Dans un monde près de chez vous :

L'évaluation des apprentissages à l'ère du *Big Data*, du *Data Mining*

et de l'intelligence artificielle

Jean-Guy Blais

Jean-Luc Gilles

Agustin Tristan

Chaque jour, chaque semaine, chaque année, des données reliées à l'évaluation des apprentissages sont récoltées par des enseignants, des écoles et des ministères. Ces données prennent différentes formes et elles peuvent être très structurées comme lors d'une épreuve standardisée ministérielle avec des questions à choix de réponse, ou beaucoup moins structurées comme lors d'ajout de pièces individuelles dans un portfolio d'apprentissage. Cependant, à l'heure actuelle l'exploitation de ces données est plutôt limitée. Elle est locale, très circonscrite et non uniforme d'un enseignant à l'autre lorsque les données sont récoltées dans le cadre des activités quotidiennes de la salle de classe. Elle répond à des impératifs d'urgence de retour des résultats lorsque les données sont récoltées dans le cadre d'une épreuve standardisée certificative sous la responsabilité d'autorités ministérielles. Entre ces deux extrêmes, plusieurs situations pourraient être évoquées comme autant d'exemples où le potentiel des données récoltées n'est pas exploité à sa juste mesure faute de temps, d'énergie, de ressources ou d'outils appropriés.

Pour envisager une exploitation adéquate il faut qu'au moins trois conditions essentielles soient réunies. D'abord, il faut que les données soient produites et que les outils utilisés à cette fin soient répertoriés, ensuite il faut que les données soient accessibles et stockées sur un support sécuritaire, finalement il faut des outils efficaces pour les analyser et fournir la rétroaction aux personnes concernées (élèves, parents, enseignants, autorités ministérielles, etc.).

Le support traditionnel pour récolter des données en vue d'évaluer les apprentissages qu'est la modalité papier-crayon se prête assez mal à une exploitation optimale. Prenons, par exemple, les cas d'une épreuve uniforme d'écriture ou plusieurs milliers d'élèves produisent un texte sur un thème imposé. Les seules données qui subsisteront après la correction des copies des élèves seront les notes attribuées par les correcteurs aux différentes copies (ou aux différents critères). Toute la richesse de l'information contenue dans les textes eux-mêmes est tributaire d'une improbable analyse future en profondeur, ce qui n'arrive jamais ou presque faute de moyens, et les textes restent stockés sur des tablettes pendant le temps légal de conservation des copies. Après cette période, ils sont tout simplement détruits. Le même constat pourrait être fait au sujet des données récoltées par les enseignants car ceux-ci disposent d'encore moins de ressources et de temps pour les exploiter dans un environnement où la rapidité du retour aux élèves est essentiel. Et que dire de tout le bénéfice d'un regard longitudinal et transversal qui n'est que très peu possible car difficile à documenter faute de données pertinentes. Cependant, le support traditionnel papier-crayon pour la récolte des données en évaluation des apprentissages est tranquillement supplanté par le support écran-clavier. Il s'agit d'un constat inéluctable et ce nouveau support ouvre la porte à une exploitation des données

qu'il était difficile d'imaginer il y a dix ans à peine, transformant radicalement l'évaluation des apprentissages (et les apprentissages eux-mêmes). À ce sujet, Blais (2011) écrivait «…il est aussi fort possible que la contribution humaine au processus d'évaluation soit toute autre que celle que les étudiants du XXe siècle ont connue.». En effet, imaginons un instant une situation qui illustrerait ce que la technologie pourrait permettre dans le rayon des opérations d'évaluation d'envergure avec plusieurs milliers de candidats. Un élève est assis dans une classe traditionnelle et il s'active à rédiger un texte dans le cadre d'une épreuve standardisée uniforme et certificative pour la fin des études secondaires. Il écrit avec le clavier qui accompagne sa tablette d'une marque bien connue et disponible à un prix raisonnable. Il exécute ce travail sur une plateforme informatique accessible en ligne, dédiée et protégée contre les intrusions extérieures. Évidemment, il n'a pas accès à Internet et à aucun autre outil que ceux disponibles sur la plateforme. Après environ deux heures de travail et après une révision minutieuse de son texte, l'élève décide qu'il a terminé et sélectionne à l'écran le bouton *J'ai terminé* et confirme une autre fois que c'est bien le cas. Il quitte la classe où l'épreuve a lieu et se dirige vers la sortie. Il n'a pas fait une dizaine de pas dans le corridor lorsqu'il reçoit un courriel sur son téléphone portable intelligent qui annonce *Votre résultat à l'épreuve d'écriture du 12 juin*. Il est soulagé et nullement surpris car il croyait qu'il aurait ce message plus rapidement, il ouvre le courriel et prend connaissance de sa note globale et de la note attribuée pour chacun des trois critères de correction. Il a obtenu une note de 74% avec des notes respectives pour chacun des critères de 68%, 72% et 80%. Il n'est pas surpris car depuis maintenant plus de dix ans les textes produits dans le cadre de cette épreuve d'écriture certificative sont corrigés automatiquement par un moteur de

correction, et sans intervention humaine. Les élèves qui réussissent reçoivent leur résultat rapidement, c'est pourquoi il est soulagé, et ceux qui échouent le reçoivent quelques jours plus tard après une révision ou le jugement humain est partiellement mis à contribution.

En quelle année pouvons-nous imaginer cette scène qui semble tenir d'une utopie futuriste un peu trop enthousiaste? Probablement entre 2025 et 2027. Donc, si cette pratique a cours depuis dix ans, cela veut dire que la correction automatisée avec un moteur de correction a été implantée entre 2015 et 2017. À toute fin pratique, 2015 c'est demain! Utopie, enthousiasme délirant, «technotopie» ou «technovangélisme»?[1] Aucune de ces réponses, car c'est bien de la réalité qu'il s'agit, une réalité qu'on pourra voir en action à grande échelle dès 2015 dans un grand nombre d'états aux États-Unis.

En 2010 le gouvernement des États-Unis lançait le programme *Race to the top* et débloquait des fonds de 350 millions de dollars pour financer des projets de développement de systèmes technologiques pour l'évaluation des apprentissages. À la suite de cette annonce, deux consortiums se sont formés pour relever le défi. Le *Smarter balanced assessement consortium* (SBAC) regroupant 24 états et le *Partnership for assessment of readiness for college and careers* (PARCC) regroupant 18 états et deux territoires.[2] Le calendrier d'implantation prévoit que des épreuves sommatives en ligne intégrant des tâches complexes et innovantes en mathématique et en langue seront opérationnelles au printemps 2015 pour plusieurs années du primaire et du secondaire. Un des objectifs de ces opérations d'évaluation est de faire en sorte que les élèves soient

[1] Expressions inspirées par Oppenheimer (2004) et résultant des contractions de technologie et utopie d'une part et de technologie et évangélisme d'autre part.

[2] http://www.smarterbalanced.org; http://www.parcconline.org

confrontés à des tâches où ils doivent écrire leur réponse et la justifier. Évidemment, si les élèves écrivent plus cela implique beaucoup plus de travail de correction lorsque celle-ci est réalisée par des humains. Donc, *de facto* plus de temps et plus de ressources à consacrer à la correction. Dans la foulée du développement de ces tâches innovantes où les élèves doivent écrire plus, la fondation William and Flora Hewlett, en partenariat avec les deux consortiums mentionnés ci-dessus, a subventionné une compétition en trois phases, le *Automated Student Assessment Prize* (ASAP) où des équipes du monde entier (certaines provenant des principaux leaders commerciaux dans le domaine du testing en éducation) rivalisent pour démontrer l'efficacité de systèmes technologiques pour la correction automatisée de textes longs (des essais), de réponses courtes et de réponses à des problèmes en mathématiques.

Au moment de l'écriture de ce texte, les deux premières phases de la compétition sont terminées et les résultats pourraient en surprendre plusieurs à tout le moins en ce qui concerne les textes longs (1er phase). Ce concours a ainsi été l'occasion d'une étude détaillée pour vérifier les prétentions de neuf moteurs de correction disponibles commercialement (Shermis et Hamner, 2013). L'étude a permis de comparer la correction automatisée de ces neuf moteurs à celle de correcteurs humains à partir d'un échantillon d'environ 22000 textes variés provenant de six états des États-Unis. Les résultats de l'étude seraient trop longs à présenter en détail dans cette introduction mais le lecteur intéressé peut consulter l'étude de Shermis et Hamner[3]. On peut résumer le tout en disant que les résultats de la correction automatisée étaient statistiquement identiques à la correction humaine. Pour les réponses courtes (2e phase), le succès de la correction

[3] Aussi disponible à l'adresse URL http://www.scoreright.org/NCME_2012_Paper3_29_12.pdf

automatisée a été plus mitigé, les moteurs de correction offrant une bonne performance mais avec moins de précision que celle des humains.[4] Naturellement, ces moteurs de correction devraient s'améliorer dans les années à venir et être encore plus performants, sans jamais se fatiguer et en corrigeant toujours de la même manière.

Bienvenue dans l'ère du *Big Data* et du *Data Mining*, avec une bonne dose d'intelligence artificielle. En effet, ces moteurs de correction, qui pour certains ont une longue histoire et ne datent pas d'hier (voir Page, 1966), ont tous la particularité de pouvoir apprendre à partir d'un échantillon de textes qui sert à les entraîner. Il s'agit de ce qu'en intelligence artificielle on nomme *Machine learning*, l'apprentissage automatisé, qui se réalise à partir d'un corpus de données existantes. L'existence d'immenses bases de données (le *Big Data*) récoltées par les plateformes Web les plus diverses (réseaux sociaux, fureteurs, téléphones intelligents, etc.) est à l'origine d'un développement important d'outils d'analyse et de croisement de données (le *Data Mining*) des plus diverses (textes, sons, images, vidéos, localisations géographiques, pour ne nommer que les plus évidentes). Les trois conditions essentielles pour l'exploitation de ces données sont donc réunies : elles existent en très grand nombre, elles sont stockées sur des supports accessibles, de puissants outils sont disponibles pour les analyser.

L'évaluation des apprentissages, lors d'opérations à grande échelle ou dans la salle de classe par les enseignants, pourrait être un grand bénéficiaire de ces développements technologiques de récolte, stockage et analyse des données. Dans les années à venir il y aura sans l'ombre d'un doute de plus en plus d'épreuves passées en ligne par les élèves et

[4] http://gettingsmart.com/wp-content/uploads/2013/02/ASAP-Case-Study-FINAL.pdf

de travaux transmis directement aux enseignants. Donc, beaucoup, beaucoup de données disponibles pour suivre les élèves, suivre l'état du système d'éducation, faire des diagnostics individuels et collectifs... Et ces données ne seront exploitables qu'à la condition de bien comprendre les mécanismes qui prévalent à leur production et à leur validité dans les systèmes technologiques mis en place.

Dans la lignée de deux ouvrages précédents sur le même thème (Blais, 2009; Blais et Gilles, 2011), le présent ouvrage expose différentes facettes propres à l'évaluation des apprentissages à l'aide des technologies de l'information et de la communication. Il est divisé en quatre parties qui abordent la complexité sous-jacente à de nouvelles approches pour récolter des données pour l'évaluation, des dispositifs numériques en ligne, des outils pour l'analyse de séquences filmées, et finalement, la modélisation de données pour l'évaluation.

Dans la première partie, Morin et Blais présentent en détail la carte conceptuelle (aussi réseau de concepts ou schéma sémantique), un outil d'apprentissage qui permet aux étudiants de déterminer les notions importantes en lien avec un thème ou un contenu de cours, et de les organiser sous une forme graphique. Leur texte s'attarde à tracer un portrait des différentes caractéristiques des tâches d'évaluation faisant appel aux principes de la cartographie conceptuelle, mais aussi à établir un bilan de la cohérence entre les divers aspects de ces tâches, notamment en ce qui concerne leur correction et leur notation, en précisant au passage les apports des technologies de l'information et de la communication. Ensuite, Diarra et Laurier s'interrogent sur la comparabilité entre une

modalité d'évaluation où l'élève utilise le traitement de texte (modalité informatisée écran clavier) et une autre où il utilise le papier-crayon (modalité manuscrite). Dans un contexte où on demandera aux élèves d'écrire plus et où la transition d'une modalité à l'autre ne sera pas uniforme pour tous les milieux, il est possible que les deux modalités subsistent en même temps et il sera important de s'assurer que les élèves bénéficient tous d'un traitement juste et équitable peu importe la modalité de passation.

Les trois chapitres de la seconde partie présentent des dispositifs numériques d'évaluation en ligne. D'abord, Deruaz s'intéresse à l'évaluation de savoirs mathématiques au début, pendant et à la fin d'un module d'apprentissage dans le cadre de la formation initiale des futurs maîtres du primaires à la Haute école pédagogique du canton de Vaud à Lausanne en Suisse. L'utilisation de questions à choix multiples lors d'un examen ne fait pas encore partie de la culture docimologique locale et elle souffre d'*a priori* négatifs tant du côté des étudiants que de celui des enseignants. Pour palier ces difficultés, un support informatique a été utilisé pour un pré-test et un post-test avec des items QCM qui intègrent la technique des degrés de certitude (voir le chapitre 11), pour des exercices d'entraînement, pour des forums de discussion et pour des votes électroniques. Ensuite, Durand, Loye, Stasse et Dupuis font état de l'expérimentation d'un dispositif de formation numérique auprès de groupes d'étudiants de 1er cycle universitaire. Des modules de formation en ligne asynchrones ont été proposés à l'intérieur d'un cours hybride et ces modules prenaient la forme d'une situation de compétences intégrant l'évaluation à l'apprentissage. Cette expérimentation a permis de constater les défis technologiques liés à la mise en application d'un dispositif de formation numérique et les

enjeux pédagogiques liés à la combinaison des approches en présentiel et en ligne. Finalement, Riopel, Potvin et Boucher-Genesse présentent le jeux Mécanika qui a été développé pour intervenir spécifiquement sur les conceptions identifiées par le test du *Force Concept Inventory*, un questionnaire standardisé sur l'application des lois de Newton en physique. Le jeu vidéo éducatif Mécanika est le résultat d'une collaboration entre le laboratoire mobile pour l'étude des cheminements d'apprentissage en sciences (LabMÉCAS) et la compagnie CRÉO. Une expérimentation à laquelle 205 élèves ont participé a mis en évidence un apprentissage significativement plus grand pour ceux qui ont utilisé le jeu dans leurs cours et avec le support de leur enseignant. La taille de cet effet sur l'apprentissage des élèves est importante et comparable à d'autres méthodes beaucoup plus exigeantes pour les enseignants.

La troisième partie de l'ouvrage propose deux textes sur l'analyse de contenu de séquences filmés. Dans le premier chapitre Derobertmasure et Demeuse s'intéressent à une compétence que doivent développer les futurs enseignants de la Communauté français de Belgique. Cette compétence consiste à amener les futurs enseignants à porter un regard réflexif sur leur pratique et constitue un défi permanent pour les formateurs et les chercheurs (comment la développer ? Comment l'évaluer ?). La réponse à ces deux questions étant primordiale si on souhaite accorder à la notion de réflexivité un statut différent de celui de « slogan » de formation. Après une brève description du contexte dans lequel la recherche présentée a été menée, les fondements théoriques sur lesquels reposent les analyses de contenu sont détaillés et les logiciels d'analyse de contenu Nvivo® et QDA Miner® sont présentés. Le chapitre se focalise ensuite sur une

fonctionnalité du logiciel QDA Miner ®: le calcul de séquences de codages. Les résultats saillants obtenus suite à l'utilisation du logiciel sont discutés et une réflexion quant à la notion d'outil d'analyse de contenu est proposée. Dans le deuxième chapitre de la section, Dehon et Canzittu posent la question du choix technique pour l'observation des classes et des interactions élèves-enseignants. Ainsi, parmi les choix disponibles sur le marché, quels outils peuvent être une aide à l'évaluation des pratiques ? Selon le statut occupé par un utilisateur – formateur, enseignant ou chercheur – quelle solution est la plus adaptée ? Le but du chapitre étant de faire état d'outils technologiques disponibles permettant d'approcher l'évaluation des pratiques de classes en dépassant la subjectivité d'une observation uniquement basée sur des perceptions que l'on pourrait étiqueter comme étant « individualistes ».

Finalement la quatrième et dernière section propose trois textes dont les préoccupations sont en lien avec l'analyse et la modélisation des données. Au premier chapitre, Malaise propose l'application de l'analyse statistique implicative avec le logiciel CHIC pour mettre en évidence les liens existant entre la maîtrise de différentes compétences et pour déterminer, parmi un ensemble de situations de compétences, lesquelles peuvent être considérées comme étant d'un même niveau de complexité. La comparaison de l'utilisation du logiciel CHIC avec une méthode davantage « classique » reposant d'une part, sur l'échelle de Guttman et, d'autre part, sur les probabilités conditionnelles met en évidence que les analyses implicatives permettent d'approcher les liens existant entre la maîtrise de différentes compétences de manière plus précise et plus fiable. Pour sa part, l'analyse des similarités révèle dans quelle mesure différents items sont ou non

semblables et permet ainsi de vérifier que les différents exercices d'un test correspondent bien au niveau de complexité identifié *a priori*. Au deuxième chapitre de la section, Béland, Magis et Raîche s'intéressent au fonctionnement différentiel d'items d'une épreuve passée selon deux modes différents, soit papier-crayon et clavier-écran. La comparaison entre les deux modalités est faite en quatre étapes : une première estimation des paramètres d'items selon le modèle logistique à trois paramètres de la théorie des réponses à l'item pour la version papier-crayon administrée en 2009 et pour la version informatisée administrée en 2011; la détection des patrons de réponses inappropriés à l'aide de l'indice l_z ; une nouvelle estimation de la difficulté des items à la suite du retrait des patrons de réponses inappropriés; une analyse du fonctionnement différentiel des items permettant de vérifier si les items sont équivalents, indépendamment de la version administrée. Les résultats des diverses analyses permettent de faire émerger un constat général : l'existence d'une différence notable entre les deux modes de passation du test de classement en anglais-langue seconde. Par contre, un manque d'informations ne permet pas de bien expliquer la nature de ces différences. Le dernier chapitre de la section et du livre, celui de Prosperi, Gilles et Blais, aborde la technique des degrés de certitude de la réponse à un item de type QCM, une technique qui permet de dépasser le caractère dichotomique de la cotation des performances des candidats (la proposition choisie est soit correcte, soit incorrecte) à condition de veiller à respecter une série de règles méthodologiques appelées *admissible probability measurement procedures*. Un nouvel indice spectral de difficulté subjective *DS90* est proposé faisant appel à la théorie des réponses aux items, et ce, dans le but d'y intégrer la dimension des degrés de certitude. L'intention n'étant pas de remplacer ce qui fonctionne déjà, mais d'intégrer les degrés de

certitude dans ce champ de la recherche en édumétrie, notamment en vue d'améliorer la calibration des items et des tests. Le calcul du *DS90* fait appel au principe de l'analyse turbo (Gilles, 2002) qui permet de calculer des indices spectraux à partir des données des candidats dont le score en réalisme est élevé ($Rs \geq 90$), ce qui offre l'avantage de calculer des indices spectraux à partir de données très fiables comportant peu d'erreurs de sur ou sous-estimation, mais qui peut aussi conduire à une forte réduction des effectifs.

Ainsi qu'il a été souligné ci-dessus, la question de la saisie des données pour l'évaluation des apprentissages avec un support technologique contemporain est une solution qui devrait à moyen terme être acceptée et utilisée dans nombre de juridictions éducatives à travers le monde. Les États-Unis font évidemment figure de leader en la matière en investissant massivement dans le développement de plateformes génériques (et certaines ouvertes) pour les opérations de récolte de données à grande échelle. Pour l'instant, on s'attend à ce que les enseignants et les autorités locales bénéficient aussi de ces développements et utilisent les ressources disponibles pour adapter à leurs besoins ce qui leur sera proposé. Cela reste à voir. Géré de façon centralisée, le stockage des données ne pose pas de problème, tant au niveau de la quantité que de l'accès sécurisé. Mais pour ce qui est des données récoltées par des enseignants en salle de classe par exemple, et malgré la présence d'un système technologique adéquat, l'isolement de chacun pourrait contribuer à une dissolution de l'information et limiterait l'apport complémentaire de la comparaison entre différents contextes. Une mise en réseau et des ressources d'analyse seraient nécessaires pour une meilleure exploitation des données et devrait régler le problème du stockage et de la sécurité. Donc, si les problèmes reliés à la capture des

données avec un support technologique adéquat et ceux reliés au stockage des données sont pratiquement résolus, que reste-t-il à régler pour que l'évaluation des apprentissages à l'aide des technologies de l'information et de la communication remplisse des promesses maintes fois réitérées depuis une cinquantaine d'années? Il manque en fait les outils d'analyse des données. Et qui dit *Big Data*, dit aussi potentiellement données peu structurées et difficiles à analyser sans programmation dédiée à cette tâche. Heureusement, le développement de ces outils est déjà en branle, comme mentionné ci-dessus, et beaucoup d'efforts sont déployés pour l'analyse des données récoltées par les plateformes Web les plus diverses (réseaux sociaux, fureteurs, téléphones intelligents, etc.). Il faudra cependant les adapter, les rendre accessibles à coûts modiques (un autre grand défi) et former des individus qui seront à même de les utiliser et de les disséminer au bénéfice du plus grand nombre.

Références

Blais, J.-G. (dir.) (2009). *Évaluation des apprentissages et technologies de l'information et de la communication : Enjeux, applications et modèles de mesure.* Québec : Les Presses de l'Université Laval.

Blais, J.-G. et Gilles, J.-L. (dir.) (2011). *Évaluation des apprentissages et technologies de l'information et de la communication : Le futur est à notre porte*. Québec : Les Presses de l'Université Laval.

Blais, J.-G. (2011). L'évaluation des apprentissages intégrée à l'enseignement avec les technologies de l'information et de la communication. Le défi du passage à l'acte. Dans J.-G. Blais et J.-L. Gilles (dir.), *Évaluation des apprentissages et technologies de l'Information et de la communication*. *Le futur est à notre porte*. Québec : Les Presses de l'Université Laval.

Page, E.B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, **48**, 238-243.

Shermis, M.D. et Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. Dans M.D. Shermis et J. Burstein (dir.), *Hanbook of automated essay evaluation : Current applications and new directions*. New York : Routledge.