Références : Laurier, M. (1998). "Méthodologie d'évaluation dans des contextes d'apprentissage des langues assistés par les environnements informatiques multimédias". Dans Chanier, T., Pothier, M. (Dirs), "Hypermédia et apprentissage des langues", *études de linguistique appliquée (éla)*, *110*. 247-255.

MÉTHODOLOGIE D'ÉVALUATION DANS DES CONTEXTES D'APPRENTISSAGE DES LANGUES ASSISTÉS PAR DES ENVIRONNEMENTS INFORMATIQUES MULTIMÉDIAS

Résumé : Souvent le parent pauvre dans l'élaboration d'environnements informatiques multimédias, l'évaluation peut grandement contribuer à la qualité de ces environnements. En effet, l'évaluation permet d'obtenir une information précieuse sur l'apprenant en vue de prendre les décisions les plus appropriées relativement à son cheminement. Dans le cadre de cet article, nous esquissons l'évolution des méthodes et approches en évaluation des apprentissages en langue seconde ou étrangère à l'aide de l'ordinateur. Nous présentons les différentes décisions que peut faciliter l'évaluation en voyant comment l'ordinateur peut contribuer à chacune d'entre elles. Enfin, à titre d'illustration des possibilités des environnements informatiques multimédias, nous présentons un prototype de test adaptatif.

1. LES QUATRE GÉNÉRATIONS DE TESTS INFORMATISÉS

Pour mieux comprendre les possibilités qui s'ouvrent avec le recours à des environnements multimédias dans le domaine de l'évaluation des apprentissages en langue, il est utile de se référer aux courants que distinguent Bunderson & al (1989). Ces auteurs identifient quatre générations d'environnements utilisant l'ordinateur :

- les tests informatisés,
- les tests adaptatifs,
- la mesure continue,
- la mesure intelligente.

1.1. Les tests informatisés

Un des principaux avantages réside dans la possibilité que permet l'ordinateur d'automatiser l'administration et la correction d'un test qui, dans sa forme traditionnelle, demanderait de réunir plusieurs apprenants en un même lieu pour faire le même test et demanderait ensuite de soumettre les copies àun correcteur pour établir le score de chaque apprenant. De fait, plusieurs logiciels dont la fonction est d'automatiser l'administration et la correction de tests sont apparus sur le marché. Il s'agit, dans la plupart des cas, de convertir des instruments existants sous une forme informatisée. On pense, par exemple, à la transposition de tests de closure sur ordinateur ou à l'utilisation de questions à choix multiple. Les avantages de l'informatisation de ce type de test ne sont pas négligeables particulièrement dans le cas d'une administration à grande échelle. On exploite la capacité de traitement de l'ordinateur pour la saisie de la réponse, pour l'établissement rapide et quasi infaillible d'un score faisant parfois appel à des calculs complexes et pour la gestion des résultats de l'évaluation à travers la production de listes ou même de statistiques. À ces avantages que permet la capacité de traitement, s'ajoute l'intérêt que peut présenter, du moins dans certaines situations, une administration individualisée.

Toutefois, il faut bien admettre que dans leur conception, ces instruments ne se démarquent guère des à choix de réponses de forme papier-crayon. Qui plus est, on peut facilement objecter que plusieurs moyens traditionnels, comme l'entrevue ou la composition, évaluent mieux les aspects plus communicatifs de la compétence langagière. En effet, on obtient ainsi des réponses ouvertes qu'on ne trouve pas dans les tests informatisés parce qu'elles présentent des difficultés de traitement considérables. Il faut noter par ailleurs que le concept de test informatisé, où l'on cherche essentiellement à établir un score pour chaque apprenant à partir de l'administration linéaire d'un ensemble commun de tâches, est celle qu'adopte la plupart des concepteurs de didacticiels qui comportent une dimension évaluative.

1.2. Les tests adaptatifs

Cette seconde génération se distingue essentiellement de la précédente du fait que l'administration du test n'est plus linéaire. Sur le modèle de ce que proposait Binet (1909) au début du siècle pour mesurer l'intelligence, chaque réponse que fournit l'apprenant devient une information supplémentaire qui sert à raffiner l'estimation du niveau général de l'apprenant. Cette estimation devient de plus en plus précise à mesure que progresse le test. Le processus est d'autant plus rapide que l'estimation elle-même sert à déterminer, parmi un ensemble de tâches possibles, celle qui est susceptible de maximiser l'information. Ainsi, l'administration du test n'est plus linéaire puisque le choix des questions qui sont soumises à l'apprenant dépend de la performance de celui-ci. On en arrive ainsi à une évaluation individualisée qui ne sacrifie en rien les exigences d'une mesure qui doit mener à une comparaison, que cette comparaison se fasse en fonction des autres apprenants (interprétation normative) ou par rapport à une performance attendue (interprétation critériée).

Nous présenterons plus loin un exemple de test adaptatif. Il faut souligner ici que la naissance des produits de cette seconde génération a été possible grâce aux développements des connaissances dans le domaine de la mesure et, plus particulièrement, grâce à l'émergence de la théorie de réponses aux items (pour une introduction, voir Baker, 1992). Dans le cadre de cette théorie, il devient plus facile de construire des banques d'items, c'est-à-dire des ensembles de tâches dont le niveau de difficulté varie mais qui mesurent une habileté commune. La théorie prévoit des techniques d'estimation grâce auxquelles on peut situer différents apprenants sur une échelle d'habileté même si ces apprenants n'ont pas tous accompli les mêmes tâches. L'utilisation de ces techniques dans la conception de tests utilisant l'ordinateur permet non seulement de tenir compte du niveau de l'apprenant de façon à réduire frustration et anxiété, mais elle permet aussi de réduire considérablement la durée du test. Certaines recherches (notamment Kingsbury & Weiss, 1980 ; Laurier, 1996) démontrent que les tests adaptatifs peuvent facilement être deux fois moins longs qu'un test conventionnel sans que la fidélité en soit affectée. Depuis les premiers tests adaptatifs en langue, mis au point par l'équipe de l'Université Brigham Young (Larsen & Madsen, 1985) jusqu'au projets plus récents (Chalhoub-Deville & al., 1996), on trouve plusieurs exemples de tests adaptatifs en langue seconde ou étrangère.

1.3. La mesure continue

En repoussant les limites des tests adaptatifs, on s'achemine vers les instruments de la troisième génération. La transition s'opère à plusieurs niveaux. En premier lieu, les modèles psychométriques se raffinent de façon à permettre un meilleur traitement des réponses reliées à des apprentissages complexes. Ces apprentissages comportent souvent plusieurs dimensions qui complexifient l'application des modèles habituels issus de la théorie de réponse aux items. On voit alors apparaître des tests qui comportent de multiples banques de façon à constituer un profil de l'apprenant plutôt qu'un score unique (Laurier, 1996). On en vient alors à pouvoir diagnostiquer en fonction d'un certain nombre d'habiletés importantes au lieu de situer l'élève sur le continuum relativement abstrait que représente le développement de la compétence langagière. Dans d'autres cas, comme dans le système proposé par Trentin (1997), la procédure adaptative est associée à une représentation hiérarchique du contenu qui permet de déterminer les lacunes chez l'élève en vue d'un diagnostic. Dans une autre optique, les modèles multi-dimensionnels issus de la théorie de réponse aux items permettent de rendre compte d'apprentissages complexes comme, par exemple, la lecture en langue étrangère (Kaya-Carton & al., 1993). Toutefois, puisque ces modèles sont lourds et requièrent un nombre de sujets prohibitif, on voit se développer des applications basées sur des modèles à facettes (McNamarra, 1996) qui permettent de prendre en considération différents aspects d'une réponse et de tenir compte notamment du contexte (Laurier, 1997).

En second lieu, on voit se développer des techniques qui libèrent les évaluateurs des contraintes qu'imposent les questions à choix multiple (Tatsuoka, 1993 ; Davey & al., 1997). Étant donné l'importance que revêt le traitement des réponses ouvertes dans l'évaluation des apprentissages en langue seconde ou étrangère, l'application de ces techniques offre des perspectives intéressantes même si elle pose encore des problèmes considérables.

En troisième lieu, on tente de rendre la prise d'information la plus fréquente possible de façon à ce que l'élève reçoive régulièrement une information pertinente sur ses apprentissages. L'évaluation est alors en interaction avec l'intervention pédagogique. Une prise d'information fréquente permet également de rendre compte de la maîtrise d'habiletés complexes qui mettent du temps à se développer et dont l'évolution peut se décrire en terme de niveaux de performance. C'est pourquoi Bunderson & al. (1989) parlent de mesure continue. Pour ces chercheurs, la mesure continue permet l'intégration de l'évaluation à l'environnement pédagogique. Dans la construction de didacticiels, cette caractéristique est fondamentale puisque l'évaluation devient un mécanisme de régulation sur lequel repose une partie importante de l'interaction avec le système. Toutefois, malgré l'intérêt que présente la mise au point de dispositifs de mesure continue, il faut convenir qu'on trouve peu d'exemples en langue seconde ou en langue étrangère.

1.4. La mesure intelligente

Bunderson & al. (1989), entrevoient l'avènement d'une quatrième génération où le lien avec l'intervention pédagogique se fait non seulement en tenant compte des facettes multiples de l'objet et du contexte, mais en simulant le jugement du maître par une analyse approfondie de la réponse. Cette quatrième génération s'inscrit dans la foulée des travaux sur la modélisation (McCalla & al., 1992 ; Shute, 1992). En langue seconde, les travaux de Bull (1994) sur la modélisation axée sur la métacognition offrent des perspectives intéressantes en vue de la mise au point de mesures intelligentes. Pour Brown (1997), il ne fait pas de doute que cette quatrième génération ne peut compter sur les seules contributions du domaine de la mesure, mais doit aussi faire appel à des techniques mises au point dans le cadre de recherches en intelligence artificielle. L'apport de celle-ci dans l'évaluation de la compétence langagière est d'autant plus important que les instruments de cette génération devront permettre l'analyse de réponses élaborées et devront par conséquent inclure des procédures de traitement des langues naturelles. L'évaluation sera alors totalement intégrée à un environnement pédagogique hautement interactif où elle jouera un rôle fondamental dans l'adaptation du système aux besoins de l'apprenant.

Dans le cas de l'évaluation en langue seconde ou étrangère, on pourrait ajouter que les instruments de cette génération devraient permettre de mieux satisfaire les critères d'authenticité et d'interactivité qui caractérisent les tests de langue (Bachman & Palmer, 1996). La création d'univers virtuels qui reproduisent des situations de communication authentiques permettrait de satisfaire ces exigences.

2. LES COMPOSANTES DE LA COMPÉTENCE LANGAGIÈRE

Un défi important pour la mise au point de moyens d'évaluation dans des contextes d'apprentissage des langues assistés par des environnements informatiques multimédias est de tenir compte de la complexité de l'objet àmesurer, à savoir la compétence langagière. Il est maintenant reconnu que cette compétence est multidimensionnelle. Dans la foulée des travaux de Canale et Swain (1980) et de Moirand (1990), Bachman (1990) distingue quatre composantes qui suppose le développement de plusieurs habiletés. Selon Bachman, la compétence langagière comporte tout d'abord une *composante linguistique* qui est reliée au degré de maîtrise des règles de la syntaxe et de la morphologie, à la qualité de la prononciation tant au niveau phonologique que prosodique, à la maîtrise des conventions graphiques de même qu'à l'étendue et la précision du vocabulaire. En second lieu, la compétence langagière comporte une *composante organisationnelle* qui est associée à des habiletés de discours visant à assurer la cohésion et l'organisation rhétorique. Troisièmement, on retrouve une *composante illocutionnaire* qui se développe avec la capacité à exécuter différentes fonctions langagières : exprimer des idées (fonction idéationnelle), faire faire (fonction manipulative), utiliser la langue comme outil de la pensée (fonction heuristique), créer avec la langue (fonction imaginative). Enfin, la compétence langagière comporte une *composante sociolinguistique* qui résulte d'une connaissance des références culturelles des locuteurs de la langue cible ainsi que d'une sensibilité aux variations dialectales, aux variations de registre et à l'aspect plus ou moins naturel de l'expression. Par ailleurs, il se greffe à la compétence langagière une compétence métacognitive qui permet à l'apprenant de planifier et d'auto-évaluer sa performance en vue de rendre la communication plus efficace et d'améliorer ses apprentissages.

Bachman et Palmer (1996) suggèrent de vérifier de façon systématique les caractéristiques des instruments afin de vérifier si les composantes qu'il faut évaluer sont effectivement représentées. L'importance que l'on doit accorder aux différentes composantes et habiletés dépend en grande partie du type de décision à prendre. La décision est fondée sur une inférence à partir d'une performance mesurable ou observable relativement à une compétence (ou une composante de cette compétence) sous-jacente. Ainsi, le processus d'évaluation suppose :

- que l'information recueillie dans un environnement multimédia puisse être représentée de façon à prendre en compte les éléments pertinents de la performance ;
- qu'on fournisse au système des procédures qui permettront d'inférer, à partir de l'information obtenue, les caractéristiques essentielles de la compétence.

En pratique, ces deux contraintes impliquent que les limites du système, tant sur le plan technologique que sur le plan théorique (les modèles de compétences et les modèles psychométriques) auront un effet sur la qualité de l'évaluation. Les environnements informatiques multimédias se prêtent bien à l'évaluation de certaines habiletés. On trouve par exemple plusieurs tests qui mesurent la maîtrise des mécanismes grammaticaux ou la capacité à comprendre de courts documents sonores. Par contre, dans l'état actuel des choses, il est clair que ces environnements sont beaucoup moins efficaces avec d'autres habiletés. Il suffit de penser aux problèmes importants que soulève l'évaluation des habiletés discursives associées à la composante organisationnelle, autant en production écrite qu'orale, ou aux problèmes tout aussi importants que soulève l'évaluation des habiletés nécessaires à interaction orale en petit groupe (Lhote et al., ce numéro).

3. LA NATURE DE LA DÉCISION

Nous avons précédemment défini l'évaluation comme une inférence faite àpartir d'une performance pour induire l'état d'une compétence sous-jacente en vue d'une prise de décision. La démarche évaluative n'a de sens que lorsqu'elle conduit à une décision, laquelle se traduit par une intervention pédagogique ou administrative. Dans cette perspective, on peut distinguer les fonctions suivantes :

- l'évaluation à des fins de sélection,
- l'évaluation à des fins de classement,
- l'évaluation certificative,
- l'évaluation formative,
- l'évaluation de rendement.

Dans cette section, nous décrirons chacune de ces décisions et nous examinerons jusqu'à quel point un environnement d'évaluation informatique multimédia est susceptible de contribuer à la qualité d'un processus décisionnel.

3.1. L'évaluation à des fins de sélection

Ce type d'évaluation s'utilise dans des situations où il s'agit d'identifier les apprenants dont la compétence est supérieure. Ces situations se trouvent dans beaucoup de systèmes éducatifs qui comportent des programmes auxquels on accède par voie de concours : par exemple, admission dans un établissement privé, entrée dans un programme d'études avancées qui peut être contingenté, obtention de bourses d'études etc. Dans le cas de l'évaluation en langue seconde ou étrangère, ces situations ne sont pas les plus fréquentes et se limitent habituellement à une section d'un test de connaissance générale que peuvent utiliser certains organismes pour choisir les meilleurs élèves. Ce type de tests implique habituellement une interprétation normative des résultats puisque les apprenants sont ordonnés selon leur rang et que l'opération consiste essentiellement à retenir les résultats supérieurs. Les avantages qu'on peut trouver à recourir à des environnements multimédias sont surtout d'ordre pratique, d'autant plus que les tâches à correction objective qu'on retrouve souvent dans ces tests se prêtent facilement à une informatisation. Les organismes qui recourent à de tels tests pourraient aussi voir un avantage pratique au fait qu'on peut éviter d'inutiles déplacements par une administration à distance qui exploitent la Toile.

L'informatisation simplifie donc l'administration, la correction et la gestion des résultats. Dans certains cas, il peut être utile de concevoir des instruments de deuxième génération dans lesquels des tâches préalablement calibrées seraient caractérisées par un indice de difficulté. Un algorithme relativement simple de correction permettrait d'interrompre l'administration lorsque le taux de réussite aux tâches tombe sous un certain seuil. On évite ainsi à des apprenants de se soumettre à des tâches dont la difficulté dépasse largement leur niveau de compétence.

3.2. L'évaluation à des fins de classement

La fonction de classement est particulièrement importante dans le déroulement de programmes d'enseignement d'une langue seconde ou étrangère. La plupart des établissements se trouvent confrontés à la difficulté d'assigner à chaque élève le groupe qui correspond le mieux à son niveau. Le classement se réalise en plaçant les apprenants sur un continuum qui correspond à la progression établie par le programme d'études de façon à les distribuer entre les groupes. En apprentissage autonome, il est important de déterminer le meilleur point d'entrée. Quelques rares établissements qui reçoivent un grand nombre d'apprenants et qui disposent d'une infrastructure organisationnelle appropriée peuvent bénéficier d'une évaluation par profil qui permet de placer les apprenants à des niveaux différents selon l'habileté, la composante ou le savoir-faire : par exemple, un apprenant sera au niveau débutant pour ce qui est de la prononciation mais au niveau intermédiaire pour ce qui est de la lecture. Des systèmes d'auto-apprentissage raffinés peuvent aussi tirer bénéfice d'une telle information.

Dans la majorité des situations de formation, la fonction de classement est réductrice puisqu'il s'agit de placer les apprenants sur un continuum pour ensuite les répartir entre les niveaux que comporte le programme. À condition que le contenu soit valide (c'est-à-dire conforme aux orientations et aux objectifs du programme), une interprétation normative est acceptable. Les tests de classement conventionnels doivent donc comporter des tâches de plusieurs niveaux pour que le classement soit aussi fiable pour les candidats tout à fait débutants que pour les plus avancés. Les tests adaptatifs présentent un avantage considérable sur les tests de classement conventionnels du fait qu'ils permettent de limiter l'épreuve aux tâches dont la difficulté convient à l'apprenant. Nous verrons dans la section suivante comment se déroulent de tels tests. Par ailleurs, il faut souligner qu'à l'instar des évaluations de sélection, le recours à des environnements informatiques multimédias offre une solution attrayante aux problèmes que pose la gestion d'une opération de classement.

3.3. L'évaluation certificative

L'évaluation certificative joue un rôle important dans la plupart des programmes de formation linguistique. C'est elle qui sanctionne les apprentissages au terme de la formation. Bien qu'elle puisse avoir une fonction pédagogique lorsqu'elle oriente des apprentissages ultérieurs, l'évaluation certificative a un caractère essentiellement administratif. Elle confirme le niveau atteint à l'étudiant lui-même, à ses parents, aux responsables de programme, aux employeurs éventuels ou à toute autre personne susceptible de requérir une attestation du niveau. Lorsque la communication de l'évaluation est adéquate, ces personnes peuvent savoir dans quels types de situations de communication réelles l'apprenant pourra fonctionner convenablement. Étant donné que l'issue de cette évaluation peut avoir des conséquences majeures sur l'avenir des individus, les instruments doivent se conformer à des exigences métrologiques élevées afin de s'assurer que les inférences sur les compétences acquises sont justes. Outre les exigences classiques de validité et de fidélité, les tests de certification doivent présenter des situations communicatives complexes qui s'apparentent à celles de la vie réelle. On voit mal, par exemple, comment on pourrait faire l'économie d'une entrevue pour attester de la capacité à interagir dans des situations réelles.

À la différence des tests de classement qui doivent couvrir une gamme très large de niveaux, les tests de certification sont souvent des instruments dont le niveau est ciblé ; cela implique que les tâches sont souvent de difficulté semblable. De ce point de vue, l'utilisation de tests adaptatifs offre peu d'avantages. Il est possible que des tests de certification élaborés dans l'optique de la quatrième génération voient le jour mais leur élaboration reste difficile. Quelques travaux en vue de créer des tests de certification faisant appel à des environnements informatiques multimédias pourraient à court terme se traduire par des tests de certification opérationnels et utilisables dans des contextes spécifiques. Une version informatisée du TOEFL est en préparation (Eignor, 1996) et une équipe de l'Université du Minnesota travaille à la mise au point d'un test de lecture (Chalhoub-Deville & al., 1996).

3.4. L'évaluation formative

L'évaluation formative joue un rôle déterminant dans le déroulement de l'apprentissage en langue seconde (Lussier, 1992). La pratique de l'évaluation formative est étroitement associée à l'enseignement au point où il est permis de douter que la distinction soit toujours réalisable (Bain & Schneuwly, 1993). Cette symbiose entre l'enseignement et l'évaluation est en accord avec les principes sous-jacents aux environnements de troisième génération. Ces environnements devraient compléter le travail du professeur de langue de façon à ce que celui-ci puisse obtenir une information continue sur l'atteinte des objectifs ou le développement des compétences. De la même manière, les didacticiels, que ce soit dans le cadre d'un apprentissage autonome ou comme support à l'enseignement, devraient intégrer des activités qui visent l'évaluation régulière des apprentissages.

Il importe de souligner que même si l'évaluation formative conduit à une décision d'une importance moindre que la certification, il est nécessaire de veiller à la qualité de ces mécanismes de façon à ce que les inférences soient justes. Il ne suffit pas de transformer un exercice de classe en lui associant un score pour que l'activité devienne un instrument pour l'évaluation formative. Ce n'est que si l'environnement informatisé apporte une dimension nouvelle au travail du professeur ou permet véritablement de réguler une situation d'apprentissage autonome qu'on pourra justifier la mise en place de tels mécanismes d'évaluation formative. À titre d'exemple, ces mécanismes pourraient fournir à l'élève, àintervalles réguliers, une fiche diagnostique décrivant ses forces et faiblesses dans l'exécution d'une tâche donnée ou en regard d'un nombre limité d'objectifs d'apprentissage.

3.5. L'évaluation de rendement

Ce type d'évaluation remplit à la fois une fonction administrative et une fonction pédagogique. L'évaluation de rendement se déroule habituellement à des moments prédéterminés. Elle joue un rôle important dans les systèmes scolaires et dans les organismes où il est important de dresser, à divers moments, un bilan des apprentissages. De par sa nature, l'évaluation de rendement doit s'aligner étroitement sur les contenus. Selon les cultures spécifiques à chaque milieu, l'interprétation peut se faire en fonction de l'atteinte des objectifs de contenus (interprétation critériée) ou en fonction du groupe (interprétation normative). Des considérations pratiques telles que la rapidité de la correction et le traitement administratif des résultats, peuvent justifier le recours à des environnements informatiques. L'utilisation des environnements informatiques peut être partielle ; par exemple, plusieurs organismes disposent de banques d'items informatisées qui peuvent servir au professeur dans la composition d'une épreuve traditionnelle papier-crayon. Dans le réseau scolaire, l'évaluation périodique sert notamment àinformer les parents du déroulement des apprentissages et à prendre des décisions sur la gestion de certains cas particuliers. L'évaluation de rendement est alors souvent associée à une note ; celle-ci ne rend d'ailleurs pas nécessairement compte des apprentissages réalisés dans la mesure où elle peut inclure des éléments de différents ordres (effort, participation, conduite, etc.). En apprentissage autonome, l'évaluation de rendement peut servir à marquer les étapes de l'apprentissage. Cependant, on peut douter de sa pertinence dans des

environnements informatiques multimédias où des dispositifs d'évaluation formative assurent la régulation des apprentissages.

Si l'on considère l'ensemble des fonctions de l'évaluation, il apparaît que les environnements informatiques multimédias se prêtent particulièrement bien à la fonction de classement et à la fonction formative. Dans le cas du classement, l'application de procédures de testing adaptatif s'avère tout à fait appropriée ce qui suppose des environnements de la deuxième génération ou même d'une génération supérieure. Dans le cas de l'évaluation formative, il faut s'orienter vers des techniques qui permettent le diagnostic et la régulation de l'apprentissage/enseignement. Il faut donc concevoir des systèmes qui analysent la réponse de l'élève, tiennent compte du contexte et interagissent avec la situation de formation, ce qui suppose alors au moins des environnements de la troisième génération.

4. UN EXEMPLE DE TEST ADAPTATIF

Afin de montrer en quoi les tests adaptatifs constituent une innovation par rapport aux environnements de première génération, nous présentons les caractéristiques de ce type de test. Nous nous référerons, à titre d'exemple, au *French CAPT*, un instrument mis au point à l'Université de Montréal (Laurier & Perron, 1996) pour le classement d'étudiants anglophones qui s'inscrivent à des cours de français dans des établissements post-secondaires au Canada.

Ainsi que nous l'avons déjà mentionné, l'élaboration d'un test adaptatif suppose d'abord la création d'une (ou de plusieurs) banque(s) d'items. Les items, habituellement des questions à choix de réponse, sont d'abord administrés de façon expérimentale à un grand nombre de sujets en vue d'une analyse. Cette analyse (la calibration) vise à établir les caractéristiques (les paramètres) de chaque item. Contrairement aux approches psychométriques classiques, la calibration permet de modéliser chaque item en fonction du niveau d'habileté du sujet. Au minimum, il faut estimer la difficulté de chaque item. Pour être suffisamment juste, cette estimation peut requérir une centaine de sujets. Toutefois, il est souvent nécessaire de recourir à des échantillons beaucoup plus grands lorsqu'on veut que l'estimation tienne compte de caractéristiques particulières des items. Ainsi, si les items ne départagent pas de la même manière les sujets plus avancés des sujets moins avancés, on aura intérêt à ajouter un paramètre de discrimination. S'il s'agit de questions à choix multiple, il peut s'avérer utile d'estimer un troisième paramètre pour tenir compte de l'effet du hasard. Une calibration avec un modèle à trois paramètres exige au moins 500 réponses par question ce qui fait de la collecte de données une opération d'envergure. La technique habituelle, qui est celle que nous avons appliquée dans le cas du *French CAPT*, consiste à expérimenter les items àpartir d'une version papier-crayon.

Le *French CAPT* étant un test de classement, il est normal qu'il comporte des items très faciles (à l'intention des étudiants débutants) et des items très difficiles (à l'intention des étudiants très avancés). Toutefois l'administration d'un item difficile à un débutant n'apporte que peu d'information puisque le résultat est prévisible. De plus, cette situation est une source de frustration chez l'apprenant. Inversement, il y a peu d'avantages à soumettre un item facile à un apprenant qui maîtrise bien la langue. Une fois les paramètres estimés, il devient possible, au cours de l'administration d'un test adaptatif, de sélectionner, à partir d'une banque, les items les plus pertinents, c'est-à-dire ceux qui ne sont ni trop faciles, ni trop difficiles. Ainsi, un test adaptatif présente à chaque apprenant une épreuve différente. Des procédures développées dans le cadre de la théorie des items permettent d'estimer le niveau de l'apprenant après chaque réponse et de calculer l'erreur de mesure. Le système choisit alors l'item qui est le plus approprié parmi les items non présentés et estime de nouveau le niveau en tenant compte de la nouvelle réponse. La boucle prend fin quand l'erreur atteint une proportion acceptable. À ce moment, le système ouvre un autre sous-test ou communique le résultat final.

Le *French CAPT* contient cinq sous-tests. Dans le premier sous-test, le sujet lit un paragraphe d'une trentaine de mots et doit répondre à une question de compréhension. Dans le second sous-test, le sujet lit la description d'une situation dans sa langue maternelle et doit choisir l'énoncé français qu'il juge le plus convenable en prenant en considération les aspects sémantiques et sociolinguistiques. Le troisième sous-test contient des phrases lacunaires classiques qui mesurent des aspects lexicaux et grammaticaux. Comme ces trois premiers sous-tests sont composés de questions à choix multiple, nous avons opté pour un modèle de calibration à trois paramètres. Les deux derniers sous-test portent sur l'oral. Dans le quatrième sous-test, l'apprenant écoute de trois à cinq courts passages semi-authentiques et répond à trois questions de compréhension sur chaque passage. Le cinquième sous-test fait appel à l'auto-évaluation : l'apprenant détermine lui-même dans quelle mesure (sur une échelle Likert) il est capable de faire face à différentes tâches langagières (résumées en un énoncé en langue maternelle). Les deux derniers sous-tests font appel à un modèle de calibration plus sophistiqué. Ce modèle, désigné comme « modèle à réponse gradué» (Samejima, 1978), permet de considérer chaque passage du quatrième sous-test comme un item complexe et, dans le cas du cinquième sous-test, de tenir compte du fait qu'une correction dichotomique est inapplicable.

5. LES PRINCIPES DE L'ÉVALUATION CONTINUE

Bien qu'il appartienne aux tests de seconde génération, le *French CAPT* se rapproche à bien des égards des instruments de la génération suivante. D'une part, il comporte plusieurs sous-tests (et donc plusieurs banques) qui mesurent des aspects qui se rapportent à différentes composantes de la compétence langagière. Ainsi, le résultat final se présente non seulement comme un niveau parmi les treize niveaux que distingue le test, mais aussi comme un profil établi selon les résultats à chacun des sous-tests. D'autre part, le test utilise des modèles psychométriques plus raffinés que le modèle à un ou trois paramètres. Par contre, le *French CAPT* reste un instrument de deuxième génération puisqu'il n'est pas intégré à une intervention pédagogique, ni en salle de classe, ni à travers l'utilisation d'un didacticiel. Sa fonction de test de classement ne s'y prête guère d'ailleurs. De plus, le critère de sélection des tâches étant essentiellement le niveau estimé de l'apprenant, le *French CAPT* ne tient pas compte d'éléments contextuels, contrairement aux instruments de troisième génération. On peut cependant penser que des systèmes basés sur des techniques avancées de testing adaptatif qui mesurent différentes dimensions de la compétence langagière et qui sont complétées par des diagnostics construits à partir d'un ensemble de règles dans le but de traiter les erreurs des apprenants seront de plus en plus intégrés aux environnements multimédias pour l'apprentissage des langues. Ces systèmes s'apparenteront de plus en plus aux tests de troisième génération.

Par ailleurs, le passage vers des instruments de troisième génération ne va pas sans une remise en cause de la théorie de réponses aux items qui a servi de fondement aux tests adaptatifs. La théorie impose une collecte de données onéreuse qui ne convient guère à des situations changeantes d'évaluation formative où l'évaluation sert de moyen de régulation aux apprentissages. Le problème est d'autant plus important que les modèles qui conviennent à des réponses plus complexes, requièrent souvent un nombre de sujets expérimentaux inaccessible. Plusieurs chercheurs ont donc été amenés à prendre leurs distances face aux approches psychométriques habituelles. Dans cette recherche de nouvelles techniques, depuis quelques années, les spécialistes de la mesure s'attaquent à des problèmes qui avaient jusqu'à récemment surtout préoccupé les chercheurs en sciences cognitives et en intelligence artificielle. L'apport des théories de la mesure permettra de résoudre certains problèmes reliés à la modélisation de l'apprenant. Il s'agit ici de construire une représentation de l'apprenant à partir de ses forces et de ses faiblesses. On voit se mettre en place des systèmes basés sur des modèles probabilistes ou des moteurs d'inférence (Mislevy, 1995 ; Corbett & al., 1995), d'autres qui tentent de reconstruire les représentations de ces derniers (Tatsuoka, 1993 ; Trentin, 1997). Bien que pour l'instant, la plupart des applications se limitent à des apprentissages simples, tous ces travaux offrent des perspectives intéressantes face à un apprentissage aussi complexe que celui d'une langue seconde ou étrangère.

Michel LAURIER
Université de Montréal
Faculté des sciences de l'éducation
C. P. 6128, succ. Centre-ville
Montréal (Québec), H3C 3J7, Canada
Mél : laurierm@scedu.umontreal.ca

Notice biographique

Michel Laurier a obtenu un Ph.D. de l'Institut d'études pédagogiques de l'Ontario (Toronto). Après avoir enseigné pendant plusieurs années le français comme langue seconde, il s'est specialisé dans l'évaluation de la compétence en langue seconde. Il s'intéresse particulièrement à l'informatisation de l'évaluation et a publié plusieurs articles relativement à l'application des principes de test adaptatif en langue seconde. Il est maintenant professeur agrégé a l'Université de Montréal dans la section de mesure et évaluation. Il fait partie du GRAEMI, un groupe de recherche sur les systèmes multimédias interactifs.

REFERENCES BIBLIOGRAPHIQUES

BACHMAN, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford : Oxford University Press.

BACHMAN, L. F. & PALMER, A. S. (1996). *Language Testing in Practice*. Oxford : Oxford University Press.BAKER, F. B. (1992). *Item Response Theory : Parameter Estimation Techniques.* New York : Marcel Dekker.

BAIN, D. & SCHNEUWLY, B. (1993). "Pour une évaluation formative intégrée dans la pédagogie du français: de la nécessité et de l'utilité de modèles de référence". In *Évaluation formative et didactique du français,* Allal L., Bain D. et Perrenoud P.(dir.). Neuchâtel : Delachaux et Niestlé. pp 51-80.

BINET, A. (1909). *Les idées modernes sur les enfants*. Paris : Flammarion.

BROWN, J. D. (1997). “Computers in Language Testing : Present Research and some Future Directions”. *Language Learning & Technology,* vol. 1, 1. (http://polyglot.cal.msu.edu/llt/) pp 44-59.

BULL, S. (1994). “Student Modelling for Second Language Acquisition”. *Computers and Education,* vol. 23, 1-2, pp 13-20.

BUNDERSON, C. V., INOUYE, D. K. & OLSEN, J. B. (1989). “The Four Generations of Computerized Testing”. In, *Educational Measurement*, 3e édition, Linn R. L. (dir.). New York : American Council on Education / Macmillan. pp 367-408.

CANALE, M. & SWAIN, M. (1980). “Theoritical Bases for Communicative Approaches to Second Language”. *Applied Linguistics,* vol. 1, 1, pp 1-47.

CHALHOUB-DEVILLE, M., ALCAYA C. & McCOLLUM LOZIER, V. (1996). *An Operational Framework for Constructing a Computer-Adaptive Test of L2 Reading Ability: Theoritical and Practical Issues.* (Document non publié), Minneapolis, MN : The Center for Advanced Research on Language Acquisition, University of Minnesota. CORBETT, A. T., ANDERSON, J. R. ET O’BRIEN, A. T. (1995). “Student Modeling in the ACT Programing Tutor”. In *Cognitive Diagnostic Assessment,* Nichols P.D., Chipman S. F. & Brennan R. L. (dir.). Tillsdale, NJ : Lawrence Erlbaum, pp 19-45.

DAVEY, T, GODWIN, J. & MITTELHOLTZ, D. (1997). “Developing and Scoring an Innovative Computerized Writing Assessment”. *Journal of Educational Measurement,* vol. 34, 1, pp 21-41.

EIGNOR, D. R. (1996). “Adaptive Assessment of Reading Comprehension for TOEFL”. Communication au symposium *Issues in Computer-Adaptive Testing of Second Language Reading Proficiency*. Bloomington, MN, 20-22 mars.

KAYA-CARTON, E., CARTON, A. S. & DANDONELLI, P. (1991). “Developing a Computer-AdaptiveTest of French Reading Proficiency”. In *Computer-Assisted Language Learning and Testing,* P. Dunkel (dir.). New York : Newbury House, pp 259-284.

KINGSBURY, G. C.& D. J. WEISS (1980). *An Alternative-Forms Reliability and Concurrent Validity Comparison of Bayesian Adaptive and Conventional Ability Tests*, Research Report 80-5. Minneapolis, MN: University of Minnesota, Dept of Psychology.

LARSEN, J. W. & H.S. MADSEN (1985). “Computerized Adaptive Testing: Moving Beyond Computer-Assisted Testing”. *CALICO Journal,* vol. 3, 3. pp 23-36, 43.

LAURIER, M. (1996). “Using the Information Curve to Assess Language CAT Efficiency”. In *Validation in Language Testing*, A. Cumming et R. Berwick (dir.). Clevedon, UK : Multilingual Matters. pp111-123.

LAURIER, M. (1997). “Pour un diagnostic informatisé en révision de texte”. *Mesure et évaluation en éducation,* vol 18, 3. pp 85-106.

LAURIER, M. & PERRON, M. (1996). “Un test adaptatif pour le classement des élèves dans un cours de langue”. In *La technologie éducative en réseau : réseaux technologiques, réseaux humains,* L. Sauvé & al. (dir.). Sainte-Foy, QC : Télé-Université / CIPTE. pp 279-286.

LHOTE, E., ABECASSIS, L & AMRANI A. (ce numéro) “Apprentissage de l’oral et environnement informatique”.

LUSSIER, D. (1992). *Évaluer les apprentissages dans une approche communicative.* Paris : Hachette.

McCALLA, G., GREER, J. & al. (1992). “Special Issue on Student Modeling : Editor’s introduction”. *Journal of Artificial Intelligence in Education*, vol. 3, 4. pp 377-380.

McNAMARRA, T. (1996). *Second Language Performance Measurement.* London : Longman.

MISLEVY, R. J. (1995). “Probability-Based Inference in Cognitive Diagnosis”. In *Cognitive Diagnostic Assessment*, Nichols P.D., Chipman S. F. & Brennan R. L. (dir.). Hillsdale, NJ : Lawrence Erlbaum. pp 46-71.

MOIRAND, S. (1990). *Enseigner à communiquer en langue étrangère.* Paris : Hachette.

SHUTE, V. J. (1992). “Aptitude-Treatment Interactions and Cognitive Skill Diagnosis”. In *Cognitive Approches to Automated Instruction*, Regian, J. W. & Shute, V. J. (dir.). Hillsdale, NJ : Lawrence Erlbaum. pp 15-48.

SAMEJIMA, F. (1978). “The Application of Graded-Response Models : The Promise of the Future”. In *Proceedings of the 1977 Conference on Computerized Adaptive Testing*, . J. Weiss , D.Jj. (dir.). pp 28-37.

TATSUOKA, K. K. (1993). “Construction versus Choice in Cognitive Measurement”. In *Item Construction and Psychometric Models Appropriate for Constructed Responses*, Bennett R. E. & Ward W. C. (dir.).. Hillsdale, NJ : Lawrence Erlbaum. pp 107-133.

TRENTIN, G. (1997). “Computerized Adaptive Tests and Formative Assessment”. Journal of Educational Multimedia and Hypermedia, vol 6, 2, pp 201-220.