Tricot, A., & Lafontaine, J. (2002). Une méthode pour évaluer ensemble l'utilisation un outil multimédia et l'apprentissage réalisé avec celui-ci. *Le Français dans le Monde*, numéro spécial, Janvier, 41-52.

Une méthode pour évaluer conjointement l'utilisation un outil multimédia et l'apprentissage réalisé avec celui-ci*

André Tricot
CERFI, IUFM de Midi-Pyrénées
56 avenue de l'URSS
31 078 Toulouse cedex
chercheur associé au Laboratoire Travail et Cognition, UMR CNRS et Université Toulouse 2
andre.tricot@toulouse.iufm.fr

Jeanine Lafontaine
Groupe recherche 82 - CDDP Tarn et Garonne
Boulevard Montauriol
82 000 Montauban cedex

Introduction

Considérons un usage comme un ensemble de d'actions réalisées, plus ou moins automatiquement et fréquemment, avec un outil considéré par un usager (ou un groupe d'usagers) comme utile et utilisable pour un ensemble de buts dans un ensemble d'environnements.

Evaluer l'impact des technologies de l'information et de la communication sur les apprentissages pose un problème délicat, celui du lien entre l'utilisation de l'outil et l'apprentissage réalisé (l'utilité supposée de l'outil). Il est en effet souvent difficile de dire si tel apprentissage est réalisé « grâce à » tel cédérom, ou si tel apprentissage a échoué «à cause » de tel site Web utilisé par l'apprenant. Il est surtout très difficile de dire si tel

* Cet article reprend, dans sa seconde partie, un chapitre d'André Tricot à paraître dans l'ouvrage coordonné par Guy Boy : *L'Ingénierie Cognitive : IHM et Cognition*, aux éditions Hermès Science. Le lecteur y trouvera une description plus détaillée des méthodes évoquées ici. La partie expérimentale rapportée dans cet article a été réalisée grâce à Janine Igual, Emmanuel Fol et Etienne Poussou, et aux élèves des écoles de Chabrié Moissac, Piquecos et Caylus du département de Tarn et Garonne. Nous les remercions tous pour leur participation.

apprentissage eut été réalisé «grâce à» une meilleure utilisation de tel cédérom, ou si tel apprentissage eut été moins bon avec une utilisation moins optimale de tel site Web. En bref, si l'on dispose de méthodes et de résultats en évaluation, de l'utilisation d'une part et de l'apprentissage d'autre part, on ne dispose pas de cadre d'interprétation des liens entre les deux. Ce problème dépasse de loin le domaine du multimédia : il provient de l'inexistence (à notre connaissance) d'un cadre général, correctement formalisé, d'interprétation des liens logiques entre des moyens (ici les outils multimédia) et des buts (ici les apprentissages).

L'objectif de cet article est de rappeler les méthodes d'évaluation de l'utilisation d'une part et de l'apprentissage d'autre part, puis de montrer comment on peut interpréter les liens entre ces évaluations. Un exemple d'application de ce cadre interprétatif est donné : il s'agit d'une évaluation de l'utilisation par des élèves de cycle III (CE2, CM1, CM2) des versions électronique et papier d'un dictionnaire encyclopédique.

L'évaluation de l'utilisation et des apprentissages : de l'approche séparée à l'approche conjointe

Le fait que l'évaluation des outils multimédia pour l'apprentissage soit difficile semble unanimement reconnu. Cela conduit certains, par excès de pessimisme, à considérer que ce type d'évaluation est impossible, que nous ne savons rien sur l'efficacité du multimédia, etc. Dans cette partie, nous voudrions rappeler qu'il existe des méthodes pour évaluer l'utilisation et les apprentissages. Nous montrons ensuite comment interpréter les liens entre ces évaluations.

L'évaluation de l'utilisation d'un outil informatique

L'évaluation de l'utilisation d'un outil informatique est réalisée par deux grands types de méthodes : d'une part, l'application de critères d'utilisabilité ou de critères ergonomiques, qui permettent de diagnostiquer en quoi l'utilisation peut être améliorée par une amélioration de l'outil ; d'autre part, l'analyse des protocoles d'interaction, qui permet d'interpréter, pour une part, le comportement de l'utilisateur.

L'évaluation de l'utilisabilité des applications informatiques est un enjeu important depuis quelques années. Depuis la publication d'un ouvrage de référence par Nielsen (1993) les

publications se comptent en milliers. Ce thème de recherche et d'ingénierie ergonomique concerne particulièrement les sites Web (Buckingham Shum & McKinght, 1997), domaine où la quantité de mauvais outils est phénoménale. Le succès de la notion d'utilisabilité est tel que tout se passe comme si l'utilisabilité devenait progressivement synonyme de qualité générale d'une application informatique.

Les cinq critères d'utilisabilité proposés par Nielsen (op. cit.) intègrent : l'efficience (c'est le fait d'atteindre sans perdre trop de temps le but recherché) ; l'apprenabilité (c'est la facilité ou la rapidité avec laquelle l'utilisateur apprend à utiliser l'outil) ; la mémorisation (c'est le fait que l'utilisateur parvienne à mémoriser « comment ça marche » et plus généralement « ce qu'il a fait ») ; la fiabilité (c'est la prévention ou la gestion des erreurs par le système) ; la satisfaction de l'utilisateur.

Les critères ergonomiques de Scapin et Bastien (1997) sont un bon exemple de définition rigoureuse de critères d'évaluation ergonomique des systèmes d'information. Nous en recommandons vivement la lecture et l'utilisation. Parmi les critères de Scapin et Bastien, l'adaptabilité est assez proche d'une certaine acception de l'utilité. L'adaptabilité est strictement dépendante des buts de l'utilisateur, alors que l'utilité telle que nous la concevons est l'adéquation entre les buts de l'utilisateur et la finalité d'une application informatique (ou d'un logiciel).

Ces méthodes d'évaluation des interactions homme-machine (IHM), fondées sur l'application de critères, ne sont pas spécifiques au domaine des apprentissages, mais utilisables quand on sait bien définir leur domaine d'application. Il n'est pas le lieu d'en discuter davantage.

Une seconde catégorie de méthodes consiste à analyser les protocoles d'interaction à l'aide de critères issus des sciences de l'information, du domaine des bases de données en particulier. Le principe de ces méthodes est de mesurer l'efficacité de l'interaction en fonction du but poursuivi par l'utilisateur. On peut mesurer l'efficience plutôt que l'efficacité, en pondérant les mesures par le temps passé à la réalisation de la tâche. Ces deux mesures ont été utilisées par Chen et Rada (1996) qui ont conduit une méta analyse de la littérature empirique sur l'interaction utilisateur / hypertexte. Pour effectuer ce type de mesure, il suffit de considérer le but recherché par l'utilisateur comme un ensemble d'items pertinents (cibles) dans le système. Les principales mesures sont :

- le taux de rappel = nombre de cibles atteintes / nombre de cibles existantes ;
- le taux de précision = nombre de cibles atteintes / nombre d'items ouverts ;
- l'économie = 1 – (nombre d'items différents ouverts / nombre total d'items ouverts).

Ces critères importés sont intéressants, mais leur interprétation pose problème. Une polémique a eu lieu à ce sujet (Rouet, 1990 ; Bernstein, 1991, 1993) certains pensant, comme Foss (1989) qu'un parcours non économique correspond à des difficultés rencontrées par l'utilisateur, tandis que d'autres pensent au contraire qu'un parcours non économique peut correspondre à une utilisation intéressée, voire approfondie, du système. Dans une recherche expérimentale, Tricot et Coste (1995) ont montré que le critère d'économie peut être corrélé parfois positivement, parfois négativement, à la réussite d'une tâche. Un groupe de sujets à qui on proposait de résoudre une série de problèmes complexes (impliquant le traitement de nombreuses informations) avait une allure paradoxale : les sujets aux parcours les moins économiques trouvaient plus de solutions correctes aux problèmes posés… que les sujets les plus économiques dans leurs parcours. Le lecteur pourra se rapporter à la synthèse que nous avons publiée à propos du domaine de validité de ces critères (Tricot et al. 1999)

Puisque la mesure de l'efficacité ou de l'efficience des parcours implique une connaissance précise de la signification de l'efficacité des parcours (connaissance non disponible à l'heure actuelle), on peut vouloir simplement décrire les parcours en fonction de questions que l'on se pose. Dans une des recherches mentionnées ci-dessus (Tricot & Coste, 1995), nous nous sommes demandés quelles relations les utilisateurs établissent entre les contenus et les fonctionnalités d'accès à ces contenus (les menus). Nous nous sommes aussi demandés comment les sujets passent d'un domaine de contenu à l'autre, la tâche prescrite impliquant ces passages. Nous avons donc distingué les nœuds de contenu et les nœuds de menu. Les nœuds de menu donnent des informations sur les relations entre les nœuds de contenus, à un niveau global (organisation générale du système) ou local (relation entre deux nœuds). Les nœuds de contenu peuvent faire partie d'un même thème ou non (il n'y a aucune restriction concernant l'ajout de catégories de nœuds différentes). Nous notons :

- αi un nœud de menu,
- ai, aj deux nœuds de contenu d'un même thème,
- ai, bi deux nœuds de contenu de thèmes différents.

Dans notre notation, il y a donc quatre types de déplacements possibles entre deux nœuds. Chacun des quatre déplacements possibles s'écrit dans les deux sens :

(a b) = - (b a) :

L (largeur) = ai bi

O (orientation) = αi αj

S (surface) = αi ai

T (thème) = ai aj

Nous définissons une séquence [x y(n)] comme une suite de déplacements (x y) comprenant n éléments. Par exemple, [ai αi aj αj] s'écrit [a α (4)] soit [S4]. Nous appliquons la règle suivante : «si deux écritures d'une même séquence sont possibles, alors choisir celle où (n) est le plus grand ».

Ce formalisme nous a par exemple permis de montrer que lors de la phase de découverte d'un nouvel outil multimédia complexe, les utilisateurs ont tendance à faire des parcours de type S (une alternance de consultation de nœuds de menu et de nœuds de contenu), et non pas, comme on aurait pu s'y attendre, d'abord des parcours O (l'utilisateur apprend à se servir de l'outils) puis des parcours L ou T (l'utilisateur se sert de l'outil : il traite des contenus).

Nous allons maintenant aborder l'évaluation des apprentissages, en dépassant partiellement les environnements informatiques.

L'évaluation des apprentissages

Les méthodes d'évaluation des apprentissages constituent un champ très important d'activités, commun à plusieurs disciplines : psychologie, didactique, éducation, formation, etc. Le principe général est assez simple, pour ce qui concerne l'approche expérimentale de la question : prescrire une tâche en rapport avec l'apprentissage visé pour évaluer l'état des connaissances de l'apprenant, lui faire apprendre ce que l'on veut lui faire apprendre, lui prescrire une seconde tâche, analogue à la première. S'il y a une différence positive de performance entre les deux tâches, alors on considère que l'apprenant a appris quelque chose. On prend généralement la précaution d'utiliser un groupe contrôle d'apprenants qui font « exactement la même chose », sauf ce qui est évalué. Donc, dans le cas où l'on veut évaluer

l'efficacité d'un outil multimédia pédagogique, le protocole expérimental consiste à faire apprendre quelque chose par le biais de l'outil au premier groupe et à faire apprendre la même chose sans cet outil au second groupe. Si l'on veut que cette évaluation ait un peu de sens, on prend la précaution de présenter au second groupe les mêmes contenus, la même démarche, le même temps, le même environnement, la même consigne, etc. L'évaluation initiale des connaissances de l'apprenant peut poser un certain nombre de problèmes (effet d'attente, anticipation, reconnaissance de l'évaluation a posteriori, etc.) qui peuvent détourner les résultats. On renonce donc parfois à évaluer l'état initial des connaissances, en étant particulièrement vigilant sur la constitution des deux groupes : le groupe expérimental et le groupe contrôle doivent pouvoir être considérés comme équivalents au niveau de leurs connaissances.

Dans une publication récente, Rouet et Passerault (1999) ont présenté et discuté les principales variables importées de la psychologie cognitive vers l'analyse des apprentissages dans les environnements informatisés. Ils abordent en particulier le problème du « grain » de l'évaluation (le degré de précision optimal de la mesure en fonction du but d'évaluation recherché). Le lecteur trouvera dans cet article et dans le volume dont il est issu (Rouet et al., 1999) des informations plus complètes sur ce sujet. Nous présentons ici les principales catégories de tâches utilisées en évaluation des apprentissages.

Les tâches de reconnaissance : il s'agit de demander à l'apprenant si un élément (un mot, une phrase, une image, etc.) était présent ou pas dans le matériel présenté lors de la phase d'apprentissage. Ce type de tâche s'utilise généralement sans évaluation préalable. On peut faire varier le degré de ressemblance entre l'item proposé et le matériel traité lors de l'apprentissage. Par exemple, dans les tâches de reconnaissance de phrases, il est devenu assez commun depuis une publication de Kintsch et ses collaborateurs (1990) de proposer des phrases identiques aux phrases présentées, des phrases différentes, mais aussi des paraphrases et des propositions absentes du matériel présenté mais « inférables » à partir de celui-ci. Ces tâches sont faciles à traiter, mais leur domaine de pertinence est restreint aux apprentissages fondés sur la compréhension littérale et intégrale d'un contenu défini.

Les tâches de rappel du contenu : on demande à l'apprenant de dire, d'écrire, de dessiner, …, ce qu'il a retenu du matériel présenté. Ces tâches de rappel peuvent être indicées, c'est-à-dire que l'on peut proposer au sujet des mots, des phrases, des images, … *i.e.* des indices pour

« lancer » sa mémoire. Bien que limitées elles aussi au domaine de la compréhension, ces tâches sont souvent plus intéressantes, car elles fournissent des données sur les relations entre les connaissances construites par le sujet. Il est souvent très intéressant de comparer ce matériel à celui produit par l'apprenant en production (libre ou indicée) lors d'une évaluation préalable de ses connaissances dans le domaine de contenu abordé. Le matériel verbal ou imagé produit par le sujet est, en revanche, souvent très délicat à interpréter. Les techniques comme l'analyse propositionnelle sont restreintes aux petits corpus verbaux (quelques centaines de mots). Des techniques récentes comme l'indexage sémantique latent (Landauer & Dumais, 1997) permettent d'envisager de façon assez intéressante le traitement du matériel verbal, y compris quand celui-ci est important en volume.

Les tâches de rappel de la structure : on demande à l'apprenant non pas de rappeler le contenu, mais l'organisation de celui-ci. Le plus souvent, on propose au sujet de dessiner cette structure, pour la comparer ensuite à la structure « réelle » (celle qui était présentée). Bien que fréquemment utilisées, ces tâches n'ont qu'un intérêt indirect : on sait bien peu de choses sur les liens entre une telle structure produite par le sujet et l'utilisation ultérieure des connaissances par celui-ci.

Les questionnaires fermés ou ouverts sont fréquemment utilisés pour évaluer un apprentissage. Avec quelques précautions, ils sont particulièrement bien adaptés aux protocoles utilisant l'évaluation initiale et l'évaluation finale. Comme les tâches précédentes, les questionnaires concernent surtout les connaissances déclaratives (factuelles et, surtout pour les questionnaires ouverts, conceptuelles). On est parfois tenté d'utiliser un questionnaire pour évaluer l'acquisition d'un savoir-faire, alors qu'il y a un décalage fondamental entre l'aspect déclaratif et explicite des réponses fournies à un questionnaire et l'aspect procédural parfois implicite d'un savoir-faire. Les questionnaires fermés présentent l'avantage d'être faciles à traiter, tandis que les questionnaires ouverts, plus difficiles à traiter, permettent au sujet d'exprimer plus librement ses connaissances.

Les tâches de résolution de problème sont les plus communément utilisées pour évaluer un apprentissage. Elles présentent l'avantage de se prêter à l'évaluation de connaissances procédurales mais aussi déclaratives, aux protocoles d'évaluation initiale / finale comme aux protocoles d'évaluation finale seule. Comme beaucoup d'évaluations de connaissances, elles sont limitées aux diagnostics positifs : si un sujet humain résout un problème qui implique

l'utilisation de la connaissance A alors on peut conclure qu'il sait utiliser la connaissance A ; si, en revanche, il ne résout pas ce problème, on ne peut rien conclure sur sa connaissance A (ou sa non connaissance A).

Les tâches de détection d'erreurs consistent à présenter un matériel (une explication, une démonstration, un schéma fonctionnel par exemple) et à demander à l'apprenant si ce matériel est correct ou non, voire à identifier l'erreur éventuellement présente dans le matériel. Ce type de tâche présente un intérêt certain, elle est particulièrement sensible aux effets d'expertise des sujets et permet souvent de distinguer les aspects superficiels de la compréhension des aspects profonds.

Les tâches de production consistent à demander à l'apprenant de réaliser un artefact : texte, image, objet technique, application informatique, etc. Ces tâches sont, dans l'idéal, les plus intéressantes : elles permettent d'évaluer des aspects très divers des connaissances (déclaratives, procédurales, méthodologiques, etc.). Ces tâches sont aussi les plus difficiles à évaluer. Elles présentent entre autre la particularité d'imposer au sujet un travail important d'élaboration d'un but opérationnel (voilà ce que je vais faire) et de planification (voilà comment je vais faire). Si bien qu'il est difficile d'évaluer si tel aspect du produit final est la traduction de la mobilisation de telle ou telle connaissance, à telle ou telle étape de l'activité.

En bref donc, les tâches d'évaluation des apprentissages sont nombreuses et variées, les principaux biais sont bien connus, mais elles sont restreintes aux diagnostics de réussite. Elles ont permis, dans le domaine du multimédia pour l'apprentissage, d'accumuler des centaines de résultats, dont certains sont cohérents entre eux. Ces résultats, qu'il faudrait plusieurs centaines de pages pour synthétiser, nous donnent des indications importantes en ce qui concerne les effets de la structure (formelle et rhétorique) des contenus sur les apprentissages, ainsi que les effets de la mise en forme matérielle du texte à l'écran, ceux de la multimodalité, ceux des fonctionnalités disponibles à l'écran, etc.

Interpréter les liens entre évaluation de l'utilisation et évaluation des apprentissages

Le défi à l'heure actuelle nous semble être de pouvoir interpréter conjointement l'évaluation de l'utilisation et l'évaluation des apprentissages : comme l'indique Grudin (1992) un bon système est à la fois utilisable et utile.

Pour évaluer un outil multimédia pédagogique il faut donc être capable d'interpréter les liens entre les variables qui mesurent l'utilisation et celles qui mesurent l'apprentissage. Tricot et Tricot (2000) ont récemment tenté d'aborder ce problème. Leur approche consiste à envisager qu'il soit possible d'évaluer l'utilisation et l'apprentissage avec des variables binaires : Vi (réussite / échec) pour l'utilisation et Va (réussite / échec) pour l'apprentissage.

Une table de vérité décrit les relations entre deux variables binaires : elle donne l'interprétation logique de la nature de la relation. Par exemple, dans la table suivante (où l'apprentissage réussi est noté Va_1, l'échec à l'apprentissage Va_0, l'utilisation réussie Vi_1 et l'utilisation ratée Vi_0), la deuxième colonne correspond au «ou inclusif » (noté $\vee$). Le «ou inclusif » décrit entre autres la situation au restaurant où l'on peut prendre du fromage, du dessert ou les deux. Cette colonne décrit de la même manière, un outil d'apprentissage que les élèves peuvent réussir à utiliser avec ou sans résultat, ou que les élèves peuvent échouer à utiliser tout en apprenant quelque chose.

Va_1Vu_1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1
Va_0Vu_1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	1
Va_1Vu_0	1	0	0	1	1	0	0	1	1	0	0	1	1	0	1
Va_0Vu_0	0	1	0	1	0	1	0	1	0	1	0	1	0	1	1
	$\vee$	=>	Vu	<=	Va	$\Leftrightarrow$	$\wedge$	/	w	Va_0	$\neg Va \Rightarrow Vu$	Vu_0	$Va \Rightarrow \neg Vu$	$\downarrow$	T

Si on réalise un ensemble d'observations d'élèves en train d'utiliser un outil pour apprendre quelque chose, on peut décrire les fréquences *f* des cooccurrences des états de Vi et Va dans une table de contingence. La somme de ces fréquences est évidemment égale à 1.

	Vi réussite (noté Vi_1)	Vi échec (noté Vi_0)	
Va réussite (noté Va_1)	$f\,Va_1Vi_1$	$f\,Va_0Vi_1$	$f\,Va_1$
Va échec (noté Va_0)	$f\,Va_1Vi_0$	$f\,Va_0Vi_0$	$f\,Va_0$
	$f\,Vi_1$	$f\,Vi_0$	

L'analyse implicative (Bernard & Charron, 1996) permet, à partir de la distribution des fréquences dans cette table de contingence, de décrire la relation logique entre ces deux variables. Il s'agit de considérer une table de contingence 2x2 comme une colonne de la table

de vérité : les fréquences = 0 correspondant aux états « faux » (disons impossibles), les fréquences > 0 correspondants à des états « vrais » (disons possibles). Dans le modèle de Tricot et Tricot (op.cit.), on considère que la somme = 1 des fréquences > 0 correspond à une équirépartition de ces fréquences. Voici quelques exemples d'interprétation des liens entre ces deux variables pour quelques valeurs standards des fréquences.

	Vi_1	Vi_0
Va_1	*f*=1	*f*=0
Va_0	*f*=0	*f*=0

Outil « parfait » : utilisable et utile

Conjonction Va Ù Vi

	Vi_1	Vi_0
Va_1	*f*=0	*f*=0.5
Va_0	*f*=0.5	*f*=0

Outil nuisible

Ou exclusif Vi w Va

	Vi_1	Vi_0
Va_1	*f*=0	*f*=0
Va_0	*f*=1	*f*=0

Outil utilisable mais inutile

Implication ¬Va Þ Vi

	Vi_1	Vi_0
Va_1	*f*=0.5	*f*=0
Va_0	*f*=0.5	*F*=0

Outil utilisable mais moyennement utile

Indépendance Vi est vrai, " Va

	Vi_1	Vi_0
Va_1	*f*=0	*f*=0
Va_0	*f*=0	*f*=1

Outil mauvais ou inadéquat

NOR (non ou) ou « ni …ni …» Va ¯ Vi

	Vi_1	Vi_0
Va_1	*f*=0	*f*=0.5
Va_0	*f*=0	*f*=0.5

Outil inutilisable

Indépendance Vi est faux, " Va

	Vi_1	Vi_0
Va_1	*f*=0	*f*=1
Va_0	*f*=0	*f*=0

Outil paradoxal

Implication Va Þ ¬ Vi

	Vi_1	Vi_0
Va_1	*f*=0.5	*f*=0.5
Va_0	*f*=0	*f*=0

Outil placebo

Indépendance Va est vrai, " Vi

	Vi_1	Vi_0
Va_1	*f*=0.5	*F*=0
Va_0	*f*=0	*f*=0.5

Outil « spécifique » (nécessaire et suffisant)

Équivalence Va Û Vi

	Vi_1	Vi_0
Va_1	*f*=0	*f*=0
Va_0	*f*=0.5	*f*=0.5

Outil inutile, bien que moyennement utilisable

Indépendance Va est faux, " Vi

	Vi_1	Vi_0
Va_1	*f*=0.33	*f*=0.33
Va_0	*f*=0	*f*=0.33

Outil suffisant mais non nécessaire.

Implication Vi Þ Va

	Vi_1	Vi_0
Va_1	*f*=0.33	*f*=0.33
Va_0	*f*=0.33	*f*=0

Outil peu spécifique (ni nécessaire ni suffisant)

Ou inclusif Vi Ú Va

	Vi_1	Vi_0
Va_1	*f*=0.33	*f*=0
Va_0	*f*=0.33	*f*=0.33

Outil nécessaire mais pas suffisant

Implication Va Þ Vi

	Vi_1	Vi_0
Va_1	*f*=0	*f*=0.33
Va_0	*f*=0.33	*f*=0.33

Outil nuisible

Fonction NAND Vi / Va : incompatibilité

Exemple d'application : l'utilisation d'un dictionnaire encyclopédique au Cycle III

Pour illustrer ce cadre d'interprétation des liens entre utilisabilité et utilité d'un outil multimédia pour l'apprentissage, nous avons réalisé une expérience auprès d'élèves de Cycle III. Notre objectif est, plus précisément, de montrer les diverses relations possibles entre utilisabilité et utilité d'un outil fréquemment présent dans les situations scolaires. Nous avons proposé à des élèves de Cycle III de rédiger des définitions de mots, en s'aidant ou non d'un dictionnaire encyclopédique (Hachette, 2001). Les mots recherchés sont soit connus des élèves (auquel cas, le dictionnaire est peu utile), soit inconnus des élèves (auquel cas, le dictionnaire est utile). Deux versions du même dictionnaire sont utilisées : une version papier et une version électronique (pour chaque élève, soit l'une, soit l'autre). Nous faisons les hypothèses suivantes :

- en début de Cycle III (classe de CE2), le dictionnaire encyclopédique proposé est inutilisable, donc inutile ;
- en fin de Cycle III (classe de CM2), le dictionnaire proposé est utilisable et utile ;
- la version papier, plus habituelle pour les élèves, est plus utilisable ;
- la présence d'un dictionnaire électronique va conduire les élèves à sur-utiliser l'outil (à l'utiliser quand ils n'en ont pas besoin), tandis que le dictionnaire papier ne va pas produire cet effet de sur-utilisation.

Nous avons considéré qu'un dictionnaire est utile quand il permet d'apprendre la définition d'un mot que l'on ne connaît pas et utilisable quand il permet de trouver la définition que l'on cherche.

Méthode

Les 49 élèves qui ont passé l'expérience sont issus de trois classes différentes de trois écoles différentes du Tarn et Garonne. Une de ces écoles est dans un quartier urbain populaire et deux en zones rurales. Parmi les élèves, 39 sont en CM2, 6 en CM1 et 4 en CE2 ; 32 sont des filles et 17 des garçons. L'expérience a eu lieu au mois de Juin.

La passation est individuelle. L'élève reçoit un carnet de 6 pages, avec, sur chaque page, un mot à définir. Ces mots sont « Epitaphe », « Ecchymose », « Didascalie », « Escargot », « Bouteille », «Escalier ». Ils sont tri ou quadrisyllabiques, de début d'alphabet, trois d'entre eux sont connus des élèves, trois d'entre eux sont inconnus (notre expérience vérifie ce fait). L'ordre de présentation des mots suit une rotation de telle sorte que au total chaque mot apparaît à chaque page du carnet avec la même fréquence (*i.e.* 5 à 6 fois).

La consigne est la suivante : « Bonjour, je te propose de jouer au jeu du dictionnaire. Je vais te proposer une liste de mots. Tu en connais certains, d'autre pas. Pour chaque mot, tu vas essayer d'écrire une définition. Ensuite, il y a deux cas. Soit tu penses avoir besoin du dictionnaire : tu pourras alors chercher la définition. Mais attention, je ne te laisserai qu'une minute et demie pour trouver la définition. Je te redemanderai la définition. Soit tu estimes ne pas avoir besoin du dictionnaire, et on passe au suivant ». Pour chaque mot, au dessous de l'espace donné pour sa première définition, il est demandé à l'élève d'évaluer s'il est sûr, peu sûr ou absolument pas sûr de cette définition.

Le plan de l'expérience est le suivant : pour la moitié des élèves, le choix est donné entre le dictionnaire électronique et le papier. Pour l'autre moitié, la version est imposée (la moitié des élèves avec la version électronique, l'autre moitié avec la version papier). Les élèves sont répartis dans chaque groupe expérimental en fonction de l'ordre alphabétique des noms. Dans les deux cas, le dictionnaire est « fermé » au bout de 1'30'' ou quand la définition est trouvée, de sorte qu'il ne s'agisse pas pour l'élève de recopier la définition.

Résultats

Les élèves de CE2 ne réussissent pas à utiliser le dictionnaire proposé dans le temps imparti. Quand on leur en laisse le temps, 8' à 18' sont nécessaires pour trouver les définitions. Il a donc été décidé de ne pas faire passer l'expérience à plus de 4 élèves, ceux-ci étant décontenancés par cette tâche « infaisable ». Remarquons qu'un élève ayant mis trop de temps pour chercher son premier mot (bouteille), a décidé de ne pas utiliser le dictionnaire pour les 5 autres mots. Il a mis 11' pour écrire les 6 définitions.

Pour les 45 élèves de CM1 et CM2 ayant cherché 6 mots (soit 270 définitions), le dictionnaire encyclopédique est utilisable, quelle que soit son utilité, soit une indépendance logique entre les deux variables. La table de contingence ci-dessous correspond à celle de « utilisable bien que moyennement utile ».

	Utilisable	non utilisable
Utile	111 (*f*=.41)	6 (*f*=.02)
Inutile	112 (*f*=.41)	41 (*f*=.15)

Quand on distingue les trois mots *a priori* « connus » des trois mots « inconnus », on comprend que cette interprétation du dictionnaire comme outil « utilisable bien que moyennement utile » cache deux réalités :

Mots inconnus	Utilisable	non utilisable
Utile	108 (*f*=.80)	0 (*f*=.00)
Inutile	22 (*f*=.16)	5 (*f*=.04)

Mots connus	Utilisable	non utilisable
Utile	3 (*f*=.02)	6 (*f*=.04)
Inutile	90 (*f*=.67)	36 (*f*=.27)

Pour les mots inconnus, le dictionnaire est un outil presque «parfait » : il est à 80% utile et utilisable. En revanche, pour les mots connus, le dictionnaire est inutile, bien qu'utilisable. Les élèves semblent sur-utiliser le dictionnaire, quelle que soit sa version (seulement 25% des mots connus ne sont pas recherchés).

Parmi les élèves qui ont le choix, 19 préfèrent utiliser le dictionnaire électronique et 4 le dictionnaire papier. Si l'on cumule les comportements des élèves qui ont eu le choix et ceux qui ne l'on pas eu, on obtient la distribution des fréquences suivante :

Electronique

inconnus	Utilisable	non utilisable
Utile	.81	.00
Inutile	.14	.05

connus	Utilisable	non utilisable
Utile	.02	.06
Inutile	.58	.33

Papier

inconnus	Utilisable	non utilisable
Utile	.78	.00
Inutile	.22	.00

connus	Utilisable	non utilisable
Utile	.02	.00
Inutile	.86	.12

Contrairement à notre hypothèse, il n'y a apparemment que peu de différence entre l'utilisation des dictionnaires papier et électronique. On notera toutefois pour les cas où le dictionnaire est *a priori* inutile, que les élèves ont tendance à plus utiliser le dictionnaire quand celui-ci est électronique que quand il est en papier. Nos observations semblent aussi indiquer que, globalement, le temps moyen mis par les élèves pour trouver une définition dans le dictionnaire électronique est inférieur au temps mis par les élèves ayant cherché dans le dictionnaire papier.

Discussion

Nous ne doutons pas du fait que notre expérience est loin d'être représentative des situations d'apprentissage typiques. Notre façon de la conduire et les mesures que nous avons faites sont critiquables. En particulier, nous avons considéré comme valables des définitions issues du dictionnaire même quand elles étaient lacunaires, ce qui était très fréquent. Nous avons simplement voulu montrer qu'il est possible d'interpréter à la fois une mesure de l'apprentissage réalisé (variable d'utilité) avec un outil multimédia et une mesure de l'utilisation de cet outil (variable d'utilisabilité). Le cadre interprétatif que nous avons proposé permet aisément, sans passer par des calculs complexes, de comprendre les relations entre l'utilisation de l'outil et l'apprentissage réalisé.

Discussion générale

Dans cet article, nous avons eu tendance à privilégier les protocoles intrusifs classiquement utilisés en psychologie cognitive expérimentale. Il s'agit des protocoles où l'on demande à un sujet de réaliser une tâche et on évalue sa ou ses performances, en contrôlant autant que faire se peut les variables présentes dans la situation. Cette approche a l'avantage d'être contrôlée, elle permet de définir assez précisément le domaine de validité et de généralité des résultats. Et c'est justement là que le bât blesse : le domaine de validité est tellement restreint, que, souvent, on ne peut pas généraliser les résultats. Et comme pour contrôler les variables de la situation on conduit souvent l'expérience en laboratoire, les résultats obtenus n'ont que très peu de sens pour ce qui se passe sur le terrain, avec de vrais outils et de vrais apprenants. La validité externe des résultats, leur généralisation, est entièrement soumise à leur réplication.

L'expérience que nous avons présentée, bien que se déroulant dans un cadre scolaire, souffre des mêmes défauts.

Une autre tradition s'est développée, à la recherche d'une validité écologique, notamment en didactique. Elle consiste à utiliser des protocoles non intrusifs, ou en tous cas peu intrusifs : observations, enregistrements, entretiens *a posteriori*, carnets de bord, etc. Parfois ces protocoles sont même utilisés sans tâche prescrite. On regarde juste «ce qui se passe ». Bien entendu, ce que l'on gagne en validité externe avec ces protocoles, est perdu en validité interne. Ce vieux problème des disciplines expérimentales ne semble trouver sa solution qu'avec la recherche d'une multiplication des types de protocoles (ce qui peut devenir lourd) ou d'un optimum entre validité interne et externe (ce qui peut devenir illusoire).

L'évaluation des outils pour l'apprentissage reste donc difficile, le pire danger étant, sous prétexte de ces difficultés, de renoncer aux évaluations les plus rigoureuses possibles, où même de renoncer à l'évaluation.

Bibliographie

BERNARD J.M., CHARRON C. (1996), «L'analyse implicative bayésienne, une méthode pour l'étude de dépendances orientées I : données binaires », *Mathématiques, Informatique et Sciences Humaines*, vol. 134, p. 5-38.

BERNSTEIN M., BROWN P.J., FRISSE M., GLUSHKO R.J., LANDOW G., & ZELLWEGER, P. (1991), « Structure, navigation and hypertext : the status of the navigation problem », *Hypertext'91 Proceedings*, ACM Press, p . 363-366.

BERNSTEIN M., (1993), «Enactment in information farming », *Hypertext'93 Proceedings*, ACM Press, p. 242-249.

BUCKINGHAM SHUM, S., & MCKINGHT, C. (Eds.), (1997), « Web usability », *International Journal of Human Computer Studies*, special issue, vol. 47, n°1, p.1-222.

CHEN C., & RADA R., (1996), « Interacting with hypertext : A meta-analysis of experimental studies », *Human-Computer Interaction*, vol. 11, n° 1, p. 125-156.

FOSS C.L., (1989), «Tools for reading and browsing hypertext », *Information Processing and Management*, vol. 25, n° 4, p. 407-418.

GRUDIN, J., (1992), « Utility and usability: research issues and development context », *Interacting with Computers*, vol. 4, n° 2, p. 209-217.

KINTSCH W., WELSCH D., SHMALHOFER F., & ZIMNY S., (1990), « Sentence memory : a theoretical analysis », *Journal of Memory and Language*, vol. 29, p. 133-159.

LANDAUER T.K., & DUMAIS, S.T. (1996), « A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge », *Psychoogical Review*, vol. 104, p. 211-240.

NIELSEN J., (1993), *Usability engineering*, Academic Press, 1993.

ROUET, J.F., DILLENBOURG, P., STEFFENS, K., & VAN OOSTENDORP, H. (1999), « Analysing learner-computer interaction », *Instructional Science*, vol. 27, special issue. (ce volume inclus l'article mentionné de Rouet et Passerault).

ROUET J.-F., (1990), « Interactive text processing in inexperienced (hyper-) readers », in A. RIZK, N. STREITZ, J. ANDRE (Eds.), *Hypertexts : Concepts, systems and applications*, Cambridge University Press, p. 250-260.

SCAPIN D.L., & BASTIEN J.M.C. (1997), « Ergonomic criteria for evaluating the ergonomic quality of interactive systems », *Behavior & Information Technology*, vol.17, n° 4/5, p. 220-231.

TRICOT A., & COSTE J.-P., (1995), «Evaluating complex learner-computer interaction : what criteria for what task ? », *EARLI'95 Conference*, Nijmegen.

TRICOT A., PUIGSERVER E., BERDUGO D., & DIALLO M., (1999), «The validity of rational criteria for the interpretation of user-hypertext interaction », *Interacting with Computers*, vol. 12, p. 23-36.

TRICOT A., & TRICOT, M., (2000), « Un cadre formel pour interpréter les liens entre utilisabilité et utilité des systèmes d'information (et généralisation à l'évaluation d'objets finalisés) », *Actes du Colloque Ergo-IHM*, Biarritz, pp. 195-202.