

Évaluation efficace des acquis mathématiques pour l'apprentissage personnalisé

Michel C. Desmarais¹, Carole Burney-Vincent², France Caron³, Michel Gagnon¹,
Olivier Gagnon¹, François Lemieux¹, Ildiko Pelckzer²

¹Génie informatique et génie logiciel, École Polytechnique de Montréal
{michel.desmarais, michel.gagnon, olivier.gagnon, francois.lemieux}@polymtl.ca

²Génie industriel, École Polytechnique de Montréal
{carole.burney-vincent, ildiko.pelckzer}@polymtl.ca

³Didactique, Université de Montréal
france.caron@umontreal.ca

Sommaire. Les évaluations formatives fournissent un retour d'information très utile à l'étudiant. Elles lui permettent de se situer par rapport aux exigences et par rapport à son groupe. Cette information est particulièrement importante dans un contexte d'apprentissage autonome et à distance. Dans ce contexte, la possibilité de fournir une évaluation plus précise avec un petit nombre de questions s'avère un avantage important, car autrement l'étudiant risque de rapidement se lasser de longues évaluations formatives. Nous évaluons le gain que l'on peut obtenir avec une approche de tests adaptatifs basée sur la Théorie des Réponses aux Items. Cette question est explorée par le biais d'une expérience de simulation avec des données de tests en mathématiques collégiales obtenues chez des nouveaux inscrits en ingénierie. Les résultats indiquent que l'approche de tests adaptatifs permet de réduire de moitié le nombre de questions pour en arriver à un diagnostic des compétences acquises équivalent à un test traditionnel, non adaptatif. L'expérience suggère aussi qu'un nombre de 20 questions en mode adaptatif apporte une évaluation relativement stable et qui converge vers les résultats d'un test plus long.

1 Introduction

Plusieurs étudiants en ingénierie éprouvent des difficultés à réussir les cours de mathématiques. Ces cours s'avèrent souvent difficiles non seulement pour la première année, mais parfois tout au long du baccalauréat. On peut même croire qu'une partie des candidats potentiels à la formation d'ingénieur choisit une autre vocation car ces candidats n'ont pas l'assurance de posséder les acquis mathématiques nécessaires, malgré l'attrait que peuvent représenter chez eux la profession d'ingénieur et celui d'une spécialité.

Pour pallier à ce problème, l'École Polytechnique offre depuis les années 1990 un pré-test de mathématique visant à détecter les étudiants à risque pour les cours de mathématiques. Pour ces étudiants, l'École offre un cours d'appoint qui, jusqu'ici, est offert en mode intensif, toute la journée durant 3 semaines à l'été, juste avant le début du trimestre. Ce cours est certainement un pas dans la bonne direction, mais le mode intensif durant l'été n'est pas l'idéal pour une grande partie des étudiants, comme ceux qui ont un emploi ou les étudiants venants de l'extérieur de Montréal qui n'ont pas encore accès à leur logement. De plus, certains étudiants maîtrisent parfois une partie importante du contenu du cours et pourraient être dispensés de certains sujets.

Compte tenu de ces inconvénients, le cours d'appoint se prêterait parfaitement à un mode plus flexible pour les étudiants. On peut, par exemple, envisager de le diviser en trois modules et de le donner en mode d'apprentissage à distance. Dans cette perspective, il est toutefois important que l'étudiant soit bien aiguillé vers les modules qui sont pertinents en fonction de ses acquis et de lui donner les moyens de savoir à quel moment il a atteint les objectifs d'apprentissage pour obtenir un niveau de performance visé dans les cours de mathématiques qui l'attendent.

L'objectif du projet de *Services d'apprentissage personnalisés*¹, réalisé dans le cadre des projets stratégiques de la MATI, est justement d'offrir un tel outil. Le projet repose en bonne partie sur une technologie de tests adaptatifs qui permet de diminuer le nombre d'items d'un test pour en arriver à un diagnostic des connaissances acquises. Cette capacité de diagnostic rapide est primordiale, car il ne fait aucun doute que les étudiants ne désirent pas subir une longue séance de test à chaque fois qu'une évaluation des acquis est

1. <http://www.matimtl.ca/projet.jsp?id=53&type=termines>

désirée.

Nous rapportons dans ce texte les résultats d'une simulation afin de déterminer quel est le nombre suffisant d'items à administrer pour obtenir un diagnostic précis des acquis en mathématiques et dans quelle proportion une approche de tests adaptatifs permet de réduire ce nombre. Cette recherche permettra d'établir certaines bases de l'outil d'évaluation formative visé par les prochaines phases du projet de *Services d'apprentissage personnalisés*.

2 Tests adaptatifs

Les tests en format électronique offrent la possibilité de choisir les items durant l'épreuve en fonction des réponses fournies. Ce mode de choix dynamique des items d'un test correspond à la notion de *tests adaptatifs*. Pour un même nombre d'items, l'adaptation du choix des items selon les réponses permet en retour de tirer davantage d'information du test que pour un mode non adapté. Par exemple, un item très difficile apportera moins d'information pour un étudiant dont l'habileté est faible qu'un item plus facile. En fait, ce sont les items de difficulté moyenne *par rapport à l'habileté acquise* qui apporteront le plus d'information, c'est-à-dire dont la probabilité de succès s'approche de 0,5.

Dans le domaine des tests adaptatifs, la théorie la mieux établie est probablement celle de la TRI (Théorie des Réponses aux Items). Dans le cadre de la TRI, le problème du choix de l'item le plus informatif repose généralement sur la mesure de l'*information de Fisher*. L'information de Fisher peut être mesurée par rapport à chaque item et elle dépend du niveau d'habileté du répondant. Sa valeur est maximale lorsque l'habileté correspond à une probabilité de succès est de 0,5 pour cet item. D'autre part, sa valeur dépend aussi de la valeur de discrimination de l'item. Par exemple, si la probabilité de succès à un item est de 0,5 mais qu'elle ne varie pas selon l'habileté, sa discrimination sera nulle. Au contraire, si elle discrimine parfaitement entre des individus dont l'habileté diffère légèrement, sa discrimination sera grande (c'est le cas d'un item pour lequel la majorité des répondants dont l'habileté est juste au dessus d'un seuil d'habileté donné le réussissent, alors ceux dont l'habileté est juste sous le seuil échouent à cet item). Pour un répondant doté d'une

habileté donnée, l'information de Fisher sera maximale pour l'item dont la combinaison de discrimination et de difficulté sera maximale. La discrimination est un attribut propre à l'item tandis que la difficulté sera optimale si la probabilité de succès s'approche de 0,5. Les détails de cette théorie peuvent être consultés dans (Baker & Kim, 2004).

En plus de fournir les bases théoriques au choix des items, la TRI fournit aussi un cadre théorique solide pour l'estimation des habiletés acquises. En fait, il s'agit d'une condition préalable pour déterminer l'information de Fisher puisque cette information dépend de l'estimé du niveau d'habileté.

La TRI présume que la probabilité de bien répondre à un item dépend d'une habileté, laquelle on réfère généralement par le symbole θ . La probabilité augmente normalement avec l'habileté et suit une courbe en forme de 'S' dont le point d'inflexion est situé à $p = 0,5$, indiquant autant de chances de succès que d'échec à ce point. Une méthode largement utilisée pour estimer l'habileté est celle du calcul de maximum de vraisemblance. L'habileté estimé correspond à la valeur de θ pour laquelle la probabilité de la séquence de succès et d'échecs observée aux items d'un test est maximale. Ici encore, nous référons le lecteur à Baker and Kim (2004) pour obtenir des détails sur l'estimation de l'habileté.

Il existe différentes variantes de modèles pour la TRI. Celui que nous utilisons pour cette étude est le modèle logistique à deux paramètres, la difficulté et la discrimination (Lord & Novick, 1968).

3 Objectifs et méthodologie de l'étude

S'il est bien établi que la possibilité de choisir des items dans un test adaptatif apporte davantage d'information, combien d'information supplémentaire est ainsi obtenue ? La réponse à cette question est encore relativement inconnue, du moins dans le contexte du projet *Services d'apprentissage personnalisés*. Nous désirons connaître plus spécifiquement combien d'items sont nécessaires pour arriver à un diagnostic des habiletés acquises relativement fiable, à la fois pour un test adaptatif et pour un test traditionnel.

Avant d'aborder cette question, apportons une note concernant la terminologie utilisée dans ce texte : nous utiliserons à partir d'ici le terme *quiz* pour référer aux tests basés sur un petit nombre d'items administrés soit en mode traditionnel (choix fixe d'items donnés à tous les répondants) ou en mode adaptatif (choix dynamique).

Une façon d'obtenir une réponse à la question soulevée est de comparer le diagnostic d'un quiz avec d'autres mesures indépendantes de l'habileté, notamment les résultats à un test plus long et les résultats à un cours de mathématique.

Ainsi, pour quantifier la différence d'information obtenue entre un quiz en mode adaptatif et un autre en mode traditionnel, nous comparons l'estimation des scores selon ces deux modes de quiz avec deux sources : (1) le score des étudiants à un cours de mathématique et (2) le score à l'Épreuve complète qui compte 60 questions.

Pour des quiz dont les estimés de l'habileté sont égaux lorsque comparés aux deux sources mentionnées, on peut présumer que la quantité d'information obtenue des quiz pour effectuer le diagnostic est équivalente, même si le nombre et le choix des questions diffèrent. De plus, lorsque la quantité d'information obtenue d'un quiz converge vers celle d'un test plus complet, on peut alors présumer que la longueur de ce quiz s'approche du maximum d'information qu'on peut obtenir.

La méthodologie pour élucider la question qui nous intéresse consiste donc à simuler des quiz de différentes longueurs avec chaque approche, puis d'évaluer la précision du diagnostic de chaque approche en le comparant avec celui des deux sources mentionnées, un test plus long et le résultat final à un cours de Calcul I.

3.1 Données et création de quiz

Les simulations sont effectuées à partir de données qui proviennent des deux sources mentionnées et dont voici quelques détails.

Épreuve profil mathématique. La première source est celle de l'épreuve de mathématique de 60 questions administrée à 157 nouveaux étudiants en ingénierie à Polytechnique à l'été 2010. Cette épreuve comportait 30 items fixes et 30 autres adaptées selon un algorithme de tests adaptatifs. Les 30 items adaptés ont été sélectionnés dans un bassin de 74 items et sont évidemment différents d'un étudiant à l'autre.

Résultats du cours Calcul I. Une majorité des étudiants qui ont participé à l'épreuve ont suivi le cours Calcul I (MTH1101 à Polytechnique) à l'automne 2010. Leurs résultats finaux permettent d'obtenir une source indépendante d'évaluation des habiletés en mathématiques.

Données historiques. Outre les sources de données ci-dessus, une troisième source est mise à contribution pour calibrer les paramètres des modèles de chaque approche. Ces données proviennent de 2650 répondants et ont été colligées depuis environ une dizaine d'années, de sorte que pour chaque item nous disposons d'un minimum de 222 réponses et d'un maximum de 1723 réponses, selon la fréquence à laquelle l'item s'est retrouvé dans les épreuves passées. La calibration des modèles de chaque approche respective est décrite plus loin.

Quiz. Les quiz de différentes longueurs ont été simulés en échantillonnant des items dans l'Épreuve profil mathématique. Ces longueurs de quiz vont de 5 à 20 items, ce qui correspond à des quiz qu'un étudiant peut effectuer relativement rapidement pour obtenir une évaluation de nature formative.

Pour l'approche traditionnelle, l'échantillon est pris dans la partie fixe du questionnaire. Ces questions avaient été présélectionnées pour couvrir différents sujets et différents niveaux de difficulté comme on le fait normalement pour des quiz fixes. Pour l'approche adaptative, l'échantillonnage est effectué à partir de l'ensemble des 60 items administrés à chaque répondant, c'est-à-dire à la fois du bassin de 30 items fixes, mais aussi dans le bassin des 30 items adaptatifs, qui provenaient eux-mêmes d'un bassin de 74 items possibles comme mentionné.

3.2 Mesures de l'habileté et corrélations

À partir des quiz de 5 à 20 items, une estimation de l'habileté est effectuée.

Pour l'approche TRI, le modèle logistique à deux paramètres est adopté comme on l'a mentionné. Les paramètres *difficulté* et *discrimination* associés à chaque item sont estimés à partir des données historiques à l'aide du module `ltm` (Rizopoulos, 2006). La prédiction de l'habileté (θ) à partir des réponses aux quiz est effectuée à partir des échantillons conformément à l'approche TRI et toujours avec le module `ltm`.

Pour l'approche traditionnelle, l'estimation du score repose sur la cote Z . Le score du quiz est reporté sur une distribution normale ($\sigma(0, 1)$). Il tient ainsi compte de la difficulté des items du quiz.

Plus spécifiquement, supposons un quiz Q_n composé de n items tirés du bassin d'items Q , $Q_n \subset Q$, et un vecteur de résultats binaires pour un répondant donné j , $R_j(Q_n)$, alors le score du répondant est :

$$S_j = \sqrt{\frac{1 + \sum_{q_i \in R_j(Q_n)} q_i - \bar{q}_i}{2 + \sum_{q_i \in R_j(Q_n)} \bar{q}_i^2}}$$

où \bar{q}_i est le taux moyen de succès à l'item i pour les données historiques et où q_i prend la valeur de 0 pour un échec et de 1 pour un succès. L'addition de 1 au numérateur et 2 au dénominateur est une correction qui permet d'éviter les erreurs numériques.

Notons que les valeurs des habiletés S_j (traditionnel) et θ (adaptatif) se situent dans le domaine $[-\infty, +\infty]$

Pour mesurer la validité de l'estimation des habiletés S_j et θ , la corrélation entre ces mesures et le score à l'Épreuve ou à la note finale en pourcentage au cours Calcul I est calculée sur la base de la transformation du pourcentage sur une échelle $[-\infty, +\infty]$ par la fonction *logit*. De plus, comme le nombre d'items est relativement faible (de 5 à 20), les résultats varient considérablement d'un échantillon de question à un autre. Les corrélations sont donc calculées sur la base 100 replis : un échantillon aléatoire d'items est répété 100 fois et la corrélation moyenne est calculée. Les replis ne s'appliquent toutefois pas dans le cas du mode adaptatif puisque les choix d'items ne sont pas aléatoires mais plutôt déterminés par la quantité d'information de chaque item.

4 Résultats

Les résultats des différentes simulations sont rapportés au tableau 1. On y retrouve la corrélation (de Pearson) entre les quiz et le score au cours de Calcul I et celui de l'Épreuve complète de 60 questions. Les conditions de quiz composés de 5, 10, 15 et 20 sont rapportées pour les quiz en mode traditionnel et adaptatif avec la TRI. Comme mentionné

Tableau 1 – Corrélations des mesures de score

Mode du quiz	Nb. items	Corrélations	
		Calcul I	Score Épreuve
Traditionnel	5	0,34	0,64
TRI	5	0,43	0,79
Traditionnel	10	0,42	0,77
TRI	10	0,49	0,88
Traditionnel	15	0,45	0,83
TRI	15	0,50	0,90
Traditionnel	20	0,48	0,87
TRI	20	0,50	0,91

dans la section précédente, les corrélations en mode traditionnel du tableau 1 sont en fait des moyennes basées sur 100 replis et sont donc très stables (écart type d'environ $\pm 0,02$), alors que les corrélations pour le mode TRI ne varient pas car la composition des quiz est déterminée par l'algorithme de test adaptatif et il n'y a pas de dimension aléatoire.

On constate que les corrélations pour 5 items en mode adaptatif (TRI) sont semblables à celles du mode traditionnel pour 10 items. De manière similaire, les corrélations de 10 items en mode adaptatif se comparent à celles de 20 items en mode traditionnel. Pour ces longueurs de quiz, on peut donc avancer que le mode adaptatif permet de réduire de moitié le nombre d'items pour en arriver à un estimé de l'habileté semblable au mode traditionnel.

D'autre part, la corrélation entre l'Épreuve complète, qui, rappelons-le, est de 60 items, et le score au cours de Calcul I est de 0,51. Selon les résultats du tableau 1, il s'avère que le score en mode adaptatif obtenu avec un quiz de 10 à 15 items, dont la corrélation avec le cours Calcul I varie de 0,49 à 0,50, s'approche de celui du test complet de 60 items.

5 Discussion

L'utilisation de tests adaptatifs pour fournir une forme d'apprentissage personnalisé n'est pas nouvelle. Chen, Lee, and Chen (2005) rapporte une récente étude portant sur l'u-

tilisation de la TRI dans cette perspective. Le système commercial ALEKS^{MC} (Falmagne, Cosyn, Doignon, & Thiéry, 2006) utilise une autre approche de tests adaptatifs pour fournir une évaluation des acquis sur laquelle un plan d'étude dynamique est basé. Cependant, l'étude rapportée dans ce texte amène de nouvelles connaissances quant au gain qu'apporte la caractéristique adaptative de tests avec l'approche TRI.

Pour l'Épreuve mathématique de Polytechnique, les résultats suggèrent que l'approche adaptative permet de réduire de moitié le nombre d'items d'un quiz et qu'un nombre de 20 items s'approche vraisemblablement du maximum d'information de diagnostic que l'on peut obtenir. Ces informations sont très précieuses. Elles permettront de déterminer la longueur des quiz formatifs que l'on devrait concevoir pour un outil qui guidera l'étudiant dans une démarche d'apprentissage autonome.

Est-ce à dire que ces tendances sont généralisables et qu'on peut présumer que les tests adaptatifs doublent l'information de diagnostic, et que l'information plafonne à 20 items pour le mode adaptatif comparativement à 40 pour un test traditionnel ? Probablement pas, mais d'autres études devront apporter plus de lumières sur le sujet. Quoi qu'il en soit, la méthodologie proposée dans ce texte demeure une approche qui semble valide pour explorer la question.

Références

- Baker, F. B., & Kim, S.-H. (2004). *Item response theory, parameter estimation techniques (2nd ed.)*. New York, NY: Marcel Dekker Inc.
- Chen, C.-M., Lee, H.-M., & Chen, Y.-H. (2005). Personalized e-learning system using item response theory. *Computers and Education*, 44(3), 237 - 255. Available from <http://www.sciencedirect.com/science/article/B6VCJ-4BYJVV5-1/2/5e063bdb6470aafe303bacdf088289dc>
- Falmagne, J.-C., Cosyn, E., Doignon, J.-P., & Thiéry, N. (2006). The assessment of knowledge, in theory and in practice. In R. Missaoui & J. Schmid (Eds.), *ICFCA* (Vol. 3874, pp. 61–79). Springer.
- Lord, F. M., & Novick, M. R. (Eds.). (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Rizopoulos, D. (2006). ltm : An r package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5), 1–25.