1. Comment évaluer les apprentissages ?

1.1 Qu'est-ce que l'évaluation ?

Évaluer, c'est situer un acte par rapport à une référence. C'est, plus précisément, juger de la différence entre cet acte et cette référence. L'acte peut-être une activité, une performance, une production d'un élève, etc. L'idée généralement acceptée est que cet acte est un indice d'une connaissance ou d'une compétence. On part de ce qui est observable (l'acte, la performance, le comportement) et on infère la connaissance ou la compétence.

La référence peut-être :

- Un acte de l'élève lui-même : on évalue alors le progrès. On peut par exemple comparer la performance de l'élève aujourd'hui à une performance obtenue il y a une semaine, un mois, un trimestre.
- Les performances du groupe classe : on évalue alors une position, un classement.
- Les performances attendues de l'âge de l'élève : on évalue l'acte d'un élève par rapport à celui qui est produit en moyenne par les élèves de son âge. C'est ce que l'on fait dans de très nombreuses évaluations standardisées, du test de Q.I. il y a 100 ans aux évaluations PISA de l'OCDE aujourd'hui.
- Les attendus du programme, du curriculum : on évalue l'acte d'un élève par rapport aux buts d'apprentissage. C'est normalement ce que l'on devrait faire en classe, exclusivement.

On va voir dans la suite de ce chapitre, notamment dans la partie 3, qu'une des difficultés principales de l'évaluation réside dans la relation entre l'acte observé (la performance par exemple) et le but d'apprentissage : peut-on inférer de tel acte que l'élève a élaboré la connaissance but de l'apprentissage ? Un des intérêts de l'approche par compétences réside dans le fait qu'elle définit le but d'apprentissage comme la réalisation d'une tâche. On peut alors évaluer avec moins de craintes : si l'élève réalise la tâche but d'apprentissage, alors il a atteint le but.

Enfin, comme l'évaluation est un jugement de la différence entre un acte et une référence, c'est une mesure. En tant qu'instrument de mesure, l'évaluation consiste donc à attribuer une valeur à cette différence entre acte et référence. Elle doit alors présenter les qualités de tout instrument de mesure : elle doit être non seulement pertinente (mesurer effectivement ce qu'elle prétend mesurer) mais fiable (précise et objective). Dans ce chapitre, nous allons voir que l'évaluation scolaire ayant énormément de mal à présenter ces deux qualités (elle n'est, la plupart du temps, ni pertinente ni fiable), et ayant par ailleurs des effets secondaires souvent néfastes (voir la très documentée synthèse de Butera et al., 2011 sur ce point), il vaut mieux être extrêmement prudent quand on l'utilise. Pour autant, une évaluation de qualité est un outil essentiel à la conception, la mise en œuvre et la régulation d'une situation d'enseignement-apprentissage.

1.2 À quoi servent les évaluations ?

1.2.1. Les principales fonctions de l'évaluation

Les évaluations scolaires sont utilisées à des fins très diverses, des plus légitimes aux plus discutables. Nous voudrions maintenant tenter de faire le tour de ces principales fonctions afin que chacun puisse expliciter de façon claire les buts qu'il poursuit quand il évalue. L'évaluation peut servir à :

Sélectionner : on évalue lors d'un concours par exemple. Il y a *n* places disponibles ; seuls les *n* premiers réussiront à rentrer dans telle école, à exercer telle profession, etc. Dans ce cas, l'évaluation ne doit pas être seulement pertinente et fiable, elle doit en outre être indiscutable car ceux qui ne font pas partie des lauréats peuvent se retourner contre l'évaluation. Une des dérives possibles est alors de rendre l'évaluation totalement indiscutable (ex. un QCM corrigé par ordinateur) quitte à perdre beaucoup de pertinence (ex. ce qu'a été le concours de l'internat en médecine dans les universités françaises, il y a quelques années). Une autre pratique consiste à mettre de très mauvaises notes aux lauréats, pour se plaindre ensuite de leur faible niveau (ex. en 2007, le dernier admissible à l'agrégation d'Allemand était à 4,58/20, en 2008 il était à 5,18, etc.).

Valider : on évalue pour vérifier que le but visé a été atteint. Souvent, de façon très étrange, on a du mal à faire quelque chose d'aussi simple. Plutôt que de dire que le but a été atteint ou non atteint, on met une note, qui semble vouloir dire que seul celle ou celui qui a obtenu la note de 20/20 a effectivement atteint le but. Ce qui voudrait dire en retour que nous avons de grandes difficultés à définir des buts atteignables. Dans les établissements du second degré de nos pays francophones (à l'exception du Canada), nous éprouvons une grande difficulté à ne plus mettre de notes, pour simplement valider l'atteinte du but.

Sanctionner : on évalue pour punir les élèves ou un élève. Comme ceux-ci apprécient généralement peu d'être évalués et notés, l'évaluation est utilisée comme menace, notamment l'évaluation « surprise » et la mauvaise note comme punition. Bien entendu cette utilisation est une perversion de l'évaluation et interdite dans la plupart des institutions scolaires.

Motiver : l'évaluation peut être utilisée comme moteur de l'apprentissage, notamment quand on focalise l'évaluation sur les progrès réalisés par l'élève. Par ailleurs, même si certains élèves parmi les plus performants sont sensibles à l'émulation, la comparaison entre élèves est au total très délétère. Reportez-vous au chapitre 8 sur la motivation pour de plus amples détails.

Aider à apprendre : on évalue pour fournir à l'élève une aide précise et personnalisée à l'apprentissage. En situant son acte par rapport à une référence, on peut dire précisément à l'élève les étapes qu'il doit encore franchir, la connaissance qu'il n'a manifestement pas élaborée, la règle qu'il a oublié de mettre en œuvre, etc. C'est là une des principales, si ce n'est la principale utilité de l'évaluation. Nous reviendrons en détail ci-dessous, dans la partie 10.7 de ce chapitre, consacrée aux diagnostics.

Prendre conscience, réguler son propre apprentissage : de façon très proche de la précédente fonction, il s'agit cette fois-ci non pas d'aider l'élève en lui indiquant ce qui n'a pas fonctionné ou ce qui lui reste à faire, mais à se focaliser sur ce qui a fonctionné, sur ce qui a été réalisé, voire sur la connaissance qui a été élaborée. Nous nous substituons alors aux processus métacognitifs de l'élève. Ceci est particulièrement utile avec les élèves qui ont moins de 9 ou 10 ans pour qui cette opération est souvent hors de portée. C'est aussi très utile avec les élèves les plus en difficulté, les moins confiants, qui ont tendance à attribuer leurs réussites au hasard plutôt qu'à eux-mêmes.

Dépister ou repérer : de plus en plus d'évaluations sont consacrées au dépistage, notamment dans le domaine des troubles du langage et de l'apprentissage. Il s'agit le plus souvent d'une approche standardisée ou semi-standardisée. Cette évaluation permet d'identifier ce que l'on appelle une population à risque, c'est-à-dire l'ensemble des élèves pour lesquels un protocole de diagnostic vaut d'être mis en œuvre. Une des erreurs malheureusement fréquente consiste à confondre dépistage et diagnostic.

Exemple, le ROC outil de repérage des difficultés de lecture
Le ROC a été élaboré par des médecins, des chercheurs et des enseignants de Grenoble, Montpellier et Rennes. Malgré son nom (Repérage Collectif Orthographique) il est surtout utile pour dépister les grandes difficultés de lecture à la fin de l'école primaire et au début de l'enseignement secondaire. C'est un outil standardisé et validé. Il permet de repérer, mais pas de diagnostiquer.

C'est un outil extrêmement utile car il est, non seulement pertinent et fiable, mais rapide à administrer et à coder (pour une classe entière, quelques dizaines de minutes suffisent en tout !). En outre il est accessible à tous et gratuit.
Le livret élève présente simplement : une tâche de jugement orthographique (trouver les erreurs orthographiques dans un texte) et une tâche de dictée.
www.cognisciences.com

Diagnostiquer : au sens strict, un diagnostic est l'établissement de la cause d'un problème, d'un dysfonctionnement. Dit autrement, diagnostiquer c'est *expliquer pourquoi* tel élève a produit tel acte, telle performance. C'est donc une activité extrêmement difficile, car ce qui est visible pour nous, c'est l'acte, c'est la performance. Nous pouvons faire des hypothèses sur les causes de cette performance, notamment des hypothèses en termes de connaissances. Mais, en toute rigueur, nous ne pouvons rien faire d'autre que des hypothèses, car il n'existe pas de modèle validé qui établirait les liens entre performances et connaissances. Cette modestie nous semble devoir caractériser toute évaluation diagnostique.

Évaluer le système, l'établissement, le dispositif : nous sommes parfois sollicités pour participer à une évaluation du système éducatif (ex. PISA), de notre établissement scolaire voire d'un dispositif (ex. une innovation). Ces évaluations sont souvent très intéressantes et conduites avec une grande rigueur méthodologique. Cependant leur interprétation est délicate, car ces dernières n'expliquent généralement rien : elles constatent, c'est tout ! Ces évaluations permettent éventuellement de faire des hypothèses, que l'on pourra peut-être tester. Ce qui la plupart du temps n'est pas fait. Pour autant, il est bien plus utile d'avoir ces évaluations que rien du tout. Et, à une échelle plus réduite, conduire de telles évaluations pour soi est souvent extrêmement utile : lorsque je mets en œuvre une expérimentation pédagogique, un nouveau dispositif dans ma classe, il est intéressant de savoir de manière un peu objective si tout ce travail apporte quelque chose aux élèves, améliore ma façon d'enseigner ou non.

Nous allons maintenant résumer ce que nous venons d'écrire selon deux autres perspectives.

1.2.2. Pour qui évalue-t-on ?

Pour l'enseignant : évaluer ce que font les élèves, leurs progrès, leurs performances, est une bonne manière de savoir si mon enseignement atteint ses objectifs. C'est même sans doute la meilleure manière d'évaluer la qualité de mon travail et donc d'envisager la régulation de celui-ci. Le chapitre 15 est entièrement consacré à l'activité de régulation de la conduite d'un enseignement.

Pour le système : évaluer le système éducatif ou un établissement scolaire, que ce soit lors de comparaisons internationales ou de mesures de taux de réussite à tel ou tel examen est une façon de mesurer la performance ou l'efficacité de celui-ci. Mais, le plus souvent, cela ne permet pas de savoir pourquoi telle performance est bonne ou mauvaise.

Pour l'élève : évaluer l'activité, la performance ou les progrès d'un élève est souvent un excellent moyen d'améliorer et de valider les apprentissages. Ce n'est jamais un bon moyen de juger un élève lui-même ou une quelconque de ses capacités, comme nous allons le voir dans la partie 10.3 de ce chapitre.

1.2.3. Quand évalue-t-on ?

En début d'apprentissage, l'évaluation porte sur l'état des connaissances avant que l'enseignement ne commence. Cette évaluation est doublement intéressante : elle est l'un des points de départ de l'activité de conception de l'enseignement car on n'enseigne jamais qu'à partir des connaissances antérieures des élèves ; elle sert de point de référence pour évaluer les progrès des élèves.

En fin d'apprentissage, l'évaluation concerne généralement la validation de l'apprentissage, et, par comparaison avec un état antérieur, c'est un moyen d'évaluer les progrès.

En cours d'apprentissage, l'évaluation permet de réguler l'apprentissage, que ce soit pour aider l'élève à prendre conscience de ce qu'il a déjà fait (appris, produit, réalisé) ou de ce qui lui reste à faire (à apprendre).

Un effet secondaire positif de l'évaluation
L'effet de l'évaluation ou testing effect *est un des plus connus dans le domaine des apprentissages, notamment car il est soutenu par une bonne centaine d'expériences de laboratoires contrôlées. Il se présente au départ comme une idée assez originale : qu'est-ce qui se passe après l'évaluation ? Roediger a découvert cet effet il y a plus de vingt ans (voir la synthèse de Roediger et al. 2011) en comparant simplement des personnes qui avaient été évaluées dans la réalisation d'une tâche avec des personnes qui n'avaient pas été évaluées. Les résultats montrent un effet positif de l'évaluation. L'explication de Roediger est très simple et très fondamentale : l'évaluation est une tâche de mobilisation et d'utilisation des connaissances, elle a donc pour effet de renforcer l'apprentissage. Au total, Koedinger recense 10 effets secondaires positifs de l'évaluation, outre celui que nous venons de mentionner :*
- l'évaluation permet d'identifier les lacunes en termes de connaissances,
- l'évaluation conduit les élèves à apprendre plus la fois suivante,
- l'évaluation améliore l'organisation des connaissances,
- l'évaluation améliore le transfert des connaissances,
- l'évaluation permet de mobiliser des connaissances qui n'ont pas été préalablement évaluées,
- l'évaluation améliore le contrôle métacognitif,
- l'évaluation prévient l'interférence avec les contenus préalables quand on aborde un nouveau contenu,
- l'évaluation fournit un retour aux enseignants,
- l'évaluation fréquente encourage les élèves à apprendre.
Toutefois nous encourageons le lecteur à consulter de façon attentive la synthèse de Roediger, les conditions d'obtention de ces résultats en laboratoire n'étant pas toujours perçues comme convaincantes quand il s'agit de transférer les résultats à la classe.

1.3 Quelle information est fournie par une évaluation ?

Après avoir vu ce qu'était l'apprentissage et les fonctions qu'il pouvait remplir, nous allons maintenant aborder l'objet de l'évaluation et tenter de répondre à la question : que peut-on évaluer ? Généralement, on cherche à évaluer des compétences ou des connaissances. Comme on ne peut pas observer directement les compétences et encore moins les connaissances, on les infère des actes des élèves. C'est pourquoi on préfère parfois évaluer directement les activités et n'en rien inférer. Nous allons maintenant essayer de comprendre quels problèmes sont liés à la réalisation de cette inférence : passer de ce que j'observe de l'activité d'un élève à une hypothèse sur les connaissances qu'il a élaborées.

1.3.1. Quelques problèmes avec l'objet de l'évaluation

Prenons un exemple célèbre, qui a révolutionné l'histoire de la psychologie cognitive (rien que ça !). En 1966, le psychologue anglais Peter Wason a publié un chapitre dans lequel il rend compte d'une expérience *a priori* très simple.

L'expérience de Wason (1966)
On présente à des adultes cultivés quatre cartes.

Et la consigne suivante : « Quatre cartes comportant un chiffre sur une face et une lettre sur l'autre, sont disposées à plat sur une table. Une seule face de chaque carte est visible. Les faces visibles sont

les suivantes : A, B, 2, 3. Quelle(s) carte(s) devez-vous retourner pour déterminer la ou les carte(s) qui ne respecte(nt) pas la règle suivante : Si une carte a un A sur une face, alors elle porte un 2 sur l'autre face. Il ne faut pas retourner de carte inutilement, ni oublier d'en retourner une. »
Quelle est votre réponse ?
Généralement, 80% des participants à cette expérience ne donnent pas une réponse correcte.
Pourtant, pour trouver la réponse, il faut faire ce simple raisonnement :
- s'il y a autre chose qu'un 2 derrière la carte A ça contredit la règle, donc il faut la retourner
- peut importe ce qu'il y a derrière la carte B, car la seule lettre concernée par la règle est un A
- peut importe ce qu'il y a derrière la carte 2, car si il y a un A ça vérifie la règle, s'il y a autre chose qu'un A, cela ne la contredit pas
- s'il y a un A derrière la carte 3, alors la règle est contredite, il est donc nécessaire de la retourner.
Pourquoi si peu de personnes réussissent à mettre en œuvre ce raisonnement, pourtant si simple ? Ceci est d'autant plus troublant que, quelques années plus tard, des élèves de Peter Wason ont proposé une nouvelle version de cette tâche.
« Quatre personnes sont en train de boire dans un bar et vous disposez des informations suivantes : la première boit une boisson alcoolisée, la seconde a moins de 18 ans, la troisième a plus de 18 ans et la dernière boit une boisson sans alcool. Quelle(s) personne(s) devez-vous interroger sur leur âge ou sur le contenu de leur verre pour vous assurer que tous respectent bien la règle suivante : Si une personne boit de l'alcool, elle doit avoir plus de 18 ans. »
Dans ce cas, la plupart des personnes interrogées donnent une réponse correcte.
Pourtant, logiquement, le raisonnement à produire est exactement le même.

Que peut-on inférer de ces performances ? Dans la première version de la tâche, si l'on fait un lien entre réussite à la tâche et connaissance, on est obligé de conclure que 80% des participants ne savent pas raisonner, même quand le raisonnement à mettre en œuvre est assez élémentaire. Avec la seconde version, on serait pourtant tenté de conclure l'inverse. Des centaines d'articles et réplications de cette expérience ont été publiés. Dans l'interprétation la plus prudente de ces travaux, le consensus actuel permet au moins de dire : des personnes qui habituellement raisonnent correctement sont, avec la première tâche, incapables de raisonner correctement ; et pourtant il n'y a aucun piège dans cette tâche. Gardons à l'esprit pour l'instant que, dans ce cas au moins, on ne peut pas inférer la connaissance des individus en observant leur performance.

Regardons maintenant un autre résultat.

L'expérience de Bastien (1987)
Quatre versions d'un même problème d'ordonnancement de fractions sont proposées à des élèves de 5ème. Ces derniers doivent ranger 62/185, 66/170 et 62/170 par ordre croissant. Une première version du problème est présentée à un premier groupe de 21 élèves avec un énoncé concernant un rapport qualité-prix : ce rapport correspond à une notion « non représentable » (i.e., il n'est pas possible d'élaborer une image mentale d'un tel rapport) et les unités au numérateur et au dénominateur sont différentes. La seconde version du problème est présentée à un autre groupe de 21 élèves avec un énoncé relatif au taux de participation de chanteurs à une chorale : ce rapport correspond à une notion « non représentable » et les unités au numérateur et au dénominateur sont identiques. Le troisième problème concerne des précipitations (représentable ; unités différentes). Le quatrième problème concerne des pentes de ski (représentable ; mêmes unités). « Représentable » signifie donc qu'une image mentale de la notion peut être élaborée. Les résultats sont les suivants et ils ont été largement répliqués depuis. On donne le nombre d'élèves parmi 21 a avoir trouvé la bonne réponse selon la version du problème.

	Représentable	Non représentable
Mêmes unités	5 élèves	11 élèves
Unités différentes	11 élèves	19 élèves

L'analyse des résultats et de la façon dont les élèves procèdent pour traiter les problèmes montre que, quand le rapport est représentable, certains élèves ne traitent pas ce problème comme un problème mathématique, mais comme un problème « concret » (par exemple, ils dessinent les pentes de ski) ; quand les unités au numérateur et au dénominateur sont identiques, certains élèves n'identifient pas la nature mathématique des objets à traiter (ils font une soustraction au lieu de faire un rapport). Dans le cas contraire, quand le problème n'est pas représentable et que les unités sont différentes au numérateur et au dénominateur (e.g., le problème des rapports qualité-prix), les élèves réussissent le problème. Autrement dit, la façon dont est rédigé l'énoncé influence la stratégie qui est mise en œuvre qui, à son tour, influence le résultat.

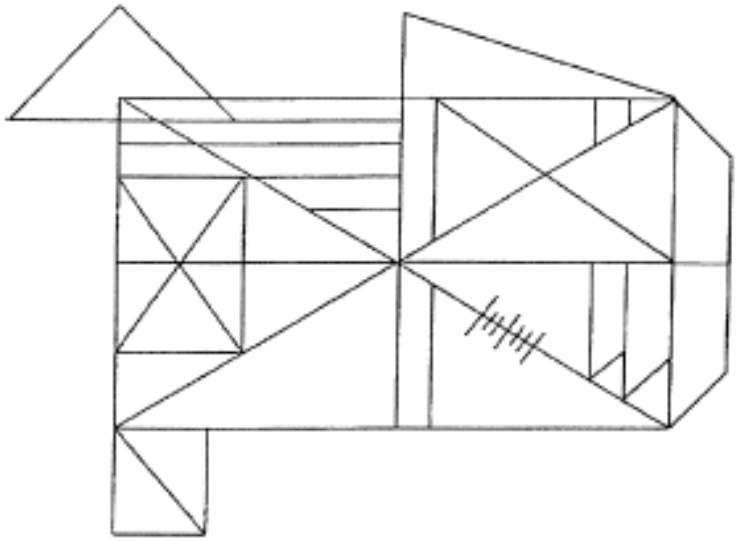
Là encore, on voit qu'il est très difficile d'inférer une connaissance (la maîtrise de l'ordonnancement de fractions présentant des numérateurs et dénominateurs communs) en observant les performances. De façon intéressante ici, on voit que dans une condition, presque tous réussissent : c'est la version où les unités sont différentes et où le rapport n'est pas représentable. Ils sont en quelque sorte conduits à traiter le problème sur le bon registre : « c'est un problème de maths et non pas un problème concret ; c'est un problème de fractions et non pas de soustraction ». Dans les autres versions, les élèves seraient gênés par la formulation du problème pour mobiliser la bonne connaissance, la stratégie correcte.

Ces deux résultats, celui de Wason et celui de Bastien, ne sont que deux exemples parmi des centaines d'autres qui montrent la même chose : on ne peut pas inférer une connaissance en observant simplement la performance. Entre la performance et la connaissance il y a la tâche. **Selon comment l'élève interprète la tâche, il va parvenir ou non à mobiliser les connaissances nécessaires à la réalisation de la tâche**.

Continuons à examiner ce que l'on peut évaluer au regard d'une troisième expérimentation.

L'expérience de Monteil et Huguet (1991)

Dans cette expérience, la figure de Rey - Osterrieth est proposée à des élèves de début de collège. Il s'agit de regarder la figure pendant 50' et de la reproduire le plus fidèlement possible.



La tâche est proposée à des élèves ayant le statut de « mauvais » ou de « bons » élèves. La tâche est présentée comme étant une épreuve de « dessin » pour la moitié des élèves et comme une épreuve de « géométrie » pour l'autre moitié. La performance est mesurée à partir d'une analyse de la figure en 22 unités et corrigées avec le barème suivant : 2 points sont accordés si l'unité est correctement reproduite et positionnée ; 1 point si elle est soit altérée mais correctement positionnée, soit intacte mais incorrectement positionnée ; 1/2 point si elle est à la fois altérée et incorrectement positionnée ; 0 point si elle est absente ou non. Les résultats montrent que les élèves ont la même performance moyenne (autour de 18/44) quand l'épreuve est présentée comme relevant du dessin. En revanche,

quand la tâche est présentée comme relevant de la géométrie, les « bons » élèves obtiennent une moyenne de 21/44 tandis que les « mauvais » obtiennent une moyenne de 16/44. Comme les autres expériences rapportées ci-dessus, cette expérience a été répliquée de nombreuses fois.

Si cette expérience présente un intérêt immense pour comprendre une source des difficultés scolaires, bornons-nous à regarder ce que ce résultat implique pour l'évaluation : **il n'y a toujours pas de lien entre performance et connaissances. Il y a toujours l'interprétation de la tâche entre elles**. Mais cette expérience montre quelque chose de plus : l'interprétation de la tâche contient la représentation que l'élève se fait de lui-même, ses croyances dans sa capacité à réussir ou non la tâche, qui dépendant de la « valeur scolaire » de la discipline. Ceci a été largement confirmé depuis par les expériences conduites autour de l'effet de « menace du stéréotype » (voir la synthèse de Croizet & Leyens, 2003).

Synthèse
- Si un élève ne réussit pas une tâche qui implique la connaissance A, cela ne permet pas de conclure qu'il n'a pas la connaissance A.
- Si un élève réussit une tâche qui implique la connaissance A, cela ne permet pas de conclure qu'il a la connaissance A.
- Seule la multiplication des tâches et des modalités de présentation de la tâche permet une approximation du fait que l'élève maîtrise ou pas la connaissance A.
- Deux élèves peuvent réussir (ou rater) la même tâche pour des raisons différentes.
- Il peut être intéressant d'évaluer par les compétences, mais avec une définition stricte de la notion de compétence : la compétence A' correspond au fait de réaliser régulièrement la tâche A dans telles conditions avec tel niveau de performance. Alors, si un élève réalise régulièrement la tâche A dans telles conditions avec tel niveau de performance, on peut conclure qu'il maîtrise la compétence A'.

1.3.2. L'importance des formats de connaissances dans l'évaluation

Dans le chapitre 2, nous avons défini la notion de format de connaissance. Nous avons voulu attirer l'attention sur le fait que, pour réaliser une même tâche, deux individus différents pouvaient mobiliser deux connaissances de format différent (par exemple, dans une tâche de diagnostic, l'un mobilise une méthode pendant que l'autre mobilise un automatisme). Dans les exemples que nous avons donnés, nous avons montré que, selon le format de la connaissance mobilisée, la tâche pouvait être réussie ou non. Il nous semble donc important, lorsque l'on conçoit une tâche pour évaluer une connaissance, d'être bien au clair sur le format de cette connaissance. Ainsi, concevoir une évaluation, c'est concevoir une tâche qui correspond précisément au contenu et au format de la connaissance que l'on veut évaluer.

Dans le chapitre 2, nous avons décrit les liens qui associaient certaines tâches à certains formats de connaissance. Ceci peut-être utilisé comme cadre pour concevoir des tâches d'évaluation. Une erreur dans la conception des évaluations consiste simplement à proposer une tâche correspondant au format *x* alors qu'on veut évaluer une connaissance du format *y*. Par exemple, on propose une tâche de rappel exact alors qu'on veut évaluer si les élèves ont compris, i.e. s'ils ont élaboré une connaissance particulière.

Si la connaissance à évaluer est un **automatisme**, c'est sans doute un exercice qui peut être utilisé comme tâche. Si l'on propose une tâche différente de celles utilisées lors de l'apprentissage, on prend le risque d'un diagnostic de type faux-négatif : on infère que l'élève n'a pas élaboré l'automatisme alors qu'il ne l'a simplement pas déclenché.

Si la connaissance à évaluer est un **savoir-faire**, la tâche est la même que pour l'automatisme. C'est l'indicateur qui sera différent : avec l'automatisme, le temps sera un critère important alors qu'avec le savoir-faire ce n'est pas le cas. Réciproquement, on ne peut raisonnablement pas demander à un élève d'élaborer une explication du déroulement d'un automatisme alors que c'est souvent intéressant avec un savoir-faire.

Si la connaissance à évaluer est une **méthode**, c'est sans doute un problème complexe ou une tâche de production qui peut être utilisée. Si l'on veut véritablement évaluer la méthode, c'est une tâche analogue à celles utilisées lors de la phase d'apprentissage qu'il faut utiliser. Dans le cas contraire, si l'on utilise une tâche trop nouvelle, on prend le risque de ne pouvoir savoir si les élèves qui ne réussissent pas la tâche ne maîtrisent pas la méthode ou n'ont pas pensé à la mobiliser. Une solution est alors de proposer une tâche nouvelle en indiquant la méthode qui doit être utilisée. Les explicitations sont souvent très pertinentes à recueillir (pourquoi cette méthode ? pourquoi cette étape ?). Une autre solution est de proposer des problèmes résolus (correctement ou pas) : des tâches de diagnostic.

Si la connaissance à évaluer est un **concept**, deux grandes voies peuvent être suivies. On peut tout d'abord proposer une tâche de description du concept : sa dénomination, ses attributs ou propriétés, sa structure, ses relations externes, son évolution, des exemples. Cependant, on court ici le risque d'un diagnostic faux-positif : un élève qui réussirait cette tâche en apprenant par cœur ce que l'on a vu en cours à propos de ce concept n'a pas forcément réussi à véritablement conceptualiser. Une autre voie peut donc être de proposer des situations de mobilisation du concept, pour comprendre une situation, analyser un cas, résoudre un problème par exemple. Si cette seconde voie permet de contrer l'obstacle de la mémorisation littérale, elle pose une autre difficulté, que nous avons déjà abordée : le risque du diagnostic faux-négatif, *i.e.* le risque de conclure que l'élève ne maîtrise pas le concept alors qu'il ne l'a simplement pas mobilisé. C'est pourquoi il est plus prudent d'indiquer quel concept doit être mis en œuvre pour comprendre telle situation lors d'une évaluation.

Si l'on veut évaluer une **connaissance spécifique** la tâche d'évaluation ressemble fort à cette de l'évaluation d'un concept, avec les mêmes obstacles. Ici cependant, la tâche de compréhension d'une autre situation que celle vue lors de l'apprentissage devient une tâche de transfert… fort risquée pour une évaluation (risque de faux-négatif très élevé).

Si l'on veut évaluer une **trace littérale**, c'est une tâche de rappel exact qui est sans doute la plus pertinente. Si l'on veut vérifier que l'élève a élaboré une trace littérale de la définition de x, la tâche de rappel exact peut correspondre à la consigne : « dites-moi exactement la définition de x ».

Nous allons revenir plus en détail sur ce point dans la suite du présent chapitre. Il nous semble que ce que nous avons souligné précédemment permet en tous cas de renoncer aux tâches de transfert pour réaliser une évaluation.

1.3.3. Quels sont les critères de réussite ?

Récapitulons le raisonnement qui est à l'œuvre dans une évaluation scolaire. Notre but est de juger si un élève a bien élaboré la connaissance-but de l'apprentissage. On conçoit une tâche pour laquelle il est nécessaire de mettre en œuvre la connaissance-but de l'apprentissage. Cette tâche ne présente pas d'autres difficultés que la mise en œuvre de la connaissance. Cette tâche présente de façon explicite le fait que la connaissance doit être mobilisée. On infère alors que si l'élève réalise la tâche c'est un bon indicateur du fait qu'il maîtrise la connaissance… sans en être complètement sûr, car il se peut que l'élève ait réalisé la tâche en utilisant une autre connaissance. En outre, on n'infère pas du fait que l'élève ne réalise pas la tâche… qu'il ne maîtrise pas la connaissance. Voyons maintenant quels critères peuvent être définis pour apprécier la réalisation de la tâche.

L'atteinte du but de la tâche ? Le premier critère utilisé est simplement le fait que l'élève réalise la tâche conçue pour l'évaluation. On se préoccupe peu de la manière, on admet qu'il a très bien pu réaliser la tâche en utilisant une autre connaissance. Seul le résultat compte ici. Ce type de critère, exclusivement focalisé sur le résultat, présente un inconvénient important, notamment dans les tâches de résolution de problème : on peut passer à côté d'un résultat « faux à cause d'un détail ». L'élève a mobilisé la connaissance visée, il l'a correctement mise en œuvre, mais une erreur de calcul, une erreur dans la simple recopie de son résultat, nous conduit à conclure qu'il n'a pas réussi la tâche (ce qui est vrai) et à en inférer qu'il ne maîtrise pas la connaissance (ce qui est totalement faux). C'est la raison pour laquelle la plupart des enseignants n'utilisent pas strictement ce critère.

La façon d'atteindre le but ? Le second critère, sans exclure l'atteinte du but de la tâche, se focalise sur la manière d'atteindre le but, sur les étapes qui ont été franchies, sur la stratégie utilisée, sur le temps mis pour réaliser la tâche, voire sur les tâtonnements, le nombre d'essais infructueux, etc. L'utilisation de ce type de critère peut être extrêmement intéressante, renseigner beaucoup sur les connaissances des élèves et sur les apprentissages qui sont encore nécessaires pour atteindre le but d'apprentissage visé. Pour utiliser ce type de critère, il est important d'être au clair avec non seulement la connaissance-but de l'apprentissage et le but de la tâche d'évaluation, mais aussi avec la manière optimale de réaliser la tâche. Sans ce référent, explicite et justifié, il devient difficile par définition de conduire un jugement évaluatif. Il est en outre nécessaire de disposer des indicateurs qui permettront de mesurer précisément l'écart entre ce modèle optimal de réalisation de la tâche et la façon dont l'élève aura effectivement réalisé la tâche. Ceci ne nous oblige pas à rester bloqués sur ce modèle optimal de réalisation de la tâche. On peut par exemple découvrir grâce à un élève une façon de faire originale, intéressante, à laquelle on n'avait pas pensé. Mais cette explicitation préalable est nécessaire, car, sans elle, cette référence existera de façon implicite. Sans cette explicitation, nous serions en quelque sorte conduits à faire un jugement comme celui-ci : « s'il fait comme moi alors c'est correct ; s'il fait différemment alors ce n'est pas correct ».

La réalisation d'une activité ? A l'opposé de ce que nous venons de décrire, certaines approches de l'évaluation se veulent ouvertes : le but de la tâche est très large, très ouvert, et il n'y a pas de modèle optimal de réalisation de la tâche. Cela peut être le cas quand la tâche relève de la production ou de la résolution de problèmes ouverts. On peut alors être conduit à choisir des critères relativement formels ou externes (par exemple : le document de présentation contiendra tant de pages, il commencera par une introduction, etc). On peut aussi renoncer à l'utilisation de critères explicites et connus *a priori*. Mais cela requiert de notre part énormément de prise de recul et de maîtrise, pour ne pas prendre le risque d'être soi-même la référence utilisée dans l'élaboration du jugement.

1.4 Quels sont les principaux biais d'évaluation ?

Nous avons présenté dans ce chapitre un certain nombre de problèmes liés à la pertinence et à la fiabilité de l'évaluation. Nous nous focalisons maintenant sur les biais de jugement et de notation qui persistent, malgré une tâche d'évaluation et des critères bien définis (Bonniol, Caverni & Noizet, 1972 ; Noizet & Caverni, 1978 ; Merle, 1996, 1998, 2007). On peut en effet voir des fluctuations importantes (de 4 à 6 points sur 20 en moyenne) dans le résultat de l'évaluation quand une même copie est évaluée avec les mêmes critères et le même barème par des enseignants différents, et parfois par le même enseignant… quand la même copie est glissée dans deux tas différents (Aymes, 1979).

La place de la copie. Quand nous évaluons successivement plusieurs copies d'élèves, notre jugement fluctue selon la place de la copie. En effet, si une même copie est placée dans le premier tiers du tas de copie elle recevra généralement une moins bonne note que si elle est placée dans le second et surtout le troisième tiers. Plus encore, quand une copie moyenne est placée juste après une excellente copie, nous avons tendance à être plus exigeants avec cette deuxième. Réciproquement, nous serons plus indulgents si cette copie moyenne est corrigée juste après une très mauvaise copie.

L'effet de contexte. Une même copie, et plus largement un même élève, ne seront pas notés de la même manière s'ils sont dans une classe (dans un établissement) où globalement les autres élèves sont meilleurs : on aura alors tendance à être plus exigeant. Réciproquement, nous serons moins exigeants avec cette copie ou cet élève s'il fait partie d'une classe (d'un établissement) où globalement les autres élèves sont moins performants.

L'effet de réputation du contexte. Quand un élève change d'établissement, il a tendance à bénéficier ou à subir la réputation de son établissement d'origine. Nous avons tendance à sur-noter

les élèves issus d'établissements réputés et à sous-noter ceux qui sont issus d'un « mauvais établissement ». Cet effet est exactement à l'intersection entre l'effet de contexte que nous venons de voir et l'effet de halo que nous allons aborder maintenant.

L'effet de halo. Le fait d'avoir une bonne opinion sur un élève pour une raison ou une autre (il ou elle est en avance sur son âge, ses frères et sœurs étaient de bons élèves, nous avons une bonne opinion de ses parents, des collègues nous on dit qu'il était bon élève, voire même il ou elle est beau) a un effet positif sur sa note. A l'inverse, avoir une mauvaise opinion sur un élève a un effet généralement négatif sur sa note. Certains d'entre nous, conscients de ce biais très connu depuis 40 ans, ont tendance à corriger voire à sur-corriger ce biais et à attribuer une meilleure note à un élève qui nous semble souffrir d'une mauvaise réputation (notons que nous corrigeons l'opinion d'autrui, pas la notre propre). Notons que cet effet est renforcé par l'effet de menace du stéréotype dont nous avons parlé plus haut : l'élève qui subit un effet de halo négatif a tendance à produire de moins bonnes performances quand il sait que son évaluateur connaît sa mauvaise réputation. Au contraire, s'il l'on parvient à inhiber la conscience que l'élève a de sa mauvaise réputation, l'effet négatif sur les performances disparaît.

L'effet du genre. Dans l'enseignement primaire et au début de l'enseignement secondaire, les filles bénéficient d'une bonne réputation qui peut produire un effet de halo positif. Cependant, quand on passe dans la seconde partie de l'enseignement secondaire puis dans l'enseignement supérieur, cet effet de halo peut se renverser sous l'effet de stéréotypes sexistes. Certains enseignants hommes attribuent moins de qualités aux filles dans le domaine des sciences en général et des mathématiques en particulier. En conséquence, une même copie attribuée à un garçon aura en moyenne une meilleure note que si elle est attribuée à une fille… quand on la donne à corriger à plusieurs correcteurs masculins. Espérons que ce résultat date d'une autre époque et qu'il ne sera plus vrai dans quelques années. En attendant ces jours meilleurs, l'effet de menace du stéréotype sexiste fonctionne très bien autour de la « capacité d'abstraction », du « niveau en maths », du « niveau en sciences », etc. à l'encontre des filles.

L'effet de l'âge de l'enseignant. En début de carrière, certains d'entre nous ont tendance à être très exigeants, cette exigence s'assouplissant quelque peu avec les années d'expérience.

La précision des critères. Un barème très précis, où chaque question est notée au point ou au demi-point, ne protège pas des biais d'évaluation. C'est même tout le contraire. Une même copie corrigée avec un barème au point (chaque question est divisée en sous-questions pour atteindre le barème : une question = un point) génère plus de différences qu'un barème large (les grandes questions sont notées avec un barème à 4 ou 5 points). La raison est liée au fait qu'avec un barème au point nous sommes plus enclins à attribuer 0 ou 1 à une question. Avec un barème large, nous hésitons plus à attribuer 0 ou 5. Or une somme de vingt 0 est égale à 0…

Le piège de la réponse attendue. Quand nous avons précisément une idée de la tâche que l'élève doit réaliser, de la manière dont il doit la réaliser, nous nous fermons parfois à la possibilité qu'une autre réponse correcte existe, qu'une autre manière de faire est pertinente. Nous avons tendance alors à juger non pas du fait que la réponse de l'élève soit correcte ou incorrecte, mais du fait qu'elle correspond ou pas à notre attente. Cet effet est particulièrement contre-productif quand nous attendons comme réponse un mot, une phrase précise, bouchant nos oreilles à tout autre mot. L'élève a bien compris que nous attendions un autre mot et que nous étions devenus subitement sourds. Parfois il ou elle se met alors à procéder au hasard, espérant qu'un mot miracle nous guérira de notre surdité.

« Etouffer sous la pression »
ou comment les bons élèves peuvent parfois pâtir encore plus que les autres des situations d'évaluation
Sian Beilock (2010) a découvert un effet paradoxal, intéressant et contre-intuitif, connu sous le nom de « choking under pressure » *: quand on compare des élèves « bons » avec des élèves « faibles »*

dans deux conditions, « sous pression » ou « sans pression », on voit que les élèves faibles ont une performance assez peu détériorée par la pression. Au contraire, les élèves « forts » voient leurs performances détériorées par la pression. La série d'expériences qu'elle a consacrée à cet effet montre que les élèves « forts », qui sont connus pour pouvoir utiliser des stratégies sophistiquées pour résoudre des problèmes difficiles, sont en situation « sans pression » tout à fait capables d'utiliser ces stratégies sophistiquées. Quand on introduit une pression dans la situation (pression sociale par exemple), ils vont avoir tendance à préférer des stratégies plus simples, y compris quand celles-ci sont en réalité inappropriées. Beilock a aussi montré que le simple fait de demander aux élèves d'écrire pendant 10 minutes un texte à propos de la pression qu'ils ressentaient… faisait disparaître l'effet négatif de la pression. Pour Beilock, on aurait là un effet d'interférence d'une émotion négative sur nos ressources attentionnelles, et, comme c'est souvent le cas avec les émotions négatives, le fait d'écrire à leur propos permet de diminuer l'intensité de l'interférence qu'elles provoquent.

1.5 Comment évaluer ?

Il existe quatre grandes façons d'évaluer les apprentissages.

De façon non intrusive. Nous pouvons évaluer les apprentissages pendant que les élèves apprennent, pendant qu'ils travaillent. L'évaluation ne correspond pas alors à une tâche spécifique attribuée à l'élève. C'est l'enseignant qui s'attribue comme tâche d'évaluer, alors que l'élève reste sur sa tâche d'apprentissage. Cette évaluation non-intrusive peut être mise en œuvre tout au long de la journée, de la semaine et de l'année. Elle est particulièrement privilégiée dans les classes où les élèves sont très jeunes, mais il n'y a aucune raison de ne pas la généraliser. On a pu penser à une certaine époque que ces évaluations étaient moins objectives que les autres, mais la connaissance que nous avons aujourd'hui des biais d'évaluation nous conduit à penser que les évaluations non intrusives ne sont pas plus biaisées que les autres. Au moins, avec cette approche, les élèves consacrent-ils plus de temps à apprendre et moins de temps à être évalués. L'enseignant consacre plus de temps à concevoir son enseignement et moins de temps à corriger des copies.

De façon intrusive. La façon la plus classique d'évaluer les apprentissages consiste à prescrire des tâches spécifiques pour l'évaluation. Les élèves arrêtent d'apprendre pour être évalués. Ce peut même être perçu comme une perte de temps pour les apprentissages. Il est alors nécessaire d'être très clair sur le bénéfice qu'il y a à tirer de cette évaluation.

De façon standardisée. Une autre pratique d'évaluation, toujours intrusive, consiste à utiliser des outils d'évaluation que nous n'avons pas nous-mêmes conçus. Par exemple, nous faisons passer un « examen blanc » à nos élèves. Cette façon d'évaluer est très utile quand elle remplit la fonction de situer chaque élève par rapport aux attendus d'une échéance prochaine. Pour les enseignants, cette approche est tout aussi utile, non pas pour situer un élève mais une classe ou pour évaluer l'efficacité de leur enseignement. Par exemple, pour sortir du domaine des strictes connaissances scolaires de l'examen blanc, je peux vouloir évaluer la motivation de mes élèves, ou bien leurs connaissances métacognitives. Il est très utile alors d'utiliser un outil non seulement déjà fait mais standardisé : je pourrais ainsi situer ma classe ou mon enseignement par rapport à une référence externe.

Un exemple d'évaluation standardisée utile

Il y a quelques années, l'un entre nous a été sollicité par les enseignants d'une école qui souhaitaient améliorer leurs façons de travailler. Ils étaient en effet très ennuyés par le fait que leurs élèves avaient globalement un faible niveau et leur école globalement une mauvaise réputation. Avant de commencer à travailler ensemble, nous avons simplement comparé les résultats des élèves sortant de cette école avec ceux des élèves sortant des autres écoles de la même circonscription. Nous avons utilisé les résultats sur cinq années à l'évaluation à l'entrée en 6ème : la même évaluation standardisée passée par l'ensemble des élèves de la même circonscription. Une belle surprise nous attendait : tous

les résultats montraient que depuis des années cette école avait les meilleures performances de la circonscription. Ce que les outils internes d'évaluation empêchaient de voir.

De l'intérêt de l'autoévaluation. La dernière forme d'évaluation consiste à impliquer les élèves eux-mêmes dans le processus d'évaluation. On peut éventuellement aider les élèves à réaliser cette autoévaluation, notamment pour les élèves les plus jeunes ou les plus en difficulté. On peut aussi proposer aux élèves de confronter leur autoévaluation à l'évaluation que nous avons faite. Mettre en œuvre des situations d'autoévaluation contribue non seulement au développement de la capacité métacognitive à s'auto-évaluer mais aussi au développement de la motivation centrée sur la maîtrise et l'apprentissage.

De l'intérêt d'un contrat clair. Dans le cas où l'on décide qu'il est nécessaire de faire une évaluation de type intrusive, il semble important d'établir un contrat clair avec les élèves, leur indiquant avant l'évaluation :

- Qu'est-ce qui va être évalué ?
- Avec quelle activité ?
- Selon quels critères ?

1.6 Les principaux diagnostics

Une fois que l'évaluation est réalisée, nous nous demandons généralement ce qui peut en être exploité, quelle information est fournie, quel diagnostic peut être fait ? Cela demande un effort de notre part, que nous oublions parfois de faire, nous arrêtant à la performance de l'élève. Voici donc une liste des principaux diagnostics qui peuvent être inférés d'une évaluation :

L'élève ne travaille pas. C'est le cas quand l'élève n'a pas réalisé la tâche, n'a pas essayé de la réaliser et ceci était manifeste avant même la tâche d'évaluation. On ne sait pas pourquoi l'élève ne travaille pas.

L'élève est persuadé qu'il ne peut pas y arriver. L'élève n'a pas essayé de réaliser la tâche, ou a abandonné très rapidement. Quand on l'interroge à ce propos, il explique qu'il ne se sent pas capable, qu'il est « nul » dans cette discipline, ou qu'il n'a pas du tout compris cette notion. On ne sait pas pourquoi il pense cela à propos de lui-même.

L'élève n'a pas la connaissance nécessaire à la réalisation de la tâche. L'élève essaie de réaliser la tâche mais n'y parvient pas. Quand on lui dit que pour réaliser cette tâche il faut mobiliser telle connaissance, cela ne change rien à la situation. Quand on l'interroge à propos de cette connaissance, il n'est pas capable d'en dire plus. On ne sait pas pourquoi il n'a pas élaboré cette connaissance, mais en continuant le dialogue, notamment sur ce qu'il sait dans ce domaine, on peut commencer à émettre des hypothèses sur la cause de cette méconnaissance.

L'élève a la connaissance mais ne parvient pas à la mobiliser. L'élève essaie de réaliser la tâche mais n'y parvient pas. Quand on lui dit que pour réaliser cette tâche il faut mobiliser telle connaissance, cela change tout. Il est maintenant capable de réaliser la tâche. Cette situation peut correspondre à des degrés différents de difficulté de mobilisation. Le degré le plus léger est représenté par « le mot sur le bout de la langue » : le simple fait de dire un mot à l'élève lui permet de mobiliser la connaissance. A l'autre extrémité se trouvent les cas où l'élève n'a pas seulement une difficulté à identifier la connaissance pertinente (le format déclaratif de la connaissance) mais aussi une difficulté à la mettre en œuvre (la forme procédurale de la connaissance n'est manifestement pas complètement élaborée).

L'élève mobilise une autre connaissance à la place. L'élève essaie de réaliser la tâche mais n'y parvient pas car la connaissance qu'il mobilise n'est pas pertinente. Dans cette situation, le diagnostic peut s'orienter vers un élément très superficiel : c'est un mot de l'énoncé, un simple aspect de la consigne, qui l'a conduit à mobiliser une connaissance inappropriée. C'est un diagnostic

extrêmement fréquent. L'autre situation est plus difficile, car il s'agit cette fois d'un élément profond : ce n'est pas un mot de l'énoncé mais bien le contenu complet de l'énoncé qui a orienté l'élève vers une connaissance inappropriée. Il se peut alors que le domaine de validité de la connaissance soit erroné.

L'élève mobilise la bonne connaissance mais fait une erreur de mise en œuvre. L'élève essaie de réaliser la tâche et mobilise une connaissance pertinente. C'est au cours de la mise en œuvre de la connaissance qu'une erreur est produite. Ceci est généralement dû à un problème attentionnel, qui peut correspondre soit à une interférence (l'attention de l'élève a été attirée par autre chose au mauvais moment) soit au fait que la mise en œuvre de la connaissance est encore très coûteuse, la connaissance n'est pas encore suffisamment automatisée, voire simplement pas encore suffisamment maîtrisée.

L'élève évalue mal l'atteinte du but. L'élève essaie de réaliser la tâche et mobilise une connaissance pertinente. Mais il arrête avant d'avoir fini, ou, au contraire, continue alors que le but a été atteint. Le diagnostic peut alors s'orienter vers le versant cognitif (on est en début d'apprentissage, la connaissance n'est pas encore bien maîtrisée, en particulier l'appariement situation – procédure) soit vers le versant métacognitif (l'élève ne sait pas encore que quand on a fini une tâche il faut relire la consigne et vérifier que ce qui a été atteint correspond bien à ce qui était attendu).

Synthèse « Comment évaluer les apprentissages ? »
Ce chapitre sur l'évaluation abordait un des aspects les plus délicats et les plus difficiles du métier d'enseignant. Voici ce qui, à notre sens, peut-être retenu de ce chapitre :
- Être très prudent avec les évaluations en général et les notes en particulier (ne jamais acquérir de certitude à ce propos si ce n'est que nous sommes encore plus faillibles dans ce domaine que dans les autres).
- Un diagnostic n'est jamais qu'une hypothèse de travail, qui ne demande qu'à être réfutée, à tout moment.
- Bien connaître les biais d'évaluation et de jugement.
- Ménager la plus grande place aux évaluations non-intrusives, par observation.
- Contractualiser clairement les évaluations intrusives.
- Utiliser l'évaluation pour aider les élèves à apprendre, pour réguler leur apprentissage et notre enseignement.
- Encourager la conception de l'erreur comme inhérente à l'apprentissage.
- Multiplier les formes d'évaluation pour véritablement diagnostiquer.
- Etre clair sur le contenu et sur le format de la connaissance visée par l'évaluation.
- Utiliser l'évaluation comme moteur de progrès et de motivation.
- Centrer le retour, fait à partir de l'évaluation vers l'élève, sur la connaissance plutôt que sur la performance.
- Faire des évaluations privées, non publiques.

Pour aller plus loin

Antibi, A. (2003). La Constante macabre. Toulouse : Math'Adore.

Bressoux, P. & Pansu, P. (Eds.), (2003). Quand les enseignants jugent leurs élèves. Paris : PUF.

Butera, F., Buchs, C., & Darnon, C. (Eds.), (2011). L'évaluation une menace ? Paris : PUF.

Merle, P. (1998). Sociologie de l'évaluation scolaire. Paris : PUF.

Talbot, L. (2009). L'évaluation formative. Comment évaluer pour remédier aux difficultés d'apprentissage ? Paris : Armand Colin.