



Conseil ontarien
de la qualité de
l'enseignement
supérieur

Un organisme du gouvernement de l'Ontario

Évaluation des résultats d'apprentissage et amélioration des programmes à l'Université Queen's

Natalie Simper, Brian Frank, Jill Scott et
Jake Kaupp



Publié par

Le Conseil ontarien de la qualité de l'enseignement

1, rue Yonge, bureau 2402
Toronto (Ontario) Canada M5E 1E5

Téléphone : 416 212-3893
Télécopieur : 416 212-3899
Web : www.heqco.ca
Adresse de courriel : info@heqco.ca

Citez cette publication selon le format suivant :

Simper, N., Frank, B., Scott, J. & Kaupp, J. (2018). *Évaluation des résultats d'apprentissage et amélioration des programmes à l'Université Queen's*. Toronto : Conseil ontarien de la qualité de l'enseignement supérieur.



Les opinions exprimées dans le présent document de recherche sont celles des auteurs et ne reflètent pas nécessairement les points de vue ou les politiques officielles du Conseil ontarien de la qualité de l'enseignement supérieur ou d'autres organismes ou organisations qui ont fourni un soutien, financier ou autre, à ce projet. © Imprimeur de la Reine pour l'Ontario, 2018

Sommaire

Le présent rapport décrit une étude longitudinale de quatre ans qui a examiné une gamme de méthodes d'évaluation des résultats d'apprentissage associés à la pensée critique, à la résolution de problèmes, à la communication écrite et à l'apprentissage continu. Cet échantillon provient des Facultés des arts et des sciences, ainsi que de l'ingénierie et des sciences appliquées. Les mesures comprennent des sondages, des entrevues, deux tests standardisés (le Collegiate Learning Assessment Plus et le Critical Thinking Assessment Test) et des rubriques à l'échelle des programmes de l'Association of American Colleges and Universities utilisées pour noter des échantillons du travail des étudiants indépendamment des notes de cours. Les chercheurs ont travaillé avec les instructeurs des cours pour harmoniser l'enseignement, l'apprentissage et l'évaluation, et pour étudier et évaluer l'utilité des instruments utilisés.

Les résultats de l'étude ont permis de quantifier les réalisations longitudinales des résultats des étudiants sur trois instruments, avec une croissance incrémentale des compétences démontrées dans les programmes de premier cycle étudiés. Les résultats de haut niveau étaient les suivants :

- Les compétences des étudiants en pensée critique, en résolution de problèmes et en communication ont augmenté sur les quatre années de leur programme. Les effets étaient détectables sur les tests standardisés ($d = 0,44$ pour le CLA+, et $d = 0,65$ pour le CAT), mais ils étaient plus évidents lors de l'utilisation des rubriques du Valid Assessment of Learning in Undergraduate Education (VALUE). La médiane pour la pensée critique pour la première année atteignait le niveau Point de référence 1, tandis que la médiane pour la deuxième année et la troisième année atteignait le niveau Jalon 2, augmentant au niveau Jalon 3 pour la quatrième année.
- Les étudiants de Queen's ont démontré un niveau plus élevé de compétences en pensée critique que les étudiants comparables de la plupart des établissements homologues qui participent au CLA+ ou au CAT. Par exemple, l'échantillon de quatrième année de Queen's a atteint des résultats correspondant au 87^e centile des établissements qui participent au CLA+.
- La motivation des étudiants était une préoccupation importante relativement aux tests standardisés. Les résultats issus des groupes de discussions d'étudiants suggèrent que pour que les étudiants investissent des efforts dans les tests, les instructeurs doivent accorder de l'importance au test, le contenu de celui-ci doit être pertinent, la planification doit être étudiée attentivement et les résultats devraient être mis à la disposition des étudiants.
- La motivation n'est pas une préoccupation lors de la notation des travaux scolaires en utilisant des rubriques à l'échelle du programme, mais l'harmonisation des travaux de cours avec les dimensions des rubriques est essentielle.
- Le coût relatif de la mise en œuvre de la notation des rubriques VALUE était environ 20 \$ canadiens de moins que le coût de la mise en œuvre des tests CLA+ ou CAT.

- La rétroaction qualitative et quantitative facilitée par l'entremise de rapports et de comptes rendus des départements a déclenché des améliorations au cours.
- Le travail doit se poursuivre afin d'augmenter l'adoption de pratiques efficaces en matière d'évaluation.

Table des matières

Évaluation des résultats d'apprentissage et amélioration des programmes à l'Université Queen's.....	1
Sommaire	2
Liste des tableaux.....	6
Liste des figures	7
Définition des termes.....	8
Introduction	Error! Bookmark not defined.
Instruments.....	13
Collegiate Learning Assessment Plus (CLA+)	13
Critical Thinking Assessment Test (CAT).....	13
Valid Assessment of Learning in Undergraduate Education (rubriques VALUE).....	13
Fiabilité et validité du CLA+, du CAT et des rubriques VALUE.....	14
Sondage sur les orientations axées sur l'apprentissage transférable (OAT).....	15
Entrevues en groupe	16
Méthode	17
Protocoles de notation des rubriques VALUE.....	18
Structure des rubriques.....	18
Utilisation de rubriques pour évaluer le travail des étudiants.....	19
Utilisation des rubriques au niveau institutionnel pour améliorer l'évaluation et les travaux de cours...	19
Échantillon	21
Mise en œuvre du CLA+ et du CAT	23
Résultats du test CLA+	23
Effort lors du test.....	24
Changement longitudinal	25
Analyse de la note de propension.....	27
Résultats du CAT	30
Changement longitudinal	32
Résultats des rubriques VALUE.....	33
Validité et cohérence interne.....	35
Alignement des travaux de cours avec les critères d'évaluation	41

Corrélations entre les instruments	41
Comparaison des coûts	43
Composants qualitatifs	43
Points de vue des étudiants	45
Objectif 1 : Comprendre les motivations des étudiants à participer et à faire des efforts pendant les tests	46
Objectif 2 : Comprendre les perceptions des étudiants d'avoir fait peu d'effort et leur motivation pour passer le test	48
Objectif 3 : Recueillir les idées des étudiants concernant des stratégies de recrutement efficaces et augmenter la motivation des étudiants pour passer les tests à l'avenir	49
Discussion.....	50
Acquisition de compétences cognitives complexes sur un grade de quatre ans.....	50
Acquisition de compétences entre les programmes et les personnes, et lien avec les notes.....	52
Coût et efficacité en matière de temps des instruments.....	53
Résultats supplémentaires et répercussions.....	54
Développement d'un outil de construction de rubriques.....	54
Limitations	55
Conclusions	56
Remerciements	57
Bibliographie	58

Liste des tableaux

Tableau 1 : Questions de recherche, méthodologie et but de l'enquête.....	12
Tableau 2 : Vue d'ensemble des constructs de l'évaluation.....	14
Tableau 3 : Vue d'ensemble de la dimension de l'OAT	16
Tableau 4 : Mappage de la notation des rubriques VALUE	21
Tableau 5 : Lien entre la tâche d'exécution et la note au CLA+.....	25
Tableau 6 : Descriptifs de la totalité de l'échantillon du CLA+, première année et quatrième année.....	25
Tableau 7 : Statistiques descriptives pour la note totale au CLA+.....	26
Tableau 8 : Descripteurs pour le groupe d'échantillons répétés et jumelés au CLA+	27
Tableau 9 : Différence entre le rendement au CLA+ avec les ampleurs de l'effet du coefficient d de Cohen.	30
Tableau 10 : Statistiques descriptives pour le groupe d'échantillons répétés et jumelés au CAT	32
Tableau 11 : Différence entre le rendement de première année et le rendement de quatrième année sur le CLA+ avec les ampleurs de l'effet d de Cohen.....	32
Tableau 12 : Type de travail et taille des échantillons pour la notation des rubriques VALUE par année.....	34
Tableau 13 : Cohérence interne et validité des notes VALUE	35
Tableau 14 : Importance des gains sur chaque rubrique de la première à la quatrième année.....	39
Tableau 15 : Corrélation entre les notes partielles des instruments pour les données biographiques et la moyenne pondérée cumulative.....	42
Tableau 16 : Coût comparatif de chaque instrument (en dollars canadiens).....	43

Liste des figures

Figure 1 : Vue d'ensemble du projet.....	11
Figure 2 : Cadre du processus	18
Figure 3 : Échantillon du projet pour chaque instrument	22
Figure 4 : Pourcentage de l'échantillon du CLA+ à chaque niveau de la norme.....	24
Figure 5 : Distribution des notes des étudiants jumelés sur la tâche d'exécution du CLA+ par programme de grade universitaire	29
Figure 6 : Distribution des notes des étudiants jumelés sur la tâche de sélection de réponses du CLA+ par programme de grade universitaire	30
Figure 7 : Diagramme de quartiles de la distribution des notes au CAT par programme	31
Figure 8 : Diagrammes de quartiles des moyennes des notes au CAT répétées et jumelées pour la première année et la quatrième année.....	33
Figure 9 : Distribution des notes par année pour chaque niveau de la rubrique Pensée critique	37
Figure 10 : Distribution des notes par année pour chaque niveau de la rubrique Résolution de problèmes..	38
Figure 11 : Distribution des notes par année pour chaque niveau de la rubrique Communication écrite	39
Figure 12 : Changement de la note médiane sur chaque dimension	40
Figure 13 : Exemple des résultats d'un rapport de projet d'un étudiant de troisième année	45

Définition des termes

Pensée critique	Pour éviter des débats continus concernant ce qui constitue exactement la pensée critique, la définition de la rubrique VALUE a été adoptée tôt dans le projet : « La pensée critique est une habitude d'esprit caractérisée par l'exploration exhaustive d'enjeux, d'idées, d'artéfacts et d'événements avant d'accepter ou de formuler une opinion ou une conclusion »(AAC&U, 2014). Elle a été opérationnalisée en utilisant les cinq dimensions suivantes : explication des enjeux, données probantes, influence du contexte et des hypothèses, situation de l'étudiant, conclusions et résultats associés.
Compétences cognitives complexes	Ce terme est emprunté à l'étude de la cognition et de la mémoire de travail. Il fait référence à l'acquisition, la rétention et l'application de connaissances et de compétences complexes « telles que celles qui sont impliquées tant dans les décisions immédiates que dans les stratégies à plus long terme » (Logie, Baddeley, Mané, Donchin & Sheptak, 1989, p. 54). Les compétences cognitives complexes (CCC) comprennent plusieurs compétences constituantes interreliées et il faut beaucoup de temps et d'effort pour les maîtriser (Van Merriënboer, 1997). Le terme CCC est utilisé très largement dans ce rapport pour englober collectivement les compétences décrites dans Deller, Brumwell and MacFarlane (2015) comme des compétences cognitives fondamentales (communication), des compétences cognitives d'ordre supérieur (pensée critique et résolution de problèmes) et des compétences transférables (apprentissage continu).
Résultats d'apprentissage	Les résultats d'apprentissage sont des énoncés mesurables des connaissances et des capacités des étudiants, décrites comme « se situant au croisement de concepts (ce que les étudiants savent et comprennent) et des compétences (ce que les étudiants sont capables de faire) » (Roksa, Arum, & Cook, 2016, p. 17).
Apprentissage continu	L'apprentissage continu est un terme qui est largement utilisé, mais pour lequel il n'existe pas de définition universelle. En termes globaux, il décrit un apprentissage autodidacte continu et il comprend la motivation requise pour acquérir constamment des connaissances et des compétences. Le terme est utilisé dans ce rapport pour décrire les attitudes et les comportements d'une personne envers l'apprentissage, spécifiquement « le tempérament (comment une personne se sent) et les stratégies d'apprentissage (compétences, processus et métacognition) » (Simper, Kaupp, Frank, & Scott, 2015, p. 1160).
Orientations axées sur l'apprentissage transférable (OAT)	Cela fait référence à un sondage de réflexion élaboré à l'Université Queen's pour obtenir une mesure ponctuelle de l'apprentissage continu. Il comprend les dimensions suivantes : orientation axée sur les buts, croyance en l'apprentissage, autoefficacité, transfert des connaissances et organisation.

Résolution de problèmes	Le terme « résolution de problèmes » est utilisé ici pour décrire la résolution de problèmes complexes et difficiles, traitant « d'un nombre important d'obstacles qui coexistent simultanément et (de la volonté) de surmonter les obstacles entre un état donné et un but recherché » (Sternberg & Frensch, 2014, p. xi). Le projet a adopté les rubriques VALUE pour l'évaluation des travaux des étudiants, pendant laquelle la résolution de problèmes a été opérationnalisée en utilisant les six dimensions suivantes : définir le problème, identifier les stratégies, proposer des solutions, évaluer les solutions potentielles, mettre en œuvre les solutions et évaluer les résultats.												
Valeur ajoutée	La différence entre le rendement de la première année et celui de la dernière, utilisée pour estimer la contribution d'un établissement d'enseignement aux résultats des étudiants.												
Rubriques VALUE	Rubriques de la Valid Assessment of Learning in Undergraduate Education (VALUE), élaborée par l'Association of American Colleges and Universities (AAC&U).												
Communication écrite	A été évaluée en anglais dans tout format textuel applicable à la discipline (p. ex., réponses courtes/réponses longues, dissertation, rapport ou réflexion). Le projet a adopté les rubriques VALUE pour l'évaluation des travaux des étudiants, pendant laquelle la communication écrite a été opérationnalisée en utilisant les cinq dimensions suivantes : contexte et but, élaboration du contenu, genre et conventions, sources et preuves, et syntaxe et mécanique.												
Codes de programmes	<table border="0"> <tr> <td style="padding-right: 20px;">BA</td> <td>Baccalauréat ès arts (inclut le BA avec spécialisation)</td> </tr> <tr> <td>BCMPH</td> <td>Baccalauréat en informatique (avec spécialisation)</td> </tr> <tr> <td>BED</td> <td>Baccalauréat en éducation (inclut l'éducation concomitante)</td> </tr> <tr> <td>BNSC</td> <td>Baccalauréat en sciences infirmières</td> </tr> <tr> <td>BSC</td> <td>Baccalauréat ès sciences (inclut le BSC avec spécialisation)</td> </tr> <tr> <td>BSCÉ</td> <td>Baccalauréat en sciences de l'ingénierie (le nom du programme, qui était initialement « baccalauréat en sciences appliquées », a changé pendant le projet)</td> </tr> </table>	BA	Baccalauréat ès arts (inclut le BA avec spécialisation)	BCMPH	Baccalauréat en informatique (avec spécialisation)	BED	Baccalauréat en éducation (inclut l'éducation concomitante)	BNSC	Baccalauréat en sciences infirmières	BSC	Baccalauréat ès sciences (inclut le BSC avec spécialisation)	BSCÉ	Baccalauréat en sciences de l'ingénierie (le nom du programme, qui était initialement « baccalauréat en sciences appliquées », a changé pendant le projet)
BA	Baccalauréat ès arts (inclut le BA avec spécialisation)												
BCMPH	Baccalauréat en informatique (avec spécialisation)												
BED	Baccalauréat en éducation (inclut l'éducation concomitante)												
BNSC	Baccalauréat en sciences infirmières												
BSC	Baccalauréat ès sciences (inclut le BSC avec spécialisation)												
BSCÉ	Baccalauréat en sciences de l'ingénierie (le nom du programme, qui était initialement « baccalauréat en sciences appliquées », a changé pendant le projet)												

Introduction

Les compétences cognitives complexes comme la pensée critique, la communication, la résolution de problèmes et l'apprentissage continu constituent des éléments fondamentaux d'une éducation de premier cycle. Ils sont au cœur de cadres larges dans les études supérieures, telles que les résultats d'apprentissage essentiels de l'Association of American Colleges and Universities (AAC&U), et le profil de qualification aux grades de la Lumina Foundation et de Partnership for 21st Century Skills (Johnson, 2009), et ils sont essentiels à la pratique de domaines professionnels tels que l'ingénierie (Kaupp, Frank, & Chen, 2014). Quarante-six pour cent des étudiants des universités canadiennes classent la pensée logique et analytique parmi les trois compétences les plus importantes que leur grade universitaire devrait les aider à acquérir, soit plus du double des domaines suivants au classement, à savoir la connaissance du contenu et la gestion du temps (Canadian University Survey Consortium, 2015). Toutefois, bien que les compétences cognitives complexes soient considérées comme un élément essentiel des études supérieures, elles « sont souvent considérées comme l'un des résultats les plus difficiles à définir, à enseigner et à évaluer » (Deller, Brumwell, & MacFarlane, 2015, p. 13).

En 2013, des chercheurs de l'Université Queen's ont entamé une étude exploratoire longitudinale visant à examiner l'acquisition et l'évaluation de compétences cognitives complexes dans le cadre du Consortium sur l'évaluation des résultats d'apprentissage financé par le Conseil ontarien de la qualité de l'enseignement supérieur (COQES). Les chercheurs ont effectué le suivi du développement de compétences dans des disciplines touchant l'ingénierie, les sciences, les sciences sociales et les humanités. La méthode consistant à tester les étudiants en utilisant des mesures standardisées et en notant les travaux des étudiants pour certains cours à l'intérieur de départements spécifiques. L'étude sur quatre ans a utilisé quatre approches de l'évaluation des compétences cognitives dans plusieurs départements (voir la figure 1) :

- Instruments et sondages standardisés
- Rubriques à l'échelle du programme utilisées pour noter les échantillons des travaux des étudiants
- Entrevues des étudiants et des instructeurs
- Association des données avec des variables démographiques et les notes des étudiants

L'étude a documenté les coûts, l'engagement en temps, les taux de participation, la motivation et les corrélations entre les mesures, et a évalué leur utilité et leur fiabilité. Pendant la première année de l'étude, une évaluation transversale double a été effectuée pour mettre à l'essai les outils auprès des étudiants de première année et de quatrième année afin de comparer le rendement. Quatre questions de recherche très large guidaient l'enquête, et un grand nombre de questions plus spécifiques ont découlé du processus. Les méthodes ont été sélectionnées comme étant appropriées à la question et au but sous-jacent de l'enquête. Le plan de recherche est résumé dans le tableau 1.

Figure 1 : Vue d'ensemble du projet

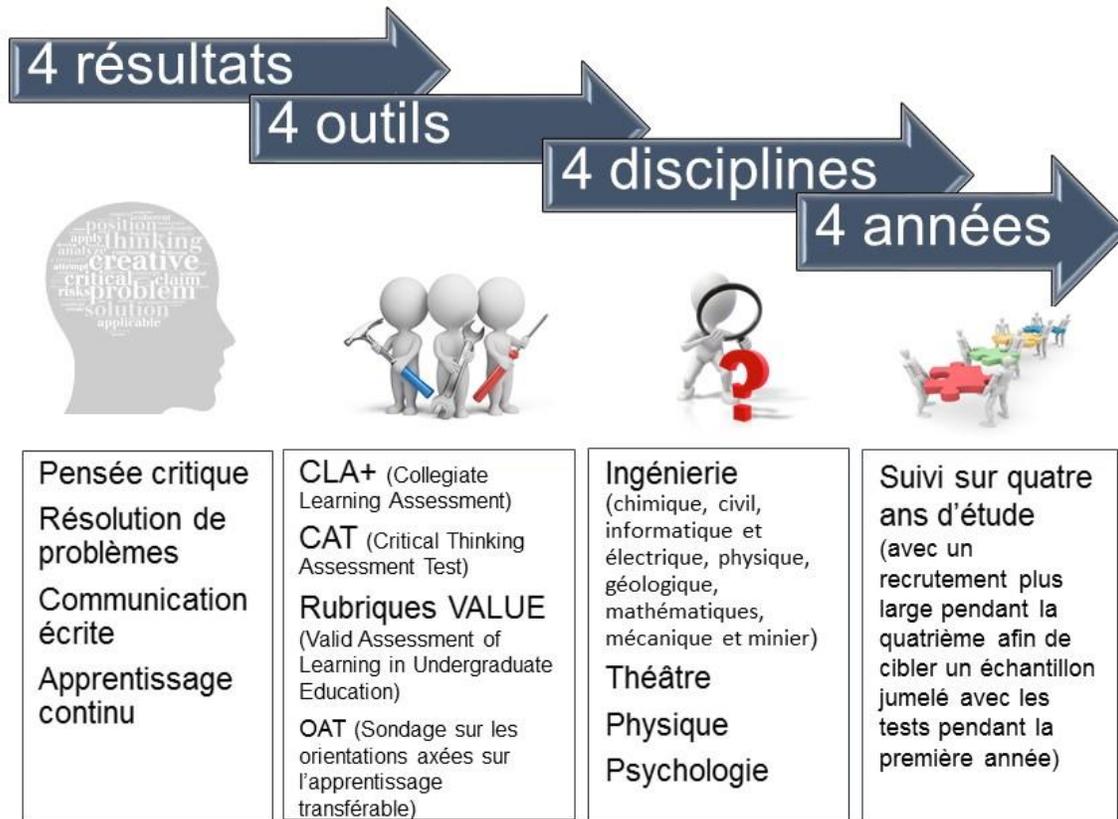


Tableau 1 : Questions de recherche, méthodologie et but de l'enquête

Question de recherche	Méthode : Modèle	Échantillonnage	Objectif
A. Dans quelle mesure les compétences cognitives complexes des étudiants changent-elles entre la première année et la quatrième année d'études de premier cycle?	Aspect quantitatif : Échantillon représentatif d'étudiants de premier cycle	<ul style="list-style-type: none"> • Évaluation CLA+ • Test CAT • Évaluation des rubriques VALUE 	Quantifier les réalisations des étudiants entre la première année et la quatrième année : Valeur ajoutée par l'établissement
B. En quoi l'acquisition de compétences cognitives complexes et l'apprentissage continu varient-ils d'un programme à l'autre et d'une personne à l'autre, et quel est le lien entre les mesures standardisées et les notes de cours?	Aspect quantitatif : Étudiants de premier cycle suivant certains cours	<ul style="list-style-type: none"> • Évaluation CLA+ • Test CAT • Évaluation des rubriques VALUE • Sondage sur les orientations axées sur l'apprentissage transférable (OAT) • Variables démographiques • Moyenne pondérée cumulative 	Pour mieux comprendre les tests standardisés en ce qui concerne leur fiabilité, leur validité et leur pertinence pour différentes disciplines
C. Les données des instruments peuvent-elles être utilisées pour appuyer le développement de compétences dans des cours?	Aspects qualitatifs : Étudiants de premier cycle Instituteurs des cours	<ul style="list-style-type: none"> • Individuellement et en groupe • Entrevues 	Encourager le corps professoral à développer et à évaluer les compétences cognitives complexes dans leurs cours et programmes
D. Dans quelle mesure l'utilisation de ces évaluations dans une université canadienne est-elle réalisable?	Aspect quantitatif : Étude documentaire Aspects qualitatifs : Instituteurs des cours	<ul style="list-style-type: none"> • Documentation financière et issue de la recherche • Comptes rendus ou entrevues 	Pour mieux comprendre les tests standardisés en ce qui concerne leur mise en œuvre, et pour guider l'investissement par les établissements dans les évaluations au cours des prochaines années

Instruments

Collegiate Learning Assessment Plus (CLA+)

Le Collegiate Learning Assessment Plus (CLA+) a été élaboré par le Council for Aid to Education (CAE) (Council for Aid to Education, n.d.). Il s'agit d'un instrument Web de 90 minutes auquel les étudiants accèdent par l'entremise d'un navigateur sécurisé. Il comprend une tâche d'exécution (maximum de 60 minutes) et d'une série de questions à sélection de réponses (maximum de 60 minutes). La tâche d'exécution présente une situation réelle où les étudiants assument un rôle pour régler le problème, identifier une solution ou présenter des conclusions et recommandations découlant d'une analyse minutieuse des faits fournis. La tâche d'exécution est utilisée pour mesurer la pensée critique, la résolution de problèmes et la communication écrite. Les réponses de l'étudiant à la tâche de performance sont notées par un système automatisé en utilisant une rubrique validée. Les questions à sélection de réponses sont des questions fondées sur des documents, à choix multiples ou nécessitant une réponse courte, et sont utilisées pour mesurer le raisonnement scientifique et quantitatif, la lecture critique et l'évaluation, et la critique d'un argument.

Critical Thinking Assessment Test (CAT)

Le CAT a été élaboré par Tennessee Tech, avec le soutien de la National Science Foundation (NSF). Il s'agit d'un test imprimé de 60 minutes. De façon semblable au CLA+, le CAT présente aux étudiants des problèmes réels. Le test a été conçu pour évaluer la pensée critique, les compétences en imagination créatrice, ainsi que la résolution de problèmes non routiniers et la communication efficace (CAT n.d.). Le fournisseur du CAT exige qu'au moins deux membres des établissements participants suivent une formation sur les protocoles de notation. Les représentants qualifiés de l'établissement utilisent alors une clé de correction détaillée pour diriger la notation du CAT effectuée à l'établissement. Le personnel du CAT note un échantillon aléatoire des tests afin d'attester la fiabilité de la notation.

Valid Assessment of Learning in Undergraduate Education (rubriques VALUE)

Les rubriques VALUE ont été développées par l'Association of American Colleges and Universities dans le but de fournir une évaluation valide de l'apprentissage dans les études de premier cycle (AAC&U, 2014). Ces rubriques sont des descriptions larges, neutres sur le plan de la discipline, présentant certains résultats d'apprentissage essentiels des études de premier cycle sélectionnés dans l'initiative Liberal Education America's Promise (LEAP). Des groupes d'experts ont recensé des thèmes communs et ont élaboré des critères de rendement pour chaque rubrique. Les efforts des experts ont été axés sur une démonstration positive des résultats, décrivant les critères de rendement devant être utilisés pour évaluer les démonstrations récapitulatives de l'apprentissage des étudiants. Il y a quatre niveaux de critère de rendement, allant du niveau Référence pour les étudiants qui entrent à l'université au niveau Point culminant pour les étudiants qui viennent de terminer leurs études de premier cycle. Il existe en tout 16 rubriques VALUE, dont quatre ont été utilisées dans l'étude. Les rubriques VALUE Pensée critique, Résolution de problèmes et Communication écrite ont été utilisées sous leur forme publiée, et la rubrique

Apprentissage continu a été adaptée dans le cadre d'un sondage (voir la section ci-dessus concernant le sondage sur les orientations axées sur l'apprentissage transférable).

Fiabilité et validité du CLA+, du CAT et des rubriques VALUE

La fiabilité et la validité de chacun des instruments susmentionnés ont été évaluées rigoureusement. En sélectionnant ces instruments, le Queen's Learning Outcomes Project a tiré parti de données empiriques antérieures qui appuyaient la validité des constructs de l'évaluation et la fiabilité du test ou de la structure des rubriques. Par exemple, lors d'une étude menée par le CLA, les étudiants ont répondu de façon positive à la question de savoir si le CLA+ est une mesure efficace de la pensée critique et de la résolution de problèmes (86,2 %), de la compréhension de lecture (88 %) et de l'écriture (75,8 %). Benjamin, Klein, Steedle, Zahner & Elliot, (2012), et Klein et al. (2009) ont déclaré que le CLA+ avait une bonne corrélation avec deux mesures de la pensée critique, à savoir la Collegiate Assessment of Academic Proficiency et la Measure of Academic Proficiency and Progress (r de la tâche de rendement du CLA = 0,73 à 0,83, r de la tâche de critique d'un argument du CLA = 0,73 à 0,94). Il a été démontré que le CAT avait une corrélation avec l'évaluation de la préparation de l'American College Testing (ACT) ($r = 0,56$ $p < 0,01$); le SAT ($r = 0,57$ $p < 0,01$); la moyenne pondérée cumulative ($r = 0,35$ $p < 0,0$); et le California Critical Thinking Skills Test (CCTST) ($r = 0,64$ $p < 0,01$) (Stein et al., 2006). Stein & Haynes (2011) a également constaté que l'instrument CAT était suffisamment sensible pour détecter les gains dans la pensée critique des étudiants dans différentes disciplines.

Après l'initiative LEAP en 2005 (AAC&U, n.d.), l'AAC&U a travaillé régulièrement pour établir la validité et la fiabilité des rubriques VALUE. Un groupe d'experts diversifié de plusieurs disciplines a été utilisé pour estimer la validité apparente du contenu des rubriques. Dans chaque rubrique, des thèmes communs ont été identifiés et des groupes d'experts ont élaboré des critères de rendement devant être utilisés pour évaluer les démonstrations récapitulatives de l'apprentissage des étudiants. Ces experts ont convenu que les rubriques étaient une mesure efficace et appropriée des constructs sous-jacents (Finley, 2011). Des travaux en cours qui appuient la validité et la fiabilité des rubriques VALUE sont présentés dans Rhodes (2011) and Rhodes & Finley (2013). Les constructs de l'évaluation pour chaque rubrique sont résumés dans le tableau 2.

Tableau 2 : Vue d'ensemble des constructs de l'évaluation

Instrument	Construct de l'évaluation			
	Pensée critique	Résolution de problèmes	Communication écrite	Autre
Collegiate Learning Assessment (CLA+)	<ul style="list-style-type: none"> Lecture critique et évaluation Critique d'un argument 	<ul style="list-style-type: none"> Analyse et résolution de problèmes 	<ul style="list-style-type: none"> Mécanique de l'écriture Efficacité de l'écriture 	<ul style="list-style-type: none"> Raisonnement scientifique et quantitatif

Critical Thinking Assessment Test (CAT)	<ul style="list-style-type: none"> • Évaluation et interprétation d'information 	<ul style="list-style-type: none"> • Résolution de problèmes 	<ul style="list-style-type: none"> • Communication efficace 	<ul style="list-style-type: none"> • Imagination créatrice
Valid Assessment of Learning in Undergraduate Education (rubriques VALUE)	<ul style="list-style-type: none"> • Explication des enjeux • Faits • Influence du contexte et des suppositions • Position de l'étudiant • Conclusions et résultats 	<ul style="list-style-type: none"> • Définition du problème • Identification de stratégies • Solution/hypothèse • Évaluation de la solution • Mise en œuvre de la solution • Évaluation des résultats 	<ul style="list-style-type: none"> • Contexte et but • Développement de contenu • Genre et conventions • Sources de faits • Syntaxe et mécanique 	

Sondage sur les orientations axées sur l'apprentissage transférable (OAT)

Un des objectifs de la recherche consiste à élaborer une méthode d'évaluation du tempérament et des comportements des élèves correspondant à l'apprentissage continu. Cela ne se limitait pas au simple suivi de la démonstration des autres résultats faisant l'objet de l'enquête. L'un des principaux défis consistait à arriver à un consensus sur la définition de l'apprentissage continu. La nature latente des constructs en cause signifiait que nous devions chercher une mesure autodéclarée appropriée. L'instrument le plus prometteur disponible sur le marché était l'Effective Lifelong Learning Inventory (ELLI) (Crick, Broadfoot, & Claxton, 2004). L'étude de recherche devait déjà engager des coûts importants pour l'utilisation du CLA+ et du CAT, et les honoraires pour l'utilisation de l'ELLI étaient hors de portée financière pour le projet. À la place, les chercheurs ont mis en œuvre un inventaire gratuit, le questionnaire Motivated Strategies for Learning (MSLQ) (Pintrich, Smith, Garcia, & Mckeachie, 1993). Nos enquêtes sur l'essai initial du MSLQ ont révélé quelques problèmes et incohérences. Dans notre analyse quantitative, nous avons éprouvé des difficultés à trouver de la cohérence dans les réponses, la structure des facteurs était problématique et les courts temps de réponse moyens suggéraient que les étudiants ne tenaient pas pleinement compte de la nature métacognitive des questions.

Ce résultat nous a incités à élaborer le sondage sur les orientations axées sur l'apprentissage transférable (OAT). Pour obtenir tous les détails relatifs à l'élaboration de l'OAT, voir Simper et al. (2015). Ce processus a nécessité de préciser les échelles du MSLQ, d'adapter la rubrique VALUE Apprentissage continu, et d'incorporer une réponse qualitative pour chaque dimension visant à augmenter l'engagement métacognitif avec l'instrument. L'OAT est une mesure triangulée qui utilise le jumelage quantitatif d'éléments de l'échelle avec une rubrique holistique autoévaluée et une réponse ouverte afin de valider l'évaluation finale. Voir le tableau 3 pour obtenir une vue d'ensemble des dimensions de l'OAT.

Tableau 3 : Vue d'ensemble de la dimension de l'OAT

		Orientation axée sur les buts (motivé)	Croyance en l'apprentissage (apprenant flexible)	Autoefficacité (confiant)	Transfert de connaissances (établit des connexions)	Organisation (apprend de façon indépendante)
Niveau ciblé	4	Explore le sujet en profondeur, motivation intrinsèque	Approche par la maîtrise, avec un contrôle total sur son propre apprentissage	Confiant envers ses propres connaissances et compétences nécessaires pour exceller	Connecte les connaissances de façons innovantes, apprend en profondeur	Techniques organisationnelles adaptatives
	3	Motivé à explorer le sujet avec une certaine profondeur	Croit que l'effort entraînera la réussite	Niveau approprié de connaissances et de compétences pour réaliser ses objectifs	Se réfère à l'apprentissage antérieur	Organisation systématique
Bas niveau	2	Explore le sujet, mais la motivation est principalement extrinsèque	Croit avoir un certain contrôle sur la réussite ou l'échec	Niveau adéquat de connaissances et de compétences	Se réfère à l'apprentissage antérieur de temps à autre	Organise ses processus de travail
	1	Engagement superficiel	Croit que les capacités sont fixes	Doute de son propre niveau de connaissances et de compétences	Mémorise l'information, apprentissage superficiel	Organisation ad hoc

L'outil a été en développement pendant une grande partie de l'étude longitudinale et notre travail dans ce domaine se poursuit : pour cette raison, aucun résultat n'a été inclus dans le présent rapport.

Entrevues en groupe

Les instructeurs participants ont fourni une rétroaction concernant les résultats du projet. Une partie de la rétroaction a été fournie de façon informelle, et quelques-uns des instructeurs ont fourni une rétroaction plus formelle sur l'utilité de chacun des instruments par l'entremise de séances de compte rendu enregistrées qui ont été transcrites à des fins d'analyse. Dans de tels cas, des rapports spécifiques aux cours ont été fournis aux instructeurs. Les rapports expliquaient les réalisations démontrées par les étudiants du cours et présentaient une comparaison entre les réalisations de cet échantillon et ceux des étudiants dans le contexte longitudinal. De plus, la rétroaction des étudiants a été sollicitée afin d'obtenir des suggestions quant aux formats de la présentation des rapports sur les résultats de l'apprentissage des étudiants. Les étudiants ont fourni des renseignements sur les expériences scolaires et les expériences parallèles au

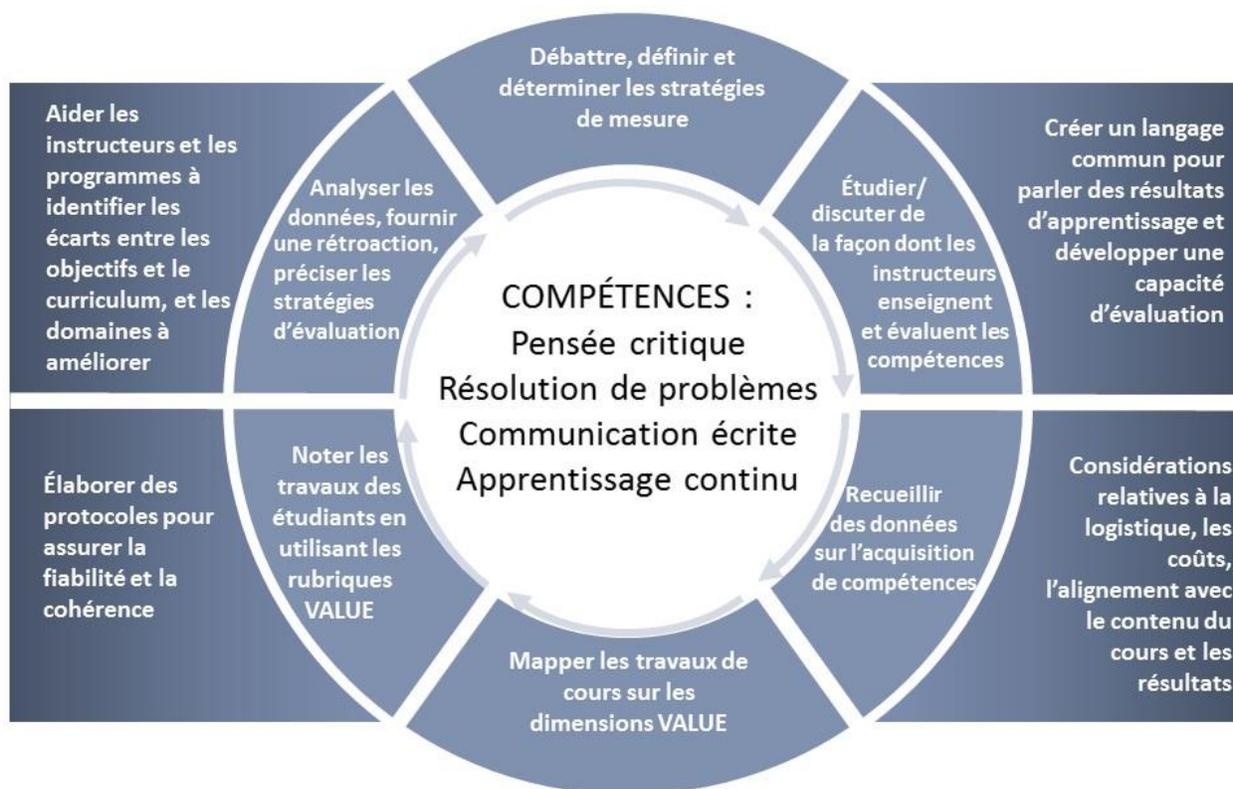
programme qui les aident à développer des compétences en pensée critique, en résolution de problèmes et en apprentissage continu.

Méthode

La recherche a été facilitée par la participation des instructeurs de cours. Chaque année du projet comprenait les six étapes, dont chacune correspondait à un objectif de recherche plus large. La figure 2 affiche les différentes étapes de la recherche et les objectifs correspondants. La première étape consistait à définir les constructs de l'évaluation et à recenser les outils qui correspondaient le mieux aux besoins de l'évaluation. Chaque année, nous avons étudié l'efficacité des outils et nous avons perfectionné l'outil ou le processus de mise en œuvre, selon les besoins. L'étape suivante consistait à travailler avec les instructeurs pour expliquer les constructs et les instruments d'évaluation utilisés pour déterminer les travaux de cours qu'il pourrait être approprié d'inclure dans la notation de la rubrique VALUE. Voici un message typique pour les instructeurs : « Dans lesquels de vos travaux de cours les étudiants démontrent-ils leurs compétences cognitives? »

Par l'entremise de consultations et de discussions, l'équipe a travaillé à l'élaboration d'un langage commun pour décrire les compétences cognitives. À cette étape, bon nombre des instructeurs ont décidé d'adapter un projet afin de mieux susciter une démonstration des compétences cognitives. Il y a eu des étapes de recrutement, de test et de collecte de données sur les cours, puis le mappage de la démonstration des compétences par les étudiants aux critères d'évaluation afin de déterminer leur alignement. L'étude des coûts et de la logistique était également l'un des objectifs de la recherche. Un investissement considérable a été effectué dans la notation des tests CAT et des exemples de travaux de cours. Cela a nécessité le développement de protocoles définis ou leur respect, la formation et l'assurance de l'uniformité longitudinale des marqueurs.

Figure 2 : Cadre du processus



Protocoles de notation des rubriques VALUE

La notation des rubriques VALUE a été effectuée en suivant le processus décrit dans une vidéo éducative mettant en vedette Ashley Finley de l'AAC&U,¹ et résumée ici :

Structure des rubriques

Bien qu'il soit tentant de lire la rubrique d'abord, il est important de consulter la première page afin de se familiariser avec le langage d'encadrement et le but de la rubrique. Le glossaire de la rubrique offre la meilleure occasion de modifier celle-ci et d'ajouter ou de modifier des termes afin de clarifier les descripteurs de rendement sur la page couverture arrière. Il énumère les critères pour les résultats d'apprentissage et souligne la multidimensionnalité nécessaire pour démontrer n'importe lequel d'entre eux : essentiellement, il s'agit d'une ventilation des composantes des compétences. Les repères en matière

¹ Ashley Finley est vice-présidente à l'enseignement et doyenne de Dominican Experience, et agrégée supérieure, AAC&U, Dominican University of California. La vidéo est accessible à <https://youtu.be/josqNZpoZnkm>

de rendement ne sont pas spécifiques à un moment donné : ils sont plutôt développementaux et continus. Le repère Point culminant est intentionnellement mis à gauche afin d'être lu en premier. Si le marqueur lit le repère en premier, cela pourrait artificiellement limiter l'évaluation de l'étudiant; la rubrique ne doit pas être lue en fonction de ce que l'on attend d'un étudiant pendant une année en particulier, mais plutôt du niveau de compétence qu'il démontre.

Utilisation de rubriques pour évaluer le travail des étudiants

Il est suggéré que la meilleure façon de comprendre la rubrique est de l'utiliser lors de séances d'étalonnage, autrement dit d'évaluer une série de travaux des étudiants en utilisant une rubrique. Les participants peuvent discuter du langage et des attentes de la rubrique. Notre expérience indique qu'il est préférable de lancer ce processus en comprenant l'artéfact, sans tenir compte de la rubrique, afin de nous familiariser avec le contenu. Ensuite, nous examinons le langage d'encadrement et les descripteurs du rendement. Pour terminer, nous parcourons systématiquement les critères et nous attribuons un niveau de rendement en nous fondant sur les données probantes contenues dans l'artéfact. Une discussion avec d'autres marqueurs des niveaux attribués, des données probantes utilisées et de l'interprétation du langage des descripteurs de rendement est la clé qui permet de maîtriser l'utilisation des rubriques.

Utilisation des rubriques au niveau institutionnel pour améliorer l'évaluation et les travaux de cours

Le corps professoral peut utiliser les rubriques pour des travaux ciblés, plutôt que pour structurer la totalité du cours. Le corps professoral a plusieurs questions à examiner lorsqu'il structure ces travaux, par exemple comment les étudiants peuvent démontrer ces résultats d'apprentissage dans le projet et la façon dont cette démonstration est fondée sur les niveaux existants de compétences ou les remet en question. Les domaines faibles des critères devraient être identifiés afin de cibler des compétences spécifiques, tout comme les domaines forts afin de maximiser le travail déjà en cours. Les rubriques sont intrinsèquement interdisciplinaires : par exemple, les compétences en écriture ne sont pas spécifiques au département d'anglais, et les compétences quantitatives ne sont pas spécifiques au département de mathématiques. Pour que les établissements encouragent cette utilisation des rubriques, le corps professoral doit être motivé et avoir l'occasion de s'appropriier les rubriques. La façon d'y arriver doit être déterminée au niveau de l'établissement, car les écoles doivent adapter leur approche à leur culture.

La rubrique VALUE a été effectuée selon le protocole suivant :

- a) Créer une compréhension commune :
 - Lire les instructions du travail et les exemples de réponses afin de comprendre la nature et le contexte du projet de cours.
 - Déterminer ce que les étudiants devaient démontrer (cela pourrait, par exemple, nécessiter de lire le rapport de recherche sur lequel les étudiants basaient leurs réponses).
 - Opérationnaliser les « enjeux », les « facteurs contextuels » et les « suppositions » qui sont pertinents pour les réponses des étudiants.

b) Évaluer un exemple de travail :

- Parcourir collectivement la réponse d'un seul étudiant (qui n'était pas incluse dans l'échantillon de recherche) afin de repérer des preuves pour chacune des dimensions à évaluer. Les assistants à la recherche discutent alors du niveau que les données probantes suggèrent que la réponse démontre.
- Évaluer individuellement 5 à 10 échantillons de travaux à la fois, en compilant une liste annotée pour appuyer la décision pour chacun des critères.
- Attribuer et consigner un niveau de rendement (pour chaque dimension) pour les échantillons de travaux.

c) Étalonnage :

- Les deux marqueurs utilisent leurs annotations pour discuter de toute différence entre les niveaux attribués.
- Dans certains cas, l'un ou l'autre des marqueurs ajuste le niveau qu'il a attribué pour une dimension. Comme le processus d'évaluation est fondé sur l'interprétation individuelle, des différences dans la détermination des niveaux étaient observées de temps à autre. Ces changements sont consignés et déclarés comme accord postétalonnage.
- Recommencer le processus d'évaluation et d'étalonnage pour les autres échantillons de travaux. Généralement, plus le nombre de travaux évalués est élevé, moins il y a de différences dans les évaluations.

Pour des raisons financières et logistiques, bon nombre des échantillons de travaux ont été notés par des étudiants de premier cycle qualifiés. Des experts dans la discipline ont été utilisés pour effectuer l'étalonnage avec les étudiants de premier cycle qui effectuaient la notation, le cas échéant. La cohérence longitudinale de la notation a été appuyée en utilisant les mêmes marqueurs pour différents artefacts de cours, à la fois longitudinalement et entre les disciplines. Il y a eu 18 marqueurs des rubriques VALUE, désignés par les lettres A à R dans le tableau 4, sur les quatre années du projet.

Tableau 4 : Mappage de la notation des rubriques VALUE

		Ingénierie				Théâtre		Physique		Psychologie			
		Année				Année		Année		Année			
		1	2	3	4	1	2	1	3	1	2	3	4
Chercheurs (experts en la matière)	A	*	*		*								
	B				*						*	*	*
	C			*									
	D				*								
Étudiants de cycle supérieur (experts en la matière)	E				*								
	F			*					*				
	G				*								
	H							*					
Étudiants de premier cycle (étalonnage avec les experts en la matière)	I					*	*			*	*	*	*
	J	*						*		*			
	K	*				*							
	L	*					*		*		*		
	M	*	*										
	N	*	*										
	O			*					*				
	P			*								*	
Q				*								*	
R				*								*	

Chaque année, après la collecte de données et la notation, l'analyse a été effectuée et les rapports générés pour chacun des cours participants. Dans de nombreux cas, cette étape a incité une enquête supplémentaire sur les anomalies. Dans la mesure du possible, les rapports sur les cours ont été fournis lors d'une séance de compte rendu avec l'instructeur, dans le but de faciliter la réception d'une rétroaction utilisable pour améliorer les cours. Sur invitation, les rapports ont été présentés à des comités d'étudiants de premier cycle dans des départements en particulier. Pour terminer, la recherche a été diffusée tous les ans lors de conférences, des colloques et de webinaires.

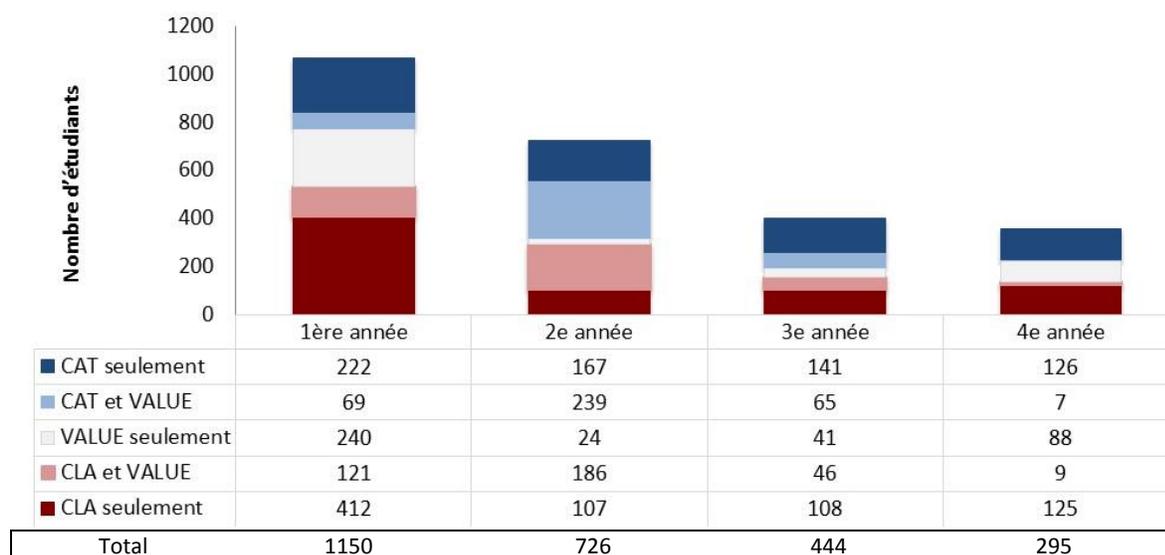
Échantillon

Les étudiants de premier cycle ont été recrutés par les assistants à la recherche dans des cours spécifiques des programmes, y compris de psychologie, de physique et de théâtre de la Faculté des arts et des sciences, et de génie chimique, de génie civil, de génie informatique et électrique, de génie géologique, de génie des

mathématiques, de génie mécanique et de génie minier de la Faculté d'ingénierie et de sciences appliquées. Conformément à nos lignes directrices en matière d'éthique, le consentement de chaque étudiant a été sollicité pendant chaque année du projet. Pendant les premières années du projet, le nombre d'étudiants ayant consenti à la recherche était supérieur à notre capacité de noter les travaux de cours. Le nombre total d'étudiants consentants ayant participé à l'étude pendant la première année s'élevait à 2 697; à 785 pendant la deuxième année, à 599 pendant la troisième année, et à 419 pendant la quatrième année. La figure 3 présente uniquement les étudiants consentants qui ont passé le test ou dont le projet de cours récapitulatif a été évalué selon les rubriques VALUE. De plus, de nombreux étudiants ont consenti et ont commencé le test, mais ne l'ont pas terminé pour différentes raisons.

Tout a été mis en œuvre pour évaluer les travaux des étudiants consentants qui ont terminé un des tests standardisés, mais cela n'a pas toujours été possible. Lorsque le projet avait une composante de groupe, l'utilisation éthique des données signifiait que chacun des membres du groupe devait donner son consentement (cela a présenté un problème particulier pour la première année). De plus, les tests pour la quatrième année ont été effectués séparément des cours inclus dans l'étude; par conséquent, un moins grand nombre des étudiants que nous avons recrutés étaient inscrits aux cours de quatrième année qui avaient des travaux que nous pouvions utiliser. Dans certains cours, les chercheurs ont noté plusieurs travaux afin d'examiner les gains d'apprentissage dans le cadre d'un cours. Pour ce rapport, les chercheurs ont uniquement inclus dans l'analyse le travail récapitulatif que les étudiants ont terminé vers la fin de leur cours.

Figure 3 : Échantillon du projet pour chaque instrument



Mise en œuvre du CLA+ et du CAT

Dans la mesure du possible, les tests CLA+ et CAT ont été effectués dans les cours participants dans le cadre des activités normales du cours. En raison des horaires et des engagements de cours, cela n'a pas toujours été possible. Dans de tels cas, les tests ont été effectués à l'extérieur de la classe, en offrant de la nourriture ou une incitation financière. Pour les tests passés en classe, il a été demandé aux étudiants qui ne devaient pas être dans un laboratoire d'apporter leur propre ordinateur, ou de se rendre à un autre endroit. Les étudiants de la cohorte de première année ont été affectés soit au CLA+, soit au CAT. Tout a été mis en œuvre pour recruter les mêmes personnes pour les mêmes tests sur les quatre années, mais en raison de l'exigence éthique relative au consentement annuel et des différents parcours de cours suivis par les étudiants, les échantillons étaient différents pour chacune des quatre années.

Résultats du test CLA+

Les fournisseurs du test CLA+ affectent des normes de réalisation « inférieures au minimum », « de base », « maîtrise », « avancé » et, depuis 2016, un niveau « accompli ».² Les notes de coupure pour chaque norme sont calculées en utilisant la note totale du test CLA+. La note totale est un composé numérique de la note de la tâche d'exécution et de la note des questions à sélection de réponses. Seuls les étudiants qui répondent à la totalité du test reçoivent leur norme de réalisation. Pendant la surveillance du test, il a été observé que certains étudiants passaient très peu de temps à répondre au test. Puisque le test n'avait pas de conséquences, les résultats étaient considérés comme étant invalides si les étudiants passaient moins de 10 minutes sur la tâche d'exécution (60 minutes prévues), ou s'ils déclaraient dans le sondage de départ qu'ils n'avaient fait « aucun effort » sur le test. En fonction de ces critères, six étudiants de première année, 12 étudiants de deuxième année, trois étudiants de troisième année et un étudiant de quatrième année ont été exclus de l'analyse. Les étudiants exclus représentent 2 % de la population testée. La figure 4 affiche le pourcentage d'étudiants à chaque niveau, avec les tailles de l'échantillon pour le nombre d'étudiants consentants qui ont répondu au test pour chaque année. La moyenne de la note totale au test CLA+ pour la première année était 1 155,3 (écart-type de 117,8), et la moyenne pour la quatrième année s'établissait à 1 211,5 (écart-type de 116,9).

Les données pour les comparaisons américaines ont été tirées du Council for Aid to Education National Report (2016). L'étude comptait 80 établissements participants aux États-Unis, avec une moyenne de 1 053 (écart-type de 74) pour la première année et de 1 126 (écart-type de 74) pour la quatrième année. Les comparaisons ont été calculées en utilisant le coefficient *d* de Cohen ($d = M_2 - M_1 / \text{écart-type cumulé}$), les résultats indiquant que les étudiants de l'Université Queen's commencent à un écart-type de plus que la moyenne américaine ($d = 1,07$), et terminent avec un peu moins d'un écart-type de plus ($d = 0,90$). La moyenne de la première année pour les étudiants de l'Université Queen's se situait au 89^e centile (comparativement aux établissements qui ont participé au CLA+ en 2013-2014), et la moyenne pour la

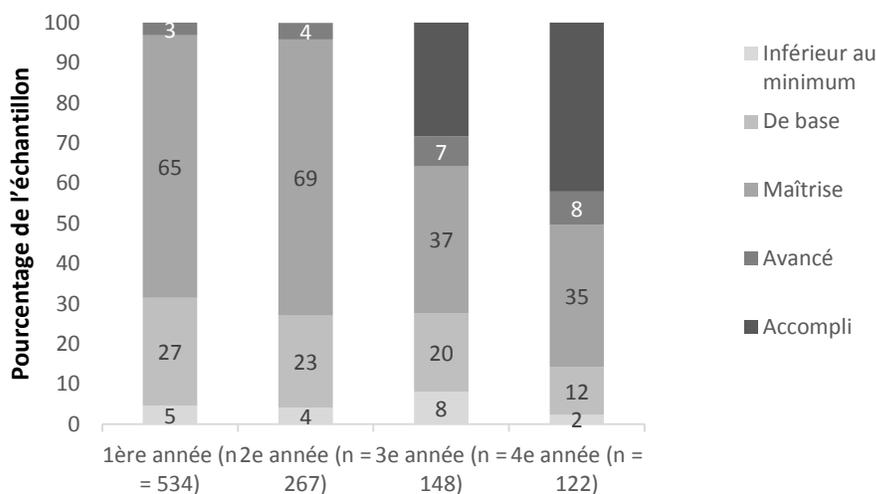
² La recherche relative à l'établissement des normes a été effectuée par la RAND Corporation. Les détails sont accessibles ici : https://www.rand.org/pubs/technical_reports/TR663.html

quatrième année se situait au 87^e centile (comparativement aux établissements qui ont participé au CLA+ en 2016-2017).³

Effort lors du test

Le sondage de départ du CLA+ demande aux étudiants de signaler séparément leur effort sur la tâche d'exécution et sur les sections à sélection de réponses. L'analyse s'est concentrée sur l'effort sur la tâche d'exécution, car celle-ci constitue la partie la plus importante du test et demande plus de travail que la section à sélection de réponses. Les notes moyennes d'effort pour la tâche d'exécution du CLA+ ont diminué au cours des trois premières années (l'effort moyen s'établissait à 3,06 pour la première année, à 2,74 pour la deuxième année et à 2,69 pour la troisième année). Ces résultats ont été calculés après l'exclusion des 22 étudiants dont les résultats étaient considérés comme étant invalides (comme mentionné dans la section ci-dessus). Le tableau 5 trace l'effort par rapport à la note à la tâche d'exécution, affichant la régression par rapport à la moyenne. Le $R^2 = 0,10$ pour l'effort lors de la première et de la deuxième année suggère que 10 % de la variance des notes correspond à l'effort lors des premières années du projet. Ces résultats ont été examinés pendant le projet et en réponse aux recommandations issues de nos entrevues avec des groupes d'étudiants (voir la section ci-dessous), nous avons fourni une incitation de 25 \$ canadiens à chaque personne qui passait les tests et un tirage pour un prix de 750 \$ canadiens. Le tableau 5 présente la moyenne de l'effort au CLA+ pour la quatrième année, à savoir 3,23, mais à la quatrième année, l'effet de l'effort sur la note au test n'était pas important.

Figure 4 : Pourcentage de l'échantillon du CLA+ à chaque niveau de la norme



³ Source pour les données comparatives sur le CLA+ dérivée du CAE. Les rapports complets sont accessibles ici : <http://cae.org/flagship-assessments-cla-cwra/cla-and-cwra-national-report-archive>

Tableau 5 : Lien entre la tâche d'exécution et la note au CLA+

	1 ^{ère} année	2 ^e année	3 ^e année	4 ^e année
Moyenne de l'effort sur la tâche d'exécution	3,06	2,74	2,69	3,23
R ² de l'effort sur la tâche d'exécution	0,102**	0,099**	0,020	0,013
Variance de la note de la tâche d'exécution correspondant à l'effort	10,2 %	9,9 %	2,0 %	1,3 %

** $p < 0,001$

Changement longitudinal

Les statistiques descriptives de la totalité des étudiants de première année et de quatrième année par programme sont présentées dans le tableau 6.

Tableau 6 : Descriptifs de la totalité de l'échantillon du CLA+, première année et quatrième année

Totalité de l'échantillon : Programme	Année	N	Sexe		Moyenne pondérée cumulative pour la première année		Effort à la tâche d'exécution	
			Hommes	Femmes	Moyenne	Écart-type	Moyenne	Écart-type
BA	1	52	17	35	3,3	0,8	3,5	1,0
BA	4	25	3	22	2,8	1,1	3,7	0,9
BCMPH	1	14	13	1	2,6	0,8	3,6	0,9
BCMPH	4	4	3	1	2,8	1,1	3,8	1,0
BCOM	1	1	1	0	3,0	S.O.	2,0	S.O.
BED	1	1	1	0	2,0	S.O.	4,0	S.O.
BED	4	2	0	2	1,3	S.O.	4,0	0,0
BFAH	1	1	0	1	3,7	S.O.	3,0	S.O.
BNSC	1	38	3	35	3,3	0,4	3,6	0,9
BNSC	4	1	1	0	3,8	S.O.	4,0	S.O.
BPHEH	1	2	0	2	3,4	0,4	3,0	1,4

Totalité de l'échantillon : Programme	Année	N	Sexe		Moyenne pondérée cumulative pour la première année		Effort à la tâche d'exécution	
			Hommes	Femmes	Moyenne	Écart-type	Moyenne	Écart-type
BSC	1	107	50	57	3,1	0,8	3,9	0,8
BSC	4	19	7	12	2,9	0,9	3,6	0,9
BSCE	1	210	174	36	3,1	0,8	3,8	0,7
BSCE	4	10	3	7	3,4	0,7	4,1	0,7
CIB	4	1	0	1	NaN	S.O.	3,0	S.O.

L'échantillon de la quatrième année inclus dans l'analyse comptait 122 étudiants. Parmi ceux-ci, 73 avaient passé le test pendant la première année de l'étude (voir les descriptifs dans le tableau 7). Des gains importants ont été observés dans la note totale au CLA+ dans cette population ($t = -4,42$, $df = 72$, $p < 0,0001$) avec une ampleur de l'effet $d = 0,52$, mais il y a également des différences importantes dans la manière dont ces étudiants ont signalé leur effort sur la tâche d'exécution pendant la première année et la quatrième année ($V = 501$, valeur $p = 0,01$), mais pas sur la tâche de sélections de réponses (non illustrée). Les différences des moyennes de l'effort sur la tâche d'exécution dans la dernière colonne du tableau 6 suggèrent sur les étudiants de quatrième année ont fait plus d'effort que ceux de première année dans tous les programmes. Toutefois, cela n'était pas le cas pour les 73 étudiants qui ont passé le test en première année et en quatrième année. Il est possible que ces étudiants fassent moins d'effort en moyenne qu'en première année, ou comme en quatrième année, ils avaient déjà passé le test à plusieurs reprises, ils aient sous-déclaré leur effort.

Les tests ont été analysés pour étudier le changement longitudinal chez les étudiants au sein de programmes individuels. La différence entre la première année et la quatrième année pour les 28 étudiants en génie ($t = -4,2$, $p < 0,001$) et les 15 étudiants en arts ($t = -3,3$, $p < 0,01$) présentait une amélioration importante. La petite taille de l'échantillon limite la généralisation de ces conclusions. Par conséquent, nous examinerons une analyse de la note de propension.

Tableau 7 : Statistiques descriptives pour la note totale au CLA+

Année	N	M	F	Moyenne	Écart-type	Moyenne de l'effort sur la tâche d'exécution	Écart-type de l'effort sur la tâche d'exécution
1	73	26	47	1170,60	122,10	3,84	0,73
4				1232,74	110,42	3,59	0,72

Analyse de la note de propension

La supposition sous-jacente à la déclaration de la différence de valeur ajoutée entre la performance au CLA+ pendant la première et la quatrième année est le fait que le groupe de première année et le groupe de quatrième année sont comparables. Toutefois, les résultats dépendent de l'échantillon et il existe des préoccupations associées à la motivation des étudiants, le temps consacré à la tâche d'écriture, et le fait que « les établissements essaient peut-être d'exploiter le système en choisissant des étudiants de quatrième année dont le rendement scolaire est élevé » (Douglass, Thomson, & Zhao, 2012, p. 320). Pour examiner de plus près les résultats du CLA+ et aborder les préoccupations relatives à la distorsion de l'échantillon, l'effort et la motivation, et les différences de potentiel scolaire, un jumelage analytique des échantillons des étudiants de première année et de quatrième année a été effectué. La première étape consistait à analyser les différences de l'effort sur le test et à étudier l'importance des caractéristiques individuelles, après quoi un processus de jumelage de la propension a été effectué, associant des étudiants de quatrième année à des étudiants de l'échantillon de première année ayant des caractéristiques similaires.

Certaines différences étaient visibles entre l'échantillon de la première année et celui de la quatrième année. La proportion d'étudiantes dans l'échantillon de la quatrième année était plus élevée, et il y avait des différences importantes de l'effort sur la tâche d'exécution pour la première année et la quatrième année et sur la moyenne pondérée cumulative moyenne de la première année (MPCA1). Un modèle linéaire général a été utilisé pour extraire les facteurs de prédiction les plus importants de la note totale au CLA+, qui ont émergé comme étant la MPCA1, l'effort à la tâche d'exécution et à la tâche de sélection de réponses, le temps passé et la langue maternelle. Les étudiants qui ont passé le test CLA+ en première année et en quatrième année ont passé en moyenne plus de temps sur le test pendant la quatrième année (42 minutes) que pendant la première (37 minutes), mais ils ont déclaré un effort inférieur pendant la quatrième année. Il est possible que ce soit l'effort perçu par les étudiants qui était moins élevé, et non l'effort réel. En raison de cette incertitude, l'effort n'a pas été utilisé pour le jumelage de la propension. À la place, le programme de grade universitaire, la moyenne pondérée cumulative pour la première année, le temps consacré à la tâche d'exécution et la langue maternelle ont été utilisés. Cette analyse a été effectuée en utilisant la bibliothèque *MatchIt* dans R (Ho, Imai, King, & Stuart, 2007) en utilisant un jumelage *optimal*. Ce groupe de dossiers a été fusionné avec les étudiants qui ont passé le test CLA+ en première année et en quatrième année, produisant ainsi un groupe final ayant les caractéristiques indiquées ci-dessous. Le tableau 8 affiche les descripteurs des étudiants combinés qui ont repris le test ($n = 73$) et des échantillons jumelés par propension ($n = 49$). La somme de l'échantillon de première année est 122, jumelée à 122 étudiants de quatrième année.

Tableau 8 : Descripteurs pour le groupe d'échantillons répétés et jumelés au CLA+

Échantillon jumelé : Programme	Année	N	Langue maternelle		MPCA1	
			Anglais	Autre	Moyenne	Écart-type
BA	1	25	19	6	3,02	0,80
BA	4	29	23	6	2,84	0,97

Échantillon jumelé : Programme	Année	N	Langue maternelle		MPCA1	
			Anglais	Autre	Moyenne	Écart-type
BCMPH	1	4	4	0	2,90	0,95
BCMPH	4	4	3	1	3,15	1,18
BED	1	6	5	1	3,29	0,65
BED	4	7	5	2	2,92	1,03
BNSC	1	6	3	3	3,12	0,43
BNSC	4	7	3	4	3,22	0,48
BSC	1	41	32	9	2,97	0,98
BSC	4	37	31	6	3,08	0,94
BSCE	1	40	36	4	3,33	0,59
BSCE	4	38	35	3	3,37	0,65

Lorsque la taille de l'échantillon le permet, la tâche d'exécution et la tâche de sélection des réponses ont été tracées dans les figures 5 et 6. Les moyennes les plus élevées en quatrième année ont été atteintes par les étudiants au baccalauréat ès sciences (BSc), mais les gains les plus importants ont été démontrés par les étudiants au baccalauréat ès arts (BA). Ces deux groupes comprennent des étudiants dans des programmes de spécialisation.

Le tableau 9 résume les tests t pour la totalité du groupe et trois grades comptant au moins 20 dossiers par année, et le coefficient d de Cohen pour l'ampleur de l'effet. Test t comparant la première année à la quatrième : $t = -3,4312$, $df = 244$, $p < 0,001$, $d = 0,44$. Comme susmentionné, le total au CLA est une note pondérée échelonnée calculée en utilisant les sous-totaux de la tâche d'exécution et de la tâche de sélection de réponses. La ventilation du CLA+ pour les échantillons jumelés a été incluse en tant qu'annexe 1.

Figure 5 : Distribution des notes des étudiants jumelés sur la tâche d'exécution du CLA+ par programme de grade universitaire

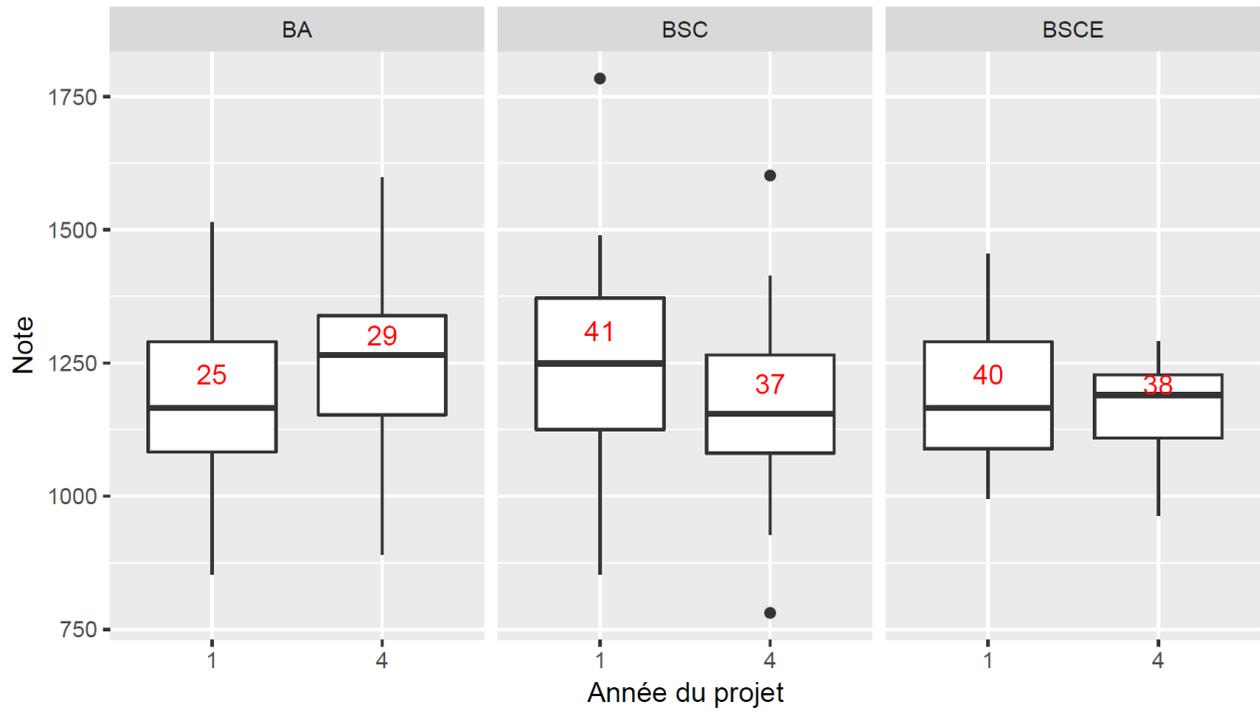


Figure 6 : Distribution des notes des étudiants jumelés sur la tâche de sélection de réponses du CLA+ par programme de grade universitaire

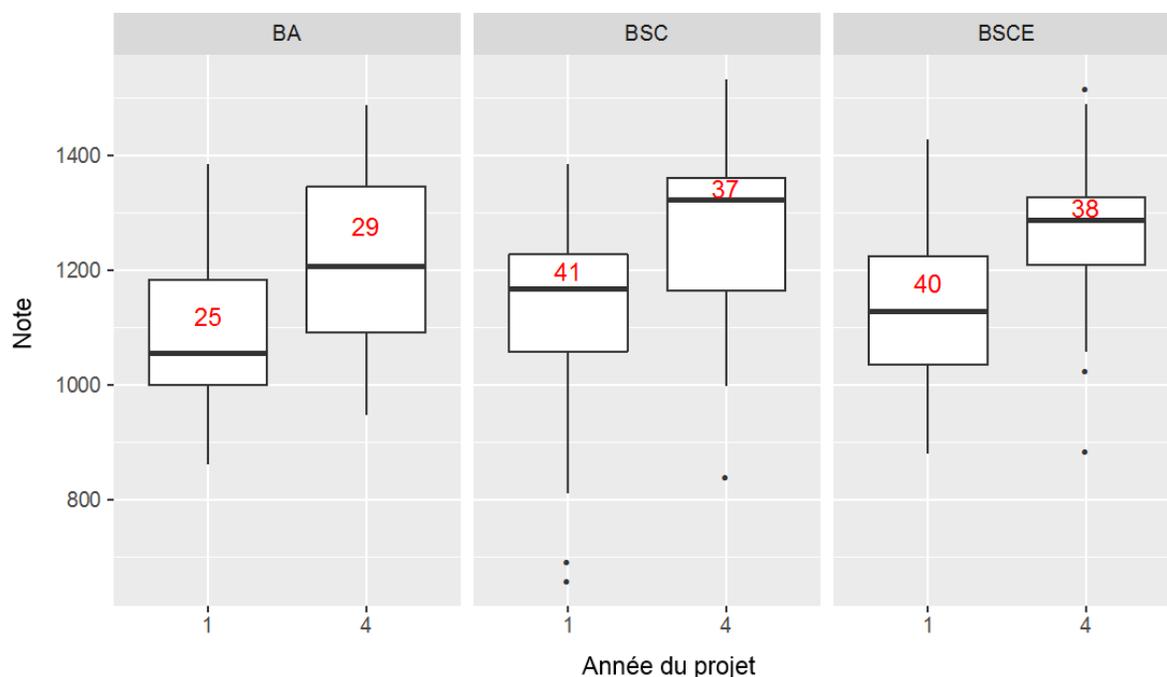


Tableau 9 : Différence entre le rendement au CLA+ avec les ampleurs de l'effet du coefficient d de Cohen.

Groupe	t	df	p	D
Totalité	3,57	242	<0,001	0,46
BA	2,63	52,0	0,011	0,71
BSc	1,15	75,4	0,25	0,26
BSc (Eng.)	2,70	74,5	0,008	0,61

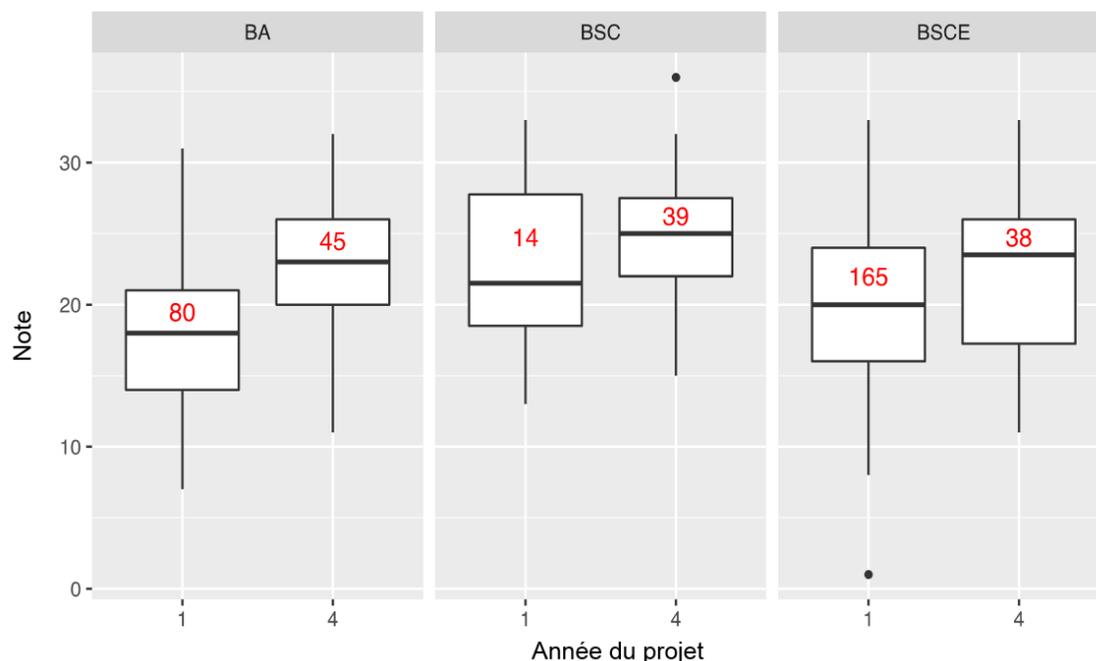
Résultats du CAT

Le CAT contient 15 questions et 1 091 résultats de test valides ont été inscrits pendant les quatre années du projet. L'analyse du coefficient alpha de Cronbach a découvert une cohérence interne moyenne ($\alpha = 0,644$). La moyenne pour la première année était 19,5 (écart-type de 5,5) ($n = 291$), et la moyenne pour la

quatrième année s'établissait à 23,0 (écart-type de 4,7) (n = 133). Les moyennes (mais pas d'écart-type) ont été déclarées par le fournisseur du CAT pour les groupes de la division inférieure et de la division supérieure dans les rapports par établissement. Les différences ont été estimées en utilisant l'écart-type de l'Université Queen's comme dénominateur. Les données de la totalité des échantillons de l'Université Queen's suggèrent que la moyenne de Queen's pour la première année correspondait à 0,84 d'un écart-type, ce qui était meilleur que la moyenne américaine pour la division inférieure; la moyenne de Queen's pour la quatrième année était 0,76 d'un écart-type, ce qui était meilleur que la moyenne américaine pour la division supérieure. La distribution des notes au CAT pour la population ayant passé le test, groupées par programme, est affichée dans la figure 7.

Comme pour l'échantillon du CLA+, l'échantillon du CAT n'est pas resté composé des mêmes personnes pendant les quatre années. L'échantillon de la quatrième année comptait 123 étudiants (qui ne sont pas les mêmes que ceux qui ont passé le CLA+ pendant la quatrième année). Comme pour les résultats du test CLA+, les moyennes les plus élevées en quatrième année ont été atteintes par les étudiants au baccalauréat ès sciences (BSc), mais les gains les plus importants ont été démontrés par les étudiants au baccalauréat ès arts (BA). Ces deux groupes comprennent des étudiants dans des programmes de spécialisation.

Figure 7 : Diagramme de quartiles de la distribution des notes au CAT par programme



Changement longitudinal

Trente étudiants de l'échantillon de quatrième année avaient passé le CAT pendant la première année. Le même processus de correspondance de programmes et de jumelage de la propension utilisé pour l'analyse du CLA+ a été appliqué aux données du CAT, pour obtenir le groupe d'étudiants répétés et jumelés illustré ci-dessous dans le tableau 10. Les résultats correspondants du test t et du coefficient d de Cohen sont présentés dans le tableau 11, et des diagrammes de quartiles présentant les distributions pour les échantillons jumelés comptant plus de 10 étudiants sont affichés dans la figure 8. L'ampleur de l'effet de l'année d'étude sur la moyenne de la note au CAT s'élevait à $d = 0,65$ ($d = M2-M1/\text{écart-type cumulé}$).

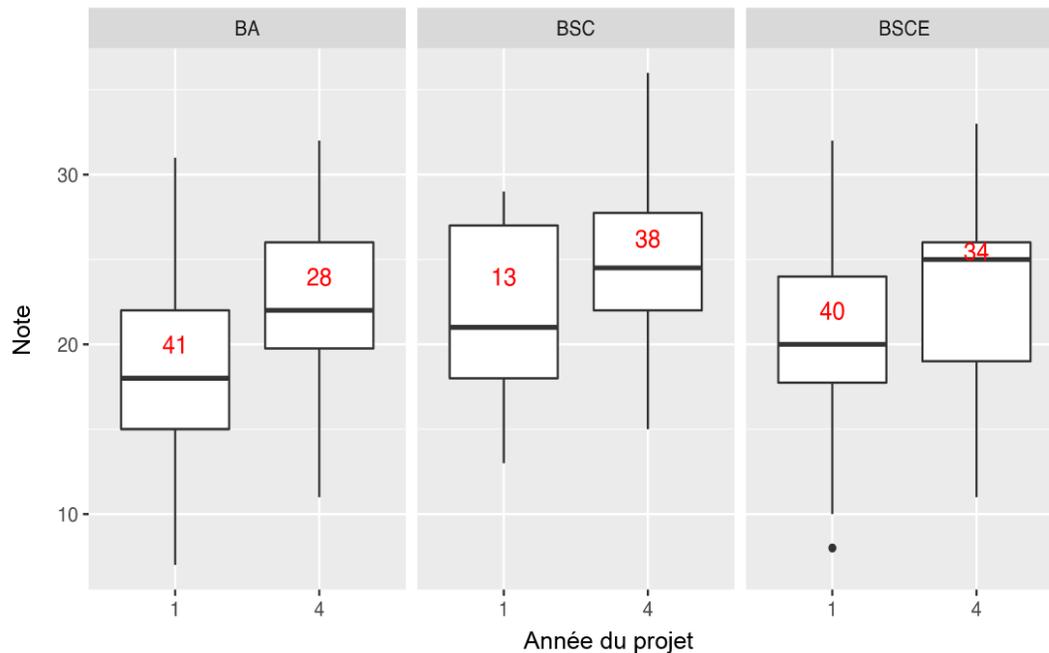
Tableau 10 : Statistiques descriptives pour le groupe d'échantillons répétés et jumelés au CAT

Grade	Année	N	Sexe		Langue		MPCA1	
			Garçons	Filles	Anglais	Autre	Moyenne	Écart-type
BA	1	45	7	38	41	4	2,99	0,73
BA	4	32	3	29	30	2	3,19	0,84
BCMPH	1	2	2	0	2	0	3,80	0,71
BCMPH	4	3	2	1	3	0	2,74	0,64
BED	1	11	2	9	11	0	3,29	0,54
BED	4	4	1	3	2	2	3,09	0,34
BNSC	1	2	0	2	2	0	3,55	0,03
BNSC	4	2	0	2	2	0	3,57	0,00
BSC	1	17	8	9	17	0	3,13	0,79
BSC	4	42	7	35	39	3	3,12	0,74
BSCE	1	58	30	28	44	14	3,15	0,59
BSCE	4	52	25	27	45	7	3,13	0,61

Tableau 11 : Différence entre le rendement de première année et le rendement de quatrième année sur le CLA+ avec les ampleurs de l'effet d de Cohen.

Groupe	t	df	p	D
Totalité	5,37	266,8	<0,001	0,65
BA	3,54	71,7	<0,001	0,80
BSc	1,74	25,7	0,093	0,54
BSc (Eng.)	3,14	105,2	0,002	0,60

Figure 8 : Diagrammes de quartiles des moyennes des notes au CAT répétées et jumelées pour la première année et la quatrième année



Résultats des rubriques VALUE

Neuf types de projets récapitulatifs ont été identifiés comme suscitant les compétences de pensée critique, de résolution de problèmes et de communication écrite. Les artefacts des étudiants issus des travaux de cours ont été recueillis et évalués en utilisant le protocole indiqué, deux évaluateurs notant chaque travail et arrivant à un consensus (pour plus de détails, voir la section sur les évaluations VALUE ci-dessus). Généralement parlant, plus le nombre d'échantillons notés pour un travail en particulier était élevé, plus l'accord entre les évaluateurs était élevé. Toutefois, la complexité de l'échantillon était également un facteur. Certains des travaux demandaient une réponse de deux pages et d'autres un rapport de 40 pages. Sont inclus dans le présent rapport les travaux récapitulatifs pour les cours dans les départements participants. Le tableau 12 affiche les différents types de projets et les tailles des échantillons correspondants.

Tableau 12 : Type de travail et taille des échantillons pour la notation des rubriques VALUE par année

Groupe d'année	Département	Type de travail	n
1 ^{ère} année	APSC	Laboratoire de conception	39
	DRAM	Dissertation	90
	PSYC	Question d'examen	106
	PHYS	Rapport de conception	207
	TOTAL		442
2 ^e année	PSYC	Tâche d'exécution	142
	APSC	Rapport de conception	56
	PSYC	Rapport de conception	46
	CIVL	Rapport de conception	26
	ELEC	Rapport de conception	19
	ENPH	Rapport de conception	67
	MECH	Rapport de conception	21
	PHYS	Rapport de conception	10
TOTAL		387	
3 ^e année	PHYS	Rapport de laboratoire	9
	PSYC	Rapport de conception	29
	CIVL	Proposition de projet	36
	DRAM	Proposition de recherche	92
	ELEC	Projet de conception	32
	TOTAL		198
4 ^e année	MECH	Conception	25
	ELEC	Proposition et examen	15
	PSYC	Thèse	15
	CIVL	Thèse	9
	TOTAL		64

L'accord du pré-étalonnage (notation indépendante) pour les travaux de première année était de 64 %, de 68 % en deuxième et en troisième année, et de 57 % en quatrième année. Après l'étalonnage, l'accord pour les travaux de première année était de 99 %, de 93 % en deuxième et 96 % en troisième année, et de 100 % en quatrième année. Une représentation graphique du pourcentage d'accord entre les évaluateurs avant et

après l'étalonnage est incluse en tant qu'annexe 2. L'écart initial plus élevé dans les notes de quatrième année a nécessité des conversations plus longues, mais l'accord final était plus élevé. Le coefficient Kappa de Cohen a été calculé pour chacun des types de projets au niveau de la dimension individuelle. Une fiabilité inhabituellement basse de $K = 0,5$ a été obtenue pour la dimension de données probantes de la proposition de recherche, mais pour le reste, $K > 0,7$. Pour toute dimension controversée (où les évaluateurs n'étaient pas d'accord), la note a été arrondie au niveau inférieur, en se disant que le niveau de la dimension n'avait pas été intégralement démontré.

Validité et cohérence interne

Les corrélations entre les dimensions VALUE et la moyenne pondérée cumulative de session ont été calculées en tant que mesure de la validité convergente. Des corrélations importantes entre la pensée critique, la résolution de problèmes et la communication écrite telles qu'elles étaient évaluées sur les rubriques VALUE et le rendement scolaire des étudiants (voir le tableau 13) suggèrent l'existence d'un lien. Toutefois, les coefficients peu élevés suggèrent que la moyenne pondérée cumulative capture plus que ces compétences cognitives complexes. Un tableau général des descripteurs a été composé afin que les chercheurs puissent discuter de dimensions des rubriques VALUE Pensée critique, Résolution de problèmes et Communication écrite avec les instructeurs, sans se laisser distraire par les détails des critères spécifiques contenus dans chacune des rubriques. Les descripteurs sont présentés dans le tableau 13.

Les figures 8, 9 et 10 affichent les distributions des notes en pourcentage de la population échantillonnée pour chacune des rubriques pour chaque année du projet. Une étude supplémentaire des tendances centrales a révélé que la médiane pour la pensée critique pour la première année atteignait le Point de référence 1, tandis que la médiane pour la deuxième année et la troisième année atteignait le Jalon 2, passant au Jalon 3 pour la quatrième année. En ce qui concerne la résolution de problèmes et la communication écrite, les médianes étaient Jalon 2 pour la première et la deuxième année et Jalon 3 pour la troisième et la quatrième année. Les changements entre la première et la quatrième année étaient significatifs pour toutes les rubriques, en utilisant le test de Wilcoxon avec correction de continuité.

Tableau 13 : Cohérence interne et validité des notes VALUE

Rubrique VALUE	Dimension	Descripteur	n	Corrélation – moyenne pondérée cumulative de session	Tableau
Pensée critique	Explication des enjeux	Enjeu/Problème pris en considération; renseignements pertinents nécessaires pour la compréhension	1174	0,260**	0,886
	Utilisation des données probantes	Sélection et utilisation de renseignements pour étudier un point de vue ou une conclusion	1096	0,263**	

Rubrique VALUE	Dimension	Descripteur	n	Corrélation – moyenne pondérée cumulative de session	Tableau
	Contexte et suppositions	Ses propres suppositions et celles des autres, et évaluation de la pertinence des contextes	1054	0,279**	
	Position des étudiants	Perspective subjective/objective – thèse/hypothèse	1004	0,281**	
	Conclusions et résultats	Évaluation des conséquences et des implications	1080	0,290**	
Résolution de problèmes	Définition du problème	Énoncé contextuel du problème	975	0,272**	0,804
	Hypothèses relatives à la solution	Approches multiples	917	0,284**	
	Identification de stratégies	Compréhension, sensibilité aux questions contextuelles	857	0,260**	
	Mise en œuvre de la solution	Tient compte de l'historique du problème, examine la logique et le raisonnement, examine la faisabilité de la solution et soupèse les répercussions de la solution	575	0,253**	
	Évaluation des résultats	Aborde les multiples facteurs contextuels (mise en œuvre)	624	0,265**	
	Évaluation de la solution	Examine les résultats par rapport au problème défini en tenant compte de la nécessité de travail supplémentaire	984	0,287**	
Communication écrite	Contexte et but	Public, but et circonstances entourant les tâches d'écriture	1160	0,341**	0,856
	Développement de contenu	Utilise un contenu approprié et pertinent	1151	0,366**	
	Sources de faits	Règles formelles et informelles inhérentes aux attentes en matière d'écriture dans des formes ou des domaines scolaires particuliers	1075	0,198**	
	Genre et conventions	Démontre l'utilisation de sources dignes de foi et pertinentes pour développer des idées qui sont appropriées pour la discipline et le genre d'écriture	1113	0,285**	

Rubrique VALUE	Dimension	Descripteur	n	Corrélation – moyenne pondérée cumulative de session	Tableau
	Syntaxe et mécanique	Utilise un langage qui communique la signification pour les lecteurs avec clarté et fluidité	1165	0,299**	

** La corrélation est significative au niveau 0,01 (bilatéral).

Remarque : Corrélation calculée en utilisant le coefficient de corrélation des rangs de Spearman.

Figure 9 : Distribution des notes par année pour chaque niveau de la rubrique Pensée critique

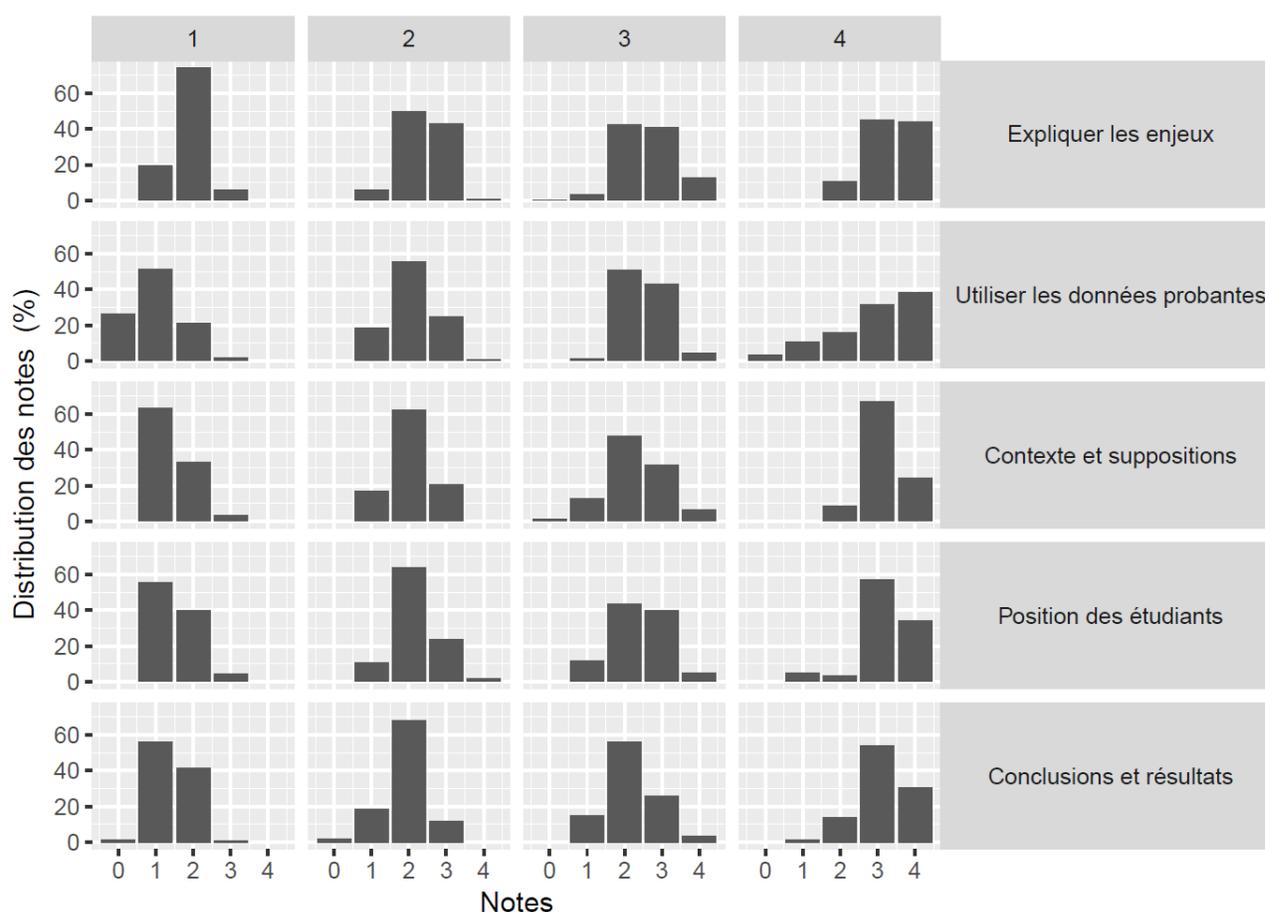


Figure 10 : Distribution des notes par année pour chaque niveau de la rubrique Résolution de problèmes

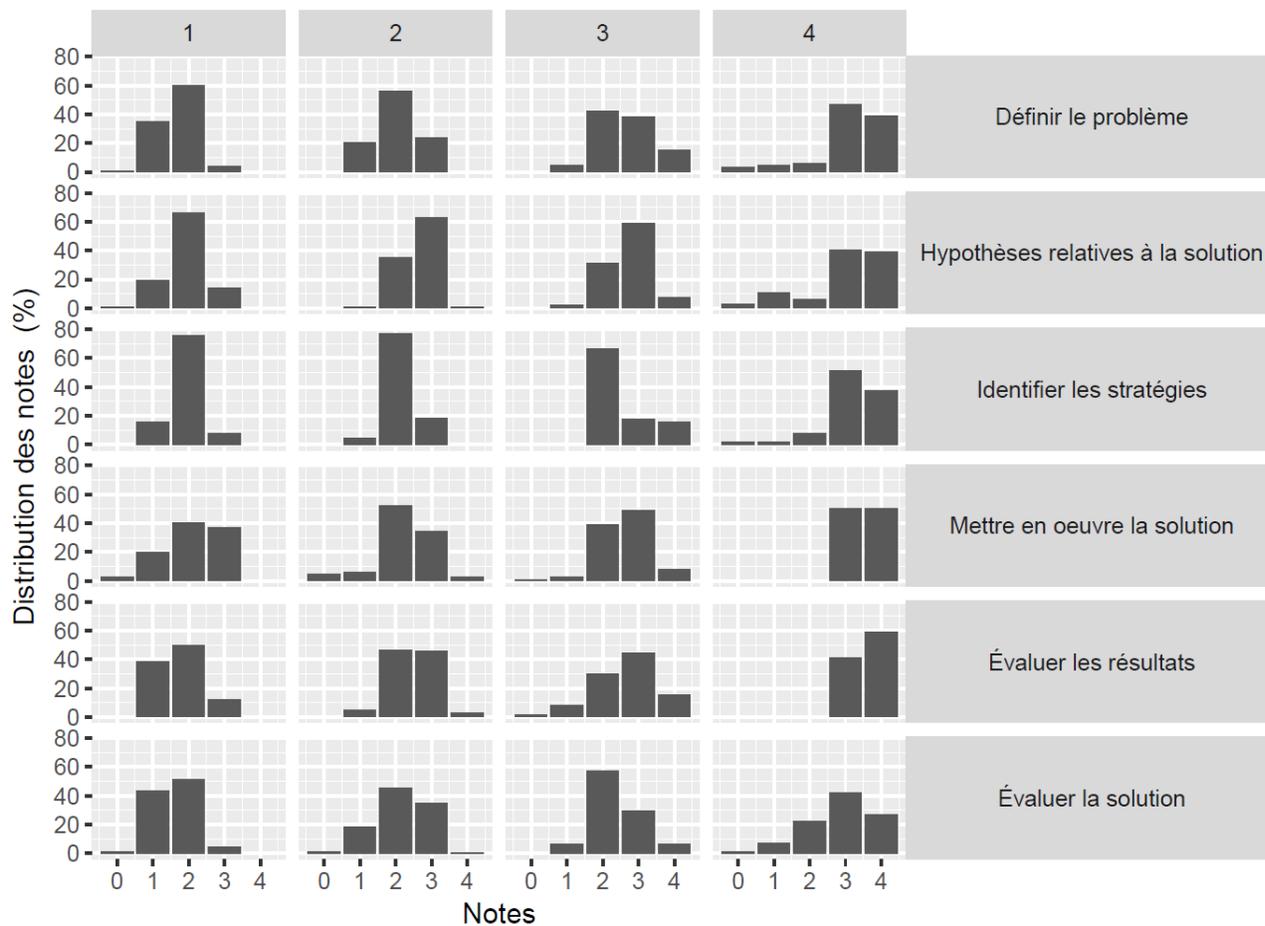
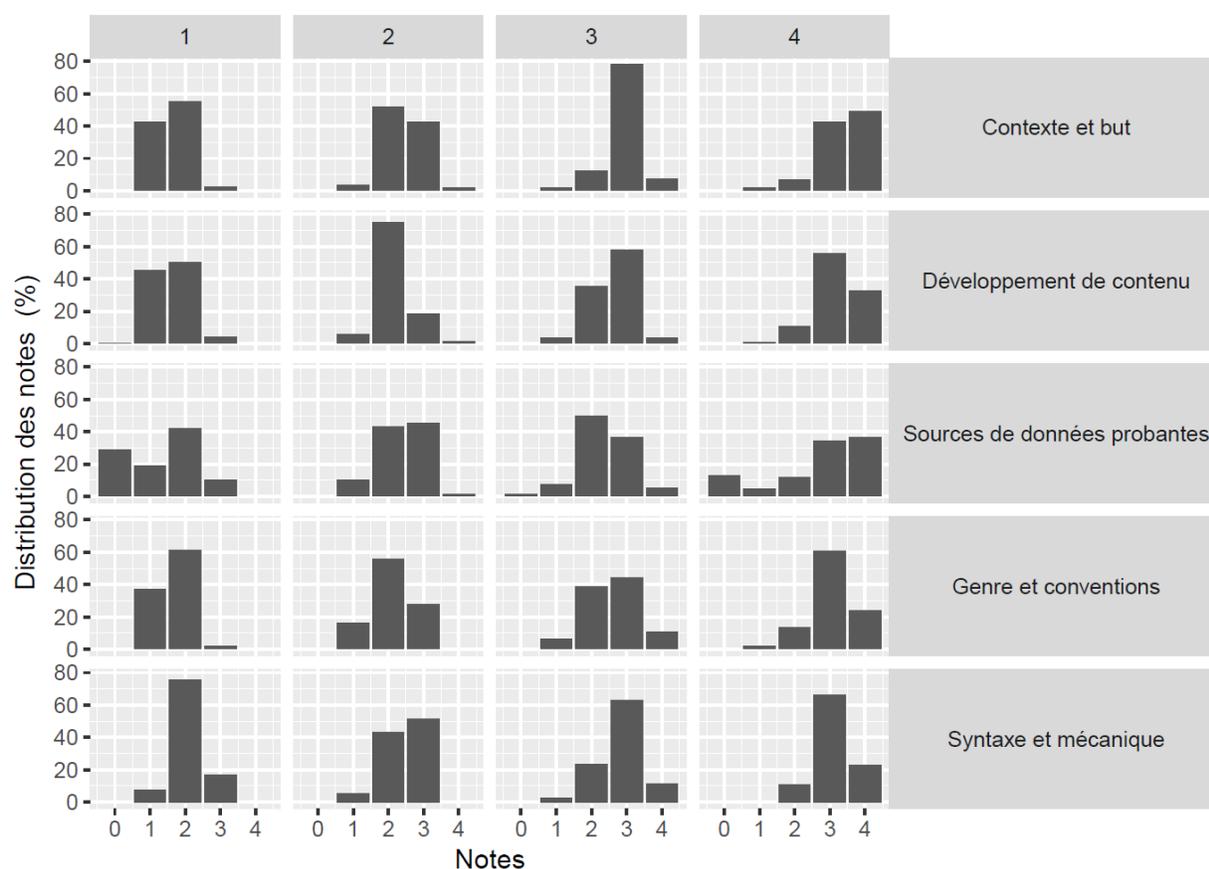


Figure 11 : Distribution des notes par année pour chaque niveau de la rubrique Communication écrite

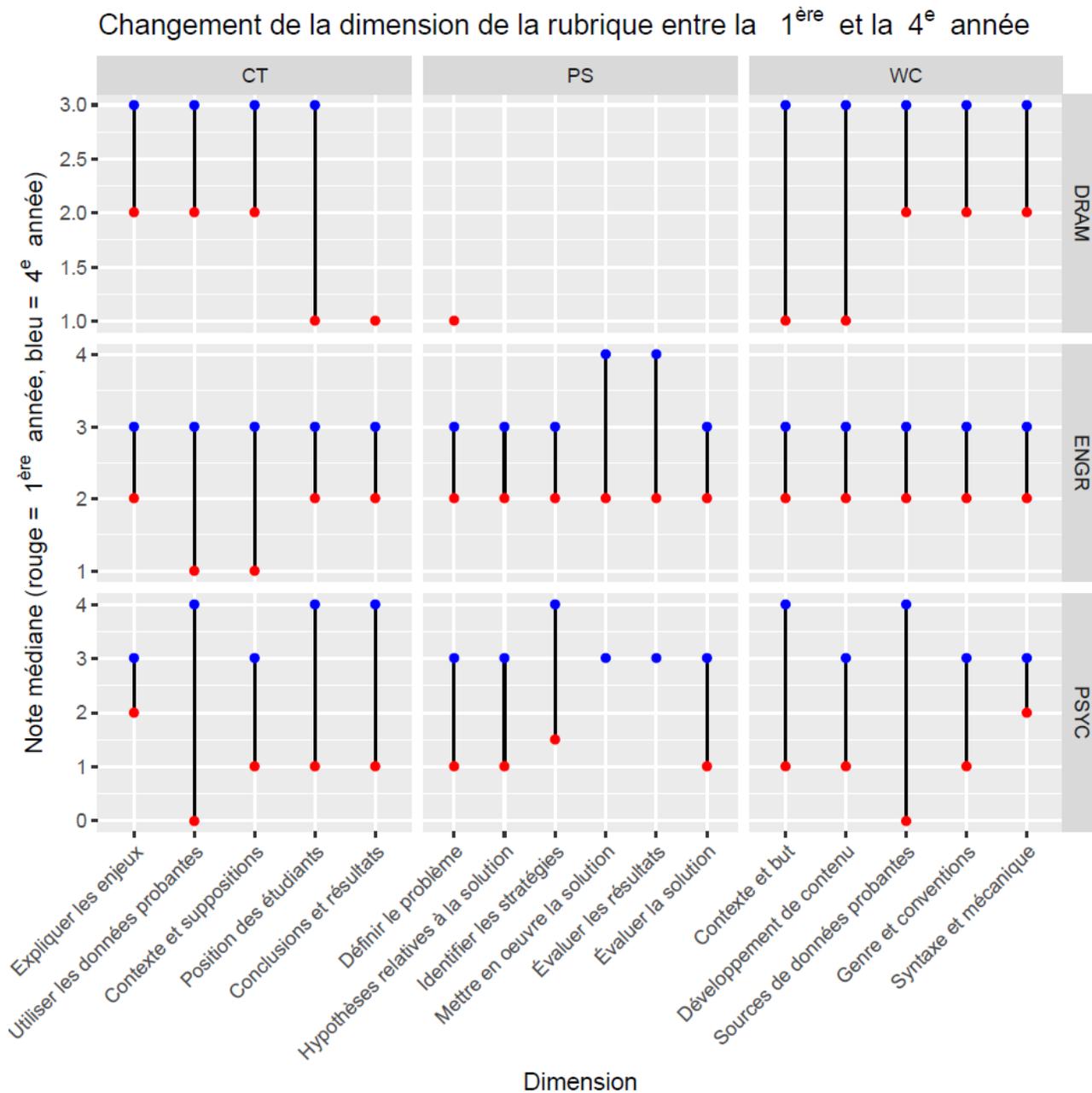


Les changements entre la première et la quatrième année étaient significatifs pour toutes les rubriques, en utilisant le test de Wilcoxon avec correction de continuité. Le tableau 14 illustre la variable à tester et l'ampleur de l'effet r par rubrique. Une comparaison des notes médianes pour la première année et la quatrième année est présentée dans la figure 12.

Tableau 14 : Importance des gains sur chaque rubrique de la première à la quatrième année

	W	p	r
Pensée critique	54424,5	<0,001	0,61
Résolution de problèmes	69040,5	<0,001	0,50
Communication écrite	92994,5	<0,001	0,60

Figure 12 : Changement de la note médiane sur chaque dimension



Alignement des travaux de cours avec les critères d'évaluation

Certains des travaux de cours étaient plus alignés que d'autres avec les dimensions évaluées. Si le travail ne suscitait pas une démonstration de données probantes pour une dimension en particulier (étape 1 du processus de notation), la dimension n'était pas évaluée. En général, plus l'alignement était élevé, plus le score de la rubrique VALUE était élevé. Les exceptions étaient la question d'examen pour la première année, lorsque le temps et le nombre de mots étaient limités, et un des rapports de conception pour la deuxième année, qui provenant d'une matière technique et ne se prêtait pas bien à une évaluation des dimensions des rubriques VALUE. Généralement parlant, plus le nombre de dimensions évaluées était élevé, plus la note médiane sur la rubrique était élevée. L'annexe 3 fournit un graphique à code de couleurs qui illustre la relation entre le pourcentage de dimensions évaluées par rapport au niveau moyen de réalisation, groupé par année et par sujet. Plus la pente de la ligne est élevée, plus le lien est significatif.

Corrélations entre les instruments

Parce que nous travaillions avec différents types de données, les corrélations du classement du coefficient de corrélation des rangs de Spearman ont été utilisées pour comparer les instruments, les données biographiques (sexe et langue) et les moyennes pondérées cumulatives. Les valeurs médianes ont été calculées pour les rubriques VALUE qui comprenaient des données ordinales. Le sexe a été codé de la façon suivante : homme = 1, femme = 2; et la langue a été codée par langue maternelle : anglais = 1, français = 2 et autre = 3. Les résultats présentés dans le tableau 15 sont fondés sur les données cumulatives des différentes années du projet et ont été calculés en utilisant une analyse par paires. Les corrélations entre la moyenne pondérée cumulative, le sexe et la langue n'ont pas été incluses dans le tableau pour éviter la convergence des comparaisons par paires dans l'échantillon (c.-à-d. les données étaient contraintes par année, jumelées par instrument, avec les points de données valides inclus).

Les notes des rubriques VALUE étaient fortement liées, mais le lien entre les notes des tests et les notes des rubriques était faible. La corrélation la plus forte était entre la note au CAT et la note de la rubrique Résolution de problèmes score ($r(303) = 0,234$ $p < 0,01$). Les notes des rubriques VALUE avaient un lien plus fort avec la moyenne pondérée cumulative des étudiants que les notes au CLA+ ou au CAT. La conclusion des corrélations négatives entre le sexe et les notes des rubriques VALUE serait que les hommes ont obtenu des résultats marginalement supérieurs à ceux des femmes sur les travaux de cours. Toutefois, le lien entre le sexe et les notes de test n'était pas significatif. La langue n'était pas significative dans la notation des rubriques VALUE, mais il y avait une corrélation faible qui suggérait que les anglophones ont obtenu un résultat légèrement supérieur aux questions à sélection de réponses du CAT et du CLA+ que les étudiants francophones ou allophones. (Remarque : huit personnes ont passé le CLA+ et le CAT pendant la même année.)

Tableau 15 : Corrélation entre les notes partielles des instruments pour les données biographiques et la moyenne pondérée cumulative

		Rubriques VALUE			CLA+			Note au CAT
		PC	TE	SR	TE	SR	T	
Rubrique VALUE –	Pensée critique (PC)	-						
valeur médiane (n)	Résolution de problèmes (RP)	0,720** (1051)	-					
	Communication écrite (CE)	0,767** (1183)	0,715** (1050)	-				
CLA+ (n)	Tâche d'exécution (TE)	0,031 (366)	0,002 (341)	0,043 (365)	-			
	Sélection de réponses (SR)	0,202** (351)	0,198** (326)	0,231** (350)	0,182* (1131)	-		
	Note totale (T)	0,155** (347)	0,149** (322)	0,179** (346)	0,714* (1131)	0,792** 1131	-	
Note au CAT (n)		0,133** (384)	0,227** (306)	0,191** (384)	- 0,829* (8)	0,902** (8)	0,659 (8)	-
Moyenne pondérée cumulative de session (n)		0,304** (1184)	0,357** (1051)	0,348** (1183)	0,145* (1181)	0,260** 1135	0,269** 1131	0,253** 1089
Sexe (n)		-0,087** (1184)	- 0,166** (1051)	- 0,101** (1183)	-0,029 (1181)	-0,012 1135	-0,024 1131	-0,024 1091
Langue (n)		-0,023 (1135)	-0,019 (1002)	-0,028 (1134)	-0,012 (1114)	-0,075* 1068	-0,052 1064	- 0,108** 1033

** La corrélation est significative au niveau 0,01 (bilatéral).

Comparaison des coûts

Les coûts associés à l'utilisation des approches du test CLA+, du test CAT et des rubriques VALUE ont été fondés sur un échantillon nominal de 100 étudiants (voir le tableau 16). Ils ont été calculés en additionnant les frais pour l'instrument, les coûts accessoires (frais de formation et salaires) et les salaires pour les marqueurs (les étudiants de premier cycle ont reçu 14 \$ canadiens l'heure et les étudiants de cycle supérieur 24 \$ canadiens l'heure). Bien que les frais par étudiant ayant passé l'examen s'élevaient à 35 \$ américains pour le CLA+ et 9,95 \$ américains pour le CAT, une fois les coûts supplémentaires pris en considération (et en utilisant un taux de change de 1,2 \$ canadien), le total des coûts s'élevait respectivement à 51 \$ canadiens et 47,54 \$ canadiens. La notation des échantillons VALUE a nécessité des durées différentes, allant de 30 minutes à trois heures, selon la complexité du travail. La majorité des échantillons de première année ont été notés par des étudiants de premier cycle, tandis que la plupart des échantillons de quatrième année ont été marqués par des chercheurs ou des étudiants de cycle supérieur (voir le tableau 4). Le coût moyen pour tous les travaux notés s'élevait à 32 \$ canadiens.

Tableau 16 : Coût comparatif de chaque instrument (en dollars canadiens)

	Formation ou soutien technique	Frais pour l'instrument par 100 étudiants	Surveillance des tests (4 séances – 100 étudiants)	Coûts de notation	TOTAL	Coût par étudiant
CLA+	100,00 \$	4 200,00 \$	800,00 \$	-	5 100,00 \$	51,00 \$
CAT	1 000,00 \$	1 194,00 \$	560,00 \$	2 000,00 \$	4 754,00 \$	47,54 \$
Notation des rubriques VALUE	200,00 \$	-	-	3 000,00 \$	3 200,00 \$	32,00 \$

Composants qualitatifs

Des méthodes qualitatives ont été utilisées pour étudier les avantages perçus de chacun des outils. Deux séances d'entrevues ont été effectuées, avec quatre instructeurs de deux départements, dont chacun donnait un cours utilisé dans le cadre du projet. Chaque instructeur a reçu un rapport sommaire et a été invité à le commenter. Lorsque c'était possible, les séances de compte rendu ont été enregistrées et transcrites à des fins de suivi et d'évaluation. Nous avons obtenu des commentaires très positifs de la part des instructeurs concernant leur implication dans le projet et la rétroaction qu'ils ont reçue :

« Du point de vue de notre département, j'ai l'impression d'avoir reçu gratuitement quelque chose qui a une immense valeur. N'est-ce pas? Comme si nous venions de recevoir une énorme quantité de renseignements utiles sur la conception de l'enseignement. C'est comme de l'or. Et cela a été

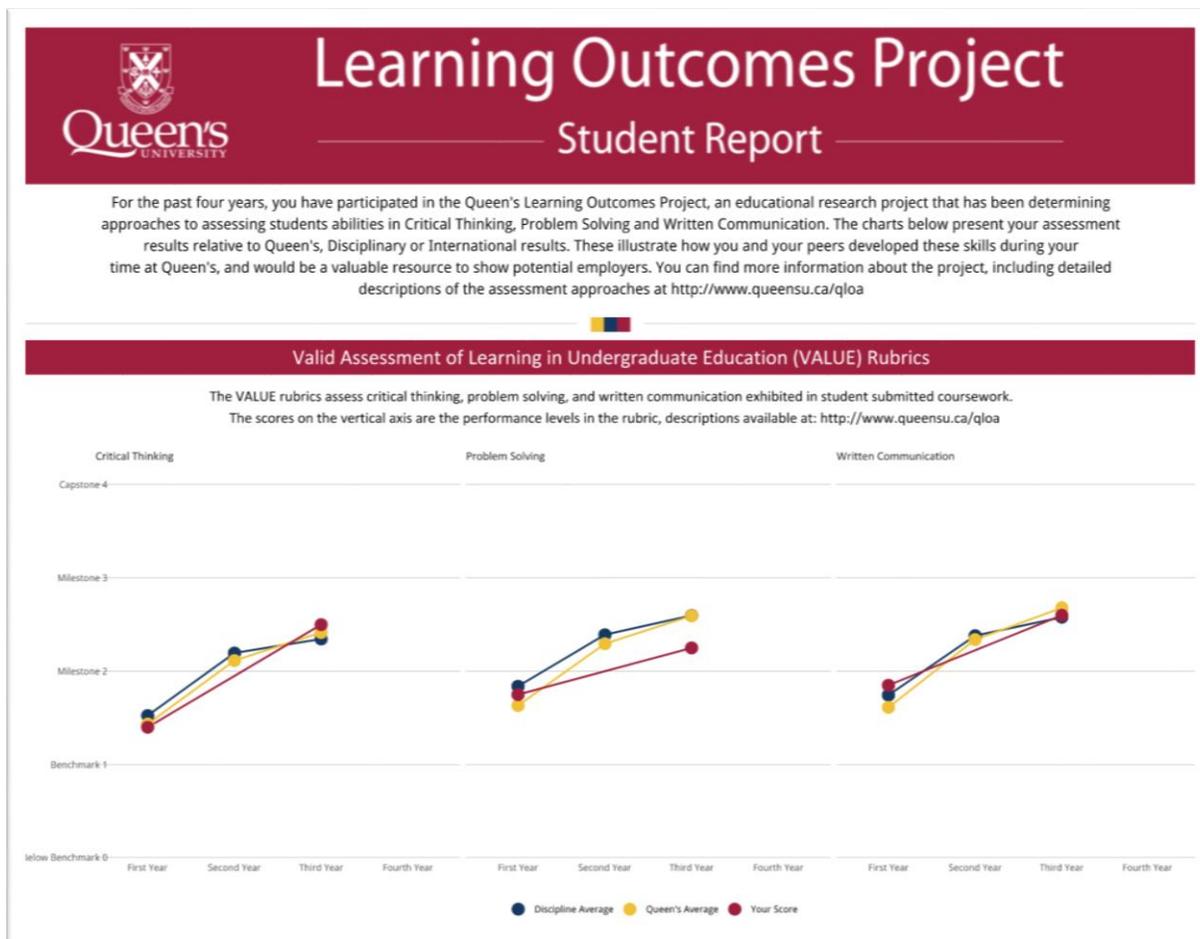
exceptionnellement utile, et vraiment intéressant. Ce qui est encore plus utile est une autre paire d'yeux là-dessus. J'évalue ce travail depuis cinq ans; il m'a fallu cinq ans pour constater ce que vous avez fait d'un seul coup. » (instructeur de cours de première année)

La prestation de rétroaction aux instructeurs a souvent eu pour effet de susciter des questions de réflexion. Parmi les questions courantes qui ont été posées, on peut citer les suivantes : « Quelles sont les caractéristiques désirables pour les étudiants? » « Est-ce que je devrais leur fournir le plan étape par étape pour la résolution (échafaudage)? » « Quelle information est utile pour les étudiants? »

Ce projet n'était pas conçu comme étude des interventions pédagogiques, mais il était tout de même important de noter les répercussions du projet sur les efforts d'amélioration de l'enseignement. Grâce à l'autoréflexion et en fournissant aux étudiants de la rétroaction provenant des travaux de cours, bon nombre des instructeurs impliqués dans le projet ont choisi d'adapter de nouvelles stratégies d'évaluation ou d'adopter de nouvelles stratégies d'enseignement afin de mieux cibler l'acquisition de compétences cognitives par les étudiants. Une liste des changements spécifiques apportés au cours est fournie dans l'annexe 4.

En plus des comptes rendus des enseignants, une série d'entrevues auprès de groupes d'étudiants a été effectuée chaque année afin d'explorer des sujets pertinents pour l'étude. Dès la deuxième année du projet, les étudiants avaient indiqué qu'ils souhaitaient avoir accès à leurs résultats individuels. Pendant la troisième année du projet, un format a été élaboré pour fournir aux étudiants leur note aux rubriques VALUE sur une page, et les résultats des tests CAT et CLA+ étaient présentés au verso. Les points de données sur le rapport affichaient la note individuelle (si elle était disponible) la moyenne pour la discipline et la moyenne pour l'établissement. Pendant la troisième année du projet, les étudiants ont fourni une rétroaction sur le format et des améliorations ont été apportées. La figure 13 affiche la page couverture du rapport.

Figure 13 : Exemple des résultats d'un rapport de projet d'un étudiant de troisième année



Points de vue des étudiants

Pendant la dernière année du projet, les étudiants qui avaient indiqué qu'ils accepteraient de participer à des entrevues en groupe ont été recrutés (il y avait une option à cet effet sur le formulaire de consentement). Trois entrevues en groupe ont été organisées avec un total de neuf étudiants, six femmes et trois hommes, des départements participants d'ingénierie, de physique et de psychologie. En général, les neuf étudiants avaient eu des expériences antérieures positives avec des tests standardisés. Le but consistait à recueillir des renseignements auprès des étudiants concernant leurs expériences relatives au Collegiate Learning Assessment Plus et au Critical Thinking Assessment Test. Spécifiquement, nous souhaitons mieux comprendre les perspectives des étudiants concernant les éléments suivants :

- Leurs motivations à participer et à faire des efforts pendant les tests
- Leurs perceptions d'avoir fait peu d'effort et de faible motivation lors des tests
- Leurs idées concernant des stratégies de recrutement efficaces et l'augmentation de la motivation des étudiants pour passer les tests à l'avenir

Dans les sections suivantes, nous décrivons les résultats des entrevues en groupe et nous formulons notre compréhension des perspectives des étudiants associées à chacun des trois objectifs présentés ci-dessus.

Objectif 1 : Comprendre les motivations des étudiants à participer et à faire des efforts pendant les tests

Pour tous les étudiants que nous avons interviewés, nous avons observé certaines indications d'une motivation intrinsèque et d'une motivation extrinsèque à passer les tests et à faire des efforts. En partie, certains étudiants ont considéré que l'acceptation par les professeurs et la promotion des tests dans leurs classes les avait motivés à participer. Selon un étudiant en psychologie, c'était un facteur :

« Je pense que ça dépend de la façon dont leur prof présente le test. Parce que pour nous, il est évident que le professeur pense que c'est très important, et aussi la façon dont elle le décrit, je pense, a un effet sur la décision des étudiants de participer ou pas ». (Groupe 1)

D'autres étudiants qui ont mentionné l'acceptation des professeurs comme source de motivation ont fourni des raisons qui comprenaient la façon dont le professeur leur avait présenté le test ou s'il avait donné l'impression qu'il pensait que les tests étaient importants. Qui plus est, la motivation des étudiants était associée à leur connaissance du contenu de leur discipline et avec leurs expériences en matière de recherche dans leur propre discipline.

Les étudiants ont dit que leur motivation à passer le test était associée à leurs expériences antérieures, ce qui, pour certains, comprenait un manque de familiarisation avec le contenu spécifique couvert dans les questions du test. Par exemple, les étudiants ont noté que les questions de test qui mettaient l'accent sur la connaissance du contenu et les processus de résolution de problèmes tirés de leur discipline étaient plus pertinentes et intéressantes pour eux. Comme l'a noté un étudiant en ingénierie,

« Je pense que cela dépend des expériences antérieures. J'ai transféré en ingénierie seulement l'an dernier. Avant, je suivais beaucoup de cours de philosophie et de choses comme ça. Alors les questions du test étaient assez pertinentes à mes intérêts. » (Groupe 2)

Un autre étudiant en psychologie a répondu : « J'ai pu appliquer ce que j'avais appris dans ces cours aux questions posées » (Groupe 1). À l'opposé, un étudiant pour qui les questions n'étaient pas aussi pertinentes a commenté : « Je n'aimais pas celles où il y avait des choses médicales et des lectures à ce sujet et à tirer différentes conclusions de ce domaine... mais c'était peut-être simplement le fait que je n'ai pas de connaissances dans ce domaine, et donc qu'il était difficile pour moi d'établir des liens » (Groupe 4). Les étudiants en psychologie, en particulier, ont indiqué que leur motivation était associée à leurs expériences

en psychologie, où la participation à des projets de recherche et des laboratoires de recherche est souvent une exigence de leur programme de grade universitaire. Comme l'a commenté un étudiant en psychologie : « Je travaille comme bénévole dans un laboratoire de recherche en psychologie, et je sais à quel point il est difficile d'obtenir des participants pour différentes choses. Après avoir eu cette expérience, cela m'a donné envie. . de participer » (Groupe 1). Ces étudiants ont dit qu'ils savaient à quel point il était difficile d'obtenir des participants à des études de recherche, et que cela les avait donc motivés à s'impliquer et à faire de l'effort lors des tests.

Pour terminer, la motivation des étudiants était aussi associée à leur appréciation des questions de test qui leur donnaient des occasions d'appliquer leurs compétences en résolution de problèmes et en pensée critique à de nouvelles situations. Dans le cadre des entrevues en groupe, on a demandé aux étudiants de parler de leurs expériences, dans le cadre de leurs études ou à l'extérieur, où ils avaient dû penser de façon critique et appliquer des stratégies de résolution de problèmes. Tous les étudiants ont, d'une façon quelconque, tiré des comparaisons entre les questions de test, leur travail scolaire qui exigeait qu'ils pensent de façon critique et qu'ils règlent des problèmes, et leurs expériences en matière de résolution de problèmes et de pensée critique à l'extérieur de leur travail scolaire. Les étudiants qui estimaient que les tests étaient intéressants et stimulants ont affirmé que c'était principalement en raison de la nature ouverte des questions et du fait qu'il n'y avait pas de bonne réponse. Par exemple, un point commun était le fait que les étudiants étaient motivés par des expériences d'apprentissage qui les encourageaient à trouver de l'information et des solutions aux problèmes à leur façon. Cela se reflétait lorsqu'ils parlaient des questions de test, de leurs expériences scolaires et de leurs expériences non scolaires. Concernant l'expérience d'une étudiante avec les tests, elle a indiqué : « Il y a beaucoup de façons de trouver le problème. Et ensuite, on peut aussi le présenter de façon intelligente » (Groupe 2). De même, un étudiant, commentant un projet de cours, a dit : « J'aime avoir la pression de pouvoir faire son propre projet. Cela motive à trouver l'information à sa façon » (Groupe 2). Un autre étudiant a commencé un projet d'été qui avait un lien avec son domaine d'ingénierie :

« La Faculté d'ingénierie m'a ajouté à l'équipe solaire cette année. J'ai été lancé dans une nouvelle idée de projet et ils voulaient que je jette un coup d'œil et que je fasse un peu de construction et de choses semblables. Une fois encore, c'est une façon différente de penser à un problème lorsqu'on essaie de préparer quelque chose de vrai, pas seulement y penser de façon théorique » (Groupe 2).

Une étudiante qui avait des majeures en psychologie et en éducation a commenté ses expériences en résolution de problèmes lors de son stage à l'extérieur de l'université :

« À l'extérieur de la salle de classe [d'université]... je fais un stage, et la résolution de problèmes et la pensée critique sont essentielles : il y a 30 étudiants, et il faut y arriver. On nous force en fait à résoudre des problèmes immédiatement. Et je pense qu'il s'agit également d'une compétence importante. Il y a l'expérience de devoir le faire immédiatement, et aussi de pouvoir le faire dans un contexte plus scolaire, moins concret » (Groupe 1).

Il est intéressant de noter que les étudiants n'ont pas initialement compris que leurs expériences en résolution de problèmes et en pensée critique à l'extérieur de leur programme étaient en fait transférées à leur façon de penser dans le cadre de leur programme (et vice-versa). Lorsqu'on l'a souligné, ces étudiants

ont pu voir le lien plus clairement. Nous avons conclu que la plupart des étudiants utilisaient des compétences en pensée critique et en résolution de problèmes dans leurs expériences à l'extérieur des programmes et dans certains cours semblables à celles qu'ils rencontraient sur les tests. Cela contribue à expliquer pourquoi ils ont exprimé leur intérêt et leur motivation pour des questions de test qui leur permettaient d'appliquer des solutions innovantes à des problèmes complexes. Toutefois, il est ressorti de ces conversations avec les étudiants qu'ils séparent souvent l'apprentissage qu'ils font dans leurs cours universitaires de leur apprentissage à l'extérieur des cours, qu'ils adoptent des processus d'apprentissage identiques ou similaires ou pas. Il est possible que les questions de test axées sur la résolution de problèmes et la pensée critique, administrées dans un milieu scolaire, mais mettant l'accent sur du contenu ne faisant pas partie de leurs disciplines universitaires (du moins, pour de nombreux étudiants), agissent comme relais entre ces deux contextes d'apprentissage. Autrement dit, les tests agissent comme un troisième espace entre le contexte universitaire et le contexte non universitaire des étudiants, qui leur permet de devenir plus conscients des compétences cognitives et des processus d'apprentissages qui sont inhérents aux deux, et que nous essayons de favoriser chez les apprenants du XXI^e siècle dans les études supérieures d'aujourd'hui.

Objectif 2 : Comprendre les perceptions des étudiants d'avoir fait peu d'effort et leur motivation pour passer le test

Les étudiants que nous avons interviewés ont fourni quelques renseignements clés permettant de déterminer les raisons pour lesquelles l'effort et la motivation des étudiants sur le CLA+ et le CAT étaient peu élevés. Parmi les raisons principales, on peut citer le fait que les étudiants devaient faire face à des exigences contradictoires, par exemple devoir se concentrer sur le reste de leur travail scolaire, et le manque de temps dans leur horaire. Une explication courante était le fait que lorsque les étudiants étaient occupés avec d'autres engagements liés à leur grade (par exemple, des projets et des laboratoires), ils n'étaient pas aussi motivés ou aussi susceptibles de participer à des tests externes. Les étudiants ont également supposé que la motivation était affectée parce que le test avait lieu tôt le matin et que la période de l'année scolaire pendant laquelle les tests étaient offerts était chargée. Un étudiant a dit, « Parce qu'il [le test] était offert un lundi matin à 8 h 30 et que je ne suis pas matinal. Je pense que le moment affecte beaucoup la performance » (Groupe 2). Une autre personne a fourni la raison suivante : « Je pense vraiment que cela dépend des autres choses qui se passent. Pendant le trimestre, nous avons eu deux ou trois semaines complètement dingues, et personne ne se serait présenté » (Groupe 1).

D'autres explications du manque de motivation concernaient la nature répétitive de l'administration des mêmes tests à plusieurs reprises, le fait que les tests exigeaient beaucoup d'écriture, le fait que les questions à choix multiple devenaient ardues, et qu'ils étaient lassés et fatigués à la fin du test. Dans le cas du test CLA+, les étudiants ont mentionné le fait que la dynamique de groupe importait beaucoup : « Si vous êtes avec un groupe de personnes qui ne veulent pas faire le projet et ne veulent pas faire d'effort, c'est beaucoup de pression », a dit un étudiant. Les étudiants avaient une vision plus positive du CAT et ils estimaient qu'il était plus représentatif de leur effort et de leurs capacités individuelles. Selon certains étudiants, la rubrique VALUE associée au CLA+ ne tenant pas compte d'une répartition inégale des efforts, ce qui a été considéré comme étant potentiellement décourageant pour certains étudiants.

Objectif 3 : Recueillir les idées des étudiants concernant des stratégies de recrutement efficaces et augmenter la motivation des étudiants pour passer les tests à l'avenir

Pour terminer, nous avons demandé aux étudiants leurs idées concernant les façons dont nous pourrions améliorer le recrutement des étudiants et leur participation, ainsi qu'augmenter l'effort lors des tests. Tous les étudiants ont souligné l'idée que les administrateurs des tests pourraient présenter de façon plus explicite les avantages des tests pour les personnes qui les passent. Par exemple, les étudiants aimaient l'idée de voir comment ils se comparaient à d'autres étudiants de leur département et, de façon plus large, dans l'ensemble de l'université afin d'obtenir un point de référence pour leur propre réussite scolaire. Un étudiant a intelligemment suggéré que le fait de connaître ces notes aiderait les étudiants à se faire valoir pour des occasions d'emploi futures : « S'il s'agissait d'une note qu'ils pouvaient publier en tant que qualification, je pense que cela motiverait beaucoup de personnes qui n'auraient normalement aucun intérêt [à passer les tests] » (Groupe 2).

Les étudiants participants ont également suggéré fortement de changer le moment de l'année scolaire pendant lequel le test est offert, et même introduire des horaires flexibles pour permettre aux étudiants de passer le test soit pendant les heures de cours, soit après. Ils ont également suggéré fortement que les étudiants reçoivent un crédit quelconque, des points de bonification ou des pourcentages supplémentaires dans le cours dans le cadre duquel ils passent le test, ou de faire en sorte que le test compte en tant que travail scolaire évalué réel. Selon un étudiant en psychologie :

« Si on veut fournir une rétroaction et réellement s'assurer que les étudiants sont motivés et font de leur mieux, il faut en faire un projet de cours. Ce n'est qu'en faisant cela qu'il est possible de s'assurer que les étudiants participent et feront tous les efforts possibles » (Groupe 3).

Des étudiants ont suggéré d'autres incitations pour augmenter la motivation et l'effort des étudiants, y compris en offrant de la nourriture gratuite et de petits cadeaux monétaires. Un des étudiants a commenté que bien que l'argent puisse être un incitatif, il ne générera peut-être pas une meilleure qualité de participation et d'effort, car les tests ne sont pas associés à leurs notes. Collectivement, les commentaires suggèrent que les facteurs les plus importants à prendre en considération pour encourager la motivation et l'effort lors des tests sont les suivants :

- L'instructeur accordant de la valeur au test
- Assurer la familiarisation avec le contenu du test
- Choisir le bon moment
- Rétroaction sur les résultats
- Que la note du test compte
- Prestation d'une incitation monétaire

Discussion

Acquisition de compétences cognitives complexes sur un grade de quatre ans

Le présent rapport met l'accent sur la mise en œuvre d'outils permettant d'évaluer l'acquisition de compétences cognitives complexes par les étudiants sur un grade de quatre ans. Bien qu'il y ait eu une motivation sous-jacente à encourager le développement de compétences, la principale préoccupation de l'étude exploratoire consistait à déterminer une façon viable d'évaluer le niveau véritable des compétences des étudiants avec le temps et dans l'ensemble de l'établissement. Les instruments choisis comprenaient des constructs d'évaluation semblables et l'étude de chacun d'entre eux a recensé des gains importants de l'apprentissage des étudiants. Bien que les tests standardisés représentaient l'approche la plus fiable d'un point de vue psychométrique, l'inconvénient était la motivation des étudiants et le mauvais alignement avec les activités qui importaient réellement aux instructeurs et aux étudiants dans les cours.

Les étudiants qui commencent leurs études à l'Université Queen's ont des résultats considérablement meilleurs que ceux qui commencent leurs études dans des universités américaines, en grande partie en raison de la nature sélective de Queen's. Compte tenu des niveaux initiaux élevés de compétences en pensée critique, en résolution de problèmes et en communication écrite, le défi consiste à démontrer leur amélioration sur un grade de quatre ans. Les chercheurs reconnaissent que de nombreux facteurs contribuent au développement de ces compétences, y compris les activités parascolaires et parallèles au programme et les aptitudes de vie acquises au fur et à mesure que les étudiants acquièrent de la maturité. Les travaux et tests notés dans cette recherche étaient tous écrits. Il y a du travail à faire pour étudier l'évaluation de ces compétences dans un format de communication orale et de communication de l'information.

L'évaluation peut être compromise par différents facteurs : dans le cas des tests standardisés, ils comprennent les contraintes relatives à la planification des tests, les efforts et les incitations, et dans le cas de l'utilisation de rubriques pour noter les travaux scolaires, ils comprennent les différences entre les attentes relatives aux travaux et l'alignement variable avec les constructs d'évaluation. Les tests en classe ont été effectués conformément aux horaires de cours; par conséquent, les séances de test ont parfois eu lieu à des heures, des journées ou des semaines sous-optimales pendant le trimestre. La fenêtre de test pour le CLA+ s'ouvrait seulement en mars, ce qui était un mauvais moment, car il chevauchait les examens de mi-session. Bon nombre des cours participants offraient l'heure de leur laboratoire ou tutoriel, d'une durée de 60 minutes, pour permettre aux étudiants de passer les tests, ce qui les empêchait d'administrer le CLA+ de 90 minutes en classe. Initialement, les chercheurs ont recruté les étudiants afin de passer le test hors des heures de classe, mais la participation était très basse et la distorsion de l'échantillon présentait donc une préoccupation. Les étudiants qui ne se trouvaient pas dans un laboratoire d'informatique à cette heure-là devaient soit se rendre dans un autre local, soit utiliser leur propre ordinateur. Le déplacement était long et perturbateur pour les étudiants, et l'utilisation de leur propre ordinateur présentait une gamme de problèmes techniques. Des problèmes comme ceux-là peuvent affecter la motivation et la performance des étudiants.

Il est également important de tenir compte du fait que les tests standardisés ne contiennent pas de contenu spécifiquement axé sur une discipline en particulier. Si le contenu du test n'intéresse pas les étudiants, ou si ceux-ci se trouvent dans un local qu'ils ne connaissent pas ou n'ont pas le temps de terminer le test, leur motivation pourrait être compromise. Pendant la première année, tous les étudiants qui ont passé le test ont reçu un pourcentage sur leur note de cours pour les récompenser de leur participation. Lorsque les niveaux d'effort ont chuté, les chercheurs se sont tournés vers d'autres méthodes pour attirer les étudiants. Les incitations financières se sont traduites par un effort accru, mais leur coût pourrait être prohibitif.

Les gains d'apprentissage étaient détectés le mieux par les rubriques VALUE. Le domaine le plus faible pour les étudiants de première année de l'Université Queen's était la pensée critique, la majorité d'entre eux atteignant le niveau Point de référence 1, tandis que le rendement médian des étudiants de première année en résolution de problèmes et en communication écrite était Jalon 2. Les résultats en matière de pensée critique sont également ceux qui ont présenté l'amélioration la plus importante. La majorité des étudiants avaient atteint le niveau Jalon 3 à la quatrième année. Sur les quatre années de leurs programmes de premier cycle, les étudiants ont présenté des gains considérables sur toutes les rubriques VALUE.

Une analyse non paramétrique a été utilisée pour les données VALUE dans le présent rapport, car les résultats n'étaient pas dérivés de données échelonnées. Autrement dit, l'amélioration de Point de référence 1 à Jalon 2 sur les rubriques VALUE est différente de celle entre Jalon 3 et Point culminant 4. Le critère de niveau Point culminant sur les rubriques VALUE est plutôt ambitieux. Par voie d'analogie, pensez à la carte du terrain d'un centre de ski, puis considérez que les niveaux VALUE correspondent à la difficulté des différentes pistes. Le Point de référence 1, par exemple, serait une piste verte : avec un peu de travail, n'importe qui peut y arriver. Ensuite, comparez le Jalon 2 à une piste bleue intermédiaire : il faut un peu d'entraînement, mais ce n'est pas trop difficile. Jalon 3 est alors l'équivalent d'une piste diamant noir, où un certain niveau d'expertise est nécessaire, mais pourvu que la piste soit damée, la descente est réalisable. Pour terminer, comparez le Point culminant 4 à une piste double diamant noir non damée. Sur un tel terrain, il faut être expert. Toutefois, si les étudiants ont un guide, par exemple un superviseur de thèse ou de projet de conception pour leur montrer la voie à suivre, et s'ils font preuve beaucoup de rigueur et d'effort, ils peuvent atteindre un niveau Point culminant 4.

Les réalisations en quatrième année variaient selon le soutien dont disposaient les étudiants. Cette différence a introduit des erreurs dans l'analyse. Un autre point d'erreur découlait de la pondération du cours pour le projet. Il n'était pas vraiment juste de comparer un rapport de laboratoire hebdomadaire à une thèse de recherche d'une durée d'un an. Nos protocoles de notation décrivaient le processus d'évaluation du projet afin de déterminer les dimensions qui pouvaient être évaluées de façon valide. Une partie de la raison pour ce faire visait à minimiser ces différences. Autrement dit, nous n'avons évalué les dimensions de la rubrique que lorsque le travail suscitait une démonstration du critère. Il était inhabituel qu'un projet ayant une faible pondération dans le cours suscite des démonstrations de toutes les dimensions à évaluer. Le lien entre les dimensions évaluées et la note indique qu'il y a encore du travail à faire pour aligner les évaluations des cours avec les rubriques VALUE.

Acquisition de compétences entre les programmes et les personnes, et lien avec les notes

Les résultats du CLA+ et du CAT ont suggéré que les étudiants du baccalauréat ès sciences (BSc) atteignaient les résultats les plus élevés et que les étudiants du baccalauréat ès arts (BA) obtenaient régulièrement les gains les plus importants. Il était plus difficile de comparer directement les programmes en se fondant sur les résultats des rubriques VALUE parce que les types de projets étaient différents. Les étudiants qui suivaient des cours à thèse de recherche (un dans un cours d'ingénierie, l'autre dans le cours de thèse en psychologie) ont atteint la majorité des résultats de niveau Point culminant 4.

Le sexe n'était pas un facteur important dans la population échantillonnée, la conclusion étant que les instruments ne contiennent pas de préjugés sexistes et que le sexe n'influence pas le développement cognitif. Comme on pourrait s'y attendre, les étudiants dont l'anglais est la langue maternelle ont obtenu des résultats marginalement supérieurs sur les tests standardisés. Ces tests n'étaient offerts qu'en anglais, et plus d'un étudiant a dit que le test aurait été beaucoup plus facile s'il avait pu le passer en mandarin ou dans une autre langue. Toutefois, la langue n'était pas un facteur pour les travaux de cours. Les étudiants avaient plus de temps pour les compléter, ce qui minimisait les effets de la langue.

Les notes de tous les outils étaient corrélées avec la moyenne pondérée cumulative de session, mais les rubriques VALUE avaient le lien le plus étroit. On pourrait soutenir que la moyenne pondérée cumulative des étudiants reflète la somme des résultats d'apprentissage du cours, des connaissances, des comportements et des habitudes d'esprit que nous considérons comme étant importantes pour préparer les étudiants pour l'avenir. Selon les structures des travaux, il est possible que cela ne soit pas complètement vrai, mais nous pouvons supposer qu'au moins la moyenne pondérée cumulative reflète une partie, sinon la totalité, des connaissances déclaratives et techniques, des compétences procédurales, du travail d'équipe ou du professionnalisme, en plus de la pensée critique, de la résolution de problèmes et de la communication. Pour cette raison, nous ne nous attendons pas à une forte corrélation entre la moyenne pondérée cumulative et la pensée critique, la résolution de problèmes et la communication écrite telles qu'elles sont évaluées sur les tests CLA+, CAT et VALUE. Toutefois, nous rechercherions des corrélations plus fortes et nous accorderions une plus grande pondération aux instruments d'évaluation qui reflètent le mieux les résultats des cours. Les coefficients pour les rubriques VALUE étaient tous supérieurs à 0,3, tandis qu'ils étaient inférieurs à 0,3 pour le CAT et le CLA+.

Il est également important de noter que les stratégies et les processus d'évaluation des cours varient considérablement, et il faudra du travail supplémentaire pour déterminer le composant de la notation du cours qui reflète les constructs spécifiques de l'évaluation. Certains des travaux offerts au projet ne pouvaient pas être notés en utilisant les rubriques VALUE Pensée critique, Résolution de problèmes ou Communication écrite parce que l'alignement avec les dimensions évaluées était faible ou inexistant. C'était particulièrement vrai dans les cours très techniques, où il n'était pas réellement possible d'adapter les évaluations du cours afin de cibler spécifiquement les résultats d'apprentissage.

Coût et efficacité en matière de temps des instruments

Les chercheurs ont fait face à des difficultés en matière de coordination des exigences techniques et logistiques nécessaires pour les tests standardisés. Même lorsque les chercheurs arrivaient à convaincre les étudiants de passer les tests, ceux-ci ne faisaient pas nécessairement de leur mieux. Cela présentait un problème du point de vue de la fiabilité, car nous ne savions pas si les étudiants pouvaient produire de meilleurs résultats. Pendant la dernière année du projet, des incitations monétaires ont été utilisées pour tenter de minimiser ce problème, mais des considérations budgétaires et logistiques relatives à de telles évaluations présentent un problème en matière de durabilité. Les tests CAT doivent être notés en utilisant un protocole rigide, qui nécessite des marqueurs qualifiés et suffisamment de temps, puis envoyés au Tennessee à des fins d'étalonnage. Les données des tests CLA+ et CAT ne sont disponibles qu'environ deux mois après le test. Les délais dans la réception des données rendent impossible l'utilisation de la note au test pour une évaluation de cours à faible conséquence, et ils rendent également difficile l'utilisation des données pour améliorer les cours. Le CLA+ inclut un système de badge numérique pour les étudiants, les rapports étant envoyés directement aux étudiants approximativement en même temps que le rapport institutionnel du CLA+ est envoyé à l'établissement. Les données issues des tests sont confidentielles, et les deux fournisseurs de tests ont mis en place des protocoles de sécurité pour éviter tout accès inapproprié aux données. Ces tests fournissent des données qui peuvent être agrégées avec confiance entre les disciplines et qui peuvent également être utilisées pour effectuer des comparaisons avec les moyennes des établissements américains publiées par le fournisseur de test.

Pour atteindre la viabilité à long terme, l'équipe de recherche concentre une grande partie de son temps et de ses efforts à travailler sur le terrain avec le corps professoral existant pour créer des instruments qui minimisent toute charge de travail pour eux et qui sont pratiquement invisibles pour les étudiants. Nous notons un niveau croissant de sensibilisation à l'évaluation des résultats d'apprentissage transférables au sein des départements qui participent au projet, et nous organisons des séances d'information et des ateliers avec le corps professoral afin de partager notre compréhension avec un public plus large. Cette compréhension constitue une base nécessaire pour un déploiement à plus grande échelle. Des discussions ont commencé concernant l'inclusion possible de départements supplémentaires afin d'obtenir une évaluation des résultats d'apprentissage généraux.

Pendant l'étude, nous avons trouvé des preuves anecdotiques indiquant que certains professeurs et instructeurs étaient ouverts à la mise en œuvre de nouveaux processus d'évaluation et étaient passionnés de le faire, mais que beaucoup d'entre eux se méfiaient du changement. Certains membres du corps professoral estimaient que ces compétences intellectuelles étaient capturées avec leurs méthodes d'évaluation actuelles, et ils n'étaient pas convaincus de la nécessité de les évaluer spécifiquement. À l'opposé, les instructeurs qui ont participé à l'étude ont découvert des avantages tangibles à la prestation de données sur les réalisations des étudiants, et bon nombre d'entre eux ont utilisé cette information pour apporter des améliorations à leurs cours.

Résultats supplémentaires et répercussions

Communication et perfectionnement professionnel

- Un site Web a été développé pour le projet (<http://www.queensu.ca/gloa/home>)
- L'équipe du projet a animé des ateliers, a fourni des webinaires et a effectué des présentations lors de congrès au Canada et aux États-Unis.
- Certaines facettes du projet ont été publiées dans *The European Journal of Engineering Education*, et *Assessment and Evaluation in Higher Education* (détails disponibles ici : <http://www.queensu.ca/gloa/resources-and-research>)
- Les ateliers sur le développement de tâches ont eu lieu chaque année pour développer une compréhension commune des attentes et des structures d'évaluation de la pensée critique et de la résolution de problèmes

Développement d'un outil de construction de rubriques

Il faut un certain temps pour se familiariser avec les rubriques VALUE. Le langage des descripteurs est général et de très haut niveau. Les chercheurs du projet les ont utilisés sous leur forme publiée, mais les instructeurs du cours ont trouvé qu'il était difficile de les associer directement à leurs travaux de cours. Pendant la deuxième année du projet, les chercheurs ont utilisé des annotations dans les travaux pour adapter les rubriques VALUE afin d'inclure un langage spécifique au département. Les rubriques résultantes étaient tout de même très générales, et elles n'étaient pas plus utiles pour les instructeurs pour une application directe à leurs travaux de cours que la forme publiée des rubriques VALUE.

L'étape suivante de la recherche consistait à développer un créateur de rubriques sur le Web, qui pose à l'instructeur des questions suggestives concernant le but du projet et ce qu'ils veulent que leurs étudiants démontrent. Une application Web contrôlée par menu a été développée comme mécanisme de soutien pour aider les éducateurs à développer des rubriques pour l'évaluation des travaux d'étudiants tels que des projets d'analyse et de recherche, des projets de conception, des enquêtes ou des démarches structurées. Le créateur de rubriques, appelé Building Assessment Scaffolds for Intellectual Cognitive Skills (BASICS), met l'accent sur les compétences cognitives et intellectuelles en pensée critique, en imagination créatrice et en résolution de problèmes. Les éléments et le langage de BASICS ont été élaborés dans le cadre d'un processus de collecte d'annotations provenant de plus de 900 échantillons de travail. Les annotations étaient groupées par niveau et les termes à fréquence élevée ont été identifiés. Les échafaudages de la rubrique représentent les éléments communs pour chacune des dimensions de la rubrique. La construction des rubriques est un processus à cinq étapes, entraînant un échafaudage de rubriques contenant des verbes d'action correspondant aux niveaux sur les rubriques VALUE que les instructeurs peuvent préciser en fonction de leurs propres besoins.

Pendant ce projet, les experts du monde entier ont continué à étudier l'évaluation et la valeur d'une éducation universitaire. Le projet Mesure de l'apprentissage universitaire (MAU) a fait la synthèse de la littérature existante et de recommandations clés tirées des livres blancs de six membres du groupe de professeurs. Voici les principes fondamentaux qui ont été présentés :

- Le corps professoral devrait être à l'avant-plan de la définition et de la mesure des résultats d'apprentissage au niveau du premier cycle.
- Les étudiants de toute origine et de tous les établissements devraient bénéficier d'une occasion équitable de démontrer leurs connaissances et leurs compétences
- Toute mesure unique de l'apprentissage des étudiants devrait faire partie d'un plan d'évaluation holistique plus large
- Les établissements devraient utiliser les outils d'évaluation volontairement
- Les mesures de l'apprentissage des étudiants devraient être rigoureuses et de haute qualité, et elles devraient générer des données qui permettent d'effectuer des comparaisons avec le temps et entre les établissements (Roksa et al., 2016, pp. 7–11)

La recherche, l'analyse et les résultats du projet CERA de l'Université Queen's étaient complètement indépendants du travail effectué dans le cadre du projet MAU, mais nos méthodes empiriques ont permis de dériver des conclusions semblables. La discussion qui suit reflète les principes ci-dessus.

Limitations

Le présent rapport décrit de nombreuses limitations de l'approche et de la méthodologie, et les contraintes d'espace ne permettent pas une discussion complète. Toutefois, en bref, nous avons conclu que les outils standardisés, non incorporés, tels que le CLA+ et le CAT, avaient des limitations importantes, y compris une faible motivation de la part des étudiants en raison du manque de contenu spécifique à leur discipline, la répétition du même test sur plusieurs années, le manquement des instructeurs à encourager les étudiants à faire des efforts et le fait que ces évaluations ne comptent pas dans la note de cours finale. Parmi les autres facteurs logistiques, on peut citer le temps requis pour répondre aux tests, les difficultés techniques associées à l'administration des tests et les délais avant la réception des résultats. De plus, CLA+ et CAT étaient beaucoup plus coûteux. Bien que les évaluations des rubriques VALUE présentaient moins de problèmes concernant la motivation des étudiants, les défis en matière de logistique et les coûts, les limitations comprenaient des résultats moins comparables en raison de différences entre les types de projets et un certain désalignement entre le projet et les critères VALUE.

Une autre limitation des données longitudinales est la diminution du nombre global d'étudiants qui ont passé le test entre la première année et la quatrième année, et le fait que le nombre d'étudiants jumelés testés avec des mesures spécifiques sur toutes les années de l'étude est relativement peu élevé. Bien que les tests pour toutes les mesures lors de toutes les années ont fourni des données longitudinales robustes, la fatigue associée aux tests aurait pu être atténuée et la motivation et l'effort des étudiants auraient pu augmenter si les évaluations avaient été effectuées uniquement pendant la première et la quatrième année. Des études de ce genre sont également vulnérables à d'autres facteurs, tels que l'absence d'un groupe de contrôle parmi la population générale qui ne fait pas des études postsecondaires, ce qui signifie qu'il est difficile de tenir compte de la croissance du développement n'ayant aucun lien avec un enseignement scolaire ciblé dans l'acquisition de compétences cognitives d'ordre supérieur. D'autres études s'avèrent nécessaires dans plusieurs domaines : la mise en œuvre et la corrélation d'autres outils qui mesurent la pensée critique; des études qui maintiennent un groupe de contrôle qui n'est pas influencé par les études postsecondaires; et les études qui établissent une corrélation entre la croissance des compétences

transférables d'ordre supérieur et les résultats sur le marché du travail et dans la vie plusieurs années après la fin des études postsecondaires.

Conclusions

Le projet du Consortium sur l'évaluation des résultats d'apprentissage (CERA I) visait à étudier des méthodes durables d'évaluation de la pensée critique, la résolution de problèmes, la communication écrite et l'apprentissage continu dans les programmes de premier cycle et à générer de l'information pour le développement des cours et des programmes. Le Collegiate Learning Assessment Plus et le Critical Thinking Assessment Test ont fourni une démonstration des gains d'apprentissage et ont permis des comparaisons entre les établissements, mais ils étaient coûteux et vulnérables aux questions de motivation. Les rubriques du Valid Assessment of Learning in Undergraduate Education se sont révélées les plus économiques et utiles pour renseigner les instructeurs au sujet d'améliorations à l'enseignement et à l'apprentissage. La plus grande variation dans l'utilisation cohérente des rubriques VALUE était l'alignement entre le cours et les dimensions VALUE. La prochaine itération du projet du Consortium sur l'évaluation des résultats d'apprentissage (CERA II) mettra l'accent sur l'utilisation d'une approche de réseau et des rubriques VALUE pour un alignement ciblé.

Pour tous les intervenants qui ont participé à ce projet à l'échelle de l'établissement (administration, chefs des départements, instructeurs, étudiants et chercheurs), il a été utile de savoir dans quelle mesure les étudiants développaient les compétences que sont la pensée critique, la résolution de problèmes et la communication écrite. L'apprentissage continu était intéressant pour certains d'entre eux, mais il était plus difficile à quantifier et il est demeuré un sujet secondaire dans l'étude. Le principal défi consistait à évaluer les étudiants d'une façon qui fournissait des données pouvant être agrégées. Les tests standardisés étaient coûteux et difficiles à mettre en œuvre. Les taux de participation chez les étudiants de quatrième année étaient considérablement inférieurs à ceux des étudiants de première année, principalement en raison du fait que les tests pendant la première année avaient lieu pendant les cours, ce qui n'était pas possible pour les étudiants de quatrième année. En plus des problèmes de motivation, il s'agit de deux des inconvénients des tests standardisés comparativement à l'utilisation des rubriques VALUE pour noter les travaux de cours. Les entrevues avec les étudiants suggèrent que les éléments suivants sont essentiels pour encourager les étudiants à faire des efforts lors des tests : les instructeurs doivent accorder de l'importance au test, le contenu de celui-ci doit être pertinent pour les étudiants, l'heure prévue pour le test ne doit pas entrer en conflit avec d'autres engagements, et les résultats devraient être mis à la disposition des étudiants. De plus, si la note au test comptait pour le cours, les étudiants feraient des efforts. Autrement, la suggestion était de fournir des incitations financières.

Nous avons constaté que la meilleure stratégie consistait à travailler directement avec les instructeurs pour les aider à intéresser les étudiants en fournissant des travaux qui encourageaient et suscitaient spécifiquement la pensée critique et la résolution de problèmes. Lors de la comparaison de méthodes, la rétroaction fournie par les instructeurs a suggéré que les rubriques VALUE pourraient être plus efficaces à court terme, mais ils ont également reconnu que le CLA+ et le CAT pourraient avoir des avantages pour une évaluation à plus long terme et des comparaisons avec d'autres établissements. Ce projet a fourni de précieux renseignements, alors que les chercheurs naviguaient les défis et les contraintes contextuelles. La

présentation de ses conclusions en quelques points ne rend pas justice à sa complexité, mais à des fins de quantification des résultats, nous les fournissons ici :

- Les notes au CLA+ et au CAT démontraient une amélioration significative entre la première et la dernière année.
- Les notes des rubriques VALUE démontraient des améliorations importantes et continues en pensée critique, en résolution de problèmes et en créativité dans les échantillons d'une année à l'autre.
- Les corrélations entre les dimensions de pensée critique et de communication mesurées sur le CLA+, le CAT et les rubriques VALUE étaient faibles, mais significatives.
- Le coût de la mise en œuvre de la notation des rubriques VALUE était environ 20 \$ canadiens de moins que le coût de la mise en œuvre des tests CLA+ ou CAT. Comme les rubriques VALUE elles-mêmes ne coûtent rien, si la notation faisait partie d'un plan d'évaluation à l'échelle de l'établissement, le coût serait négligeable.
- Les résultats du test CLA+ permettent de comparer le rendement de l'établissement avec celui d'autres écoles. Cela n'est pas possible avec la notation des rubriques VALUE sans procédures communes de formation et d'étalonnage.

Remerciements

L'équipe de recherche remercie sincèrement les nombreux étudiants de premier cycle et de cycle supérieur qui ont travaillé au projet du Consortium sur l'évaluation des résultats d'apprentissage pour leur engagement envers la notation des rubriques VALUE, l'analyse et les rapports pour les départements. Nous remercions les assistantes à la recherche Sima Zakini, Launa Gauthier et Nerissa Mulligan, dont le nom ne figure pas sur les présentes en tant qu'auteurs, mais dont le travail sur le projet a été essentiel à la réalisation de ses objectifs. Le soutien des établissements a également été essentiel à la réalisation des objectifs du projet, et l'équipe souhaite remercier Brenda Ravenscroft pour son leadership pendant le projet, en enseignement et en apprentissage au sein de la Faculté des arts et des sciences; Chris Conway et le personnel de recherche institutionnelle et de planification, et le personnel du bureau d'éthique de la recherche de l'Université Queen's, qui ont débattu et traité 16 modifications à l'étude pendant les quatre années du projet. Nous remercions également tous les instructeurs des cours, dont bon nombre ont cédé de leur temps pour nous permettre de parler à leurs étudiants et d'administrer les tests. De plus, nous sommes très reconnaissants au Conseil ontarien de la qualité de l'enseignement supérieur, avec une mention spéciale à Alexandra MacFarlane, pour son soutien généreux.

Bibliographie

- Association of American Colleges & Universities (AAC&U). (n.d.). About LEAP. <https://www.aacu.org/leap>
- AAC&U. (2014). Programs | VALUE: Valid Assessment of Learning in Undergraduate Education. <http://www.aacu.org/value/index.cfm>
- Benjamin, R., Klein, S., Steedle, J., Zahner, D., Elliot, S., & Patterson, J. (2012). The Case for Generic Skills and Performance Assessment in the United States and International Setting. http://cae.org/images/uploads/pdf/The_Case_for_Generic_Skills_and_Performance_Assessment.pdf
- Canadian University Survey Consortium. (2015). 2015 University Student Survey: Master Report. http://www.cusc-ccreu.ca/CUSC_2015_Graduating_Master%20Report_English.pdf
- Critical Thinking Assessment Test (CAT). (n.d.). <https://www.tntech.edu/cat/>
- Council for Aid to Education. (2016). *CLA+ National Results, 2015–16*. New York, NY. http://cae.org/images/uploads/pdf/CLA_National_Results_2015-16.pdf
- Council for Aid to Education. (n.d.). CLA+ Overview. <http://cae.org/participating-institutions/cla-institution-users-portal/cla-overview/>
- Crick, R. D., Broadfoot, P., & Claxton, G. (2004). Developing an Effective Lifelong Learning Inventory: The ELLI Project. *Assessment in Education: Principles, Policy & Practice*, 11(3), 247–272. <https://doi.org/10.1080/0969594042000304582>
- Deller, F., Brumwell, S., & MacFarlane, A. (2015). *The Language of Learning Outcomes: Definitions and Assessments*. Toronto, ON: Higher Education Quality Council of Ontario. <http://www.heqco.ca/SiteCollectionDocuments/The%20Language%20of%20Learning%20Outcomes-Definitions%20and%20Assessments.pdf>
- Douglass, J. A., Thomson, G., & Zhao, C. M. (2012). The Learning Outcomes Race: The Value of Self-reported Gains in Large Research Universities. *Higher Education*, 64(3), 317–335. <https://doi.org/10.1007/s10734-011-9496-x>
- Finley, A. P. (2011). How Reliable Are the VALUE Rubrics? *Peer Review*, 13(4/1), 31.
- Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, 15(3), 199–236.
- Johnson, P. (2009). The 21st Century Skills Movement. *Educational Leadership*, 67(1), 11.
- Kaupp, J., Frank, B., & Chen, A. (2014). *Evaluating Critical Thinking and Problem solving in Large Classes: Model Eliciting Activities for Critical Thinking Development*. Toronto, Canada: Higher Education Quality Council of Ontario. Toronto: Higher Education Quality Council of Ontario. http://www.heqco.ca/SiteCollectionDocuments/Formatted%20Queen%27s_Frank.pdf

- Klein, S., Liu, O. L., Sconing, J., Bolus, R., Bridgeman, B., Kugelmass, H., Nemeth, A., Robbins, S., & Steedle, J. (2009). *Test Validity Study (TVS) Report*. The Fund for the Improvement of Postsecondary Education.
- Logie, R., Baddeley, A., Mané, A., Donchin, E., & Sheptak, R. (1989). Working Memory in the Acquisition of Complex Cognitive Skills. *Acta Psychologica*, 71(1), 53–87.
- Pintrich, P. R., Smith, D. A. F., Garcia, T., & Mckeachie, W. J. (1993). Reliability and Predictive Validity of the Motivated Strategies for Learning Questionnaire (Mslq). *Educational and Psychological Measurement*, 53(3), 801–813. <https://doi.org/10.1177/0013164493053003024>
- Rhodes, T. L. (2011). Emerging Evidence on Using Rubrics. *Peer Review*, 13(4/1), 4–5.
- Rhodes, T. L., & Finley, A. P. (2013). *Using the VALUE Rubrics for Improvement of Learning and Authentic Assessment*. Washington, DC: Association of American Colleges and Universities.
- Roksa, J., Arum, R., & Cook, A. (2016). Defining and Assessing Learning in Higher Education. In R. Arum, J. Roksa, & A. Cook (Eds.) *Improving Quality in American Higher Education: Learning Outcomes and Assessments for the 21st Century*. Jossey-Bass.
- Simper, N., Kaupp, J., Frank, B., & Scott, J. (2015). Development of the Transferable Learning Orientations Tool: Providing Metacognitive Opportunities and Meaningful Feedback for Students and Instructors. *Assessment & Evaluation in Higher Education*, 41(8), 1159–1175. <http://www.tandfonline.com/eprint/vYJm8CYjrD4I7ThUwDrn/full>
- Stein, B., & Haynes, A. (2011). Engaging Faculty in the Assessment and Improvement of Students' Critical Thinking Using the Critical Thinking Assessment Test. *Change: The Magazine of Higher Learning*, 43(2), 44–49.
- Stein, B., Haynes, A., & Redding, M. (2006). Project CAT: Assessing Critical Thinking Skills (pp. 290–299). Presented at the Proceedings of the National STEM Assessment Conference. Retrieved from [http://www.openwatermedia.com/downloads/STEM\(for-posting\).pdf#page=294](http://www.openwatermedia.com/downloads/STEM(for-posting).pdf#page=294)
- Sternberg, R. J., & Frensch, P. A. (2014). *Complex Problem Solving: Principles and Mechanisms*. New York: Psychology Press.
- Van Merriënboer, J. J. (1997). *Training Complex Cognitive Skills: A Four-component Instructional Design Model for Technical Training*. Englewood Cliffs, NJ: Educational Technology Publications.

