Détection de déficits d'auto-évaluation et d'auto-efficacité dans un logiciel enseignant la lecture et l'écriture

Thomas Sergent, François Bouchet, Morgane Daniel, Thibault Carron

Détection de déficits d'auto-évaluation et d'auto-efficacité dans un logiciel enseignant la lecture et l'écriture

Thomas Sergent[1,2], François Bouchet[1], Morgane Daniel[2] et Thibault Carron[1]

[1] Sorbonne Université, CNRS, LIP6, F-75005 Paris, France {prénom.nom}@lip6.fr
[2] Lalilo, Paris, France {prénom}@lalilo.com

Résumé. Plusieurs travaux montrent que la capacité à auto-réguler son apprentissage a un impact significatif sur les résultats scolaires. Nous présentons ici une étude visant à détecter les déficits d'auto-régulation de l'apprentissage liés à l'auto-évaluation et à l'auto-efficacité pour de jeunes (5-7 ans) élèves, dans le contexte d'une application web d'apprentissage de la lecture. Nous avons recueilli les réponses de plus de 15 000 enfants travaillant sur une telle application en classe. À partir de ces réponses, nous proposons une définition opérationnelle de différentes formes de déficits dont nous évaluons la prévalence auprès des élèves.

Mots-clé: Auto-régulation de l'apprentissage · Fouille de données éducatives · École primaire · Apprentissage de la lecture

Abstract. The ability to self-regulate one's learning is considered to have a significant impact on educational outcomes. We present here a study to detect self-regulated learning (SRL) deficits related to self-evaluation and self-efficacy for young (5-7 years old) students, in the context of a literacy web application. We collected answers to SRL statements of over 15,000 children working on such an application while in the classroom. From these SRL answers we propose an operational definition of different forms of deficits whose prevalence among students we assess.

Keywords: Self-regulated learning · Educational Data Mining · Primary school · Learning to read

1 Introduction

L'amélioration des compétences des enfants en matière d'apprentissage auto-régulé est critique pour améliorer les performances scolaires car les élèves auto-régulés savent globalement mieux comment apprendre, ce qui peut avoir un impact positif dans toutes les disciplines [16]. Plus tôt les enfants commencent à développer ces compétences, plus l'impact sur l'ensemble de leur scolarité peut se faire sentir, et des programmes de formation à l'auto-régulation pour les élèves de l'école primaire ont déjà été élaborés dans ce but [6]. 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'auto-régulation.

L'auto-régulation de l'apprentissage est un cycle en trois phases qui se répète à chaque nouvelle tâche à laquelle l'apprenant est confronté [16]. D'abord la phase d'anticipation pendant laquelle l'apprenant se prépare à la tâche (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 peut suivre ses progrès vers son objectif d'apprentissage, et enfin la phase d'auto-réflexion qui consiste notamment à évaluer son efficacité d'apprentissage afin de tirer des conclusions pour l'apprentissage futur. Nous nous concentrons ici sur cette phase d'auto-réflexion qui permet de travailler les compétences d'auto-régulation sans interférer avec la tâche à accomplir. Dans ce contexte, nous visons plus particulièrement deux aspects : l'**auto-efficacité** qui concerne la perception de ses propres compétences à accomplir une tâche [3], et l'**auto-évaluation** qui concerne les jugements relatifs à sa propre performance et les réactions à ces jugements [12]. En effet, l'amélioration du sentiment d'auto-efficacité est corrélée avec des gains d'apprentissage accrus [7] et l'auto-évaluation est un processus clé de l'auto-régulation [13]. De plus l'auto-évaluation est une capacité qui se développe progressivement mais dont disposent de jeunes enfants de 5 ans et plus [14], tout comme l'auto-efficacité dans des domaines liés à l'apprentissage de l'écriture [8].

Une méta-analyse des dispositifs informatiques mis en place jusqu'à 2016 pour aider l'auto-régulation de l'apprentissage montre leur effet positif significatif sur la progression [15]. Cependant, ces dispositifs concernent uniquement des élèves plus âgés (au-delà du CM2), ils se concentrent sur la phase de performance, et on mesure la progression de l'élève grâce aux dispositifs favorisant l'auto-régulation plus que la progression des capacités d'auto-régulation elles-mêmes. En effet, l'auto-régulation est surtout vue comme soutenant l'apprentissage, plutôt que comme une compétence à évaluer et entraîner en tant que telle. Dans cet article nous envisageons un autre angle : celui de la mesure directe de capacités d'auto-régulation, en vue à terme d'améliorer celles-ci.

Lalilo est l'une des nombreuses applications web utilisées par les enseignants en classe pour les aider à mettre en place une pédagogie différenciée. 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 ce phénomène (pour ensuite tenter d'y remédier) et estimer correctement sa fréquence (pour savoir quels déficits viser en priorité). Plus précisément, nous examinerons deux questions de recherche : (QR1) Peut-on mesurer les capacités d'auto-évaluation et d'auto-efficacité des jeunes élèves qui apprennent avec une

application web ? (QR2) Les déficits d'auto-évaluation et d'auto-efficacité sont-ils des problèmes courants pour les jeunes élèves qui apprennent à lire ?

Dans la suite de cet article, nous commencerons par examiner les travaux connexes sur la mesure et l'entraînement des capacités d'auto-régulation dans le cas des jeunes enfants. En section 3, nous présenterons le fonctionnement de Lalilo et l'intégration des évaluations des compétences d'auto-régulation. La section 4 présentera les données recueillies, les déficits d'auto-régulation détectés et leur fréquence, avant de discuter des limites et perspectives en section 5.

2 Travaux connexes

Dans le contexte de l'apprentissage sur ordinateur, l'auto-régulation peut être soutenue par différents types d'étayages (*scaffolding*) [1] comme des invites (*prompts*) [5] ou des rétroactions (*feedback*) automatisées [4].

Pour mesurer les capacités d'auto-régulation, on a parfois recours (hors cadre informatique) à des journaux d'apprentissage *(learning diaries)* [11], car les traces de systèmes informatiques sont souvent difficile à interpréter en termes d'auto-régulation [9].

Des programmes d'entraînement à l'auto-régulation ont montré leur effet positif significatif chez des enfants de primaire, mais hors contexte informatique [6]. D'autres, comme MetaTutor, mesurent et entraîne l'auto-régulation mais pour des étudiants du supérieur [2]. De plus, une des conclusions de [2] était la nécessité d'un temps long pour mesurer un impact, d'où la pertinence de mesurer l'auto-régulation dans un logiciel tel que Lalilo utilisé pendant une voire plusieurs années scolaires (de la Grande Section au CE1). Molenaar et al. [10] vise à entraîner les capacités d'auto-régulation d'élèves de CM2 via des tableaux de bord. Ils montrent une amélioration des capacités d'auto-régulation des élèves ayant accès au tableau de bord (mieux régulés), notamment par l'usage de la forme des *Moment by Moment Learning Curves*, démontrant ainsi également qu'on peut mesurer l'auto-régulation chez des enfants assez jeunes avec une approche informatique. Dans notre cas, les élèves sont encore plus jeunes (5-7 ans vs. 9-10 ans) et notre métrique se base directement sur la réponse d'un élève à des questions d'auto-régulation.

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

3.1 Contexte

Lalilo possède deux interfaces : une interface élève (cf. Figure 1) dans laquelle l'élève répond à des exercices et un tableau de bord enseignant sur lequel l'enseignant peut visualiser la progression de ses élèves. Cet outil étant destiné aux élèves de 5 à 7 ans, il dispose d'une interface simple. En général, les instructions sont uniquement lues (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.

Le déroulement typique d'une session (20 minutes en moyenne) 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

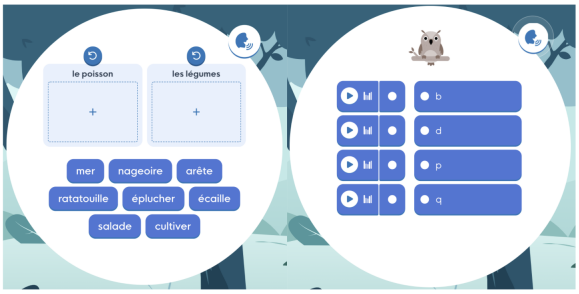

Fig. 1: Trier des mots en fonction de leur champ sémantique (gauche) et relier un son à la lettre correspondante (droite)

certains types d'exercices, l'élève peut essayer plusieurs fois une question. Les activités des élèves (par exemple, connexion, temps passé sur une question/un exercice, erreurs) sont tracées. Ici, nous nous concentrerons particulièrement sur les réponses des élèves à un exercice, nous appellerons donc désormais **trace** uniquement les réponses à cet ensemble de questions du même type.

3.2 Méthodes

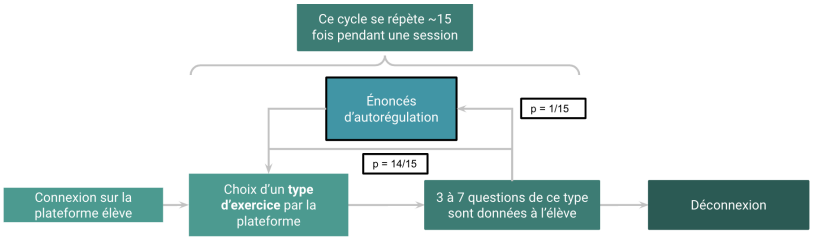

Fig. 2: Session type d'un élève sur Lalilo (durée moyenne : 20 minutes)

Fig. 3: Énoncé demandant la difficulté perçue (gauche) puis la difficulté voulue (droite). Une fois qu'une réponse est sélectionnée, un bouton de confirmation s'affiche dessous. Par exemple, à droite, l'élève a sélectionné "de même niveau".

Collecte des données, filtrage et nettoyage Pour évaluer certains aspects des compétences d'auto-régulation des élèves, nous avons introduit deux énoncés (cf. 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. Chaque élève y répond donc en moyenne une fois pendant une session d'apprentissage type (cf. Figure 2). 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 **dif-**

ficulté 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 représentation de la difficulté. 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 Lalilo normalement pendant qu'un expérimentateur était assis derrière eux et observait leur réaction devant les 2 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 complet 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 Etats-Unis apprenant le français (FR) ou l'anglais (EN) entre le 1er 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 pour lesquelles les élèves avaient répondu aux questions d'auto-régulation et nous allons désormais appeler *trace* l'ensemble regroupant les réponses à l'exercice avec les réponses associées aux énoncés d'auto-régulation.

Nous avons limité les sources potentielles de biais dans nos données en identifiant plusieurs phénomènes qui pourraient avoir un impact sur elles, notamment :

1. pas assez de réponses aux énoncés d'auto-régulation pour un élève donné
2. élèves ayant ignoré ou mal compris les questions, ce qui devrait entraîner une tendance à répondre presque au hasard, puisqu'il est impossible de sauter ou de ne pas répondre aux énoncés d'auto-régulation

Pour le premier point, nous avons filtrons les élèves ayant eu moins de $N_{min} = 12$ réponses aux énoncés d'auto-régulation. 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 souffrant de ce déficit d'auto-régulation. 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 (cf. figure 4 - gauche). 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 faire une erreur une fois sans que ça soit significatif d'une tendance à répondre au hasard) et a vocation à limiter le nombre faux positifs. L'impact des 2 filtrages est résumé sur la figure 4 (droite) : le filtre le plus important est le

difficulté perçue	difficulté voulue
trop facile	plus facile
trop difficile	plus difficile

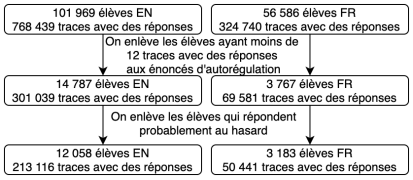

Fig. 4: Labellisation des réponses incohérentes (gauche) et filtrage des élèves (droite)

premier qui divise par plus de 10 la taille de l'échantillon initial. Toutefois cela ne signifie pas une incapacité à identifier des déficits chez ces élèves. On pourrait ainsi (1) soit abaisser le seuil de 12 (élevé, comme susmentionné), (2) 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, (3) soit augmenter la fréquence de prompt ($Freq_{autoreg} = \frac{1}{15}$ actuellement).

3.3 Caractérisation des déficits

Au niveau des traces. 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'auto-régulation. Une trace enregistre donc les réponses à chaque question de l'exercice ainsi qu'aux deux énoncés d'auto-régulation. 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 toutes les réponses étaient correctes, on note ce seuil $Perf_+$ (100% ici)
- mauvaise : si $Perf_- = 34\%$ ou moins des réponses étaient correctes
- 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 QCU avec 3 choix, 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 on pourrait estimer qu'un élève qui a répondu correctement à 6 questions sur 7 pourrait être considéré comme ayant également de très bonnes performances.

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 (cf. 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

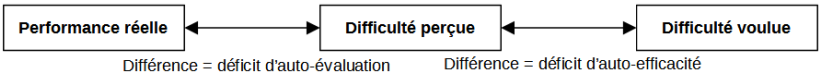

Fig. 5: Caractérisation des déficits d'auto-évaluation et d'auto-efficacité

entre leur performance et leur perception de la difficulté, on en déduit un **libellé d'auto-évaluation** (cf. tableau 1). Ensuite, à partir de la différence entre la difficulté perçue et la difficulté souhaitée, on en déduit un **libellé d'auto-efficacité** (cf. tableau 2). Notons que les deux couples de la figure 4 peuvent être considérés soit comme incohérents, soit comme étant les signes d'une auto-efficacité très élevée - pour (trop dur, plus dur) - ou très faible - pour (trop facile, plus facile).

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 l'enseignant). Nous pensons qu'il est d'abord nécessaire de résoudre les éventuels déficits d'auto-évaluation avant de s'attaquer aux déficits d'auto-efficacité. Ainsi, si le libellé d'auto-évaluation de la trace montre un déficit d'auto-évaluation, alors le déficit de la trace est ce déficit d'auto-évaluation. 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".

Libellé de perf.	Diff. 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 sur-évaluation
mauvaise	trop facile	sur-évaluation

Tableau 1: Libellés d'auto-évaluation

Diff. perçue	Diff. voulue	Libellé d'auto-efficacité
trop difficile	plus facile	
trop difficile	de même niveau	élevé
trop difficile	plus difficile	très élevé/incohérent
bien	plus facile	faible
bien	de même niveau	
bien	plus difficile	élevé
trop facile	plus facile	très faible/incohérent
trop facile	de même niveau	faible
trop facile	plus difficile	

Tableau 2: Libellés d'auto-efficacité

Libellé de perf.	Diff. perçue	Diff. voulue	Libellé de déficit global
excellente	trop difficile	plus facile	sous-évaluation
excellente	trop difficile	de même niveau	sous-évaluation
excellente	trop difficile	plus difficile	sous-évaluation
excellente	bien	plus facile	légère sous-éval.
excellente	bien	de même niveau	légère sous-éval.
excellente	bien	plus difficile	légère sous-éval.
mauvaise	trop facile	plus facile	sur-évaluation
mauvaise	trop facile	de même niveau	sur-évaluation
mauvaise	trop facile	difficile	sur-évaluation
mauvaise	bien	plus facile	légère sur-éval.
mauvaise	bien	de même niveau	légère sur-éval.
mauvaise	bien	plus difficile	légère sur-éval.
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
excellente	trop difficile	plus difficile	incohérent
mauvaise	trop facile	plus facile	incohérent

Tableau 3: Triplets de trace libellés comme ayant un déficit

Au niveau de l'élève 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'auto-régulation.

L'algorithme 1 est utilisé pour déterminer si un élève donné sera labellisé comme ayant un déficit ou non. Il dépend de 2 paramètres: $FrDef_{min}$ la fréquence minimale d'un libellé de déficit dans les traces, et $NDef_{min}$ pour nombre minimum de traces devant présenter un déficit pour labelliser 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 ($NDef_{min}$). Connaissant le nombre de libellés de déficit d'auto-régulation d'un élève avec le tableau 3 et le nombre de performances faibles et excellentes de l'élève, le ***déficitRatio*** est calculé comme 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 être significatif à 50% avec au moins 2 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 3 réponses sont proposées sur chaque énoncé) et nous voulions exclure les erreurs ponctuelles des élèves.

3.4 Résultats

Le tableau 4 résume nos résultats. Nous rappelons le choix de nos différents seuils ($Perf_+ = 100\%$, $Perf_- = 34\%, NDef_{min} = 2$ et $FrDef_{min} = 50\%$) est

Algorithm 1: Caractérisation des élèves

Result: Déficit de l'élève
for *déficit* **in** *déficits* **do**
  nbTracesAyantUnDéficit = count(traceSRLTags[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[déficit]

Déficit(s)	EN % (N=12,058)	FR % (N=3,183)
pas de déficit détecté	71.2% (42.5%)	63.1% (38.4%)
sur-évaluation	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 - sur-évaluation [1]	3.6%	4.4%
cherchant la difficulté	1.6%	3.2%
cherchant la difficulté - sous-évaluation [2]	0.5%	1.0%
évitant la difficulté - cherchant la difficulté [3]	0.2%	0.6%
sur-évaluation - sous-évaluation [3]	0.2%	0.1%
cherchant la difficulté - sur-évaluation [3]	0.0%	0.0%
évitant la difficulté - sous-évaluation [3]	0.0%	0.1%
évitant la diff. - cherchant la diff. - sous-éval. [3]	0.0%	0.0%
cherchant la diff. - sur-évaluation - sous-éval. [3]	0.0%	0.0%

[1]L'élève clique toujours sur "trop facile", quelle que soit sa performance
[2]L'élève clique toujours sur "trop difficile", quelle que soit sa performance
[3]L'élève répond probablement au hasard

Tableau 4: Pourcentage global d'élèves ayant un ou plusieurs déficits (un élève ne peut appartenir qu'à une seule ligne). Les valeurs entre parenthèses correspondent à l'inclusion des libellés *légère* sous-évaluation ou sur-évaluation

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 3 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 de la limite inférieure et supérieure 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 trace associés à ces deux déficits dans le tableau 3, 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'auto-régulation 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".

Enfin, les combinaisons de déficit dans les six dernières lignes du tableau 4 représentent très peu d'élèves, nous pouvons donc probablement nous permettre de ne pas les considérer comme des phénomènes distincts. En regardant les triplets de réponses associés, nous remarquons que ces combinaisons de déficit sont incohérentes, ce qui pourrait correspondre à des élèves répondant de manière aléatoire mais non éliminés par le filtre 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 cependant prudents car il pourrait y avoir des biais liés aux différences de didactique des langues et au profil des utilisateurs selon la langue utilisée.

Ces résultats montrent la possibilité de détecter de manière assez fiable des déficits d'auto-évaluation et d'auto-efficacité chez de jeunes élèves, et la fréquence des phénomènes détectés suggère qu'il serait nécessaire d'essayer de les traiter.

4 Discussion, limites et perspectives

Notre premier objectif était de vérifier la capacité à proposer un modèle permettant à un logiciel d'estimer des déficits d'auto-évaluation et d'auto-efficacité. Comme toute intervention, l'intervention proposée peut influencer les performances des élèves et leurs compétences réelles d'auto-régulation. 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'auto-régulation 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 intéressons à l'auto-évaluation et l'auto-efficacité d'une part car il était possible de le faire techniquement avec Lalilo, et d'autre part parce qu'elles ont rarement été étudiées dans le contexte informatique pour des élèves aussi jeunes. D'autres aspects liés à l'auto-régulation comme la planification supposent une maîtrise par l'élève des compétences à travailler (ce qui n'est pas le cas dans Lalilo qui guide l'élève avec un algorithme d'apprentissage adaptatif).

Nous détectons des réponses aléatoires via des filtres de pré-analyse 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 (cf. tableau 4) 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 est ensuite déclenchée sur la base de ce classement.

Lors de la définition des déficits, nous avons délibérément choisi le libellé d'auto-évaluation pour qu'il ait la priorité sur le libellé d'auto-efficacité. 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.

Une autre limite potentielle de ce travail est liée à la valeur des seuils. Dans notre modèle, nous avons identifié 5 ($N_{min}, Perf_{+}, Perf_{-}, FrDef_{min}, NDef_{min}$) paramètres qui correspondent à différentes valeurs seuils. Nous avons essayé différents seuils pour le marquage des élèves (fixés à 50 % et au moins deux réponses labellisées comme ayant le déficit dans l'algorithme 1) sans impact sur la fréquence relative des déficits, bien qu'elle varie en valeur absolue (résultats non affichés ici). 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 mentionnait, 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 auto-régulation 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 la plus objective possible de mesure des déficits.

Enfin, les déficits détectés devront être confrontés aux opinions des enseignants sur les déficits de leurs élèves afin de valider *a posteriori* la manière dont nous les avons définis. Les capter en même temps aurait demandé des développements supplémentaires côté enseignant impossibles pour cette étude.

5 Conclusion et travaux futurs

Pour répondre aux QR1 et QR2, nous avons présenté une méthode de détection des déficits potentiels d'auto-efficacité et d'auto-régulation 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 détecté des déficits près d'un tiers des élèves. Nous avons également proposé une définition opérationnelle des déficits permettant la labellisation.

Les travaux futurs comprennent la comparaison des données sur les déficits détectés avec la perception qu'en ont les enseignants pour leurs élèves afin de confirmer la pertinence des déficits détectés. On pourra notamment le faire en demandant à des enseignants volontaires s'ils sont d'accord avec les labellisations de sous- et sur-évaluation de leurs élèves en utilisant des échelles type Likert. On pourra également caractériser les sous-populations présentant un déficit partic-

ulier par de la fouille de séquence (*differential sequence mining*). Enfin, une fois ces déficits confirmés, une remédiation devrait être mise en œuvre, en étudiant l'évolution temporelle des déficits avec ou sans elle pour mesurer son efficience.

Références

1. Azevedo, R., Hadwin, A.F.: Scaffolding Self-regulated Learning and Metacognition – Implications for the Design of Computer-based Scaffolds. Instructional Science **33**(5), 367–379 (Nov 2005)
2. 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., Hossain, G.: The Effectiveness of Pedagogical Agents' Prompting and Feedback in Facilitating Co-adapted Learning with MetaTutor. In: Cerri, S.A., Clancey, W.J., Papadourakis, G., Panourgia, K. (eds.) Intelligent Tutoring Systems. pp. 212–221. LNCS, Springer, Berlin, Heidelberg (2012)
3. Bandura, A.: Self-Efficacy. In: The Corsini Encyclopedia of Psychology, pp. 1–3. American Cancer Society (2010)
4. Bimba, A.T., Idris, N., Al-Hunaiyyan, A., Mahmud, R.B., Shuib, N.L.B.M.: Adaptive feedback in computer-based learning environments: a review. Adaptive Behavior **25**(5), 217–234 (Oct 2017)
5. Bouchet, F., Harley, J.M., Azevedo, R.: Can Adaptive Pedagogical Agents' Prompting Strategies Improve Students' Learning and Self-Regulation? In: Micarelli, A., Stamper, J., Panourgia, K. (eds.) Intelligent Tutoring Systems. pp. 368–374. Lecture Notes in Computer Science, Springer International Publishing, Cham (2016)
6. Dignath, C., Buettner, G., Langfeldt, H.P.: How can primary school students learn self-regulated learning strategies most effectively? Educational Research Review **3**(2), 101–129 (Jan 2008)
7. Jackson, J.W.: Enhancing Self-Efficacy and Learning Performance. The Journal of Experimental Education **70**(3), 243–254 (Jan 2002)
8. Kim, J.A., Lorsbach, A.W.: Writing Self-Efficacy in Young Children: Issues for the Early Grades Environment. Learning Environments Research **8**(2), 157–175 (2005)
9. Molenaar, I., Horvers, A., Dijkstra, R., Baker, R.: Designing Dashboards to support learners' Self-Regulated Learning (Feb 2019)
10. Molenaar, I., Horvers, A., Dijkstra, R., Baker, R.S.: Personalized visualizations to promote young learners' SRL: the learning path app. In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge. pp. 330–339 (2020)
11. Schmitz, B., Perels, F.: Self-monitoring of self-regulation during math homework behaviour using standardized diaries. Metacognition and Learning **6**(3), 255–273 (2011)
12. Schunk, D.H.: Goal and Self-Evaluative Influences During Children's Cognitive Skill Learning. American Educational Research Journal **33**(2), 359–382 (Jun 1996)
13. Schunk, D.H., Zimmerman, B.J.: Self-Regulation and Learning. In: Handbook of Psychology, Second Edition. American Cancer Society (2012)
14. Stipek, D., Recchia, S., McClintic, S.: Self-evaluation in young children. Monographs of the Society for Research in Child Development **57**(1), 1–98 (1992)
15. Zheng, L.: 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 (Jun 2016)
16. Zimmerman, B.J.: Investigating Self-Regulation and Motivation: Historical Background, Methodological Developments, and Future Prospects. American Educational Research Journal **45**(1), 166–183 (Mar 2008)