



Détection de déficits d'auto-évaluation et d'auto-efficacité et remédiation dans un EIAH

► **Thomas SERGENT** (Sorbonne Université, CNRS, LIP6, F-75005 Paris, France — Lalilo), **Morgane DANIEL** (Lalilo), **François BOUCHET**, **Thibault CARRON** (Sorbonne Université, CNRS, LIP6, F-75005 Paris, France)

■ **RÉSUMÉ** • Des travaux de recherche montrent que la capacité à autoréguler son apprentissage a un impact significatif positif sur les résultats scolaires. Nous présentons ici une étude visant à détecter des déficits d'autorégulation de l'apprentissage pour de jeunes élèves, dans le contexte d'une application web d'apprentissage de la lecture. À partir des réponses de 467 116 élèves à deux questions évaluant la difficulté perçue et la difficulté voulue, nous proposons une définition opérationnelle de différentes formes de déficits et mesurons ensuite l'impact de deux stratégies de remédiation pour les réduire. Les résultats soulignent la possibilité d'étayer les compétences d'apprentissage autorégulé dans une application web dès le plus jeune âge, tout en apprenant une autre compétence.

■ **MOTS-CLÉS** • apprentissage autorégulé, *Big data*, fouille de données éducatives, école primaire, apprentissage de la lecture.

■ **ABSTRACT** • *Research shows that the ability to self-regulate one's learning has a significant positive impact on academic outcomes. Here, we present a study to detect some self-regulated learning (SRL) deficits for young students, in the context of a web-based literacy application. Based on the responses from 467,116 student to questions assessing perceived and desired difficulty, we propose an operational definition of different forms of deficits and then measure the impact of two remediation strategies to address them. The results demonstrate that scaffolding SRL skills in a web-based application at an early age, while learning another skill, is indeed possible.*

■ **KEYWORDS** • *Self-regulated learning, Big data, educational data mining, primary school, reading learning.*

1. Introduction

L'apprentissage autorégulé (AAR) est un cycle en trois phases qui se répète à chaque nouvelle tâche à laquelle l'apprenant est confronté (Zimmerman, 2008). Zimmerman identifie la phase d'anticipation pendant laquelle l'apprenant se prépare à la tâche (par ex. choix d'objectifs d'apprentissage ou activation de connaissances antérieures relatives à la tâche), puis la phase de performance pendant laquelle l'apprenant exécute la tâche et où il poursuit ses progrès vers son objectif d'apprentissage, et enfin, la phase d'autoréflexion qui consiste notamment à évaluer son efficacité d'apprentissage afin de tirer des conclusions pour l'apprentissage futur.

L'amélioration des compétences des enfants en matière d'AAR est essentielle pour améliorer les performances scolaires, car les élèves autorégulés savent globalement mieux comment apprendre, ce qui peut avoir un impact positif dans toutes les disciplines (Zimmerman, 2008). Plus les enfants commencent à développer ces compétences tôt, plus l'impact sur l'ensemble de leur scolarité peut se faire sentir. Ainsi, des programmes de formation à l'autorégulation pour les élèves de l'école primaire ont déjà été élaborés dans ce but (Dignath *et al.*, 2008). Néanmoins, il peut être difficile pour les enseignants de se concentrer sur l'aide individualisée à apporter à chaque élève, à la fois sur la tâche à accomplir (par exemple, apprendre à lire) et sur leurs compétences d'autorégulation. Il serait donc préférable que cet apprentissage puisse se faire de manière transverse, en parallèle de l'acquisition des compétences du socle commun.

Nous nous concentrons ici sur la phase d'autoréflexion qui permet de travailler les compétences d'autorégulation sans interférer avec la tâche à accomplir. Dans ce contexte, nous visons plus particulièrement deux aspects : l'auto-évaluation qui concerne les jugements relatifs à sa propre performance (Schunk, 1996) et l'auto-efficacité qui concerne les réactions à ces jugements et la perception de ses propres compétences à accomplir une tâche (Bandura, 2010). En effet, l'amélioration du sentiment d'auto-efficacité est corrélée avec des gains d'apprentissage accrus (Jackson, 2002) et l'auto-évaluation est un processus clé de l'autorégulation (Schunk et Zimmerman, 2012). De plus, l'auto-évaluation est une capacité qui se développe progressivement, mais dont peuvent déjà disposer des jeunes enfants dès l'âge de 5 ans (Stipek *et al.*, 1992), tout comme l'auto-efficacité dans des domaines liés à l'apprentissage de l'écriture (Kim et Lorschbach, 2005).

Lalilo est l'une des nombreuses applications accessibles en ligne utilisées par les enseignants en classe pour les aider à mettre en place une pédagogie différenciée lors de l'apprentissage de la lecture - pour d'autres applications soutenant l'apprentissage de la lecture en français, voir par exemple Ecalte *et al.* (2016). Elle est actuellement utilisée par 40 000 classes de maternelle et élémentaire anglophones et francophones chaque semaine pour renforcer l'alphabétisation en proposant une série d'exercices adaptés au niveau des élèves, tout en offrant à l'enseignant un

tableau de bord pour suivre les activités et les progrès des élèves. Il s'agit donc d'un terrain d'essai pertinent pour évaluer puis essayer de corriger les capacités d'auto-évaluation et d'auto-efficacité. Un défi supplémentaire est qu'il n'existe à notre connaissance pas d'études sur l'application de ces approches aux enfants de cet âge (5-7 ans), d'où le besoin de savoir si l'on peut identifier en contexte ces phénomènes (pour ensuite tenter d'y remédier) et estimer correctement leur fréquence (pour savoir quels déficits viser en priorité).

Plus précisément, nous examinerons quatre questions de recherche :

- (QR1) Comment mesurer les capacités d'auto-évaluation et d'auto-efficacité des jeunes élèves qui apprennent avec une application informatique ?
- (QR2) Les déficits d'auto-évaluation et d'auto-efficacité sont-ils des problèmes courants pour les jeunes élèves qui apprennent à lire ?
- (QR3) Peut-on, par un étayage, aider les élèves à améliorer leur capacité d'auto-évaluation ?
- (QR4) Une remédiation orale peut-elle aider les élèves à réduire l'apparition de déficits d'auto-évaluation et d'auto-efficacité précédemment identifiés ?

Dans la suite de cet article, nous commencerons par examiner les travaux connexes sur la mesure et l'entraînement des capacités d'AAR dans le cas des jeunes enfants, en particulier dans le contexte d'environnements informatiques pour l'apprentissage humain (EIAH). En section 3, nous présenterons le fonctionnement de l'application d'apprentissage de la lecture Lalilo, l'intégration des évaluations des compétences d'auto-évaluation et d'auto-efficacité ainsi que les déficits d'autorégulation détectés et leur fréquence. En section 4, nous présenterons les étayages et rétroactions et discuterons de leur impact sur la fréquence des déficits détectés. Enfin, nous concluons et présenterons les perspectives de ce travail en section 5.

2. Travaux connexes

Dans le contexte de l'apprentissage sur ordinateur, l'autorégulation de l'apprentissage peut être soutenue par différents types d'étayages (*scaffolding*) (Azevedo et Hadwin, 2005) comme des invites (*prompts*) (Bouchet *et al.*, 2016) ou des rétroactions (*feedback*) automatisées (Bimba *et al.*, 2017); une méta-analyse des dispositifs informatiques mis en place jusqu'à 2016 pour aider l'autorégulation de l'apprentissage montre leur effet positif significatif sur la progression (Zheng, 2016). Cependant, ces aides à l'autorégulation sont incluses dans des logiciels pour des élèves plus âgés (au-delà du CM2); elles soutiennent l'autorégulation uniquement pendant la phase de performance; enfin la mesure utilisée est celle de la progression dans la tâche cognitive soutenue ou non, et non la progression des capacités d'autorégulation elles-mêmes. En effet, l'autorégulation de l'apprentissage est surtout vue comme soutenant l'apprentissage, plutôt que comme une compétence à évaluer et à entraîner en tant que telle. Dans cet article, nous envisageons un autre angle : celui de la mesure directe de capacités d'AAR et de l'amélioration de celles-ci.

Pour mesurer ces capacités, on a parfois recours – hors cadre informatique – à des journaux d'apprentissage (*learning diaries*) (Schmitz et Perels, 2011), car les traces de systèmes informatiques sont souvent difficiles à interpréter en matière d'autorégulation (Molenaar *et al.*, 2019). Comme précédemment mentionné, des programmes d'entraînement à l'autorégulation de l'apprentissage ont montré leur effet positif significatif chez des enfants à l'école primaire, mais ces travaux ont été menés en dehors d'un contexte informatique (Dignath *et al.*, 2008). D'autres, comme MetaTutor, mesurent et entraînent l'autorégulation mais pour des étudiants du supérieur (Azevedo *et al.*, 2012). De plus, une des conclusions d'Azevedo *et al.* souligne la nécessité d'un temps long pour mesurer un impact, d'où la pertinence de mesurer l'autorégulation dans un logiciel tel que Lalilo utilisé pendant une, voire plusieurs années scolaires (de la Grande Section au CE1 c'est-à-dire de 5 à 7 ans). Le travail de recherche mené par Molenaar *et al.* (2020) vise à entraîner les capacités d'autorégulation d'élèves de CM2 *via* des tableaux de bord. Les auteurs montrent une amélioration des capacités d'autorégulation des élèves ayant accès au tableau de bord, notamment par l'usage de la forme des *Moment by Moment Learning Curves* – une courbe représentant la progression de l'élève au cours du temps, démontrant ainsi également qu'on peut mesurer l'autorégulation chez des enfants assez jeunes avec une approche informatique. Ils proposent ainsi une métrique basée sur les courbes d'apprentissage.

L'auto-évaluation est définie comme la capacité à évaluer correctement ses performances. Elle permet à un élève de comparer la représentation de sa performance avec les résultats attendus avant l'exercice, et de réagir pertinemment à sa performance. La sous-évaluation peut empêcher un élève d'obtenir une récompense intrinsèque lorsqu'il a bien réussi un exercice sans en être conscient, et donc ralentir indirectement ses progrès. La surévaluation peut également avoir des effets négatifs en incitant une personne à essayer des exercices plus difficiles sans avoir encore maîtrisé les exercices faciles, augmentant ainsi le risque d'échecs futurs. L'objectif pour un apprenant est d'être capable de s'autoévaluer correctement avec aussi peu d'informations que possible sur sa performance dans la tâche, c'est-à-dire d'internaliser le processus d'auto-évaluation avec aussi peu d'échafaudage externe que possible. Dans la littérature, l'auto-évaluation a été étudiée comme la possibilité pour les élèves de se noter eux-mêmes (Brown *et al.*, 2015) ou comme une métrique qualitative pour soutenir la motivation des élèves (Chang, 2005). Nous avons ici une autre perspective, en nous concentrant sur la capacité à s'autoévaluer, c'est-à-dire en s'assurant que la performance d'un élève sera proche de la représentation qu'il s'en fait, afin que l'élève soit conscient de son véritable niveau.

L'auto-efficacité est définie par la façon dont une personne réagirait à sa performance ou à la représentation de sa performance en termes de confrontation à la difficulté (Dweck, 2014). Des recherches antérieures (Hattie et Clarke, 2018) font référence aux trois zones qu'une enseignante d'école primaire utilisait avec ses élèves : la zone de confort, la zone d'apprentissage et la zone de panique. Le fait de

demander à ses élèves de réfléchir à la zone dans laquelle ils se trouvaient après avoir abordé un exercice a entraîné : (a) une diminution du nombre d'élèves choisissant volontairement des exercices faciles parce qu'ils seraient capables d'en faire beaucoup et (b) une augmentation du nombre d'élèves déclarant qu'ils ne feraient pas de progrès s'ils avaient des exercices trop faciles. Par conséquent, demander aux élèves de réfléchir à la difficulté des exercices qu'ils abordent semble pertinent pour améliorer leur auto-efficacité.

Certains travaux ont déjà tenté d'évaluer l'effet des invites d'autorégulation pour montrer leurs effets positifs sur l'auto-efficacité (Schmitz et Wiese, 2006). Par exemple, Müller et Seufert (2018) ont montré que poser des questions d'autorégulation à des étudiants universitaires avait un impact immédiat qui ne se transférait pas dans le temps. Hoffman et Spataru (2008) ont également montré un impact positif des invites sur l'auto-efficacité, mais avant l'accomplissement de la tâche, et a mesuré l'impact sur la performance plus que sur l'auto-efficacité elle-même. Plus généralement, une méta-analyse (Panadero *et al.*, 2017) a montré qu'encourager l'auto-évaluation des apprenants a un impact positif sur leur auto-efficacité.

3. Évaluer l'auto-évaluation et l'auto-efficacité des élèves

3.1. Contexte

Le logiciel d'apprentissage de la lecture Lalilo possède deux interfaces : une interface élève dans laquelle l'élève répond à des exercices et un tableau de bord enseignant sur lequel ce dernier peut visualiser la progression de ses élèves (non présenté ici). Cet outil étant destiné aux élèves de 5 à 7 ans, il dispose d'une interface volontairement facile à prendre en main. Les consignes sont données à l'oral (et non écrites) et réécoutables. L'application couvre une grande variété d'exercices sur un champ de difficulté étendu, allant de la maternelle avec des exercices d'association graphèmes-phonèmes, jusqu'à des exercices de conjugaison et de vocabulaire pour les CE1-CE2. La figure 1 fournit deux exemples d'exercices, l'un où l'élève doit trier des mots en fonction de leur champ sémantique (gauche) et l'autre où il doit relier un son à la lettre correspondante (droite).



Figure 1 • Exemples d'exercices dans Lalilo en français

Le déroulement typique d'une session (figure 2) est la réalisation par l'élève d'une quinzaine d'exercices courts de 3 à 7 questions chacun, choisis par un algorithme d'apprentissage adaptatif (non détaillé ici). Pour certains types d'exercices, l'élève peut essayer plusieurs fois de répondre à la même question jusqu'à obtenir la bonne réponse. Les énoncés d'autorégulation (décrits en détail dans les parties suivantes) sont déclenchés avec une probabilité fixée (ici 1/15).

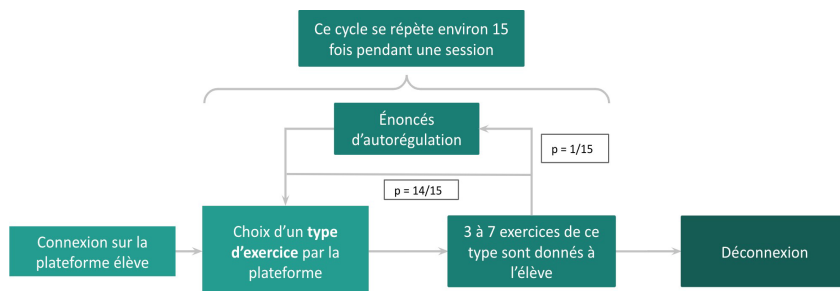


Figure 2 • Session type d'un élève sur Lalilo (durée moyenne : 20 minutes)

Les activités des élèves (par exemple: connexion, temps passé sur une question/un exercice, erreurs) sont tracées. Dans cet article, nous nous concentrerons particulièrement sur les réponses des élèves à un exercice, nous appellerons donc, à partir de maintenant, « trace » uniquement les réponses à cet ensemble de questions du même type.

3.2. Méthodes

Pour évaluer certains aspects des compétences d'autorégulation des élèves, nous avons introduit deux énoncés (figure 3) affichés l'un après l'autre à la fin d'un exercice à la fréquence $Freq_{autoreg}$ d'une fois tous les quinze exercices (Bannert et Reimann, 2012). Chaque élève y répond donc en moyenne une fois pendant une session d'apprentissage type. Quand ils sont affichés, la réponse aux deux énoncés est obligatoire. Une fois qu'une réponse est sélectionnée, un bouton de confirmation s'affiche dessous: par exemple, à droite de la figure 3, l'élève a sélectionné « de même niveau ».

Tout d'abord, l'énoncé de difficulté perçue demande à l'élève: « Quelle était la difficulté de cet exercice pour toi ? ». Ensuite, l'élève doit compléter l'énoncé de difficulté souhaitée « Tu voudrais des exercices... ». L'énoncé de difficulté perçue vise à mesurer la capacité d'auto-évaluation des élèves, c'est-à-dire leur capacité à estimer correctement la difficulté des questions auxquelles ils viennent de répondre. L'énoncé de difficulté souhaitée vise à mesurer leur sentiment d'auto-efficacité, c'est-à-dire la façon dont ils réagiraient à leur propre estimation de la difficulté.



Figure 3 • Énoncés d'autorégulation demandant la difficulté perçue (gauche) puis la difficulté voulue (droite)

Avant d'introduire les évaluations, nous avons vérifié qualitativement dans une classe utilisant Lalilo que les énoncés étaient compris par les élèves de CP. Les élèves interagissaient avec le logiciel normalement pendant qu'un expérimentateur était assis derrière eux et observait leur réaction devant les deux questions. Ensuite une discussion permettait de tester leur compréhension du concept de « difficulté ». Bien qu'informel et sur un échantillon réduit, ce travail a permis de vérifier que les énoncés compris par les élèves ne présentaient pas de décalage par rapport à l'intention sous-jacente. Elle a aussi permis de choisir la formulation la plus claire pour les élèves lorsque plusieurs options étaient envisagées (détails non présentés ici).

Nous avons recueilli des traces de classes de maternelle, de CP et de CE1 basées en France, au Canada et aux États-Unis apprenant le français (FR) ou l'anglais (EN) entre le 1^{er} août et le 26 octobre 2020 sur la plateforme Lalilo. Cette période correspond à un moment où les élèves se trouvaient principalement dans les classes et non à la maison.

Nous n'avons conservé que les traces dans lesquelles les questions d'autorégulation avaient été posées et nous allons, jusqu'à nouvel ordre, appeler trace l'ensemble regroupant les réponses à l'exercice et les réponses associées aux énoncés d'autorégulation.

Nous avons limité les sources potentielles de biais dans nos données en identifiant plusieurs phénomènes pouvant les biaiser, notamment :

- insuffisance de réponses aux énoncés d'autorégulation pour un élève donné ;
- ignorance ou incompréhension des questions par les élèves, ce qui devrait entraîner une tendance à répondre presque au hasard, puisqu'il est impossible de sauter les énoncés d'autorégulation ou de ne pas y répondre.

Pour le premier point, nous avons filtré les élèves ayant eu moins de $N_{min} = 12$ réponses aux énoncés d'autorégulation. En effet, observer des déficits établis nécessite de mesurer suffisamment de données par élève. Ce seuil peut sembler élevé : il est probable qu'un élève présentant 5 fois un même déficit sur des questions puisse être raisonnablement considéré comme présentant ce déficit d'autorégulation. Néanmoins, ce choix conservateur renforce la certitude du diagnostic quand il est posé, et limite donc le nombre de faux positifs.

En ce qui concerne le deuxième point, nous avons supprimé les élèves semblant répondre au hasard. Pour déterminer si une réponse implique une part de hasard, nous considérons que deux combinaisons de réponses sont particulièrement incohérentes (tableau 1). Nous considérons qu'un élève répond potentiellement au hasard si les **deux** combinaisons de réponses incohérentes apparaissent au moins une fois dans ses traces. Là encore, ce choix est conservateur (un élève peut commettre une erreur une fois sans que cela soit significatif d'une tendance à répondre au hasard) et a vocation à limiter le nombre de faux positifs.

Tableau 1 • Labellisation des réponses potentiellement incohérentes

Difficulté perçue	Difficulté voulue
trop facile	plus facile
trop difficile	plus difficile

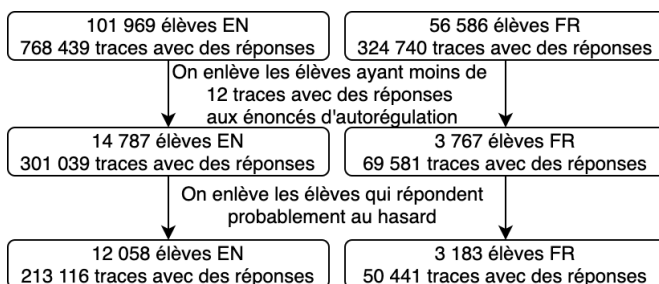


Figure 4 • Filtrage des élèves

L'impact des 2 filtrages est résumé sur la figure 4 : le filtre le plus important est le premier qui divise par 8 en anglais et par 18 en français la taille de l'échantillon initial. Toutefois cela ne signifie pas une incapacité à identifier des déficits chez les élèves exclus par le filtrage. On pourrait en effet :

- soit abaisser le seuil de 12 (élevé, comme susmentionné),
- soit étendre la période de collecte (2 mois actuellement) pour qu'ils aient répondu 12 fois aux 2 questions, ce qui devrait arriver à tout élève utilisant régulièrement Lalilo,
- soit augmenter la fréquence de prompt ($Freq_{autoreg} = 1/15$ actuellement).

3.3. Caractérisation des déficits

Nous rappelons que, dans notre contexte, les traces considérées sont celles produites lorsqu'un exercice (trois à sept questions du même type) est suivi des deux énoncés d'autorégulation. Une trace enregistre donc les réponses à chaque question de l'exercice, ainsi qu'aux deux énoncés d'autorégulation. On peut alors calculer le taux de réussite d'une trace, défini comme le nombre de réponses correctes sur le nombre total de questions de la trace. Comme pour certains types d'exercices, l'élève peut répondre plusieurs fois (parfois avec des indications fournies entre les essais), nous ne prenons en compte que la première réponse pour le calcul du taux de réussite. À partir du taux de réussite, nous pouvons déterminer un libellé de performance d'une trace avec l'une des trois valeurs suivantes :

- Excellente : si la proportion de réponses correctes est supérieure ou égale à un seuil $Perf_+$ que l'on a choisi de fixer ici à 100 % ;
- Mauvaise : si la proportion de réponses correctes est inférieure ou égale à un seuil $Perf_-$ fixé ici à 34 % ;
- Moyenne : si $Perf \in]Perf_-, Perf_+ [$.

Nous avons choisi un seuil de 34 % pour $Perf_-$, de sorte que les traces qui n'ont qu'une seule bonne réponse sur 3 soient considérées comme faibles. En effet, pour un questionnaire à choix unique avec 3 réponses possibles, la probabilité attendue de réussite aux questions est toujours d'au moins 1/3, ce qui signifie que les élèves ayant un taux de réussite d'1/3 ou moins n'obtiennent pas de meilleurs résultats que le hasard. Il convient également de noter que le seuil $Perf_+$ est, ici encore, assez conservateur, car un élève qui a répondu correctement à 6 questions sur 7 pourrait être considéré comme ayant également de très bonnes performances.

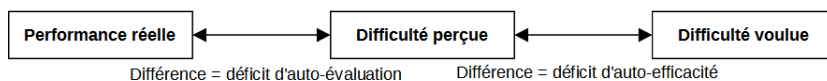


Figure 5 • Caractérisation des déficits d'auto-évaluation et d'auto-efficacité

Notre objectif est de comparer la performance réelle d'un élève avec la difficulté qu'il a perçue, puis sa difficulté perçue - qui est une représentation subjective - avec la difficulté qu'il aimerait avoir pour les prochains exercices (figure 5). En effet, la comparaison entre la difficulté souhaitée et la performance réelle peut ne pas être pertinente si les élèves sont biaisés dans leur perception de leur performance réelle ou dans la difficulté de la tâche. À partir de la différence entre leur performance et leur perception de la difficulté, on en déduit un libellé d'auto-évaluation (tableau 2). Ensuite, à partir de la différence entre la difficulté perçue et la difficulté souhaitée, on en déduit un libellé d'auto-efficacité (tableau 3).

Tableau 2 • Libellés d’auto-évaluation

Libellé de performance	Difficulté perçue	Libellé d’auto-évaluation
excellente	trop difficile	sous-évaluation
excellente	bien	légère sous-évaluation
excellente	trop facile	cohérent
mauvaise	trop difficile	cohérent
mauvaise	bien	légère surévaluation
mauvaise	trop facile	surévaluation

Tableau 3 • Libellés d’auto-efficacité

Difficulté perçue	Difficulté voulue	Libellé d’auto-efficacité
trop difficile	plus facile	cohérent
trop difficile	de même niveau	élevé
trop difficile	plus difficile	très élevé
bien	plus facile	faible
bien	de même niveau	cohérent
bien	plus difficile	élevé
trop facile	plus facile	très faible
trop facile	de même niveau	faible
trop facile	plus difficile	cohérent

Dans la perspective d’apporter une remédiation aux déficits présentés, il est nécessaire d’introduire une notion de priorité entre les déficits, car un élève peut avoir un déficit d’auto-évaluation et d’auto-efficacité. On introduit donc un libellé supplémentaire - le libellé de déficit global - dont l’intérêt est d’être orienté vers le feedback à donner (à l’élève ou à l’enseignant). Nous pensons qu’il est d’abord nécessaire de résoudre les éventuels déficits d’auto-évaluation avant de résoudre les déficits d’auto-efficacité. Ainsi, si le libellé d’auto-évaluation de la trace montre un déficit d’auto-évaluation, alors le déficit global de la trace est ce déficit d’auto-évaluation (tableau 4). En effet, nous avons estimé que la difficulté souhaitée n’était pas pertinente si les élèves n’avaient pas une représentation correcte de la difficulté de l’exercice qu’ils venaient de résoudre. Ce choix limite notre intervention à essayer de résoudre un déficit à la fois, plutôt que de s’attaquer à deux déficits potentiels à la fois.

Tableau 4 • Triplets de traces libellés comme ayant un déficit

Libellé de performance	Difficulté perçue	Difficulté voulue	Libellé de déficit global
excellente	trop difficile	toute	sous-évaluation
excellente	bien	toute	légère sous-évaluation
mauvaise	trop facile	toute	surévaluation
mauvaise	bien	toute	légère surévaluation
excellente	trop facile	plus facile	évitant la difficulté
excellente	trop facile	de même niveau	évitant la difficulté
mauvaise	trop difficile	plus difficile	cherchant la difficulté
mauvaise	trop difficile	de même niveau	cherchant la difficulté

S'il n'y a pas de déficit d'auto-évaluation, mais que le libellé d'auto-efficacité montre un certain déficit, alors deux nouveaux libellés peuvent apparaître : « évitant la difficulté » et « recherchant la difficulté ».

Comme nous avons défini la labellisation des traces, nous pouvons maintenant considérer l'ensemble des N traces d'un élève telles que $N \geq N_{min}$. Notre objectif est alors de détecter certaines tendances dans les réponses de l'élève afin de caractériser globalement son profil d'autorégulation de l'apprentissage.

Un algorithme (figure 6) est utilisé pour déterminer si un élève donné sera libellé comme ayant un déficit ou non. Il dépend de deux paramètres : $FrDef_{min}$ la fréquence minimale d'un libellé de déficit dans les traces et $NDef_{min}$ le nombre minimum de traces devant présenter un déficit pour libelliser l'élève avec ce déficit. Ces deux paramètres sont nécessaires pour caractériser uniquement les élèves ayant un déficit marqué ($FrDef_{min}$) et limiter les faux positifs qui auraient le déficit dans leurs traces par hasard ($NDef_{min}$).

$FrDef_{min}$: fréquence minimale du déficit

$NDef_{min}$: nombre minimum d'occurrences du déficit

Result : élèveADéficit // liste des déficits de l'élève, le cas échéant

for déficit **in** déficits **do**

 nbTracesAyantUnDéficit = compter(libellésDeTrace.contenant(déficit));

if déficit **in** [sous-évaluation, évitant la difficulté] **then**

 | déficitRatio = nbTracesAyantUnDéficit / nbExcellentesPerformances;

else if déficit **in** [sur-évaluation, cherchant la difficulté] **then**

 | déficitRatio = nbTracesAyantUnDéficit / nbFaiblesPerformances;

if déficitRatio $\geq FrDef_{min}$ **and** nbTracesAyantUnDéficit $\geq NDef_{min}$

then

 | élèveADéficit.ajouter(déficit)

return élèveADéficit

Figure 6 • Algorithme de détection des déficits

Sur la base du nombre de libellés de déficit d'autorégulation d'un élève (tableau 4) et du nombre de performances faibles et excellentes de l'élève, le **déficitRatio** est le rapport entre le nombre de réponses labellisées comme ayant un déficit et le nombre de performances associées : « surévaluation » et « cherchant la difficulté » sont liées à de mauvaises performances tandis que « sous-évaluation » et « éviter la difficulté » sont liées à d'excellentes performances.

Nous avons choisi les seuils du **déficitRatio** pour qu'un déficit soit considéré présent lorsque sa fréquence est supérieure à 50 % avec au moins deux occurrences du déficit c'est-à-dire $NDef_{min} = 2$ et $FrDef_{min} = 50\%$. Ces valeurs sont bien supérieures au choix aléatoire qui est de 33 % (puisque trois réponses sont proposées sur chaque énoncé) et nous voulions exclure les erreurs ponctuelles des élèves.

3.4. Résultats

Le tableau 5 résume nos résultats. Nous rappelons que le choix de nos différents seuils ($Perf_+ = 100\%$, $Perf_- = 34\%$, $NDef_{min} = 2$ et $FrDef_{min} = 50\%$) est assez conservateur, de sorte que nous détectons probablement moins de déficits qu'il n'y en a en réalité. Nous avons défini deux possibilités de calcul des déficits de surévaluation et de sous-évaluation. Pour le calcul de la première valeur de ces déficits dans le tableau, seules les traces marquées comme « surévaluation » et « sous-évaluation » dans le tableau 4 sont incluses; tandis que pour la valeur entre parenthèses, les traces marquées comme « légère sous-évaluation » et « légère surévaluation » sont également incluses. Cela nous permet d'avoir une estimation des limites inférieure et supérieure pour chacun de ces deux déficits. Les trois déficits les plus fréquents correspondent à trois des quatre déficits définis dans la section 3.3 : « surévaluation », « évitant la difficulté » et « sous-évaluation ». L'ordre entre les trois varie si l'on tient compte des libellés « légère surévaluation » et « légère sous-évaluation ». En effet, la limite supérieure de la prévalence de la sous-évaluation est supérieure à 25 % pour les élèves francophones (FR) et anglophones (EN), ce qui suggère qu'un nombre important d'élèves se sous-évaluent dans une certaine mesure. Le quatrième schéma le plus fréquemment détecté concerne les élèves qui présentent à la fois des déficits de « évitant la difficulté » et de « surévaluation ». Bien que contradictoires à première vue, en regardant les triplets de traces associés à ces deux déficits dans le tableau 4, on remarque qu'ils sont associés à une difficulté perçue « trop facile ». Nous supposons donc que les élèves labellisés comme ayant ces deux déficits ne comprennent pas correctement les énoncés d'autorégulation, mais cliquent toujours sur « trop facile »; de même, pour le sixième déficit le plus fréquemment détecté (« cherchant la difficulté - sous-évaluation »), où nous supposons que ces élèves cliquent toujours sur « trop difficile ».

Tableau 5 • Pourcentage global d'élèves ayant un ou plusieurs déficits (un élève ne pouvant appartenir qu'à une seule ligne)

Déficit(s)	EN % (12 058 élèves)	FR % (3 183 élèves)
pas de déficit détecté	72,1 % (42,5 %)	63,1 % (38,4 %)
surévaluation	8,9 % (10,2 %)	9,8 % (13,6 %)
évitant la difficulté	8,6 %	12,4 %
sous-évaluation	5,2 % (32,6 %)	5,3 % (26,2 %)
évitant la difficulté - surévaluation	3,6 %	4,4 %
cherchant la difficulté	1,6 %	3,2 %
cherchant la difficulté - sous-évaluation	0,5 %	1,0 %
autres combinaisons	0,4 %	0,8 %

Enfin, les *autres combinaisons* de déficits dans la dernière ligne du tableau 5 représentent très peu d'élèves. Les triplets de réponses associés forment des combinaisons incohérentes, ce qui pourrait correspondre à des élèves répondant de manière aléatoire, mais non éliminés par le filtrage initial.

L'ordre de prévalence des déficits est assez similaire chez les élèves FR et EN. La fréquence des réponses « évitant la difficulté » et « cherchant la difficulté » est plus élevée chez les élèves FR. Cela pourrait indiquer que les élèves FR et EN ont des capacités d'auto-évaluation comparables, mais que les élèves FR ont plus de difficultés à se positionner par rapport à la difficulté attendue. Bien que ce résultat puisse être lié à des approches pédagogiques différentes, nous restons prudents, car il pourrait y avoir des biais liés aux différences de didactique des langues (la structure didactique de la version française du logiciel a été conçue par des experts français et la structure didactique de la version anglaise a été conçue par des experts américains, ce qui peut conduire à une utilisation différente du logiciel) et au profil des utilisateurs (âge, fréquence d'utilisation, proportion d'élèves à besoins particuliers, nombre d'élèves par classe, maîtrise des outils informatiques) selon la langue utilisée.

Ces résultats montrent la possibilité de détecter de manière fiable des déficits d'auto-évaluation et d'auto-efficacité chez de jeunes élèves, répondant ainsi à la question de recherche 1 (comment mesurer les capacités d'auto-évaluation et d'auto-efficacité des jeunes élèves qui apprennent avec une application informatique ?).

De plus, ils permettent d'établir une estimation de la fréquence de ces différents phénomènes, répondant ainsi à la question de recherche 2 (les déficits d'auto-évaluation et d'auto-efficacité sont-ils des problèmes courants pour les jeunes élèves qui apprennent à lire ?). Les fréquences relativement importantes détectées - bien que les fréquences réelles puissent différer, du fait du filtrage important pour les

traces qui ont été traitées - suggèrent qu'il serait nécessaire d'essayer de remédier à ces déficits, validant ainsi l'intérêt des questions de recherche 3 et 4.

3.5. Discussion et limites

Pendant un exercice, un élève répond à des questions de niveaux similaires et proches temporellement (une minute en moyenne par exercice), chaque exercice possédant une consigne qui est énoncée oralement avant un groupe de questions. Lorsque l'on donne l'énoncé de difficulté perçue à l'issue d'un exercice, un élève répond à la question « quelle était la difficulté de cet exercice pour toi ? » en fonction de sa représentation de la difficulté de l'exercice et donne ainsi une indication sur sa capacité à s'autoévaluer. Ainsi, le « libellé d'auto-évaluation » porte bien un nom correspondant à ce qu'il est censé caractériser. En revanche, le « libellé d'auto-efficacité » porte un nom nécessitant une justification. En effet, Masson et Fenouillet (2013) ont conçu une échelle (figure 7) pour mesurer le sentiment d'efficacité personnelle chez des élèves de CM2 (10-11 ans). Elle se présente sous la forme d'une échelle de Likert contenant des affirmations pour lesquelles les élèves doivent exprimer dans quelle mesure ils sont d'accord (de « pas du tout vrai » à « totalement vrai ») et ce avec six niveaux de réponse.

-
- SEP Gén 1 - J'ai des bonnes notes à l'école.
 - SEP Gén 2 - Le travail à l'école n'est pas très difficile.
 - SEP Gén 3 - Comme je suis bon(ne) à l'école, je peux résoudre tous les exercices que l'on me pose.
 - SEP Gén 4 - Lorsque j'ai un problème dans un exercice, je me débrouille toujours pour trouver la solution.
 - SEP Gén 5 - Même si c'est très difficile, j'essaie plusieurs fois d'y arriver.
 - SEP Gén 6 - Si l'exercice est très difficile, je cherche un moyen pour réussir quand même à trouver la solution
 - SEP fr1 - J'arrive toujours à finir mes exercices de français.
 - SEP fr2 - J'arrive à me concentrer sur mes exercices de français à l'école.
 - SEP fr3 - Je comprends les exercices de français.
 - SEP fr4 - J'arrive à me motiver pour faire mes exercices de français.
 - SEP fr5 - Je suis capable de m'organiser pour faire mes exercices de français en classe.
 - SEP fr6 - Je suis capable d'écrire mes leçons de français tout seul.
 - SEP1 maths - J'arrive toujours à finir mes exercices de maths.
 - SEP2 maths - J'arrive à me concentrer sur mes exercices de maths à l'école.
 - SEP3 maths - Je comprends les exercices de mon fichier de maths.
 - SEP4 maths - J'arrive à me motiver pour faire mes exercices de maths.
 - SEP5 maths - Je suis capable de m'organiser pour faire mes exercices de maths en classe.
-

Figure 7 • Échelle du sentiment d'efficacité personnelle (SEP) scolaire (Masson et Fenouillet, 2013)

Les auteurs montrent que le SEP est différent selon le contexte (mathématique, français ou général) et que les réponses à ce questionnaire sont relativement corrélées à la performance réelle des élèves. La question posée dans le logiciel « Tu voudrais des exercices... plus faciles, de même niveau, plus difficiles » est indirectement liée à la capacité de l'élève à vouloir les résoudre, là où l'échelle proposée par (Masson et Fenouillet, 2013) demande à l'élève d'exprimer

explicitement son accord ou non avec des affirmations telles que « Même si c'est très difficile, j'essaie plusieurs fois d'y arriver ». La question de difficulté voulue ne porte pas directement sur cette capacité et l'élève pourrait cocher « plus difficiles » pour voir d'autres types d'exercices. Plusieurs critères rentraient en jeu pour le choix de la question de difficulté perçue :

- l'intelligibilité pour des enfants de 5 à 7 ans,
- l'intégration dans un EIAH,
- le temps passé à répondre aux questions d'AAR par rapport aux exercices de lecture.

L'intégration dans un EIAH pour des enfants aussi jeunes nous a conduits à simplifier au maximum les questions et à avoir trois réponses possibles. Dans ce travail, nous avons fait l'hypothèse que la répétition de certains motifs de performance et de réponses aux énoncés d'AAR nous permet de caractériser des déficits d'autorégulation corrélés au SEP. Une perspective de ce travail est une vérification externe de cette corrélation en comparant le SEP mesuré de manière externe au profil établi dans un EIAH *via* les questions de difficulté voulue et perçue.

Notre premier objectif était de vérifier la capacité à proposer une méthode permettant à un logiciel d'estimer des déficits d'auto-évaluation et d'auto-efficacité. Comme toute intervention, celle proposée peut influencer les performances des élèves et leurs compétences réelles d'autorégulation. Une première limite est donc liée au compromis nécessaire entre la valeur du seuil *Freq_{autoreg}*, fréquence des mesures, et l'impact potentiel des mesures sur l'autorégulation des élèves. L'équilibre trouvé ici garantit qu'en général, un élève ne devrait pas recevoir plus d'un énoncé par session d'apprentissage, mais nous n'avons pas évalué une éventuelle « fatigue de l'énoncé » qui peut conduire à des réponses peu fiables.

Nous nous sommes intéressés à l'auto-évaluation et l'auto-efficacité, d'une part car elles ont rarement été étudiées dans le contexte informatique pour des élèves aussi jeunes et, d'autre part, parce qu'il était possible de le faire techniquement avec Lalilo grâce au nombre important d'utilisateurs. D'autres aspects liés à l'autorégulation, comme la planification, supposent une maîtrise par l'élève des compétences à travailler (ce qui n'est pas le cas dans ce logiciel d'apprentissage de la lecture qui guide l'élève avec un algorithme d'apprentissage adaptatif). Nous détectons des réponses aléatoires *via* des filtres de préanalyse afin d'avoir des fréquences de déficit aussi proches que possible de la réalité. Cependant, nous avons pu observer *a posteriori* que des élèves ayant probablement répondu au hasard sont encore présents (tableau 5) et ont donc pu affecter la distribution des déficits des élèves. Nous avons également obtenu des déficits inattendus (répétition de réponses « trop facile » ou « trop difficile » à la difficulté perçue) : des analyses supplémentaires seraient nécessaires avant de classer les élèves dans une catégorie, si une intervention automatique ou semi-automatique était ensuite déclenchée sur la base de ce classement.

Une limite potentielle de ce travail est liée à la valeur des seuils. Dans notre méthode, nous avons identifié cinq paramètres (N_{min} , $Perf_+$, $Perf$, $FrDef_{min}$ et $NDef_{min}$) qui correspondent à différentes valeurs seuils. Nous avons essayé différents seuils pour le marquage des élèves sans impact sur la fréquence relative des déficits, bien qu'elle varie en valeur absolue (résultats non reportés ici). Les seuils choisis sont une fréquence du déficit minimale de 50 % et un nombre minimum d'occurrences de 2. Une adaptation précise des seuils devra garantir la sensibilité et la spécificité des déficits détectés, ainsi que le marquage des seuls déficits les plus saillants pour ne pas surcharger le tableau de bord de l'enseignant.

Bien que l'état de l'art le mentionne, nous n'avons pas mobilisé la possibilité d'ajouter un tableau de bord. En effet, ici on s'intéresse uniquement à la mesure, alors qu'un tableau de bord serait surtout adapté à la remédiation. De plus, les traces d'un élève liées à son autorégulation sur un tableau de bord se heurtent à une forte difficulté d'interprétation, surtout chez de jeunes enfants. Nous essayons ici d'avoir une méthode de mesure des déficits la plus objective possible.

3.6. Généralisation de la méthode

Les sections précédentes ont présenté une méthode permettant de mesurer chez un élève la présence de déficits d'auto-évaluation et d'auto-efficacité. On peut généraliser cette méthode à tout EIAH en introduisant des questions portant sur l'autorégulation de l'apprentissage adaptées à l'âge de l'élève (Molenaar *et al.*, 2020).

La figure 8 représente le potentiel diagramme de séquence dans un tel EIAH. Celui d'un élève cliquant sur « jouer » dans Lalilo est le même excepté la partie en orange, car le modèle de recommandation des exercices dans Lalilo n'utilise pas encore les réponses aux questions d'AAR. Après que l'élève a répondu aux exercices liés au domaine d'apprentissage, un modèle détermine la pertinence de donner des questions liées à l'AAR en fonction de ses réponses à l'exercice et de ses réponses précédentes aux questions d'AAR. Si la liste de questions retournées n'est pas vide, celles-ci sont données à l'élève. En fonction des réponses à ces nouvelles questions et du profil passé d'AAR, le modèle détermine la rétroaction liée à l'AAR à donner.

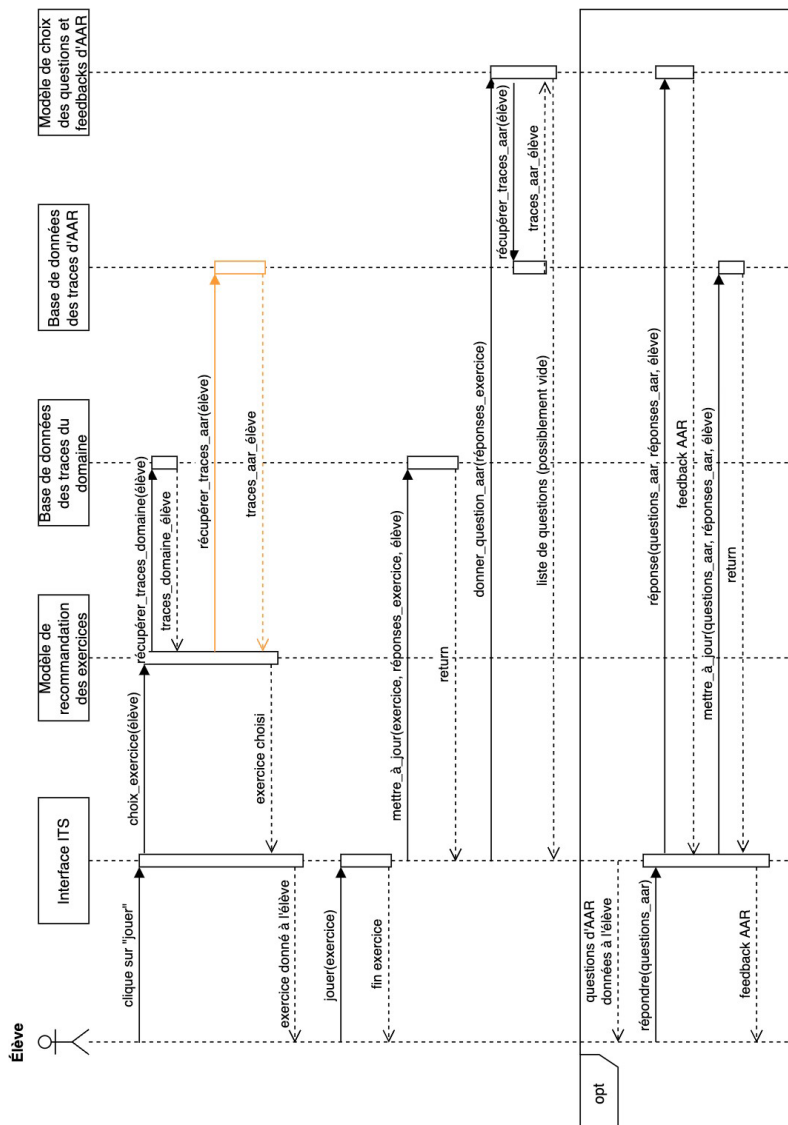


Figure 8 • Diagramme de séquence envisageable pour un EIAH souhaitant ajouter des questions d’AAR et utiliser le profil d’AAR pour la recommandation des exercices

4. Étayage et remédiation des déficits

Dans cette seconde étude, on ne vise plus seulement à mesurer, mais à tenter de corriger certains déficits identifiés, tout d’abord à l’aide d’étayage de type indice (*hint*) via une jauge ou un enregistrement audio (section 4.2), puis d’une rétroaction via un message audio donné après la réponse aux énoncés d’autorégulation (section 4.3).

4.1. Collecte des données

Dans les deux cas, nous travaillons sur un nouvel ensemble de traces provenant de classes de maternelle, de CP et de CE1 basées en France, au Canada et aux États-Unis apprenant en français (FR) ou en anglais (EN) entre le 18 janvier et le 8 avril 2021 sur la plateforme Lalilo.

Nous n'avons conservé que les traces pour lesquelles les élèves avaient répondu aux énoncés d'autorégulation (soit en moyenne 1/15e de toutes les traces) et nous appelons ci-après « traces » les réponses à l'exercice avec les réponses associées aux énoncés d'autorégulation.

4.2. Impact d'une jauge ou d'un enregistrement audio pendant la réponse aux questions d'auto-évaluation

4.2.1. Méthodes

Pour répondre à notre troisième question de recherche sur la façon dont l'étayage peut aider les élèves présentant des déficits d'auto-évaluation, nous nous sommes concentrés sur les déficits impliquant uniquement l'auto-évaluation, c'est-à-dire uniquement les quatre premières lignes du tableau 4. Afin de mesurer l'impact des indices visuels sur les réponses à l'énoncé de difficulté perçue (énoncé d'auto-évaluation), les élèves ont reçu au hasard l'un des deux visuels pour les énoncés : un visuel similaire aux énoncés initiaux (figure 9 gauche) et un autre avec une jauge supplémentaire affichant le nombre de réponses correctes et incorrectes dans l'exercice précédent (figure 9 droite). Il n'y a donc pas ici de groupe de contrôle puisqu'un même élève pouvait parfois voir la jauge et parfois non.

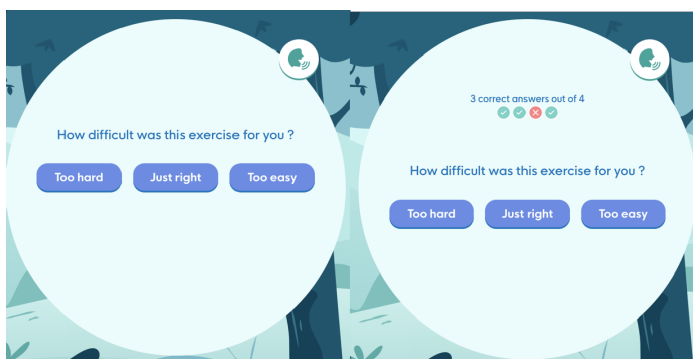


Figure 9 • Énoncés d'auto-évaluation sans et avec étayage visuel

En outre, afin de mesurer l'impact des indices auditifs, les élèves ont reçu de manière aléatoire un enregistrement audio indiquant leur nombre de réponses correctes et le nombre total de réponses dans leur dernier exercice : par exemple, « Dans le dernier exercice, tu as eu trois bonnes réponses sur quatre questions ». Cette phrase est lue au lieu d'être affichée, parce que les jeunes élèves peuvent ne pas être capables de bien la lire. Pour la même raison, chaque texte affiché sur les captures

d'écran est également lu à haute voix à l'élève, et il peut demander à réentendre la question en utilisant le bouton en haut à droite. Le choix de modalités alternatives est donc uniquement dû au public particulier (jeunes lecteurs qui ne sont pas nécessairement encore totalement autonomes), et non à une hypothèse sur les styles d'apprentissage qui sont qualifiés de neuromythe par un certain nombre de chercheurs (Newton, 2015).

Dans l'ensemble, lorsqu'un élève recevait un message d'auto-évaluation, il était assigné au hasard dans l'une des quatre conditions suivantes : (a) pas de jauge et pas d'enregistrement audio (condition de contrôle), (b) jauge et pas d'enregistrement audio, (c) pas de jauge mais enregistrement audio, (d) jauge et enregistrement audio.

Ces indices visuels et audios peuvent être vus par les élèves à la fois comme une rétroaction suite à leur performance aux questions de lecture et comme un étayage de type « indice » pour les aider à répondre aux questions d'autorégulation. Nous avons choisi le terme d'étayage de type indice, car nous nous plaçons dans la perspective de la réponse à ces questions d'autorégulation et non de la réponse aux exercices de lecture.

Notre hypothèse était que les étayages visuels ou audios pouvaient soutenir plus ou moins efficacement l'auto-évaluation des élèves.

4.2.2. Résultats

Nous avons utilisé le logiciel JASP pour nos analyses statistiques (JASP Team, 2022). Comme nous comparions des fréquences d'apparition d'un comportement dans deux versions différentes, nous avons utilisé le module d'A/B test bayésien décrit dans Gronau *et al.* (2020) avec ses valeurs par défaut. En effet, la mauvaise interprétation de *p-values* peut conduire à de mauvaises décisions (Robinson, 2019). Dans la présentation des résultats nous utilisons la probabilité *a posteriori* de l'hypothèse, sachant les données. Par exemple dans cet article, $P(\text{augmentation}|\text{données}) > 0,999$ signifie que la probabilité *a posteriori* que la modalité augmente la fréquence d'un événement, sachant les données, est supérieure à 99,9 %.

Dans cette expérience, nous n'avons conservé que la première réponse des élèves au questionnaire d'auto-évaluation, afin que nos données ne soient pas biaisées par les élèves répondant de manière répétée à l'énoncé d'auto-évaluation et disposant ou non d'un support visuel ou audio. Cela nous permet également d'isoler l'effet de la jauge et de l'enregistrement audio seuls. Nous avons ensuite calculé la distribution du déficit de réponse à l'aide du tableau 4 en fonction de la présence de la jauge et de l'enregistrement audio. Les résultats sont résumés dans les figures 10 et 11. Dans la figure 10 on regarde l'impact de la présence de jauge (b + d vs a + c), et dans la figure 11 on regarde l'impact de la présence de l'audio (c + d vs a + b). Pour chaque figure, le diagramme de gauche concerne les excellentes performances et le diagramme de droite les mauvaises performances.

La figure 10 gauche montre qu'il y a une différence dans la distribution du déficit de réponse selon qu'il y a un étayage visuel ou non : les élèves ayant une excellente performance et ayant une jauge comme support visuel sont moins susceptibles de montrer une sous-évaluation. La réduction absolue (RA) du nombre d'élèves montrant une sous-évaluation est de 0,6 %, soit une réduction relative (RR) de 8 % ($P(\text{réduction}|\text{données}) > 0,999$). Bien que la réduction absolue soit très faible, le nombre de traces (plus de 270 000) et le faible pourcentage initial de sous-évaluation fait que cette réduction permet d'affirmer que la jauge conduit très probablement à moins de sous-évaluation par les élèves ayant une mauvaise performance.

La figure 10 droite montre qu'il existe des différences dans la distribution du déficit de réponse lorsqu'il y a une jauge pour les mauvaises performances. La réduction absolue (RA) du nombre d'élèves montrant une surévaluation, est de 2 %, soit une réduction relative (RR) de 8 % ($P(\text{réduction}|\text{données}) > 0,999$). La jauge conduit très probablement à moins de surévaluation par les élèves ayant une mauvaise performance.

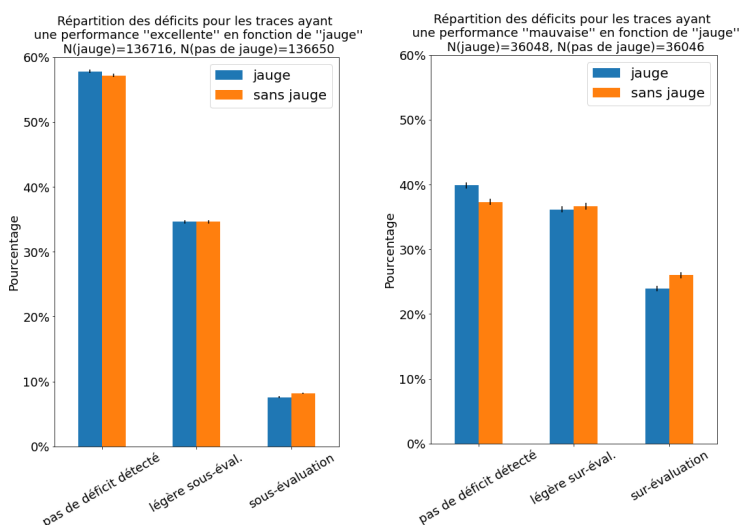


Figure 10 • Distribution du déficit d'auto-évaluation en fonction de la présence d'un étayage visuel suite à une excellente performance (gauche) ou une mauvaise performance (droite), intervalle de confiance à 95 %

Nous pouvons noter que le pourcentage de traces présentant un déficit est beaucoup plus élevé pour les énoncés suivant des traces de mauvaises performances que pour les énoncés suivant des traces d'excellentes performances (63 %, voir figure 10 droite, contre 33 %, voir figure 10 gauche). Ceci est cohérent avec le fait que les élèves ayant obtenu une excellente performance sont plus susceptibles de s'autoévaluer correctement. De plus, la réduction relative du nombre de traces présentant un déficit est indépendante de la performance, environ 8 %.

La figure 11 gauche montre qu'il y a une différence dans la distribution du déficit de réponse selon qu'il y a un étayage auditif ou non : les élèves ayant une excellente performance et ayant un enregistrement du nombre de réponses sont **plus** susceptibles de montrer une sous-évaluation. L'augmentation absolue (AA) du nombre d'élèves montrant une sous-évaluation est de 0,4 %, soit une augmentation relative (AR) de 5 % ($P(\text{augmentation}|\text{données}) > 0,999$). Bien que l'augmentation absolue soit très faible, le nombre de traces (plus de 270 000) et le faible pourcentage initial de sous-évaluation permet d'affirmer que l'invite audio conduit très probablement à **plus** de sous-évaluation par les élèves ayant une mauvaise performance.

La figure 11 droite montre qu'il existe des différences dans la distribution du déficit de réponse lorsqu'il y a un étayage auditif après une mauvaise performance. La réduction absolue (RA) du nombre d'élèves montrant une surévaluation est de 1,5 %, soit une réduction relative (RR) de 6 % ($P(\text{réduction}|\text{données}) > 0,999$). L'invite audio conduit très probablement à moins de surévaluation par les élèves ayant une mauvaise performance.

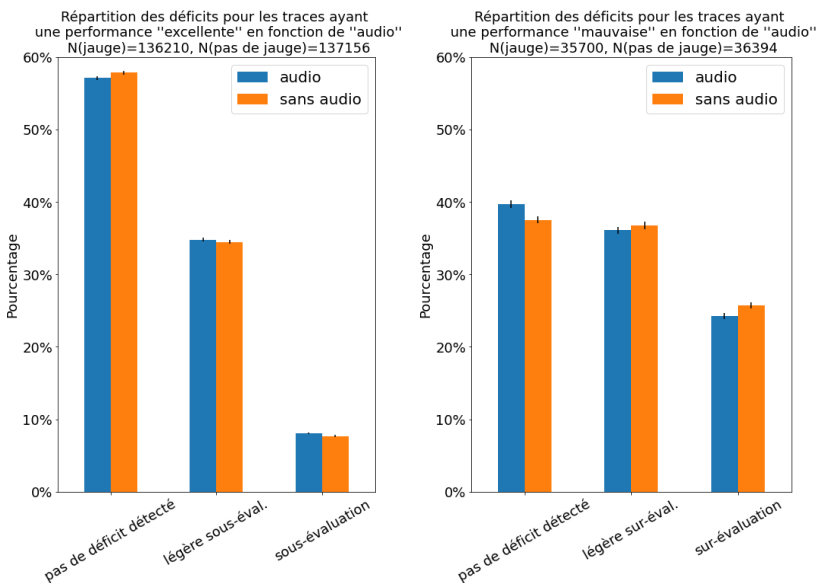


Figure 11 • Distribution du déficit d'auto-évaluation en fonction de la présence d'un étayage auditif suite à une excellente performance (gauche) ou une mauvaise performance (droite), intervalle de confiance à 95 %

Comme l'impact de la jauge semble positif pour les traces d'excellente ou de mauvaise performance, alors que l'impact de l'audio semble dépendre de la performance, nous avons mesuré l'impact de la présence ou non de l'audio lorsqu'il y avait une jauge pour soutenir l'auto-évaluation de l'élève (condition c vs b). Les résultats sont résumés dans la figure 12. Ils indiquent que le fait de disposer d'un enregistrement audio indiquant le nombre de bonnes réponses par rapport au nombre total de questions (par exemple, « Dans le dernier exercice, tu as trouvé trois bonnes réponses sur quatre questions ») a très probablement un impact sur la diminution du nombre d'élèves qui se surévaluent lorsque la performance de l'élève est mauvaise dans le dernier exercice : la réduction absolue du nombre d'élèves montrant une surévaluation est de 1,7 %, soit une réduction relative de 7 % ($P(\text{réduction}|\text{données}) > 0,999$). En revanche, il y a très probablement un impact sur l'augmentation du nombre d'élèves se sous-évaluant lorsque leur performance est excellente : l'augmentation absolue du nombre d'élèves montrant une sous-évaluation est de 0,4 %, soit une augmentation relative de 5 % ($P(\text{augmentation}|\text{données}) > 0,999$).

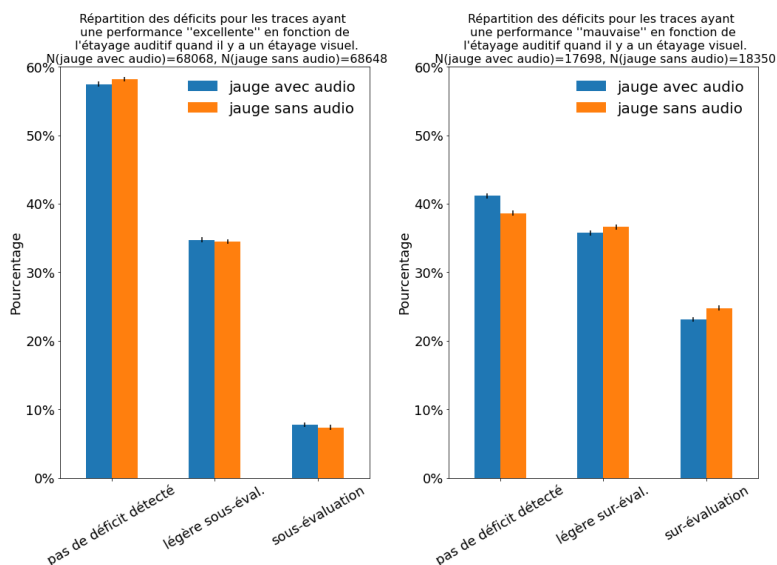


Figure 12 • Impact de l'étayage visuel et auditif sur les déficits de sous-évaluation et de surévaluation

Dans l'ensemble, nous pouvons donc répondre positivement à la question de recherche 3 puisque les étayages visuels fournis semblent avoir aidé les élèves à réduire leurs déficits d'auto-évaluation. En revanche, les étayages auditifs ont eu un impact mitigé : positif pour les élèves ayant une mauvaise performance et négatif pour les élèves ayant une excellente performance. On peut noter que l'étayage visuel est instantané, car il est affiché à l'élève, alors que l'enregistrement audio a une temporalité : pendant quelques secondes, l'élève entend son nombre de bonnes

et mauvaises réponses. L'enregistrement audio pourrait permettre à un élève potentiellement distrait de se concentrer à nouveau après une mauvaise performance. Inversement, il pourrait distraire un élève qui était concentré après une excellente performance. L'étayage visuel a un coût en temps quasi nul pour les élèves et permet une réduction d'entre 5 et 8 % du nombre de déficits, c'est pourquoi nous avons décidé à l'issue de cette étude de l'utiliser pour l'intégralité des élèves et des réponses.

4.3. Impact d'une rétroaction audio de remédiation sur les déficits d'auto-évaluation et d'auto-efficacité

4.3.1. Méthodes

Pour répondre à notre quatrième question de recherche, nous avons conçu quatre enregistrements possibles de rétroaction de remédiation (tableau 6) à écouter après que l'élève a montré l'un des quatre déficits du tableau 4. Les élèves qui ont présenté une légère surévaluation ou une légère sous-évaluation n'ont pas reçu de rétroaction de remédiation.

Dans cette expérimentation, la moitié des élèves reçoit toujours une remédiation lorsqu'ils présentent une auto-évaluation ou un déficit d'auto-efficacité (groupe de remédiation) et l'autre moitié n'en reçoit jamais (groupe témoin).

Tableau 6 • Les quatre rétroactions audio de remédiation

Libellé de déficit	Enregistrement audio
sous-évaluation	Tu as dit que c'était trop difficile alors que tu as très bien réussi. L'exercice était probablement trop facile pour toi.
surévaluation	Tu as dit que c'était trop facile alors que tu n'as pas très bien réussi. L'exercice était probablement trop difficile pour toi.
évitant la difficulté	Tu as dit que c'était trop facile, mais pourtant tu as dit vouloir des exercices plus faciles. Il faudrait plutôt des exercices plus difficiles, car tu as raison les exercices étaient en effet trop faciles.
cherchant la difficulté	Tu as dit que c'était trop difficile, mais pourtant tu as dit vouloir des exercices plus difficiles. Il faudrait plutôt des exercices plus faciles pour le moment, car tu as raison les exercices étaient en effet trop difficiles.

4.3.2. Résultats

4.3.2.1. Impact de la remédiation de la sous-évaluation

Afin de mesurer l'impact de la remédiation de la sous-évaluation, nous avons sélectionné les traces avec une excellente performance pour les élèves du groupe de contrôle et du groupe de remédiation. Dans ces traces, nous avons sélectionné les

élèves dont la première trace avec une excellente performance présentait une sous-évaluation (voir tableau 4 pour la définition des déficits).

Nous avons ensuite calculé la distribution des déficits sur leur trace suivante avec une excellente performance où ils ont eu une question d’auto-évaluation. Les résultats pour les deux groupes sont présentés dans la figure 13 (à gauche). Nous rappelons que tous les élèves ont montré un déficit de sous-évaluation dans leur première réponse à l’énoncé d’auto-évaluation. Nous remarquons que le nombre d’élèves montrant *à nouveau* une sous-évaluation est plus faible pour les élèves du groupe de remédiation que pour les élèves du groupe contrôle (RA = 6,6 %, RR = 20 %, $P(\text{réduction}|\text{données}) > 0,999$). À l’inverse, il y a plus d’élèves pour lesquels nous ne détectons aucun déficit (AA = 4,3 %, AR = 12 %, $P(\text{augmentation}|\text{données}) > 0,999$). Il y a aussi plus d’élèves pour lesquels nous avons détecté une légère sous-évaluation (AA = 2,3 %, AR = 7 %, $P(\text{augmentation}|\text{données}) > 0,999$), ce qui correspond aux élèves pour lesquels le déficit n’a été que partiellement traité (en effet, les élèves avec une légère sous-évaluation n’ont pas reçu de feedback, il n’y a donc aucune raison de s’attendre à un changement).

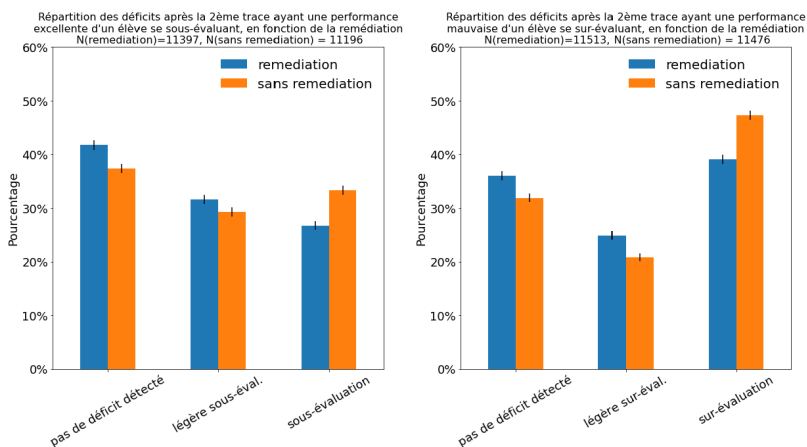


Figure 13 • Impact des remédiations auditives sur les déficits de sous-évaluation et de surévaluation

4.3.2.2. Impact de la remédiation de la surévaluation

Nous avons effectué la même analyse pour mesurer l’impact de la remédiation de la surévaluation. Dans ces traces, nous avons sélectionné les élèves des deux groupes dont la première trace avec une mauvaise performance montrait une surévaluation. Nous avons ensuite calculé la distribution des déficits sur leur trace suivante avec une mauvaise performance où ils ont eu une nouvelle question d’auto-évaluation. Les résultats pour les deux groupes sont présentés dans la figure 13 (à droite). Nous remarquons que le nombre d’élèves montrant *à nouveau* une surévaluation est plus faible pour les élèves du groupe de remédiation que pour

les élèves du groupe contrôle (RA = 8,2 %, RR = 17 %, $P(\text{réduction}|\text{données}) > 0,999$). À l'inverse, il y a plus d'élèves pour lesquels nous ne détectons aucun déficit (AA = 4,1 %, AR = 13 %, $P(\text{augmentation}|\text{données}) > 0,999$). De même, il y a plus d'élèves pour lesquels nous détectons une « légère surévaluation », ce qui signifie qu'ils ont eu une mauvaise performance, mais ont déclaré que la difficulté de l'exercice qu'ils ont obtenu était adaptée (AA = 4,1 %, AR = 20 %, $P(\text{augmentation}|\text{données}) > 0,999$). Nous pouvons supposer qu'il s'agit d'élèves qui peuvent se sentir proches de la réussite à cet exercice, même si leur performance actuelle n'est pas encore bonne. Par exemple, un élève qui a eu 3 questions dans le dernier exercice et qui a obtenu deux mauvaises réponses puis une bonne réponse peut penser que la difficulté est adaptée (« bien »), car sa dernière réponse était correcte, bien que sa performance soit considérée comme « mauvaise ».

4.3.2.3. Impact de la remédiation du déficit « évitant la difficulté »

Nous avons effectué une analyse similaire pour mesurer l'impact de la remédiation pour les élèves cherchant à éviter les difficultés. Nous avons sélectionné dans les deux groupes les traces avec une excellente performance et dans ces traces, nous avons sélectionné les élèves dont la première trace avec une excellente performance montrait qu'ils voulaient éviter la difficulté. Nous avons ensuite calculé la distribution du déficit de réponse sur leur trace suivante avec une excellente performance où ils ont obtenu les énoncés d'autorégulation de l'apprentissage. Les résultats sont présentés dans la figure 14 (à gauche). Nous pouvons remarquer que le nombre d'élèves détectés comme voulant éviter la difficulté diminue lorsqu'ils bénéficient d'une remédiation (RA = 11,1 %, RR = 33 %, $P(\text{réduction}|\text{données}) > 0,999$). Cependant, nous détectons également une augmentation du nombre d'élèves montrant une certaine sous-évaluation. Ce point sera abordé dans la section Discussion.

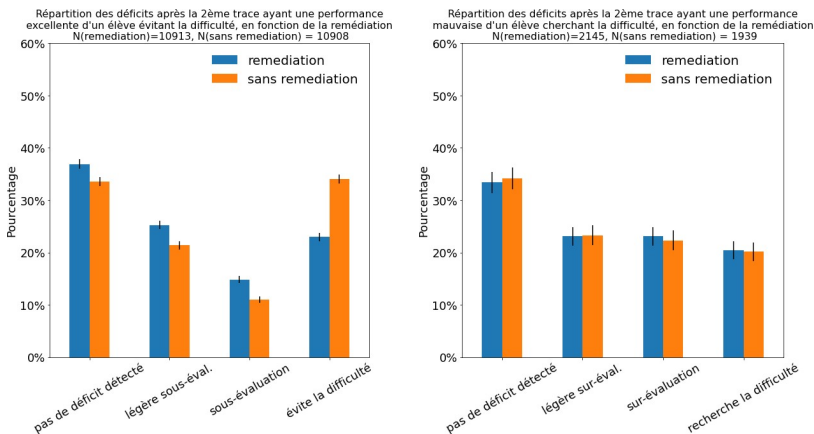


Figure 14 • Impact des remédiations auditives sur les déficits « évitant la difficulté » et « cherchant la difficulté »

4.3.2.4. Impact de la remédiation du déficit « cherchant la difficulté »

Enfin, nous avons effectué une analyse similaire pour mesurer l'impact de la remédiation pour les élèves cherchant la difficulté. Dans les deux groupes, nous avons sélectionné les traces présentant une performance « mauvaise » et, dans ces traces, nous avons sélectionné les élèves dont la première trace montrait qu'ils voulaient rechercher la difficulté (tableau 4). Nous avons ensuite calculé la distribution des déficits sur la trace suivante où les énoncés d'autorégulation ont été donnés. Les résultats sont présentés dans la figure 14 (à droite). Nous pouvons observer qu'il n'y a pas de différence suffisante entre la distribution des réponses du groupe de remédiation et la distribution des réponses du groupe de contrôle, nous ne pouvons donc pas conclure que la remédiation de la recherche de difficulté que nous avons conçue a eu un quelconque effet.

Globalement, grâce à la remédiation, il y a une réduction du nombre d'élèves montrant une surévaluation ou une sous-évaluation. Comme nous n'effectuons l'analyse que sur la deuxième trace avec une performance similaire et que les élèves ont été placés aléatoirement dans l'une des deux conditions possibles, nous pouvons déduire une relation causale entre la présence ou non de la remédiation et la différence dans la distribution des réponses à l'invite d'auto-évaluation. De plus, comme les énoncés d'autorégulation ne sont donnés avec une probabilité 1/15 qu'après avoir terminé un exercice, l'impact d'une remédiation n'est pas vu immédiatement après qu'elle a été donnée, ce qui suggère des effets durables de la remédiation. Nous pouvons donc répondre partiellement positivement à la quatrième question de recherche sur l'effet de la remédiation donnée à l'élève.

4.4. Discussion et limites

La façon la plus fiable d'évaluer les déficits d'autorégulation de l'apprentissage est de poser des questions directes aux étudiants (Barnard *et al.*, 2009). Notons qu'actuellement la fréquence des énoncés d'autorégulation de l'apprentissage est d'un tous les 15 exercices. D'une part, l'augmentation de la fréquence de la remédiation pourrait améliorer les compétences d'autorégulation de l'apprentissage, cependant, le fait d'être constamment sollicité peut conduire à une perception globale dégradée de l'environnement d'apprentissage (Bouchet *et al.*, 2018). La formation aux compétences d'autorégulation de l'apprentissage ne doit en effet pas se faire au détriment de la formation à l'alphabétisation qui reste l'objectif principal du logiciel. D'autre part, une fois que nous sommes en mesure de détecter qu'un élève possède de bonnes compétences en matière d'auto-évaluation et d'auto-efficacité, nous pourrions envisager de réduire la fréquence des énoncés, précisément pour ne pas perdre de temps inutilement, ouvrant ainsi la voie à du sondage adaptatif (*adaptive polling*). Inversement, la fréquence pourrait être augmentée sur les cas pour lesquels un déficit a été précédemment observé mais pas encore corrigé : la fréquence globale resterait donc en moyenne d'un tous les 15 exercices mais en privilégiant les cas les plus pertinents pour chaque élève.

Les résultats de notre deuxième expérience ont montré que les déficits d'auto-efficacité (éviter la difficulté et rechercher la difficulté) ne sont pas idéalement traités par la remédiation audio que nous avons conçue. Pour la remédiation visant à éviter la difficulté, le nombre d'élèves montrant une sous-évaluation est plus élevé dans le groupe de remédiation que dans le groupe de contrôle. Par conséquent, il devrait y avoir un compromis dans la mise en œuvre de cette remédiation afin qu'elle n'ait pas d'impact négatif sur l'auto-évaluation des élèves. En d'autres termes, il semble que certains élèves qui s'autoévaluaient bien, mais évitaient les difficultés, aient résolu cette contradiction non pas en demandant des exercices plus difficiles, mais en déclarant qu'ils n'avaient pas l'impression de si bien faire après tout. Ce comportement pourrait indiquer soit qu'ils n'étaient pas si sûrs de leur auto-évaluation au départ, et que le fait de remettre en question leur évaluation les a fait hésiter, soit qu'ils « se jouent du système » (i.e. le contournent) en pensant que répondre différemment à la première question empêcherait le système d'augmenter trop la difficulté, ce qui confirmerait le diagnostic de leur comportement d'évitement de la difficulté.

Pour le déficit de « recherche de difficulté », nous n'avons pas détecté d'effet. Nous pouvons cependant noter que la taille de l'échantillon est plus petite ici que dans les analyses précédentes, il est donc possible qu'avec un nombre similaire d'échantillons, un effet apparaisse. Néanmoins, il peut également y avoir des explications pour lesquelles ce comportement n'est pas aussi facile à corriger que d'autres, car « rechercher la difficulté » est un comportement cohérent avec une forme d'excès de confiance, et rejeter un feedback du système est également cohérent avec un excès de confiance. Pour ces élèves, une intervention de l'enseignant pourrait être plus appropriée. Des stratégies alternatives pourraient consister à laisser l'élève sentir qu'il s'est trompé en lui donnant un exercice beaucoup plus difficile, ou encore lui demander avant un exercice dans quelle mesure il pense réussir afin de confronter sa performance réelle à sa propre auto-évaluation *a priori* (et pas seulement *a posteriori*).

Enfin, nous nous sommes concentrés uniquement sur la première et la deuxième réponse des élèves aux énoncés de mesures de difficulté perçue et voulue, qui sont des mesures très locales. Cela nous permet de mesurer précisément l'impact des remédiations sur les réponses des élèves. Cependant, nous n'avons pas décrit l'état d'autorégulation de l'apprentissage global de chaque élève et son évolution dans le temps après les deux premières réponses aux énoncés.

5. Conclusion et travaux futurs

Pour répondre aux QR1 et QR2, nous avons proposé une définition opérationnelle de certains déficits d'autorégulation de l'apprentissage. Cela a permis la conception d'une méthode de détection des déficits potentiels d'auto-efficacité et d'autorégulation d'élèves de 5 à 7 ans en posant aux élèves deux questions sur leur difficulté perçue et la difficulté souhaitée pour les exercices suivants. En analysant

et en caractérisant les schémas de réponses d'environ 12 000 élèves nord-américains et 3 000 élèves français, nous avons détecté des déficits chez près d'un tiers des élèves. Nous avons ensuite présenté un diagramme de séquence pour généraliser l'introduction de questions liées à l'AAR dans un EIAH.

Pour répondre aux QR3 et QR4, nous avons déterminé, à l'aide de métriques locales, l'impact des étayages visuels et auditifs lors des réponses aux questions d'auto-évaluation. L'impact des étayages visuels (une jauge) est toujours positif avec une diminution du nombre de déficits d'auto-évaluation. L'impact des étayages auditifs est mitigé : ils permettent de diminuer le nombre de surévaluations, mais augmente le nombre de sous-évaluations. Par conséquent, ils ne devraient être déclenchés que pour les élèves aux performances médiocres.

Nous avons également mesuré l'effet d'une remédiation qui était déclenchée lorsque des déficits d'auto-évaluation et d'auto-efficacité étaient détectés. Nous avons réussi à réduire ces déficits pour certains élèves avec nos actions locales (pour l'auto-évaluation plus que pour l'auto-efficacité). Les travaux futurs incluent l'étude de l'impact de la remédiation dans le temps et de la nécessité de la renforcer régulièrement.

Nous avons limité notre champ d'action à l'auto-évaluation et à l'auto-efficacité dans les compétences d'autorégulation de l'apprentissage, car nous avons considéré que ces compétences étaient mesurables et pouvaient éventuellement être améliorées pour les élèves de la grande section au CE1. Les travaux futurs incluent donc l'étude d'autres compétences d'autorégulation chez les élèves de l'école primaire.

Ce travail souligne la possibilité d'étayer les compétences d'apprentissage autorégulé dans une application en ligne dès le plus jeune âge tout en apprenant une autre compétence.

6. Financement et précautions éthiques

Ce travail a été réalisé dans le cadre d'une thèse CIFRE au sein de l'équipe MOCAH de Sorbonne Université et de l'entreprise Lalilo. L'utilisation du logiciel éponyme est gratuite pour les enseignants français dans le cadre d'un contrat de licence avec le ministère de l'Éducation nationale et de la Jeunesse suite au P2IA (<https://eduscol.education.fr/1911/partenerariat-d-innovation-et-intelligence-artificielle-p2ia>). Tous les traitements de données issues de Lalilo respectent le RGPD, conformément à la réglementation (<https://ressources.lalilo.com/privacy-fr.pdf>).

REMERCIEMENTS

Les auteurs remercient les relecteurs pour leurs retours riches et constructifs.

RÉFÉRENCES

- Azevedo, R. et Hadwin, A. F. (2005). Scaffolding self-regulated learning and metacognition - Implications for the design of computer-based scaffolds. *Instructional Science*, 33(5), 367-379. <https://doi.org/10.1007/s11251-005-1272-9>
- Azevedo, R., Landis, R. S., Feyzi-Behnagh, R., Duffy, M., Trevors, G., Harley, J. M., Bouchet, F., Burlison, J., Taub, M., Pacampara, N., Yeasin, M., Rahman, A. K. M. M., Tanveer, M. I. et Hossain, G. (2012). The effectiveness of pedagogical agents' prompting and feedback in facilitating co-adapted learning with MetaTutor. Dans S. A. Cerri, W. J. Clancey, G. Papadourakis et K. Panourgia (dir.), *Proceedings of the 11th International Conference Intelligent tutoring systems (ITS 2012)* (p. 212-221). Springer. https://doi.org/10.1007/978-3-642-30950-2_27
- Bandura, A. (2010). Self-Efficacy. Dans *The corsini encyclopedia of psychology*. John Wiley & Sons. <https://doi.org/10.1002/9780470479216.corpsy0836>
- Bannert, M. et Reimann, P. (2012). Supporting self-regulated hypermedia learning through prompts. *Instructional Science*, 40(1), 193-211. <https://doi.org/10.1007/s11251-011-9167-4>
- Barnard, L., Lan, W. Y., To, Y. M., Paton, V. O. et Lai, S.-L. (2009). Measuring self-regulation in online and blended learning environments. *Internet and Higher Education*, 12(1), 1-6.
- Bimba, A. T., Idris, N., Al-Hunaiyyan, A., Mahmud, R. B. et Shuib, N. L. B. M. (2017). Adaptive feedback in computer-based learning environments: A review. *Adaptive Behavior*, 25(5), 217-234. <https://doi.org/10.1177/1059712317727590>
- Bouchet, F., Harley, J. M. et Azevedo, R. (2016). Can adaptive pedagogical agents' prompting strategies improve students' learning and self-regulation? Dans A. Micarelli, J. Stamper et K. Panourgia (dir.), *Proceedings of the 13th International Conference Intelligent Tutoring Systems (ITS 2016)* (p. 368-374). Springer. https://doi.org/10.1007/978-3-319-39583-8_43
- Bouchet, F., Harley, J. M. et Azevedo, R. (2018). Evaluating adaptive pedagogical agents' prompting strategies effect on students' emotions. Dans R. Nkambou, R. Azevedo et J. Vassileva (dir.), *Proceedings of the 14th International Conference Intelligent Tutoring Systems (ITS 2018)* (p. 33-43). Springer. https://doi.org/10.1007/978-3-319-91464-0_4
- Brown, G. T. L., Andrade, H. L. et Chen, F. (2015). Accuracy in student self-assessment: Directions and cautions for research. *Assessment in Education: Principles, Policy & Practice*, 22(4), 444-457. <https://doi.org/10.1080/0969594X.2014.996523>
- Chang, M.-M. (2005). Applying self-regulated learning strategies in a web-based instruction—an investigation of motivation perception. *Computer Assisted Language Learning*, 18(3), 217-230. <https://doi.org/10.1080/09588220500178939>
- Dignath, C., Buettner, G. et Langfeldt, H.-P. (2008). How can primary school students learn self-regulated learning strategies most effectively? *Educational Research Review*, 3(2), 101-129. <https://doi.org/10.1016/j.edurev.2008.02.003>
- Dweck, C. S. (2014). *Self-theories: their role in motivation, personality, and development*. Psychology Press. <https://doi.org/10.4324/9781315783048>
- Ecalte, J., Navarro, M., Labat, H., Gomes, C., Cros, L. et Magnan, A. (2016). Concevoir des applications sur tablettes tactiles pour stimuler l'apprentissage de la lecture : avec quelles hypothèses scientifiques? *STICEF (Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation)*, 23(2), 33-56. <http://sticef.org/num/vol2016/23.2.2.ecalle/23.2.2.ecalle.htm>
- Gronau, Q. F., Raj, K. N. A. et Wagenmakers, E.-J. (2020). *Informed Bayesian Inference for the A/B Test* (arXiv:1905.02068). arXiv. <https://arxiv.org/abs/1905.02068>
- Hattie, J. et Clarke, S. (2018). *Visible learning: Feedback*. Routledge. <https://doi.org/10.4324/9780429485480>
- Hoffman, B. et Spataru, A. (2008). The influence of self-efficacy and metacognitive prompting on math problem-solving efficiency. *Contemporary Educational Psychology*, 33(4), 875-893. <https://doi.org/10.1016/j.cedpsych.2007.07.002>

Jackson, J. W. (2002). Enhancing self-efficacy and learning performance. *The Journal of Experimental Education*, 70(3), 243-254. <https://doi.org/10.1080/00220970209599508>

JASP Team. (2022). *JASP (Version 0.16.3)* [logiciel]. <https://jasp-stats.org/>

Kim, J.-A. et Lorschbach, A. W. (2005). Writing self-efficacy in young children: Issues for the early grades environment. *Learning Environments Research*, 8(2), 157-175. <https://doi.org/10.1007/s10984-005-7248-5>

Masson, J. et Fenouillet, F. (2013). Self-efficacy and academic results at the primary school: Development and validation of a scale. *Enfance*, 4(4), 374-392.

Molenaar, I., Horvers, A., Dijkstra, R. et Baker, R. (2019). Designing dashboards to support learners' self-regulated learning. Dans *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK '19)* (p. 1-12). ACM.

Molenaar, I., Horvers, A., Dijkstra, R. et Baker, R. S. (2020). Personalized visualizations to promote young learners' SRL: The learning path app. Dans *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge (LAK '20)* (p. 330-339). <https://doi.org/10.1145/3375462.3375465>

Müller, N. M. et Seufert, T. (2018). Effects of self-regulation prompts in hypermedia learning on learning performance and self-efficacy. *Learning and Instruction*, 58, 1-11. <https://doi.org/10.1016/j.learninstruc.2018.04.011>

Newton, P. M. (2015). The learning styles myth is thriving in higher education. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.01908>

Panadero, E., Jonsson, A. et Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review*, 22, 74-98. <https://doi.org/10.1016/j.edurev.2017.08.004>

Robinson, G. K. (2019). What properties might statistical inferences reasonably be expected to have?—Crisis and resolution in statistical inference. *The American Statistician*, 73(3), 243-252. <https://doi.org/10.1080/00031305.2017.1415971>

Schmitz, B. et Perels, F. (2011). Self-monitoring of self-regulation during math homework behaviour using standardized diaries. *Metacognition and Learning*, 6(3), 255-273. <https://doi.org/10.1007/s11409-011-9076-6>

Schmitz, B. et Wiese, B. S. (2006). New perspectives for the evaluation of training sessions in self-regulated learning: Time-series analyses of diary data. *Contemporary Educational Psychology*, 31(1), 64-96. <https://doi.org/10.1016/j.cedpsych.2005.02.002>

Schunk, D. H. (1996). Goal and self-evaluative influences during children's cognitive skill learning. *American Educational Research Journal*, 33(2), 359-382. <https://doi.org/10.3102/00028312033002359>

Schunk, D. H. et Zimmerman, B. J. (2012). Self-regulation and learning. *Educational Psychology*, 7. <https://doi.org/10.1002/9781118133880.hop207003>

Stipek, D., Recchia, S. et McClintic, S. (1992). Self-evaluation in young children. *Monographs of the Society for Research in Child Development*, 57(1), 1-98.

Zheng, L. (2016). The effectiveness of self-regulated learning scaffolds on academic performance in computer-based learning environments: A meta-analysis. *Asia Pacific Education Review*, 17(2), 187-202. <https://doi.org/10.1007/s12564-016-9426-9>

Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, 45(1), 166-183. <https://doi.org/10.3102/0002831207312909>