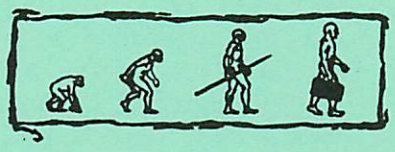
Actes du 14e colloque de l'AQPC

ÉVALUATION ! ÉVOLUTION ?



Où s'en va le collégial ?

7A19

L'évaluation nationale individualisée et assistée par ordinateur

par
RAÎCHE, Gilles
conseiller pédagogique
Collège Joliette - De Lanaudière
BÉLAND, Anne
chercheure
Institut Philippe-Pinel de Montréal

Association québécoise de pédagogie collégiale

L'ÉVALUATION NATIONALE INDIVIDUALISÉE ET ASSISTÉE PAR ORDINATEUR

Gilles Raîche
Conseiller pédagogique
Cégep Joliette - De Lanaudière

Anne Béland
Chercheure
Institut Philippe-Pinel de Montréal

INTRODUCTION

Lors de la 16e session d'étude de l'Association pour le développement de la mesure et de l'évaluation en éducation (ADMÉÉ), dont le thème était l'évaluation de demain, Micheline Lavallée, représentante du Ministère de l'éducation présentait les priorités à venir, selon elle, en éducation.

Ces priorités se divisent en cinq catégories: l'évaluation formative, l'évaluation authentique, l'évaluation nationale, la recherche appliquée en évaluation et l'imputabilité.

Ainsi, selon elle, l'évaluation formative devra prendre de plus en plus de place dans l'enseignement. Le dossier de l'évaluation sommative est fermé. Le temps est arrivé de mettre vraiment en place la pratique de l'évaluation formative.

L'évaluation de demain devra être authentique. Elle devra permettre d'évaluer la pensée critique et être adaptée à la réalité. L'évaluation devra être fondée sur les productions des étudiants et des étudiantes: présentation orale, texte d'opinion, port-folio, etc.. Nous emboîterons donc le pas à nos voisins du Sud qui ont déjà pris ce tournant.

La recherche appliquée en évaluation sera favorisée par un partenariat avec les universités et les institutions seront soumises à des règles d'imputabilité.

Enfin, dans le futur, l'état mettra en place un système plus important d'évaluation nationale. Les tests seront standardisés selon des modélisations plus modernes.

Ce texte a comme objectif de présenter une application de l'évaluation nationale aux tests de classement en anglais langue seconde. De plus, cette application sera effectuée dans un contexte d'évaluation individualisée et assistée par ordinateur.

La prochaine section répond aux trois questions suivantes. *Pourquoi une évaluation nationale? Pourquoi une évaluation individualisée? Pourquoi une évaluation assistée par ordinateur?*

Ensuite, une étude critique d'une application d'un test de classement en anglais langue seconde est abordée. Il s'agit du test TCALS, fort répandu à l'ordre collégial. Les implications logistiques, financières et métrologiques seront discutées.

Suivra une section où le problème de rendre les scores comparables lorsqu'un test est individualisé sera abordé. La théorie de la réponse à l'item sera suggérée en tant que solution à ce problème. Une brève analyse métrologique de TCALS dans le contexte de la théorie de la réponse à l'item sera ensuite présentée.

Enfin, des conclusions et des recommandatoins viendront clore le texte.

JUSTIFICATION

Pourquoi une évaluation nationale?

Une évaluation nationale est une évaluation qui est administrée de façon uniforme à toute une population. Un test de classement en anglais langue seconde qui serait administré à tous les élèves nouvellement inscrits dans tous les collèges du Québec en est un exemple. Le résultat d'un élève est alors comparable à celui d'un autre car le même instrument de mesure est utilisé pour tous. Ceci confère au résultat un caractère d'équité. De plus, ces résultats peuvent ensuite être utilisés aussi bien à des fins de recherche appliquée qu'à des fins d'imputabilité institutionnelle.

Pourquoi une évaluation individualisée?

Une évaluation individualisée est une évaluation qui est adaptée à chaque élève auquel elle est administrée. Elle est donc sur mesure. Le niveau de difficulté des questions proposées à l'élève est ajusté en fonction d'informations antérieures à son sujet. Par exemple, si un élève admis à l'ordre collégial affichait des résultats élevés en anglais à l'ordre secondaire, il serait plus raisonnable de lui administrer des questions difficiles. S'il fournit une bonne réponse à un item

d'un test, la question suivante est plus difficile. À l'inverse, s'il fournit une mauvaise réponse, la prochaine question est plus facile. Le test est donc adaptatif.

Les recherches effectuées jusqu'à présent sur l'évaluation individualisée indiquent que le nombre de questions devant être administrées à un élève est diminué de moitié, tout en obtenant le même degré de précision dans la mesure de l'habileté de l'élève. En fait, il est fréquent qu'après seulement 25 questions un résultat fiable soit obtenu. Cette situation comporte aussi des avantages secondaires. Ainsi, le temps de testing est diminué. C'est un facteur non négligeable lorsque beaucoup d'élèves sont impliqués. Les élèves apprécient généralement ce type de situation car le temps de passation étant moindre, la fatigue aussi est moindre. Le stress de la passation est ainsi diminué car les questions sont adaptées au niveau d'habileté de l'élève. Enfin, les résultats obtenus sont aussi plus fidèles lorsque vient le moment de les utiliser dans un contexte de recherche ou encore d'imputabilité.

Pourquoi une évaluation assistée par ordinateur?

Une évaluation assistée par ordinateur offre l'avantage de permettre une meilleure gestion de l'administration d'un test individualisé. Les questions proposées peuvent être sélectionnées automatiquement et attribuées de façon aléatoire tout en tenant compte du niveau d'habileté estimé de l'élève. Une évaluation assistée par ordinateur, cependant, permet surtout d'optimiser le temps de passation et la fidélité du résultat en utilisant des algorithmes de calcul de pointes.

L'élève peut alors se présenter seul devant l'ordinateur à tout moment de la journée où l'appareil est disponible. Il n'est plus nécessaire de regrouper les élèves en grand groupe. Il est même possible pour lui de passer le test à la maison par modem. La correction de l'épreuve est immédiate et l'élève reçoit son résultat directement. Des économies en frais de personnel et de communication sont alors possibles.

La prochaine section présente une application d'une évaluation nationale individualisée assistée par ordinateur.

APPLICATION

En enseignement de l'anglais langue seconde au collégial, les cégeps doivent généralement évaluer le niveau d'habileté en anglais des élèves à leur entrée dans les programmes d'études. Ces évaluations ont pour objectif de diriger les élèves vers des niveaux de cours adaptés aux compétences linguistiques de cette clientèle. Ce sont donc des tests de classement. En 1992, Paul Fournier dressait un bilan des tests de classement en anglais langue seconde à l'ordre collégial. Du même coup, il formulait des recommandations pour pallier au fait que les tests existants présentent des lacunes. Ainsi, l'interprétation des résultats varie localement puisque des tests différents sont administrés dans chaque collège. Dans d'autres cas, les critères de classification à un même test varient d'un cégep à un autre. Selon Fournier, une certaine standardisation serait souhaitable. Il est alors question d'une certaine nationalisation. Ensuite, de nombreux élèves doivent être reclassés à un niveau supérieur ou inférieur à chaque session. Des lacunes au niveau de la fidélité et de la validité de prédiction des résultats sont donc présentes. Une évaluation individualisée permettrait ainsi d'augmenter la précision des résultats.

Une réalité nouvelle apportée par la réforme de l'enseignement collégial rend la formation académique en langue seconde obligatoire pour tous les élèves dans leur programme d'études. Le nombre d'élèves à qui seront administrés les tests de classement dans chaque collège sera alors considérable. Il devient donc pertinent d'envisager des conditions d'administration de ces tests qui seront plus efficaces et moins coûteuses. L'utilisation de la technologie informatique est ainsi une voie à explorer.

Une réalité supplémentaire est aussi à considérer. Le nombre d'élèves dans chaque cégep sera maintenant assez élevé pour permettre la calibration des tests et la standardisation des résultats de façon locale. Il n'est plus nécessaire d'attendre que cette opération se fasse à l'échelle de la province. Des cégeps innovateurs peuvent prendre les devants.

Cependant, puisque l'évaluation des élèves dans leur habileté en anglais langue seconde est individualisée et que, de suite, chacun reçoit donc un test différent, comment rendre les résultats de ces étudiants comparables entre eux? Comment faire en sorte que le

résultat d'un élève fort qui a réussi 50% des items à un test composé d'items difficiles puisse être comparé au résultat d'un élève faible qui a réussi lui aussi 50% des items à un test composé d'items très faciles? La prochaine section présente brièvement une solution à ce problème par le biais de la théorie de la réponse à l'item. De plus, une analyse métrologique de TCAL-S, un test de classement en anglais langue seconde, sera abordée de façon succinte dans le cadre de la théorie de la réponse à l'item.

Comment rendre les résultats comparables lorsqu'un test est adaptatif?

La théorie de la réponse à l'item offre une solution pour permettre de rendre comparables les résultats obtenus à des tests composés d'items différents (Raîche, 1994). Cette théorie repose sur le postulat que le résultat obtenu à un test est le reflet d'un trait, d'une habileté, non observable (Laurier, 1993). La performance observée est manifeste, tandis que le trait latent, en l'occurence le niveau d'habileté en anglais langue seconde, est un construit hypothétique non observable (Béland, 1989). De plus cette habileté est unidimensionnelle. Elle est donc un trait prédominant du test. Le test ne mesure, à toute fin pratique, que cette habileté.

La théorie de la réponse à l'item est construite autour d'une modélisation de la probabilité qu'a un élève de réussir un item du test lorsque son habileté est à un niveau déterminé. Plusieurs modélisations de cette probabilité ont été proposées. L'une d'elles ne tient compte que d'un paramètre: la difficulté de l'item. Dans ce modèle la difficulté de l'item correspond au niveau d'habileté où un élève a une chance sur deux de réussir l'item. Tous les items d'un test, comme TCALS, ont été calibrés selon ce modèle. Il est ensuite possible d'estimer le niveau d'habileté d'un élève en combinant les fonctions de probabilité de chacun des items répondus au test. Le lecteur intéressé à pousser plus loin sa compréhension de ce modèle pourra consulter l'ouvrage de Wainer (1990). Le niveau d'habileté est, selon ce modèle, indépendant des items retenus. Cependant, à nombre d'items égal, les tests dont les items ont un niveau de difficulté voisin du niveau d'habileté de l'élève offrent une plus grande précision que les tests où la difficulté des items diffère considérablement du niveau d'habileté de l'élève.

A titre d'exemple une brève analyse métrologique de TCALS a été effectuée. Les indices de difficulté des 100 items, ou questions, du test ont été calculés.

L'indice de difficulté moyen des items de TCALS est de -.88 avec un écart-type de 1.02. Ce score doit être interprété comme une cote z. Autour de zéro, il correspond à un niveau moyen faible. Un score positif supérieur à environ 1.5 est rattaché à un item jugé difficile, tandis que l'item est considéré très difficile s'il est supérieur à 3. A l'inverse, l'item est facile si le score est inférieur à -1.5 et très facile s'il est inférieur à -3. Le tableau 1 permet de remarquer que quelques items sont très faciles et beaucoup sont faciles. Cette brève analyse confirme les conclusions de Fournier à l'effet que TCALS est un test construit pour classer des élèves dont le niveau d'habileté est inférieur à la moyenne.

Il n'est toutefois pas nécessaire de jeter TCALS à la poubelle et de recréer de toute pièce un instrument de classement. Il suffirait simplement d'ajouter des items plus difficiles au test. Ceci permettrait d'avoir en main une banque d'items comportant des indices de difficulté variée et de pouvoir l'utiliser dans une version individualisée assistée par ordinateur.

La prochaine section dresse quelque conclusions et recommandations en ce qui a trait à la création d'une version individualisée et assistée par ordinateur de TCALS.

CONCLUSIONS ET RECOMMANDATIONS

Avec l'implantation de nouveaux cours obligatoires en langue seconde à l'ordre collégial et l'utilisation à grande échelle de tests de classement à ces cours, il semble important de se pencher sur les problèmes logistiques, financiers et métrologiques engendrés. L'adaptation de tests de classement individualisés et assistés par ordinateur pourrait participer à la résolution de ce problème.

Une première implantation pourrait être effectuée dans un contexte d'expérimentation. La démarche pourrait être locale, dans un seul établissement, ou collective, et regrouper quelques collèges expérimentateurs. Tous les étudiants et étudiantes nouvellement admis dans un programme d'études participeraient à ce projet d'implantation. Ainsi dans un collège de taille moyen-

ne environ 1,000 étudiants et étudiantes seraient mis à contribution. Ce nombre est suffisant afin d'estimer les qualités psychométriques de l'instrument de mesure et de calibrer les indices de difficulté des items. Cette opération peut devenir, dans ce contexte, locale.

Voici les étapes suggérées pour l'implantation d'une version individualisée et assistée par ordinateur du test de classement en anglais langue seconde TCALS.

1. Un ajout d'items plus difficiles doit être effectué dans le but de mesurer avec plus de précision le niveau d'habileté des élèves plus forts en anglais langue seconde.

2. Pendant la première année de l'implantation une version papier-crayon pourrait être administrée à tous les étudiants et étudiantes du ou des collèges expérimentateurs.

3. L'étape précédente permettra d'aborder la calibration des items du test. Les items non pertinents pourraient ainsi être rejetés et des items de différents niveaux de difficulté pourraient être ajoutés.

4. Des décisions devront être prises en ce qui concerne les seuils de classification dans différents niveaux de groupe-classe.

5. Une étude de validité prédictive pourrait être effectuée à la fin de la première année d'expérimentation dans le but de vérifier l'adéquation de la classification.

6. Si les étapes précédentes mènent à une évaluation positive du projet d'implantation, la mise en place d'une version informatisée et individualisée du test TCALS serait envisagée.

7. A la fin de la seconde année d'expérimentation une étude de validité prédictive serait de nouveau réalisée. Il est d'ailleurs suggéré de répéter ce type d'étude de validité à chaque année.

Il serait intéressant d'utiliser une démarche similaire pour créer des tests de certification à la fin de chaque cours d'anglais langue seconde à l'ordre collégial. Le même test individualisé et informatisé pourrait être administré à tous les niveaux de cours comme épreuve d'évaluation terminale. Cette démarche permettrait alors de comparer réellement les résultats aux différents niveaux de cours. Elle permettrait aussi de comparer les résultats entre les collèges puisque les épreuves seraient les même.

RÉFÉRENCES

Béland, A. (1989). *Discontinuité qualitative du raisonnement proportionnel sur le continuum de la théorie de réponses aux items.* Québec: 12ième session d'étude de l'Association pour le développement de la mesure et de l'évaluation en éducation (ADMÉÉ).

Fournier, P. (1992). *Pour un test incontestable: rapport de recherche sur les tests de classement en anglais (langue seconde) au collégial.* Québec: Ministère de l'enseignement supérieur et de la science, Gouvernement du Québec.

Laurier, M. (1993). *L'informatisation d'un test de classement en langue seconde.* Québec: Université Laval.

Raîche, G. (1994). *La simulation de modèle sur ordinateur en tant que méthode de recherche: le cas concret de l'étude de la distribution échantillonnale des estimateurs en testing adaptatif en fonction des règles d'arrêt.* Joliette: 6ième colloque de l'Association pour la recherche au collégial.

Wainer, H. (1990). *Computerized adaptive testing: A primer.* Hillsdale, NJ: Lawrence Erlbaum Asssociates.

Tableau 1
Indices de difficulté des 100 items du test TCALS
(calibration effectuée à partir des résultats de 307 étudiants et étudiantes du Cégep Joliette - De Lanaudière en 1992)

1	0.76	26	-0.61	51	-2.32	76	-0.31
2	-0.05	27	-1.91	52	-2.13	77	-2.10
3	-3.90	28	0.99	53	-1.64	78	1.16
4	-0.34	29	0.64	54	-1.67	79	-0.05
5	0.79	30	-0.50	55	-0.36	80	-2.10
6	-1.72	31	0.64	56	0.14	81	-1.36
7	-0.81	32	-3.15	57	0.30	82	-1.79
8	-2.02	33	-2.13	58	0.76	83	-0.13
9	-0.04	34	0.60	59	-1.03	84	-0.71
10	-2.81	35	0.24	60	-0.80	85	-1.08
11	-0.05	36	-0.48	61	0.16	86	-1.25
12	-1.34	37	0.65	62	-0.39	87	-1.26
13	-2.05	38	-0.42	63	-0.67	88	-1.46
14	-1.38	39	-2.00	64	-1.34	89	-1.23
15	-1.21	40	-0.98	65	-1.54	90	-1.00
16	-0.39	41	-0.94	66	0.51	91	-1.32
17	-1.23	42	-0.75	67	-0.67	92	-1.21
18	-2.02	43	-0.91	68	0.05	93	-1.46
19	-0.85	44	-0.33	69	-0.81	94	-0.41
20	-2.71	45	-0.42	70	-2.28	95	-0.42
21	-0.39	46	-0.62	71	-0.64	96	-0.22
22	1.49	47	-1.49	72	-1.19	97	0.10
23	-2.58	48	-1.54	73	-0.34	98	-0.25
24	-2.26	49	-1.08	74	-2.28	99	-1.14
25	-2.47	50	-1.67	75	0.84	100	-0.72