EVALUATION ET CONTROLE DES ACQUIS[1]

Paul Black, King's College, Londres, Grande Bretagne

Introduction

Les buts

Dans l'enseignement, l'évaluation a trois fonctions principales. La première est de repérer les acquis des élèves individuellement dans le but de la certification. La deuxième est de repérer les acquis de groupes, de classes ou d'écoles pour des buts politiques plus généraux. La troisième est de servir à l'enseignement et l'apprentissage.

La première fonction permet d'établir des procès-verbaux jouant le rôle de passeports - pour de meilleurs emplois, ou pour aller dans enseignement supérieur quand un élève quitte l'école. Pour remplir cette fonction, l'évaluation doit susciter la confiance du public. Une telle évaluation a aussi l'objectif d'une vue d'ensemble du travail de l'élève, ce qui fait qu'elle peut être qualifiée de sommative.

La deuxième fonction est caractérisée par l'importance de rendre compte au public à la fois des écoles prises individuellement et du système éducatif au niveau d'un état ou d'une région. L'objectif est ici d'informer pour établir une politique, par le recueil et l'analyse de données. Différents systèmes de contrôles régionaux et nationaux ainsi que les études comparatives internationales remplissent cet objectif.

La troisième fonction provient de ce que tout système d'apprentissage nécessite des informations en retour. Pour atteindre cet objectif, les données doivent fournir des renseignements sur l'apprentissage de chaque élève sur la base desquels on peut mettre en oeuvre des actions répondant aux besoins d'apprentissage de chaque élève. Une telle évaluation peut être appelée formative ou diagnostique.

De manière idéale, chacune de ces trois fonctions demande des informations d'un type différent. En pratique, il est souvent nécessaire d'utiliser la même information pour remplir des fonctions différentes. Un tel usage multiple est séduisant car il est économique, mais il n'est pas toujours réalisable et il y a toujours une tension entre les besoins de ces différentes fonctions.

Contexte politique et social

Les différences d'histoire, de traditions, de besoins politiques et sociaux se sont combinés pour produire des systèmes dont la comparaison entre pays fait apparaître des schémas de pratique très différents (Black 1992, Britton et Raizen 1996). Par exemple, dans certains pays, les certifications de fin d'étude sont entièrement sous la responsabilité de chaque école, alors que dans d'autres, une telle confiance dans les enseignants est impensable et les certifications sont fondées entièrement sur des tests externes.

De telles différences ont des origines historiques. Des mouvements sociaux ayant pour objectif de produire une société égalitaire ont trouvé dans le passé les moyens de compenser les avantages des écoles privilégiées en utilisant des procédures d'évaluation. Dans une telle entreprise, la nature externe du système d'évaluation, et son absence de lien organique avec le système d'enseignement dans les écoles étaient des éléments essentiels. Inversement, dans les sociétés où l'égalité de financement et de statut des écoles était réelle, il devenait possible de ne plus utiliser des systèmes externes et objectifs et de compter sur chaque école pour réaliser l'évaluation. Toutefois, un système d'éducation public bien développé et universel s'avère être un utilisateur important des fonds publics, ce qui explique que la justification de son budget

devienne un problème politique important. Cette pression a conduit à des exigences de surveillance et de responsabilité.

De telles caractéristiques affectent la manière dont les examens externes contraignent l'élaboration de l'enseignement et de l'apprentissage. Aux plaintes prétendant qu'ils étouffent et déforment l'apprentissage, on doit opposer le fait que, développés et appliqués de façon adéquate, ils peuvent être des agents de réforme puissants.

La place de la physique dans les évaluations pour les certifications dont les enjeux sont liés à l'entrée en université, est également très variable. Dans les procédures des universités de la plupart des pays, la performance dans plusieurs matières, plutôt que pour la physique seule, est prise en compte dans l'admission à des cours spécialisés de physique. Dans le cas où l'examen d'entrée à l'université prend en compte un large éventail de matières, il peut y avoir des règles de regroupement qui conduisent à ce qu'un élève soit admis à étudier la physique même si ses performances en physique ont été faibles - comme au Brésil. Lorsque l'examen d'entrée est fondé seulement sur un petit nombre de matières - comme en Grande-Bretagne - il est possible de donner le plus de poids aux résultats de physique, en partie grâce à la possibilité d'un exercice approfondi, ce qui ne serait pas réalisable si les candidats avaient plus de matières.

Structure de ce chapitre

Ce chapitre traitera successivement de ces trois fonctions, c'est-à-dire la certification, la responsabilité vis-à-vis du public, l'aide à l'apprentissage. Inévitablement, par rapport à la première, de nombreuses questions générales s'appliquant à l'ensemble des trois devront être explorées. Une section finale passera en revue les interactions entre ces fonctions en essayant de faire un panorama des systèmes dans différents états ou nations.

Evaluation pour la certification

Méthodes - tests écrits

Les tests papier-crayon ont longtemps été utilisés comme moyen principal pour l'évaluation dans l'enseignement. Les tests écrits conventionnels en physique comportent une liste de questions exigeant des définitions, des explications standards ou des calculs, des comptes-rendus d'expérience ou des applications, et quelques problèmes habituels. L'exemple qui suit illustre cette combinaison :

> *- Un échantillon de l'élément thallium 207/81 Tl de 16 mg émet des radiations bêta et est transformé en un isotope du plomb (Pb). La demi durée de vie du processus de désintégration est de 5 minutes. Répondez aux questions suivantes :*
>
> *- Que signifient les termes : isotopes - isobares - isotones ?*
>
> *donnez un exemple de chaque.*
>
> *- Expliquez ce que veut dire "une demi durée de vie de 5 minutes pour une telle désintégration"*
>
> *- Trouver la masse de l'élément thallium 207/81 Tl résiduelle après 1/3 d'heure*
>
> *- Calculez la constante de désintégration de l'élément thallium 207/81 Tl*
>
> *- Donner (sans explication) deux méthodes de stockage et de traitement des déchets radioactifs.*
>
> *(Egypte - Wassef pp. 57-68 in Black 1992)*

Les tests peuvent être composés de questions plus courtes, traitant séparément les aspects liés aux connaissances et à la résolution de problèmes. L'exemple suivant est typique du style de problème court :

> *Un ressort comprimé maintient à distance deux chariots ayant respectivement pour masses 0,2 kg et 0,3 kg, de telle manière que les chariots qui étaient initialement au repos sont séparés de 60 cm au bout de 5 s. La masse du ressort et les frottements sont négligeables. Quelle est la vitesse des chariots ?*
>
> *(Hongrie - Radnai pp. 84-100 in Black 1992)*

De telles questions dans des problèmes courts ne sont pas nécessairement quantitatives, comme dans cet exemple :

> *Comment un dipôle électrique se comporte-t-il dans un champ électrique uniforme et dans un champ électrique radial ?*
>
> *(Pologne - Plazak et Mazur pp. 125-139 in Black 1992)*

Un nombre plus important de questions dans le même temps et une fiabilité plus importante de la notation peuvent être obtenus à partir de l'utilisation de tests à choix multiple, qui évaluent les capacités à résoudre les problèmes, tout comme les connaissances, comme le montre l'exemple suivant :

> *Les points O, P, Q, R, S et T sont alignés verticalement, à intervalles réguliers. Un objet est lâché en chute libre de O avec une vitesse initiale nulle.*
>
> *La vitesse moyenne entre O et T est égale à la vitesse instantanée qui se situe dans l'intervalle suivant :*
>
> *a) entre O et P b) entre P et Q c) entre Q et R*
>
> *d) entre R et S e) entre S et T*
>
> *(Japon - Ryu pp. 101-123 in Black 1992)*

Du fait du nombre important de questions auxquelles les élèves doivent essayer de s'attaquer dans un temps donné, de telles questions peuvent parvenir à couvrir un champ large et être plus fiables. Leur inconvénient est qu'elles ne montrent pas les raisons du choix des élèves et que des études ont montré que près d'un tiers des élèves qui ont choisi la réponse correcte l'ont fait pour des raisons erronées (Tamir 1990). L'utilisation exclusive de ces tests peut conduire à négliger, dans l'enseignement, la discussion et l'argumentation qui sont des aspects précieux et valides de la recherche scientifique. Dans les examens nationaux, ils ont été longtemps prédominants dans les évaluations aux USA, ils ont été des composants minoritaires dans certains pays (par exemple au Royaume Uni, en Suède) et ne sont pas du tout utilisés par d'autres (par exemple en France).

Tous les exemples précédemment cités mettent en jeu une combinaison de connaissances de physique et de capacités à sélectionner et appliquer ces connaissances. Lorsque les questions longues ont une structure - comme dans le premier exemple, cette structure aide celui qui répond, à la fois en présentant la sélection du savoir nécessaire pour le problème, et en guidant vers la stratégie qui permet de l'aborder.

Il n'est pas nécessaire de se limiter aux connaissances mémorisées. Certains tests écrits présentent à l'élève des informations sous la forme d'un petit article sur un sujet de physique, et évaluent ensuite les capacités à comprendre et appliquer en posant des questions sur ce texte - par des questions ouvertes courtes ou des questions à choix multiple (voir Black 1992).

Evaluation des "savoir-faire procéduraux"

Il y a eu des tentatives pour évaluer les "savoir-faire" séparément du contenu physique associé. Les problèmes qui cela soulève peuvent être illustrés par des tâches pratiques ayant pour objet d'évaluer les savoir-faire associés à la mesure. Un des problèmes est relatif au contexte - de nombreux élèves qui semblent être capables d'utiliser les instruments de mesure et les procédures lorsqu'ils sont directement demandés dans des contextes artificiels et isolés, ne se servent pas de cette même capacité lorsqu'il leur est demandé de prendre en charge une investigation, même lorsque les instruments de mesure leur sont fournis ; ils utilisent seulement

une comparaison qualitative, même lorsqu'ils ont montré, dans un contexte différent, leur capacité à utiliser les instruments. Il est ici essentiel d'avoir une vue plus complète des savoir-faire en jeu. La capacité à se servir de manière précise d'une échelle et de disposer et régler les instruments n'est pas suffisante. Un scientifique se doit d'être précis sur ce qu'il veut mesurer - par exemple, mesurer un "débit" nécessite de comprendre qu'il faut prendre des couples de mesures reliées entre elles. De plus, l'expérimentateur doit juger quand mesurer, ce qui requière un jugement sur le fait que le puissant outil de quantification peut et doit être appliqué au problème (Black 1990).

Ainsi, une évaluation des savoir-faire relatifs aux instruments et aux échelles n'a pas beaucoup de valeur en soi, du fait que de tels savoir-faire ne se révèlent utiles qu'à la lumière d'un modèle conceptuel du système étudié et des variables en jeu. Il en est de même pour les autres "savoir-faire" - par exemple l'observation qui n'est pas une réception passive mais qui est essentiellement une activité sélective guidée par des hypothèses sur ce qui doit être sélectionné. Il en résulte que les questions qui testent des "savoir-faire" spécifiques de manière isolée ne peuvent pas apporter d'informations utiles sur les capacités d'un élève à utiliser des savoir-faire dans un travail scientifique.

Autres méthodes - orales et pratiques

Les tests oraux sont des composantes importantes des examens nationaux dans plusieurs pays européens de l'Est, mais sont rarement utilisés ailleurs. Dans l'évaluation des travaux pratiques, la logistique et les autres problèmes d'utilisation des équipements matériels pour les exercices d'évaluation ont conduit à essayer d'utiliser en remplacement les tests papier crayon. Une telle approche peut avoir en retour des effets indésirables sur l'enseignement, et les corrélations entre les performances des élèves avec le matériel réel et leur réponse avec leur "équivalent" dans les tests écrits sont faibles. Toutefois, très peu de pays, notamment le Royaume Uni et Israël, utilisent des tests avec du matériel à titre de procédure d'examens (Britton et Raizen 1996).

Une forme habituelle de test pratique a été de spécifier un ensemble de procédures avec des équipements donnés et de demander certains types de réponses en termes de mesures effectuées et d'analyse de résultats. Les contraintes précises alors imposées améliorent la fiabilité, mais des savoir-faire tels que la conception d'expériences et le choix des appareils sont négligés. L'"expérience" est en fait un moyen pour tester certains savoir-faire spécifiques ayant rapport à la science expérimentale. Une approche alternative a consisté à organiser de très courts exercices, chacun testant un savoir-faire spécifique dans un contexte donné. Les savoir-faire évalués peuvent être par exemple prendre des mesures précises avec des instruments préréglés, faire rendre compte d'observations qualitatives sur un phénomène inhabituel.

Les études concernant les tests hors contexte qui peuvent être mal interprétés ont conduit à entreprendre l'évaluation à partir de tâches expérimentales ouvertes et complètes. Si la comparabilité et le respect de conditions d'examen conventionnelles sont requises, alors ce qu'il est possible de faire est très limité. Une contrainte supplémentaire à de telles conditions est que, puisque tous les membres d'un groupe doivent travailler simultanément en un temps limité, l'évaluation doit se fonder sur les rapports écrits des élèves. Leurs actions effectives et leur raisons d'agir ne peuvent pas être observées ou interrogées directement. De même, du fait qu'aucun système ne peut manipuler plus d'un très petit nombre (probablement une ou deux) de tâches complètes, la généralisation des résultats est un problème sérieux (Shavelson et al. 1993).

Des limitations aussi importantes peuvent être dépassées par l'évaluation de travaux menés sur une longue durée et entrepris dans les conditions normales de l'enseignement. Un rapport écrit de tels travaux peut cacher quelques-uns des aspects importants des capacités dont un élève peut faire preuve, aspects qui peuvent seulement être estimés si les procédures avec lesquelles les élèves ont travaillé ont été observées et comprises. La seule solution possible face à de tels problèmes est, pour les enseignants, de s'impliquer comme évaluateur, ce qui permet d'avoir accès à des caractéristiques telles que la manière précise dont le problème a été posé, les effets de la collaboration dans le groupe, les contraintes avec lesquelles les élèves ont dû travailler et

les raisons des décisions des élèves concernant le choix d'une stratégie. Une telle attention requière une formation approfondie des enseignants ainsi que du temps et la possibilité de pouvoir observer attentivement le travail individuel ou le travail de groupe.

Bien entendu, il y a ici un paradoxe. Comme les objectifs de l'enseignement de la physique deviennent plus réels et moins artificiels, ils conduisent à des activités qui reflètent plus la complexité et le désordre de la vie réelle, et ces dernières deviennent ainsi moins susceptibles de faire l'objet d'un quelconque processus reproductible d'évaluation.

De telles difficultés ne s'appliquent pas seulement à l'évaluation des travaux pratiques. Quelques-unes des nombreuses tâches écrites qui sont utilisées sont sérieusement limitées par les contraintes de temps des tests externes usuels. Si on se donne pour but la discussion des idées et leur exploitation de façon à ouvrir à de nouvelles questions, on parviendra à des tests conduisant à la production de textes en un temps suffisant pour que l'élève puisse réfléchir et structurer ses idées.

Combiner les méthodes

Toutes les méthodes d'évaluation ont des défauts, et le meilleur choix peut souvent dépendre du contexte éducatif ou social particulier. Par exemple, la mise à disposition du matériel ou la possibilité d'une surveillance appropriée pour éviter les pratiques malhonnêtes peuvent varier énormément d'une situation à une autre. Il est également nécessaire de considérer qu'une combinaison de méthodes peut être nécessaire, à la fois pour renforcer la fiabilité et pour compenser le biais que toute technique introduit nécessairement.

La panoplie des méthodes utilisées est très variée dans certains pays, et très restreinte dans d'autres. A un extrême, on trouve l'utilisation exclusive des questions à choix multiple, comme aux USA où on utilise des agences spécialisées dans l'évaluation. A l'autre extrême, on trouve l'éventail des huit types de tâches pour l'examen Nuffield au Royaume Uni (Black 1992). L'ensemble le plus largement approuvé comporte un mélange de questions à choix multiple et de problèmes courts.

Une des raisons d'une telle variété résulte du fait que la familiarité avec certaines méthodes et la tradition de leur utilisation peut inhiber la possibilité de rechercher sérieusement des solutions alternatives. Par exemple, il existe une tradition dans certains pays selon laquelle la question théorique mathématique est le test le plus significatif de la compétence d'un physicien. Ainsi, ces questions se voient donner une importance plus grande que, dirons-nous, les compétences à aborder un problème expérimentalement ou à écrire une critique sur un sujet.

Une autre raison peut être l'impression que certaines méthodes ne peuvent pas donner de résultats fiables : on peut penser que les questions à choix multiple, avec leur notation objective, leurs pré-tests de vérification et les possibilités d'analyse statistique sur un grand nombre de questions donnent des résultats qui sont bien plus fiables que les autres types d'évaluation. Cela conduit à les considérer comme la seule méthode à utiliser.

Une troisième raison peut être le coût. Pour les examens à forts, la dépense pour préparer de bonnes questions à choix multiple est justifiée par la possibilité d'obtenir des notes fiables à un faible coût. Le choix et la notation de beaucoup d'autres types de questions sont coûteux en temps et en expertise des examinateurs. En particulier, l'évaluation des travaux pratiques est fréquemment exclue à cause de son coût et de sa faisabilité.

Fiabilité

La question de savoir si oui ou non un éventail de questions est réellement important est encore complexe. Les principaux facteurs impliqués sont la fiabilité, la validité, le feed-back, la qualité et le biais. La fiabilité est le plus simple à considérer. Un examen peut être fiable si on est certain que les mêmes résultats seraient obtenus à partir d'un examen parallèle, c'est-à-dire à partir d'un ensemble de questions et leur barème avec les mêmes objectifs et les mêmes

méthodes. Il est possible d'obtenir une mesure de la fiabilité en testant la cohérence interne des réponses, mais cela n'est possible que lorsque l'on dispose d'un nombre raisonnablement important de questions et lorsque l'on peut supposer qu'elles doivent donner des résultats homogènes. Un test plus strict est de donner des ensembles de questions en parallèle à des mêmes étudiants ou au moins de donner à quelques étudiants un grand nombre de questions dans le même domaine et de déterminer à partir des résultats, le nombre minimal de questions nécessaires pour réduire l'erreur au-dessous d'un certain seuil. Il est difficile de réaliser de manière systématique de tels tests - la confiance des examinateurs à l'université et à l'école en la fiabilité des tests externes courts est généralement injustifiée du fait qu'elle n'est pas fondée sur des preuves. La fiabilité de la notation est également un enjeu dans les examens nationaux - une formation soigneuse des examinateurs et des doubles corrections sont essentielles.

Validité

Un concept suffisamment général de validité est à la fois nécessaire et convaincant dans toute tentative d'amélioration de la qualité des évaluations. La proposition introductive de la revue réalisée par Messick (1989) donne une définition qui fait autorité :

> *La validité est un jugement évaluatif global du degré auquel les données empiriques et les justifications théoriques confirment l'adéquation et la justesse des inférences et des actions fondées sur les scores des tests ou d'autres modes d'évaluation.*

Ainsi, si un test est supposé montrer les compétences d'un élève pour la mesure, c'est l'objet du jugement de l'expert de savoir si le test fait appel à l'utilisation des capacités de mesures qui sont importantes en science. Si un test est conçu pour aider à prévoir des aptitudes à de futures acquisitions, cela peut être estimé après coup, en étudiant la corrélation des résultats au test avec ceux que le test devait prédire. Plus une activité d'évaluation est proche de l'activité réelle pour laquelle les résultats du test sont considérés comme pertinents, plus le test aura tendance à satisfaire les critères de validité. Dans cette perspective, l'évaluation dans la classe a plus de chance de succès que des tests écrits formels en temps limité.

Effets sur l'apprentissage

Un des centres d'intérêts concernant l'évaluation est son effet sur l'apprentissage. Les recherches ont montré que préparer les élèves à des tests à choix multiple peut être défavorable à une bonne pratique d'apprentissage. Ceci est exprimé par Resnik et Resnik (1992) de la manière suivante :

> *Les élèves qui pratiquent la lecture principalement de la manière dont elle intervient dans les tests - et il existe des preuves que c'est ce qui se produit dans beaucoup de classes – risquent d'être peu confrontés aux exigences et aux possibilités du raisonnement qui sont dans l'esprit d'un programme tourné vers la réflexion*
>
> *Les élèves qui pratiquent les mathématiques dans la forme correspondant à celle des tests standardisés risquent de n'être jamais confrontés au type de pensée mathématique recherchée par tous ceux qui sont concernés par la réforme de l'enseignement des mathématiques...*
>
> *Les évaluations doivent être conçues pour que, lorsque vous faites ce qui est naturel - c'est-à-dire, préparer les élèves à bien réussir – ceux-ci exerceront les types de compétences et développeront les types de savoir-faire qui sont les véritables objectifs de la réforme de l'enseignement.*

Qualité

Des différents facteurs qui affectent la qualité des tests et des évaluations, un de ceux qui ressortent le plus est le temps pris par les tests externes. Si un pays investit près de huit heures

dans des tests dans une matière pour déterminer les perspectives d'une future carrière, comment un autre peut-il remplir la même fonction en moins de deux heures ? Tout test vise à échantillonner les différents domaines de performance qui sont importants dans la matière - le contrôle de science de l'A.P.U. à l'âge de 13 ans en 1984 a nécessité 35 tests d'une heure pour obtenir des résultats fiables sur les domaines jugés importants pour la science (Johnson 1988).

Une tactique pour dépasser cette difficulté est de restreindre l'éventail des domaines évalués, cependant cela réduit d'autant les buts que l'examen vise. Par là même, en raison de l'enjeu qu'ils représentent dans la plupart des systèmes, ils influent sur les objectifs d'apprentissage à l'école.

Une question proche se rapportant à la qualité porte sur l'équilibre entre les questions auxquelles il est possible de répondre par des procédures de routine en utilisant des algorithmes, objets d'apprentissage, et les questions qui exigent une traduction réfléchie et une application de principes et de procédures. Dans les systèmes de nombreux pays, l'analyse montre que l'équilibre penche souvent en faveur des premières. Ceci doit en partie être dû aux limitations de temps pour les tests. La tâche très difficile pour la plupart des évaluateurs est de concevoir des questions qui exigent une réflexion plus qu'un apprentissage machinal, à un niveau pour lequel la moyenne des élèves pourront avoir une bonne chance de réussite.

Biais

Toute procédure d'évaluation est une interaction entre certaines questions, items et/ou procédures, et les élèves qui sont évalués. Beaucoup de raisons peuvent faire que interaction fonctionne de manière défectueuse, produisant ainsi des biais ou des imperfections dans les résultats (Gipps et Murphy 1994).

Un exemple qui a été bien étudié est celui du biais du sexe. Un problème ayant une importance comparable est celui du biais culturel ou ethnique. Les difficultés se situent dans le langage utilisé dans l'enseignement des matières, dans l'éventail et la nature des exemples de la vie de tous les jours utilisés et dans l'hypothèse culturelle inhérente à la science occidentale.

Il y a également de nombreuses façons de discriminer injustement les élèves en tant qu'individus, même au sein de la même catégorie de sexe et de culture, par la présentation, le contexte ou le langage des évaluations. Il peut s'avérer que des élèves qui produisent des réponses étranges et apparemment dénuées de qualité montrent grâce à des discussions qu'il s'agissait de réponses sensées liées à un mauvaise interprétation de la question (Gauld 1980).

Référence à la norme et au critère : regroupement et profil

La tradition dominante à tous les niveaux d'enseignement a été d'utiliser des tests avec une norme référencée. Lorsque ceux-ci sont utilisés, l'accent porte sur la comparaison d'un élève quelconque avec l'ensemble de la population testée ou évaluée. L'approche alternative est de donner la priorité à des critères, de manière à ce qu'un résultat d'évaluation signifie qu'un élève a satisfait des critères donnés sans rapport direct avec les réussites des autres élèves (Gipps 1994).

Une question proche est celle du regroupement des résultats. Une méthode courante est de prendre les notes aux questions traitant de différents sujets et évaluant différentes capacités, et de sommer ces résultats pour donner à chaque élève un score total. Cette approche peut être comparée aux tentatives d'additionner des choux et des carottes. Une solution est alors de reporter les résultats sous forme d'un profil composé d'éléments pris en compte séparément. Il est alors possible de dire que le score de chaque composante a une signification en relation avec la réussite sur des critères donnés.

Une conséquence d'une telle discussion est d'émettre des doutes sur la validité de la pratique conventionnelle des tests comportant des questions qui s'étalent sur une variété de contextes, de connaissances et de savoir-faire et d'additionner les notes pour produire un seul nombre avec

une norme référencée. Cela revient à émettre des doutes sur la signification des résultats et sur la justification de cette pratique.

Evaluations par les enseignants

En principe, un enseignant qui peut décrire la réussite d'un élève au cours du temps et dans des contextes différents et qui peut discuter des réponses particulières de manière à comprendre la pensée qui les sous-tend, peut construire une description de bien meilleure fiabilité qu'aucun test externe ne pourra le faire. De même, il semble que seules les évaluations faites par les enseignants assurent la validé quand il s'agit d'évaluer des savoir-faire des élèves lors de la résolution de problèmes réalistes et cela dans différents contextes.

Les difficultés évidentes sont le manque d'expertise des professeurs concernant l'évaluation, le manque de possibilités de comparaison de standards entre différentes écoles et bien sûr entre les professeurs d'une même école, ainsi que les dangers de préjugés et de malhonnêteté. Aucun de ces obstacles ne peut être surmonté rapidement ni à faible coût.

La comparaison entre écoles peut être obtenue lors de discussion dans des réunions de groupes d'écoles où les critères sont discutés et où des exemples de travaux d'élèves sont échangés. De telles réunions ont été considérées comme très valables pour la formation des enseignants impliqués, révélant souvent combien les enseignants sont isolés en ce qui concerne leurs standards et leurs attentes. Il existe d'autres méthodes d'homogénéisation des standards entre écoles, notamment par l'utilisation de visiteurs externes inspectant les procédures ainsi que des échantillons de travail (Black 1993).

Le Royaume-Uni semble être le seul pays à donner, lors les examens de certification nationaux, de l'importance à la notation, par les propres professeurs des élèves, de travaux effectués en dehors des conditions formelles d'examen. Pour que cela soit acceptable, en disposant de la confiance publique, il faut des règles précises et attentives et un système qui, en extrayant des échantillons des écoles et en vérifiant leurs standards, peut s'assurer de l'intégrité et de la comparabilité des résultats. La Suède est également unique d'une toute autre manière. Dans ce pays, les évaluations des enseignants sont les éléments principaux dans la détermination du résultat global. L'examen externe sert à calibrer la répartition des écoles dans son ensemble, mais laisse l'enseignant libre de prendre des décisions sur chacun des élèves (Black 1992).

3 - L'évaluation en vue de rendre des comptes

Le but est ici de renseigner les prises de décisions politiques par le recueil et l'utilisation d'informations provenant d'évaluations. Alors que les résultats publics de certifications sont utilisés comme indicateurs de performance des écoles, les détails de ces données font défaut. Toutefois, pour cet objectif, il n'est pas nécessaire de produire des résultats fiables et complets pour chaque individu. En donnant aux élèves des tests différents, la performance globale peut être étudiée plus en détail.

Si on souhaite étudier la relation entre les performances et d'autres facteurs qui peuvent être adaptés par la politique publique ou par celle de l'école, les informations sur ces facteurs doivent être recueillies. Ainsi, la tâche implique la sélection et le recueil de données sur des facteurs tels que la taille de la classe, le contexte familial de l'élève, le temps passé à apprendre, les équipement de laboratoire, et ainsi de suite, de manière à augmenter les données sur les performances des élèves. L'analyse des relations possibles entre les données devient alors complexe en raison des multiples corrélations qui doivent être explorées avec attention. Chacun des facteurs (par exemple le type d'école) peut être corrélé avec, et apparaît ainsi comme cause des variations de performance, alors qu'il est seulement le représentant d'un facteur différent (par exemple les résultats des élèves pour l'entrée dans une école) auquel il est associé. L'interprétation des relations est également difficile du fait des corrélations, même correctement

isolées par des moyens statistiques, car elles ne peuvent pas offrir par elles-mêmes des preuves de causes (Woodhouse et Goldstein 1996).

Quand il y a des études évaluatives au niveau national, celles-ci doivent refléter une structure d'objectifs et de critères. Dans la mesure où cela produit des items de test de qualité et des données détaillées concernant les objectifs et les critères, les travaux attirent l'attention des enseignants et influencent leur travail. Ainsi, dans le "National Science Monitoring" au Royaume-Uni, la structure choisie a été d'insister sur les buts en termes de processus en science. Elle a conduit les enseignants à insister davantage sur l'observation et sur la conception d'investigations pratiques (Black 1990). Ainsi, les travaux correspondant ont pu être prescriptifs, et on peut dire que l'opération a été autant un projet de développement de programme qu'un projet d'évaluation du fait qu'il a permis d'aller au delà des pratiques du moment. Ainsi, des études évaluatives peuvent fournir l'occasion de promouvoir l'innovation, tout comme elles peuvent représenter une force puissante de conservatisme si elles reflètent seulement les objectifs et les procédures établies.

4 - Soutien pour l'apprentissage - évaluation formative

Introduction aux enjeux

L'évaluation formative suppose l'utilisation d'un ensemble de plusieurs types de données pour un seul objectif. Cet objectif est la modification du travail d'apprentissage pour l'adapter aux besoins qui sont mis à jour par les données. C'est seulement lorsque l'évaluation suit cette voie qu'elle devient formative. La réaction en retour peut aller d'une réaction immédiate de la classe à une question, jusqu'à une revue détaillée d'une variété de données permettant d'estimer les progrès sur l'ensemble d'un sujet ou d'un thème. En termes de a théorie du contrôle, l'utilisation du feed-back dans ce sens peut être considéré comme une nécessité évidente.

Des caractéristiques communes des évaluations formatives de nombreux pays ressortent des recherches utilisant des enquêtes (Black 1993). En particulier il est bien établi que les programmes d'évaluation formative conçus soigneusement conduisent effectivement à une amélioration de l'apprentissage de l'élève. Une caractéristique commune est que le renouvellement des pratiques d'évaluation fait partie des modifications plus larges des stratégies d'enseignement et n'est pas seulement un ajout supplémentaire.

Une autre caractéristique est que, généralement, et cela est vrai en particulier pour l'évaluation formative, l'évaluation n'est pas valorisée parmi les pratiques et les priorités des enseignants. Plus encore, alors qu'il avait été préconisé une plus grande insistance pour l'évaluation par les enseignants, comme dans le programme national de l'Angleterre et du Pays de Galles, ces instructions ont été généralement mal interprétées. Les recherches menées en vue d'évaluations ont établi que la plupart des enseignants, et particulièrement les enseignants des écoles primaires, ont restreint le sens d'évaluation par les enseignants à celui d'évaluation sommative. Des enquêtes récentes ont également montré que dans les travaux de science, très peu d'évaluations formatives sont réalisées (Russell et al. 1994).

Une raison prépondérante de cette faiblesse est que les évaluations sommatives, et notamment les tests externes, dominent fréquemment l'enseignement du fait que leurs résultats sont tenus en meilleure estime que ceux des autres formes d'évaluation. Ceci a de nombreuses conséquences préjudiciables. Les tests externes créent des modèles et des images trompeurs des évaluations et des tests. Par exemple, une pratique courante d'enseignement est d'organiser un test de fin d'unité d'enseignement ou de fin de trimestre qui ressemble le plus possible aux tests externes, de consigner les résultats, peut-être même de les rendre publics. Du fait que les données ne sont pas utilisées pour modifier l'enseignement et l'apprentissage, il ne s'agit pas d'évaluation formative. Ainsi une pratique plus ou moins fréquente d'évaluation sommative est établie, ce qui conduit l'évaluation à être assimilée à un test, porteur de préjugés négatifs, soit comme une épreuve du feu pour les élèves, soit comme un travail lourd et improductif pour leurs enseignants.

Quoi qu'il en soit, il existe d'autres raisons au faible développement des pratiques d'évaluation formative. Elles sont relatives aux nombreuses difficultés pratiques pour recueillir et enregistrer les résultats parmi toutes les autres exigences quotidiennes de l'enseignement. Elles sont aussi un défi en termes de modification, de reprise ou de différenciation de l'enseignement pour répondre aux résultats de l'évaluation.

Deux exemples

Deux exemples spécifiques, chacun concernant le développement de la pratique dans une école anglaise, serviront à illustrer quelques-uns de ces problèmes. Dans la première école (Parkin et Richards 1995), les enseignants de science désiraient utiliser l'auto-évaluation des élèves et les discussions qui ont suivi entre professeur et élèves comme base de leur évaluation. Pour chaque module de cours, des critères constitués en objectifs étaient exprimés dans un langage accessible aux élèves. Pour chaque leçon, chacun des élèves disposait d'une feuille donnant les critères avec une place à côté dans laquelle les élèves devaient écrire si le critère avait été clair et si il avait été rempli ; il était également demandé aux élèves d'écrire d'autres commentaires - par exemple leur plaisir ou leur intérêt.

Plus tard, le professeur annotait chacune des réponses aux critères avec un des trois codes suivant : A - pour une compréhension totale, P pour une compréhension partielle, V - lorsque le travail n'a pas été plus que "visité" par l'élève.

A partir du moment où cette méthode a été introduite, il a fallu un an pour que les élèves l'utilisent de manière productive - au début, beaucoup d'élèves écrivaient des commentaires très brefs et très vagues, mais après une année, ces commentaires ont changé, ils sont devenus plus explicites et perspicaces et ainsi plus utiles. Les élèves n'étaient pas accoutumés à réfléchir sur leur propre apprentissage comme cible vers la réalisation des objectifs. Ils devaient également rompre avec l'idée de l'évaluation comme étant un test formel.

Quelques élèves, surtout les plus faibles, n'aimaient pas admettre l'échec, et disaient parfois avoir compris alors que cela n'était pas vrai. Les enseignants ont insisté auprès de chaque élève pour que leurs feuilles restent des documents privés destinés à aider l'enseignant à voir les problèmes des élèves et à fournir une aide en cas de besoin.

Dans la deuxième école (Fairbrother 1995), un enseignant de physique d'une classe d'élèves de 12-13 ans voulait qu'ils abordent le cours sur l'électricité et le magnétisme de façon plus responsable. Il avait pour objectif de les aider à :
- situer chaque leçon dans le contexte général du cours ;
- avoir un résumé de ce qu'ils avaient fait en vue des révisions ;
- voir ce qui venait juste après dans le cours.

Il a donné à chaque élève une "feuille de révision" pour la séquence qui contenait 25 déclarations d'objectifs, par exemple :

> *Savoir comment faire un électroaimant et comment faire varier sa force*
>
> *Savoir qu'il est nécessaire de disposer d'un circuit complet pour tout appareil électrique*
>
> *Savoir qu'un fil électrique placé dans un champ magnétique va essayer de bouger lorsqu'il est traversé par un courant.*
>
> *Savoir comment les interrupteurs, les relais, les résistances, les capteurs et les portes logiques peuvent être utilisés pour résoudre un problème simple, par exemple les sonneries d'alarme, les avertisseurs de gel, les éclairages urbains automatiques.*

La plupart des élèves avaient une petite idée de la façon d'utiliser cette liste, par exemple en vérifiant leurs notes dans leur cahier à partir du contenu, ou vérifiant s'ils connaissaient ce qui

leur était demandé. Certains des élèves les moins organisés ont simplement perdu cette liste, d'autres l'ont simplement mise de côté et n'ont pas fait référence à celle-ci.

L'explication du professeur concernant cet échec était que l'enseignement portait trop sur le contenu seul et pas sur la façon d'apprendre. La feuille de révision avait pour intention de traiter ce problème, mais l'enseignant n'avait pas pris conscience au début combien un véritable enseignement sur l'utilisation de cette feuille aurait été nécessaire. Par exemple, lorsqu'il demandait aux élèves de réviser chez eux pour le test, la plupart d'entre eux pataugeaient. Il semble y avoir deux raisons principales à cela. La première était que les élèves ne savaient pas comment extraire, de tout ce qu'ils faisaient, ce qu'ils étaient supposés savoir et comprendre. Les enseignants connaissent la différence entre les *fins* auxquelles ils veulent aboutir, et les *moyens* avec lesquels ils essayent d'y parvenir. Les élèves ne voient pas cette différence. Une deuxième raison était que les élèves ne savaient pas ce que l'enseignant attendait d'eux en ce qui concerne leurs connaissances et leur compréhension. La plupart d'entre eux apprennent avec l'expérience ce qu'on attend d'eux, et pour beaucoup d'élèves, cette expérience est dure et décourageante. Certains d'entre eux, surtout les plus faibles, n'apprennent jamais.

Développer une bonne pratique

La prépondérance traditionnelle de la fonction sommative se traduit par le fait qu'il y a lutte pour que l'évaluation formative existe et se développe (Fairbrother et al. 1995). Des tentatives pour mettre en valeur l'évaluation du professeur peuvent se réduire en pratique trop facilement à une plus grande utilisation de cette évaluation pour des objectifs sommatifs, et à une application plus fréquente des évaluations des enseignants, avec un recueil et un stockage des résultats qui deviennent un fardeau. La fonction sommative peut inhiber de plusieurs manières le développement de la fonction formative par les enseignants. La pratique sommative peut induire en erreur, les tests externes étant considérés comme modèles pour les évaluations par les enseignants, ce qui les conduit vers des techniques appropriées seulement pour une utilisation sommative. Les tests externes sont des modèles pauvres pour l'évaluation sommative puisque :
- dans les tests sommatifs, la nécessité d'avoir un seul résultat global signifie que des données relativement différentes (par exemple pour la pratique et la théorie) doivent être additionnées de manière souvent arbitraire : l'évaluation formative n'a pas à faire la même chose ;
- l'évaluation sommative a des problèmes particuliers de référence à des critères, en partie du fait de la nécessité d'addition des notes, en partie parce qu'on ne peut pas se fier à un jugement personnel lors de la décision de l'application de critères généraux au travail individuel des élèves ; de tels problèmes sont bien moins importants dans la pratique de l'évaluation formative ;
- le travail sommatif doit insister sur les standards d'uniformité et de fiabilité dans la récolte et l'enregistrement des résultats, ce qui n'est pas nécessaire dans le travail formatif et qui inhibe la liberté et l'attention aux besoins individuels qu'exige le travail formatif ;
- alors que les processus sommatifs doivent être considérés comme étant équitables, la pratique formative, avec ses priorités d'identification et d'aide adapter aux besoins d'apprentissage de chaque élève, peut se conduire envers différents élèves de manières très variées ;
- les objectifs sommatifs peuvent exiger des preuves argumentées des résultats - par exemple un audit - et cela ajoute du travail et déforme la pratique formative, alors que le travail formatif demande plus d'action sur les données que sur leur stockage.

Une source prépondérante de difficultés pour le développement de l'évaluation formative est qu'elle ne peut pas être simplement plaquée sur des procédures de travail déjà existantes, mais qu'elle doit être élaborée avec la procédure elle-même. Ceci seulement parce que son utilisation pour guider l'apprentissage selon les besoins ne peut être réalisée que si les projets d'enseignement prévoient un temps considérable pour la planification et l'organisation que cela suppose.

L'utilisation effective des feed-back de l'évaluation demandent un jugement du professeur, ainsi que confiance et souplesse dans la gestion du projet de programme, ce qui ne peut se produire que lorsque le professeur se sent responsable de la planification du programme. Ainsi, il semble

qu'idéalement, tout procédé pour incorporer des aspects formatifs doit être construit par les enseignants pour eux-mêmes. Dans une telle construction, les enseignants doivent gérer deux innovations - la nécessité d'appliquer de nouvelles méthodes pour la différenciation et la souplesse de l'apprentissage, et la nécessité d'apprendre, peut-être d'inventer, une nouvelle technologie pour avoir des données convaincantes sur la réussite des élèves.

L'utilisation des résultats d'évaluations formatives est peut-être l'aspect le plus stimulant. Il existe des "macro" réponses, en terme de répartition des groupes par niveaux, mais celles-ci ne concernent pas des besoins immédiats. Des enseignants ont répondu en organisant des unités de travail autour d'un noyau et de ramifications. Là le travail est très varié, depuis le traitement de nouveaux sujets plus approfondis pour ceux qui veulent aller plus loin jusqu'à la répétition des bases pour ceux ayant des besoins plus fondamentaux (Black 1993). D'autres indiquent des approches moins formelles et plus souples, avec des révisions profitant d'occasions favorables lors de travaux ultérieurs. Il y a un enjeu de souplesse ou de rigidité des programmes.

La technologie de recueil de données sur les progrès des élèves commence juste à se développer. La plupart des enseignants ont toujours utilisé une variété de sources de manière informelle – il faut affiner cette pratique dans la perspective d'obtenir des données mieux utilisables. Les feuilles décrites dans le premier exemple ci-dessus montre un moyen pour le faire ; les résultats se distinguent en ce qu'ils produisent des informations détaillées en relation avec des formulations d'objectifs spécifiques – c'est-à-dire que la référence à des critères s'applique naturellement, puisque c'est une nécessité pour l'évaluation. De plus, puisque cela produit des résultats écrits de manière systématique, l'enseignant est soulagé de la pression de la notation et de l'enregistrement exhaustifs des événements de la classe. Les données sur les circonstances peuvent toutefois avoir une importance en elle-même : certains enseignants ont trouvé particulièrement utile - et c'est surprenant - de suspendre leur enseignement actif pendant un moment - en exprimant clairement à la classe ce qu'ils faisaient et pourquoi - et de se concentrer sur l'observation et l'écoute d'un petit nombre d'élèves (voir Cavendish et al. 1990, Connor 1991).

Quand l'activité d'évaluation est construite en étroite relation avec un programme d'apprentissage, il serait insensé d'empêcher les élèves de commenter leurs résultats, de se remettre en question et de refaire l'évaluation s'ils désiraient améliorer leurs performances. Ainsi, une conséquence de leur rôle de soutien pour l'apprentissage est que les évaluations formatives deviennent à la fois informelles, et conduites par les élèves. L'importance donnée à l'auto-évaluation des élèves est une caractéristique notable. L'expérience montre que les élèves ne peuvent pas jouer un rôle efficace dans leur propre évaluation, sauf lors de programmes à long terme conçus pour les aider à réussir et à garder une vue d'ensemble de leurs objectifs d'apprentissage, de façon à appliquer les critères correspondants à leurs propres progrès. Comme le montrent ces deux exemples, on doit enseigner aux élèves comment évaluer leurs propres progrès. Un part importante de ce travail est la traduction des objectifs du programme dans un langage que les élèves peuvent comprendre, et vers un niveau de détails qui peut les aider à mettre directement en relation leurs efforts d'apprentissage. Il s'ensuit également que les objectifs doivent être à la fois accessibles à court terme et suffisamment modestes en relation avec les perspectives de réussite des élèves. Ces exigences concernent particulièrement les élèves qui rencontrent difficultés spéciales d'apprentissage - mais elles sont importantes pour tous.

Les enseignants qui ont élaboré l'auto-évaluation des élèves signalent plusieurs avantages - les élèves peuvent diriger leurs propres efforts plus clairement et efficacement, ils peuvent être impliqués plus activement et plus motivés en relation avec leurs propres progrès, ils peuvent alors suggérer leur propre manière d'améliorer leurs acquisitions, et même ils peuvent remettre en question des évaluations qu'ils estiment être injustes.

Manifestement, l'implication des élèves peut rendre la conduite d'un programme d'évaluation formative plus facile aux enseignants. Toutefois, cette implication change également le rôle des élèves comme apprenants et la nature de la relation entre enseignant et élèves, ce qui fait reposer sur les épaules de ceux-ci la responsabilité de l'apprentissage. En dehors même de la nécessité

d'améliorer l'évaluation, le besoin premier d'amélioration de l'apprentissage exige de tels changements. En effet, on a argumenté que la métacognition, en tant que sensibilisation et auto-direction sur la nature de l'apprentissage, est essentielle au développement des élèves pour l'apprentissage d'un concept. Le travail décrit ci-dessus sert clairement cet objectif (voir Brown 1987, White et Gunstone 1989, Baird et Northfield 1992). L'évaluation formative ainsi améliorée, peut conduire à des changements qui sont d'une bien plus grande signification - des changements qui devraient être une aide puissante pour le développement personnel des élèves et qui devraient également faire partie de tout programme pour les aider à être des apprenants plus efficaces.

Les systèmes et les rôles

Une bonne évaluation sommative exige l'implication des enseignants. Il semble ainsi qu'il n'y ait d'autre solution que d'engager les enseignants dans des rôles à la fois sommatifs et formatifs, en utilisant pour ces deux fonctions quelques-unes de leurs données, mais pas nécessairement toutes. Il leur faut alors distinguer soigneusement les méthodes et les nécessités en relation avec chacun des deux objectifs. Assumer les deux rôles de cette manière serait très exigeant. D'un côté, il y a les besoins d'apprentissage de leurs élèves, ce qui devrait être leur première préoccupation. De l'autre côté, il y a les pressions et les contraintes provenant de l'extérieur. Les systèmes nationaux et régionaux ont de forts enjeux et créent des pressions sur les enseignants, ce qui les oblige à travailler au sein d'une structure qui guide à la fois les décisions de leurs écoles et les attentes des parents. L'enseignant doit trouver le juste milieu entre les pressions provenant des deux côtés.

La raison principale de l'insistance sur ce point est que quelques-uns des objectifs importants en physique ne peuvent pas être reflétés, et être ainsi soutenus, par les systèmes d'évaluation qui reposent seulement sur les tests externes courts. La réforme des contrôles sommatifs nationaux est une sérieuse nécessité. En 1992, à la fin d'une revue des tests nationaux de physique dans onze pays, j'ai écrit le résumé suivant en guise de conclusion :

> *Une conclusion que je tire de cette étude est que la variété des méthodes utilisées et la variété des capacités évaluées par ces examens de physique sont trop faibles. Elles ont certainement un effet restrictif sérieux sur le développement de la physique à l'école et sur le recrutement de physiciens. Il y a plusieurs raisons à cela. L'insuffisance des ressources, avec d'autres contraintes du système, peuvent expliquer l'empressement des examinateurs de physique à travailler avec des systèmes qu'ils estiment être, au mieux, loin de l'idéal et peut-être nuisibles à l'avenir de la physique. Peut-être cette situation est-elle acceptée trop volontiers par nous tous. (Black 1992)*

Les examens publics ou officiels ont un pouvoir particulier sur l'avenir de la physique. En déterminant les objectifs et la structure avec lesquels les enseignants des écoles secondaires pensent qu'ils doivent travailler, ils déterminent la structure et l'image du sujet aux yeux des jeunes. Si de tels examens ne suscitent pas ou n'encouragent pas des activités qui sont importantes et attirantes pour le physicien et si elles transmettent une image très restrictive du domaine, ils attireront trop peu de spécialistes, et donneront à tout adulte une vue très négative de la physique.

Références

Baird, J.R. et Northfield, J.R. (eds.) (1992) *Learning from the PEEL experience*. Melbourne: Monash University.

Black, H (1993) Assessment: A Scottish Model pps.91-94 in Fairbrother, R , Black, P.J. and Gill, P. (eds.) *TAPAS : Teacher Assessment of Pupils: Active Support. King's Education Papers No.3*. London: C.E.S. King's College.

Black, P.J. (1990) APU Science - the past and the future. *School Science Review* 72. 13-28

Black, P.J. (1992) *Physics Examinations for University Entrance : an International Study*. Science and Technology Education - Document No. 45. Paris : UNESCO.

Black, P.J. (1993), Formative and Summative Assessment by Teachers. *Studies in Science Education.* 21. 49 - 97.

Britton, E.D. et Raizen, S.A. (eds.) (1996) *Examining the Examinations : An International Comparison of Science and Mathematics Examinations for College-Bound Students.* Boston : Kluwer

Brown, A. (1987) Metacognition, executive control, self-regulation and other mysterious mechanisms. pps 65 - 116 in Weinert, F.E and Kluwe, R.H. (eds.) *Metacognition, Motivation, and Understanding.* Hillsdale, New Jersey: Lawrence Erlbaum.

Cavendish, S., Galton, M., Hargreaves, L. et Harlen, W. (1990) *Observing Activities.* London: Paul Chapman.

Connor, C. (1991) *Assessment and Testing in the Primary School.* London: Falmer Press.

Fairbrother, R. (1995) *Pupils as Learners.* pp.105-124 in Fairbrother et al. op.cit.

Fairbrother, R., Black, P.J., et Gill, P. (eds.) (1995) *Teachers Assessing Pupils : Lessons from Science Classrooms.* Hatfield UK : Association for Science Education.

Gipps, C.V. (1994) *Beyond Testing : Towards a Theory of Educational Assessment*, London : Falmer, .

Gipps, C.V. et Murphy, P. (1994) *A fair test ? Assessment, achievement and equity,* Milton Keynes : Open University Press.

Gauld, C.F. (1980) Subject oriented test construction, *Research in Science Education,* 10, 77--82.

Johnson S. (1988) *National Assessment : the APU Science Approach.* London : Her Majesty's Stationery Office.

Messick, S. (1989) Validity pp. 12 - 103 in Linn, R.L. (ed.), *Educational Measurement (3rd. Edition)',* London : Collier Macmillan.

Parkin, C et Richards, N (1995) *Introducing Formative Assessment at KS3 : an attempt using pupils' self-assessment.* pp 13-28 in Fairbrother et al. op.cit.

Resnick, L.B. et Resnick, D.P.(1992) Assessing the Thinking Curriculum: New Tools for Educational Reform pp. 37 - 75 in Gifford, B.R. & O'Connor, M.C.(eds.), *Changing Assessments : Alternative Views of Aptitude, Achievement and Instruction,* Boston : Kluwer.

Russell, T., Qualter, A., McGuigan, L. et Hughes, A. (1994), *Evaluation of the implementation of Science in the National Curriculum at Key Stages 1, 2 and 3.* London: School Curriculum and Assessment Authority.

Shavelson, R.J., Baxter, G.P. et Gao, X.(1993) Sampling variability of performance measurements *Journal of Educational Measurement* 30. 215--232.

Tamir, P. (1990) Justifying the selection of answers in multiple-choice questions. *International Journal of Science Education* 12. 563-573.

White, R.T. et Gunstone, R.F. (1989) Meta-learning and conceptual change. *International Journal of Science Education.* 11. 577-586.

Woodhouse, G. et Goldstein, H. (1996) The Statistical Analysis of Institution-based Data pp.135-144 in Goldstein, H. and Lewis, T. (eds.) *Assessment: Problems, Developments and Statistical*