettent déjà en place les enseignants.

3. Obstacles rencontrés lors de la mise en œuvre d'une expérimentation randomisée

Les problèmes rencontrés lors de la mise en place d'un début d'expérience randomisée ont été de plusieurs ordres. Nous n'avons pas pu constituer des groupes expérimentaux de manière randomisée, c'est-à-dire en créant deux groupes au sein d'une classe qui réaliseront des tâches différentes, et en attribuant aléatoirement les élèves de la classe à l'un des deux groupes. Ensuite, la comparabilité des conditions de l'expérience entre groupes témoins et groupes expérimentaux a été mise à mal par un certain nombre d'aléas : problèmes de connexion et de connectivité, aléas de la vie scolaire. Enfin, nous avons été mis en difficulté pour mesurer des performances authentiquement individuelles, étant incapables d'empêcher totalement les interactions entre élèves.

3.1. La randomisation, un casse-tête logistique

Les contraintes posées par la randomisation soulèvent plusieurs problèmes de logistique, traités dans les paragraphes qui suivent.

3.1.1. De l'importance de la randomisation

Nous avons été confrontés en premier lieu au problème de la constitution des groupes. Utiliser une classe comme témoin et l'autre classe comme groupe expérimental est contraire au principe de randomisation que l'on doit suivre lors de la constitution de l'échantillon, quel que soit le nombre de classes suivies. La répartition des élèves par seconde langue illustre le propos. Les classes ayant pour seconde langue l'allemand rassemblent généralement les meilleurs éléments, ce qui peut biaiser l'analyse. C'était d'ailleurs la situation à laquelle nous faisions face en tant qu'enseignant, avec une classe dominée par les germanistes et une autre par les hispanisants. Pour des contrôles équivalents, il existait deux points de différence entre les moyennes des deux classes au premier trimestre. Si nous avions utilisé le CNEC avec la meilleure des deux classes, les performances auraient sans doute été très supérieures à celles de la classe témoin, sans que l'on puisse déterminer si cette différence de performance était due à la pratique mise en place, qu'elle implique ou non du numérique, ou si elle était simplement due au fait que l'expérience avait été menée avec la meilleure des deux classes.

Il fallait donc trouver un moyen de faire des demi-groupes de manière aléatoire au sein d'une classe. Or, quand bien même aurions-nous eu le budget pour travailler en demi-groupe, la division aléatoire en demi-groupes n'a rien de naturel. Celle-ci pose en effet beaucoup de problèmes pour l'établissement, problèmes que nous évoquons dans la sous-section suivante.

3.1.2. Articuler contraintes des protocoles de recherche et contraintes institutionnelles

Les classes sont généralement divisées en deux sur la base soit du classement alphabétique, soit d'un autre critère : par exemple, les germanistes sont regroupés ensemble tandis que les hispanisants forment un second groupe, si la langue vivante choisie constitue le critère de partition. Compte tenu de ce mode de fonctionnement, cela peut poser de nombreux problèmes logistiques qu'un algorithme coupe aléatoirement en deux la classe pour créer des demi-groupes. Si pendant qu'un demi-groupe suit un cours, l'autre moitié de la classe n'a pas cours, cela ne pose pas de problème. Mais cela fonctionne

rarement ainsi. Pendant qu'un demi-groupe est en SVT, par exemple, l'autre est généralement dans un autre cours, optimisation de l'emploi du temps des élèves oblige. Par suite, il aurait fallu expliquer à l'autre enseignant pourquoi nous avions coupé la classe en deux de manière aléatoire, sans suivre les divisions pratiquées habituellement. Par ailleurs, il aurait également été nécessaire que nous intervenions plus en amont, en juin, pour expliquer la démarche à la direction de l'établissement, afin que les protocoles de recherche soient pris en compte dans la mise en place des emplois du temps. Il aurait fallu au moins leur demander de produire des emplois du temps complexes et susceptibles d'entraîner un certain nombre de désagréments pour les élèves.

Tout en travaillant en classe entière, on aurait pu couper aléatoirement la classe en deux, chacune des deux moitiés se consacrant à une activité différente. C'est, après tout, le principe de la différenciation pédagogique. Néanmoins, même pour un enseignant expérimenté, gérer deux activités distinctes dans un même environnement peut représenter une difficulté. Cela complexifie singulièrement le bon déroulement du protocole expérimental. En particulier, du fait de la difficulté à contrôler la temporalité de la séance, il devient difficile de s'assurer que les élèves des deux groupes réalisent l'activité dans des conditions comparables. Cette considération va nous amener, dans la section qui suit, à développer davantage la question de la comparabilité des conditions de l'expérience.

3.2. Comparabilité des conditions de l'expérience

Pour des raisons pratiques, nous avons renoncé à créer des sous-groupes au sein d'une classe, préférant une classe comme témoin, et l'autre comme groupe expérimental. L'objet de l'étude de faisabilité était d'identifier les principaux problèmes potentiels, et une telle approche quasi expérimentale suffisait pour ce faire. Pour pouvoir être en mesure de comparer les performances des deux classes, la première étape consistait à concevoir des activités équivalentes entre le groupe témoin travaillant avec du papier. Nous avions produit des consignes identiques pour nous en assurer. Il fallait ensuite mener les activités dans des conditions comparables, et notamment respecter scrupuleusement des questions de temporalités – durée des activités, mais aussi parfois écart temporel entre deux activités. Sur ces points, les aléas de la vie scolaire ont constitué des obstacles de taille.

3.2.1. Le temps de connexion à l'application, un exemple de biais récurrent

Le temps de connexion à l'application a constitué une variable qui interférait avec la comparabilité des conditions de l'expérience entre groupes. Il est nécessaire pour se connecter d'utiliser des identifiants spécifiques. Il faut généralement plus d'un quart d'heure pour que tous les élèves soient connectés et le désordre qu'impliquent les difficultés récurrentes de connexion impacte le bon déroulé de l'activité. Tous les élèves ne mettent pas le même temps pour se connecter, tantôt pour des raisons techniques, tantôt pour des raisons de comportement et de maîtrise des outils.

Théoriquement, il faudrait attendre que les derniers élèves se soient connectés pour laisser les premiers commencer l'activité, sans quoi les conditions dans lesquelles les élèves réalisent l'activité ne sont pas comparables. Les uns auront travaillé une demi-heure là où les mêmes y auront passé près du double. Il n'est dès lors pas étonnant que les élèves n'aient pas avancé autant à la fin de la séance et ne présentent pas les mêmes performances. Il faudrait également s'assurer que les élèves qui travaillent sur papier aient le même temps de travail et le modifier en fonction du temps dont auront disposé les élèves du groupe expérimental. Il a été impossible de contrôler de manière rigoureuse cette variable. Pour que les problèmes de connexion soient équivalents entre groupes, on aurait également pu faire travailler le groupe témoin avec une version du Brouillon de Recherche sans étayage, et faire travailler le groupe expérimental avec le module portant les étayages. Mais ce choix nous aurait amenés à changer l'objectif de notre expérience, comparer avec les activités équivalentes sur papier.

Second point, la performance des élèves varie selon les jours de la semaine ; elle est sensiblement différente avant les vacances ou après les vacances, en début ou en fin de matinée, et pour cette raison le calendrier de l'expérience doit être adapté en fonction. Tous les événements qui changeaient le calendrier de l'expérience pour un seul des deux groupes remettaient en question la validité des résultats. Plusieurs aléas peuvent imposer de modifier le calendrier de l'expérience. La première fois, ce fut un défaut de connectivité. La bande passante de l'établissement était encore limitée au moment de l'étude et un autre enseignant avait également décidé d'utiliser les tablettes pour l'une de ses activités ; cela affectait la capacité de certains des élèves à naviguer sur Internet. Pendant vingt minutes, l'application a été inutilisable. Le test dut être reporté, modifiant le calendrier

de l'expérience. Parfois, il arrive que ce type de problème ne concerne qu'une poignée d'élèves. Existe alors la possibilité de les sortir de l'étude, mais ce choix réduit peu à peu la taille de l'échantillon considéré.

3.2.2. Diminution rapide du nombre d'élèves incorporables dans l'étude

Il existe plusieurs raisons pour lesquelles un élève ne peut pas passer au moment requis l'expérience dans de bonnes conditions : retard, problème technique avec la tablette, événement impromptu. Nous avons fait le choix de sortir ces élèves de l'expérience pour ne pas biaiser les résultats et maintenir la comparabilité des conditions de passation de l'expérience. La multiplication des incidents a conduit à une diminution importante de la taille des échantillons. En définitive, seule une moitié des élèves de l'étude a pu être prise en compte.

Cette diminution serait sans doute moins problématique si l'on mobilisait des dizaines de classes, pour compenser, mais la mise en place des protocoles s'en trouverait d'autant plus complexifiée. On peut raisonnablement s'attendre à ce que la présence physique d'un chercheur dans la salle augmente la rigueur du suivi des protocoles. Or plus une expérimentation gagne en ampleur, plus il est difficile de maintenir cette présence dans toutes les classes, et donc de s'assurer que les conditions de passation des tests sont rigoureusement similaires. Ceci est aussi vrai pour la question de la mesure de performances authentiquement individuelles, comme nous allons le voir dans les paragraphes qui suivent.

3.3. Une mesure de performances véritablement individuelles ?

3.3.1. Le contrôle, une stratégie pour limiter les interactions individuelles

Nous avons choisi de nous servir des contrôles comme lieu privilégié de mesure des performances. Ce choix tient au fait que l'on peut y limiter les interactions entre élèves, ce qui permet de mesurer plus rigoureusement des performances individuelles. Sans quoi, on mesure souvent la performance d'une rangée de trois élèves qui s'entraident. S'il existe d'autres manières de limiter les interactions entre élèves, cette approche a eu le mérite de la simplicité dans le contexte qui était le nôtre. Par ailleurs, il y a un enjeu pour l'élève, qui, *a priori,* est plus engagé lorsqu'il sait qu'il va être noté. Si l'on organise une activité qui ne pèse pas sur sa scolarité, l'influence de son rapport aux activités non notées entre dans l'équation.

Typiquement, les élèves devaient soumettre individuellement des hypothèses quant aux causes de l'érosion de la biodiversité, et proposer des mécanismes pour expliquer la disparition de telle ou telle espèce : blanchiment des coraux, raréfaction des moineaux en région parisienne, surmortalité des batraciens en Amazonie. Les bénéfices apparents de la stratégie du contrôle ont néanmoins été partiellement annulés par les caractéristiques de l'établissement dans lequel nous exercions. Les salles dans lesquelles sont organisés les contrôles sont trop petites pour que tous les élèves puissent travailler avec le même sujet. Or le fait de proposer des sujets sensiblement différents nuit à la comparabilité des performances entre élèves, qui devraient avoir dans l'idéal des tâches rigoureusement identiques. Quand bien même les salles auraient été suffisamment spacieuses, d'autres problèmes se seraient posés. Les élèves se seraient régulièrement communiqué les sujets des contrôles entre classes dès qu'ils l'auraient pu, sur le temps d'interclasse, ce qui aurait faussé les résultats. Et même en expliquant que le test n'était pas noté et qu'il ne s'agissait que d'une expérience à portée scientifique, ce type d'interactions aurait pu avoir lieu, étant donné les habitudes de discussion des élèves.

Dans la mesure où le choix d'utiliser les contrôles est particulier à notre cas d'étude et ne représente pas une pratique courante dans les protocoles d'évaluation, nous nous cantonnons ici à éclairer les conséquences de ce choix. Si cette approche représente en principe une stratégie efficace pour limiter les interactions entre élèves et mesurer des performances véritablement individuelles, elle n'est pas parfaite. D'une part, il est des interactions qui échappent au contrôle de l'enseignant. D'autre part, l'enseignant peut lui aussi vouloir interagir de manière privilégiée avec tel ou tel élève, pour des raisons légitimes sur le plan pédagogique. Nous présentons ainsi un dilemme personnel dans le dernier paragraphe de cette section, qui pourrait trouver des échos chez les praticiens engagés dans ce type d'expérimentation.

3.3.2. Une tension entre deux consciences professionnelles : la question de l'assistance aux élèves

Le problème qui suit illustre le tiraillement qu'implique parfois la « double casquette » d'expérimentateur et d'enseignant : il s'agit de la question de l'aide aux élèves en difficulté. En principe, si l'on veut que les conditions de passation de l'expérience soient strictement les mêmes pour tous les élèves, il n'est pas question d'intervenir de manière différenciée, que cela soit pendant les pré-tests, les post-tests, ou au cours de la pratique à évaluer.

Par conséquent, si un élève est en difficulté et qu'il a besoin d'une brève explication pour se lancer, on peut être amené à lui refuser au nom du suivi strict du protocole d'évaluation. Pendant les séances où nous utilisions le CNEC pour faire travailler les élèves sur la rédaction d'hypothèses, hors pré-test et post-test, il aurait fallu apporter strictement la même aide à tous les élèves si l'on voulait suivre le protocole à la lettre. Cette contrainte est particulièrement délicate, car c'est dans ces moments-là que les aides personnalisées sont les plus nécessaires. Ce tiraillement découle certes de la double contrainte associée à notre double casquette, position très particulière au demeurant. Néanmoins, on peut également arguer du fait que les enseignants qui s'engagent dans de tels protocoles ont aussi une forme de double casquette. S'ils sont avant tout praticiens, ils endossent dans une large mesure un rôle d'expérimentateur lorsqu'ils suivent un protocole à la lettre. Si nous ne prétendons pas tirer des conclusions générales à partir de ce cas d'étude, il pourrait être intéressant de se pencher sur la manière dont les enseignants qui ont été mis dans des situations analogues perçoivent cette double contrainte, et surtout, comment ils y réagissent.

4. Discussion

En endossant simultanément le rôle de l'enseignant et celui du chercheur, nous avions en théorie un large contrôle sur le déroulement des séances. Cette approche permet d'éviter d'avoir à composer avec des enseignants dont il aurait fallu modifier les pratiques, au nom d'un protocole qu'ils n'avaient pas conçu eux-mêmes. Cette approche induit des biais, comme la tension entre posture de chercheur et celle d'enseignant. Du fait de cette tension, il arrive bien souvent que l'on déroge à certains éléments d'un protocole pour des raisons pédagogiques, quitte à mettre en péril la validité de l'expérience dans son ensemble. Ces difficultés affectent la validité interne de notre travail, raison pour laquelle nous nous sommes prononcés contre une extension du protocole à de nombreuses classes.

Le dernier élément ayant contribué au choix de renoncer à une approche expérimentale quantitative est celui de la généralité des résultats que nous aurions pu obtenir *via* une telle expérimentation randomisée, ou en d'autres termes, le problème de la validité externe de notre travail. Avant de se lancer dans une expérimentation randomisée, il convient de se pencher sur les conditions à remplir pour que le propos puisse être généralisé, ce qui renvoie à la question dite de la validité externe des résultats (Onwuegbuzie, 2000 ; Onwuegbuzie et Johnson, 2006).

En premier lieu, les propriétés de l'EIAH doivent être exposées clairement et la méthode d'évaluation adaptée à une question de recherche, dont la portée se doit de dépasser le cas d'étude considéré. En second lieu, les performances des élèves doivent être influencées par les choix de conception plus que par des problèmes techniques. C'est en soi un défi lorsque l'on travaille avec le numérique et en particulier avec des prototypes, notamment car les bugs interfèrent régulièrement avec le bon déroulé des expériences. Dans la mesure où les problèmes techniques ou des détails de l'ergonomie jouent sur les performances des élèves, il devient difficile d'identifier les relations de causalité entre propriétés de l'EIAH et ces dernières. Si le propos semble relever du truisme, il mérite d'être rappelé au regard du contenu des débats contemporains. Les échanges informels dans des rencontres comme les colloques eFRAN suggèrent notamment que les décideurs politiques, recteurs et financeurs, encouragent l'expérimentation randomisée avec des outils numériques peu stabilisés. Ceci vise vraisemblablement à soutenir l'innovation technologique d'acteurs parfois privés, tout en finançant la recherche par la même occasion.

Les obstacles que nous avons recensés ont mis à l'épreuve la pertinence de la mise en œuvre d'une approche quantitative pour l'évaluation de l'utilité du Brouillon de Recherche. Nous avons dû abandonner l'idée en l'état pour préférer des approches plus qualitatives lors de la phase d'évaluation. Des entretiens avec les enseignants (Cisel, Barbier et Baron, 2019) nous ont par exemple permis de mieux appréhender la manière dont les enseignants se sont approprié les étayages. Entre autres axes de travail, nous avons analysé l'impact des modalités de médiation des étayages par l'enseignant sur la manière dont la terminologie qui leur était associée a pu être détournée. Par exemple, la mise à disposition des étayages a encouragé certains enseignants à utiliser le terme *expérience* pour qualifier une recherche documentaire sur Internet. Cela soulève des questions quant au caractère possiblement contre-productif des étayages, en l'absence d'une formation dédiée pour les enseignants quant à leur utilisation. Des entretiens semi-directifs nous ont permis de mieux comprendre les réticences que les enseignants peuvent exprimer quant à l'apport, par le CNEC, de termes dont ils voudraient contrôler l'introduction. L'exposition des élèves au terme *Hypothèse* dans le Brouillon de Recherche gênait par exemple des praticiens exerçant en CE2, qui préféraient tantôt dédier une séance à l'explicitation de ce terme, tantôt ne pas l'utiliser. Ces résultats ont été obtenus par des approches qualitatives, pour un investissement en

ressources humaines somme toute bien inférieur à celui qui aurait été nécessaire pour une expérimentation randomisée de grande ampleur, même s'il va sans dire que les objectifs de ces deux types de recherche sont distincts.

> Nul ne peut être contre l'application de protocoles de recherche précis, débouchant sur des résultats quantitatifs. Néanmoins, cette expérience de post-doctorat nous a amené à rejoindre la position de ceux (Cook, 2002 ; Baron et Bruillard, 2007 ; Biesta, 2010) qui pointent le fait, qu'en classe, les protocoles d'expérimentations randomisées sont souvent difficiles à appliquer de manière suffisamment rigoureuse pour permettre de produire des résultats scientifiques robustes. Ceci est *a priori* d'autant plus vrai que le nombre de classes concernées est important. Un suivi rapproché est nécessaire pour tendre vers la rigueur recherchée ; la difficulté à suivre chaque séance est telle qu'il est nécessaire que les enseignants acceptent de suivre rigoureusement les protocoles, même en l'absence du chercheur. Certes, ces protocoles sont parfois négociés avec les praticiens afin de bénéficier de leur expertise et accroître leur implication en évitant d'imposer une démarche depuis l'extérieur. Néanmoins, le changement d'échelle voulu notamment par les institutions politiques pourrait réduire la marge de négociation entre chercheurs et enseignants, et accroître les difficultés à assurer un suivi rigoureux des protocoles mis au point. L'injonction à développer davantage les expérimentations randomisées, qui se traduit dans la nature des projets financés, pourrait échouer à susciter l'adhésion de la majorité des chercheurs et des enseignants concernés, ce qui, en définitive, pourrait se révéler contre-productif sur le plan scientifique..

RÉFÉRENCES

Abedi, J. (2004). The no child left behind act and English language learners: Assessment and accountability issues. *Educational Researcher, 33*(1), 4-14.

Artigue, M. (1989). Ingénierie didactique. *Publications mathématiques et informatiques de Rennes, 6*, 124-128.

Azevedo, R., Cromley, J. G. et Seibert, D. (2004). Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia? *Contemporary Educational Psychology, 29*(3), 344-370. https://doi.org/10.1016/j.cedpsych.2003.09.002

Barbier, C. (2019). *Vers l'appropriation de nouveaux instruments par des enseignants : le cas de la démarche d'Éducation par la Recherche et du Carnet Numérique de l'Élève-Chercheur* [mémoire de master non publié]. Université Paris-Descartes, Paris, France.

Baron, G.-L., Barbier, C. et Cisel, M. (2019). *Synthèse sur la recherche Carnet numérique de l'élève-chercheur* [rapport de recherche, Université Paris-Descartes, Paris, France]. https://halshs.archives-ouvertes.fr/halshs-02406707

Baron, G.-L. et Bruillard, E. (2007). ICT, educational technology and educational instruments: Will what has worked work again elsewhere in the future? *Education and Information Technologies, 12*(2), 71–81. https://doi.org/10.1007/s10639-007-9033-9

Biesta, G. J. J. (2010). Why 'What works' still won't work: From evidence-based education to value-based education. *Studies in Philosophy and Education, 29*(5), 491-503. https://doi.org/10.1007/s11217-010-9191-x

Bonnat, C. (2017). *Étayage de l'activité de conception expérimentale par un EIAH pour apprendre la notion de métabolisme cellulaire en terminale scientifique* [thèse de doctorat, Université Grenoble Alpes, Grenoble, France]. https://tel.archives-ouvertes.fr/tel-01562709/

Chaptal, A. (2003). *L'efficacité des technologies éducatives dans l'enseignement scolaire : analyse critique des approches française et américaine.* L'Harmattan.

Cisel, M., Barbier, C. et Baron, G.-L. (2019). *Rapport scientifique de synthèse de la recherche Cahier numérique de l'élève chercheur (CNEC).* Université Paris-Descartes, laboratoire EDA. https://hal.archives-ouvertes.fr/hal-02278348

Cook, T. D. (2002). Randomized experiments in education: Why are they so rare? *Educational Evaluation and Policy Analysis, 24*(3), 175-199. https://www.jstor.org/stable/3594164?seq=1

Girault, I. et d'Ham, C. (2014). Scaffolding a complex task of experimental design in chemistry with a computer environment. *Journal of Science Education and technology, 23*(4), 514-526. https://link.springer.com/article/10.1007/s10956-013-9481-5

Jamet, E. (2006). Une présentation des principales méthodes d'évaluation des EIAH en psychologie cognitive. *Sticef, 13,* 129-146.

Joolingen, W. R. Van et Jong, T. D. (1991). Supporting hypothesis generation by learners exploring an interactive computer simulation. *Instructional Science, 20*(5-6), 389-404. https://doi.org/10.1007/BF00116355

Nogry, S., Jean-Daubias, S. et Ollagnier-Beldame, M. (2004). Évaluation des EIAH : une nécessaire diversité des méthodes. Dans F. Peccoud, C. Moreau et C. Frasson (dir.), *Actes du colloque Technologies de l'information et de la connaissance dans l'enseignement supérieur et l'industrie (TICE 2004*) (p. 265-271). Université de Technologie Compiègne. https://edutice.archives-ouvertes.fr/edutice-00000729/document

Onwuegbuzie, A. J. (2000). *Expanding the framework of internal and external validity in quantitative research.* ERIC. http://eric.ed.gov/?id=ED448205

Onwuegbuzie, A. J. et Johnson, R. B. (2006). The validity issues in mixed research. *ResearchGate*, *13*(48), 48-63.

Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G. et Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *Journal of the Learning Sciences, 13*(3), 337-386. https://doi.org/10.1207/s15327809jls1303_4

Saavedra, R. (2015*). Étayer le travail des élèves avec la plateforme LabBook pour donner davantage de sens aux activités expérimentales réalisées par des élèves de première S* [thèse de doctorat, Université Grenoble Alpes, Grenoble, France]. https://tel.archives-ouvertes.fr/tel-01280377

Scardamalia, M. et Bereiter, C. (2003). Knowledge building environments: Extending the limits of the possible in education and knowledge work. *Encyclopedia of Distributed Learning*, 269-272.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational practice and research. *Educational Researcher, 31*(7), 15-21.

Slotta, J. D. et Linn, M. C. (2009). *WISE science: Web-based inquiry in the classroom.* Teachers College Press.

Tricot, A., Plégat-Soutjis, F., Camps, J.-F., Amiel, A., Lutz, G. et Morcillo, A. (2003). Utilité, utilisabilité, acceptabilité : interpréter les relations entre trois dimensions de l'évaluation des EIAH. Dans C. Desmoulins, P. Marquet, D. Bouhineau (dir.), *Environnements informatiques pour l'apprentissage humain. Actes de la conférence EIAH 2003* (p. 391-402). INRP. https://edutice.archives-ouvertes.fr/edutice-00000154/document

Wajeman, C., Girault, I., d'Ham, C. et Marzin-Janvier, P. (2015). Students' reflection on experimental design during an innovative teaching sequence with Labbook. Dans J. Lavonen, K. Juuti, J. Lampiselkä, A. Uitto et K. Hahl (dir.), *Science education research: Engaging learners for a sustainable future. Proceedings of the 11th European Science Education Research Association Conference (ESERA 2015)* (p. 12-24). University of Helsinki. https://hal.archives-ouvertes.fr/hal-01278747

Yerly, G. (2017). Évaluation des apprentissages en classe et évaluation à grande échelle : quels sont les effets des épreuves externes sur les pratiques évaluatives des enseignants ? *Mesure et évaluation en éducation, 40*(1), 33-60.