Mesure et évaluation en éducation

La validation des épreuves d'évaluation selon l'approche par les compétences

Jean-Marie De Ketele and François-Marie Gerard

Volume 28, Number 3, 2005

URI: https://id.erudit.org/iderudit/1087028ar
DOI: https://doi.org/10.7202/1087028ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (print)
2368-2000 (digital)

Explore this journal

Cite this article

De Ketele, J.-M. & Gerard, F.-M. (2005). La validation des épreuves d'évaluation selon l'approche par les compétences. *Mesure et évaluation en éducation*, *28*(3), 1–26. https://doi.org/10.7202/1087028ar

Article abstract

The competence approach leads to the assessment of students achievements through complex situations that require a complex production (response) from them. The classic techniques in evaluation tests validation can no longer be used as it has been done until now. Other validations approaches, either a priori or a posteriori, should be implemented respecting the validity and fiability requirements.

La validation des épreuves d'évaluation selon l'approche par les compétences

Jean-Marie De Ketele
UCL – BIEF

François-Marie Gerard
BIEF

Mots clés : Évaluation des acquis, évaluation des compétences, pertinence, validité, fiabilité, approche par les compétences

L'approche par les compétences conduit à évaluer les acquis des élèves à travers des situations complexes nécessitant une production complexe de la part de l'élève. Les techniques classiques de validation des épreuves d'évaluation ne peuvent dès lors pas être utilisées telles quelles. D'autres approches de validation, qu'elles soient a priori ou a posteriori, devraient être adoptées, dans le respect des exigences de validité et de fiabilité, mais aussi et avant tout de pertinence.

Key words : Evaluation of achievements, evaluation of competences, relevance, validity, fiability, competence approach

The competence approach leads to the assessment of students achievements through complex situations that require a complex production (response) from them. The classic techniques in evaluation tests validation can no longer be used as it has been done until now. Other validations approaches, either a priori or a posteriori, should be implemented respecting the validity and fiability requirements.

Palavras-chave : Avaliação dos conhecimentos, avaliação das competências, pertinência, validade, fiabilidade, abordagem por competências

A abordagem por competências conduz a avaliar os conhecimentos dos alunos através das situações complexas sendo necessária uma produção complexa por parte do aluno. Isto implica que as técnicas clássicas de validação das provas de avaliação não possam continuar ser utilizadas como até então. Outras abordagens de validação, seja a priori, seja a posteriori, deverão ser adoptadas, no respeito pelas exigências de validade e fiabilidade, mas também e acima de tudo pela pertinência.

Note des auteurs : Toute correspondance peut être adressée par courriel à l'adresse suivante : [jean-marie.deketele@psp.ucl.be]

Un peu partout dans le monde, les pratiques pédagogiques se réfèrent de plus en plus à une approche par compétences se différenciant d'une approche par contenus ou par objectifs :

- l'approche par contenus considère l'enseignement en termes de liste de matières et de contenus-matières à enseigner, c'est-à-dire à transmettre ;
- l'approche par objectifs a comme porte d'entrée des comportements observables structurés, mais séparés les uns des autres, qui sont à développer chez les apprenants ;
- l'approche par compétences cherche à développer la possibilité par les apprenants de mobiliser un ensemble intégré de ressources[1] pour résoudre une situation-problème appartenant à une famille de situations (De Ketele, 2000, 2001 a et b ; Dolz & Ollagnier, 2002 ; Fourez, 1999 ; Gerard, 2005 ; Jonnaert, 2002 ; Lasnier, 2000 ; Le Boterf, 1994 ; Legendre, 2001 ; Perrenoud, 1997 ; Poirier Proulx, 1999 ; Rey, 1996 ; Rey, Carette, Defrance & Kahn, 2003 ; Roegiers, 2000, 2003 ; Scallon, 2004 ; Tardif & al., 1995 ; Tilman, 2000).

Ces différentes approches entraînent des pratiques d'évaluation différentes (De Ketele & Dufays, 2003) :

- dans l'optique des contenus, évaluer consiste à prélever un échantillon de contenus représentatif de l'univers de référence des contenus enseignés ;
- dans l'optique des objectifs, évaluer consiste à prendre un échantillon représentatif d'objectifs spécifiques et opérationnels et à générer un échantillon de questions qui traduisent au mieux cet échantillon d'objectifs ;
- dans l'optique des compétences, évaluer consiste à proposer une ou des situations complexes, appartenant à la famille de situations définie par la compétence, qui nécessiteront, de la part de l'élève, une production elle-même complexe pour résoudre la situation.

Entre pertinence, validité et fiabilité

Les épreuves élaborées à des fins d'évaluation selon chacune de ces approches seront inévitablement différentes. Néanmoins, les questions qui se posent par rapport à ces épreuves sont les questions classiques relatives à tout recueil d'information : sont-elles pertinentes, valides et fiables (De Ketele & Roegiers, 1993) ?

- La pertinence est le caractère plus ou moins approprié de l'épreuve, selon qu'elle s'inscrit dans la ligne des objectifs visés (De Ketele, Chastrette, Cros, Mettelin & Thomas, 1989). C'est son degré de « compatibilité » avec les autres éléments du système auquel elle appartient (Raynal & Rieunier, 2003).
- La validité est le degré d'adéquation entre ce que l'on déclare faire (évaluer telle ou telle dimension) et ce que l'on fait réellement, entre ce que l'outil mesure et ce qu'il prétend mesurer (Laveault & Grégoire, 2002).
- La fiabilité est le degré de confiance que l'on peut accorder aux résultats observés : seront-ils les mêmes si on recueille l'information à un autre moment, avec un autre outil, par une autre personne, etc. ? Elle nous renseigne sur le degré de relation qui existe entre la note obtenue et la note vraie (Cardinet & Tourneur, 1985 ; Laveault & Grégoire, 2002). Il ne faut cependant pas perdre de vue que la note vraie est une abstraction, un point de convergence souhaité indépendant des évaluateurs et des circonstances.

Ces dimensions, théoriquement indépendantes les unes des autres, forment ensemble des conditions nécessaires pour disposer d'une épreuve d'évaluation digne de ce nom. Mais, prises isolément, elles ne sont jamais suffisantes : un outil fiable peut très bien n'être ni valide, ni pertinent, ou un outil pertinent ni valide ni fiable, etc.

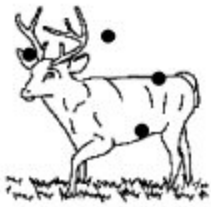
En adaptant l'exemple du tir à l'arc proposé par Laveault et Grégoire (2002), imaginons qu'on souhaite évaluer la compétence d'un trappeur à chasser du gibier grâce à son arc. Il doit pour ce faire être capable de réaliser un tir groupé sur une cible mouvante. Deux épreuves sont mises en place : d'une part, un tir de cinq flèches sur une cible classique, et, d'autre part, un tir de cinq flèches sur une cible mouvante semblable à du gibier. Les situations suivantes peuvent dès lors exister :

A. Non pertinent
Non valide
Non fiable

B. Non pertinent
Non valide
Fiable

C. Non pertinent
Valide
Non fiable

D. Non pertinent
Valide
Fiable



E. Pertinent
Non valide
Non fiable

F. Pertinent
Non valide
Fiable

G. Pertinent
Valide
Non fiable

H. Pertinent
Valide
Fiable

Dans la situation A, le tir est non fiable parce que dispersé sur toute la surface de la cible fixe. Il est non valide, parce que la mesure manque de précision et ne peut attester l'atteinte de l'objectif. Il est enfin non pertinent puisqu'il ne correspond pas à l'objectif recherché et ne mesure donc pas, de toute façon, ce que je déclare vouloir mesurer.

Dans la situation B, le tir est groupé mais rate systématiquement la cible. Il est fiable (la mesure montre que le tireur sait faire un tir groupé), mais non valide parce que cette mesure ne me permet pas de vérifier qu'il est capable d'atteindre la cible.

Dans la situation C, le centre de la cible, c'est-à-dire l'objectif, est atteint et la mesure est donc valide, mais elle n'est pas fiable parce qu'on ne peut certifier que ce sera toujours le cas.

Dans la situation D, le tir est groupé et touche le mille, mais la cible n'est pas la bonne (par rapport à la compétence visée). La mesure est donc fiable et valide, mais non pertinente.

Les situations E à H présentent des situations pertinentes par rapport à la compétence visée. Les situations E, F et G ont cependant des problèmes de validité ou de fiabilité, alors que la situation H détermine une mesure à la fois fiable, valide et pertinente. Elle seule permet de certifier que le trappeur est compétent.

La situation D est intéressante parce qu'elle reflète ce que sont les épreuves classiques, notamment celles utilisées dans les grandes enquêtes internationales du type TIMMS ou PISA, du moins lorsqu'elles s'inscrivent dans un système éducatif fondé sur l'approche par les compétences. Ces épreuves proposent un ensemble d'items selon une structure très élaborée issue principalement d'une approche par les contenus ou par les objectifs. Elles sont fiables et valides, mais ne sont pas pertinentes par rapport à l'approche par compétences et donc par rapport aux finalités des systèmes éducatifs conçus dans cet esprit. Si elles permettent bien d'évaluer les ressources jugées nécessaires (savoir-reproduire et savoir-faire), elles ne permettent pas (ou peu) d'évaluer la faculté de mobiliser celles qui sont pertinentes pour résoudre des problèmes ou effectuer des tâches complexes.

Il faudrait donc pouvoir disposer d'épreuves correspondant à la situation H, mais il est vraisemblable que de nombreuses épreuves actuellement élaborées sont plutôt proches des situations E ou F : elles devraient être pertinentes parce qu'elles proposent une situation complexe, mais il est difficile d'en attester la validité parce que les méthodes pour assurer cette validation ne sont pas évidentes ou tout simplement pas connues. Il n'est, en tous cas, pas possible de se servir telles quelles des méthodes utilisées dans la théorie classique des scores. Nous sommes conscients que cette position, et certaines propositions émises dans la suite de cet article, risquent de soulever quelques polémiques. Mais nous espérons que les questions que nous soulevons feront l'objet de travaux de la part des experts en évaluation.

Les limites de la théorie classique des scores

Comme l'a montré Cardinet, dès 1973, un premier élément à prendre en compte à cet égard est qu'il existe, de toute façon, une limite méthodologique importante lors de la validation d'épreuves d'évaluation pédagogique, quelle que soit l'approche à laquelle elles se réfèrent. Cette difficulté est liée au fait que l'on recourt à des techniques utilisées en *psychométrie* alors que l'on est dans le champ de l'*édumétrie* (Carver, 1974), qui ne dispose pas d'outils suffisamment mis au point, en cohérence avec la spécificité de l'éducation.

La différence est notamment liée à la distribution attendue des résultats, qui se concrétise dans une courbe de référence. Lorsqu'on utilise un test psychométrique, on s'attend à ce que la population cible soit distribuée selon une loi normale, ou courbe de Gauss, avec une minorité de notes basses et une autre minorité de notes élevées pour une majorité de notes moyennes, la « moyenne » étant elle-même confondue avec le mode et la médiane. Cette

logique « psychologique » est déterminée par le fait que l'on souhaite décrire une population et situer un individu dans la distribution, qui semble effectivement correspondre à une courbe de Gauss. Une telle distribution apparaît lorsque le phénomène étudié est la résultante d'un nombre important de facteurs en interaction ; elle disparaît lorsque le phénomène étudié est le fruit d'un effet systématique fort : en l'occurrence, l'intervention pédagogique est plus forte que les nombreux facteurs aléatoires individuels ou situationnels. La logique de l'éducation devrait donc être très différente parce qu'elle ne vise pas à décrire une population, mais à agir sur elle. L'éducation scolaire a pour objectif que les élèves apprennent et que **tous les élèves** apprennent. La distribution attendue au terme d'un processus d'enseignement-apprentissage ne devrait donc pas – en bonne logique – être « normale », mais devrait correspondre à ce qu'on appelle une courbe en J, c'est-à-dire où il y a une majorité d'élèves qui ont acquis les objectifs fondamentaux visés et une minorité d'élèves qui n'ont pas atteint ces objectifs. En quelque sorte, le « pédagogue » est celui qui va, grâce à son action éducative, faire *mentir* le « psychologue ». Cette logique n'est malheureusement pas toujours comprise ni prise en compte par les pédagogues eux-mêmes. Plusieurs recherches ont ainsi confirmé la loi dite de Posthumus, formulée dès 1947, selon laquelle les enseignants ont tendance, inconsciemment la plupart du temps, à recréer au sein de leur classe une distribution en cloche, ce qui traduit une logique sélective (Crahay, 1996 ; de Landsheere, 1980). Cependant, ce n'est pas la logique prônée, tant dans les mouvements pédagogiques actuels que dans de nombreux textes officiels. En effet, la logique de la « réussite » est clairement privilégiée, par exemple dans le cadre de la **pédagogie de la maîtrise** pour laquelle un objectif doit être maîtrisé par une forte proportion d'élèves (80 % à 90 %) avant de passer à l'objectif d'apprentissage suivant ou encore dans la cadre de l'approche par les **compétences de base** (De Ketele, 1993a et b ; Roegiers, 2000), qui définit un nombre très limité de compétences essentielles à la vie quotidienne et qui doivent dès lors absolument être maîtrisées par tous les élèves.

Les techniques classiques de validation des épreuves d'évaluation scolaire reposent, pour la plupart d'entre elles, sur une approche psychométrique, c'est-à-dire sur une distribution normale, en forme de courbe de Gauss. Par exemple, un indice fréquemment utilisé est l'indice de discrimination (Findley, 1956), qui vise à vérifier si les items de l'épreuve sont à même de discriminer les élèves, c'est-à-dire de distinguer les élèves faibles des élèves forts. Cet indice est la différence entre l'indice de difficulté de l'item pour le groupe dit « fort » et l'indice de difficulté pour le groupe dit « faible » : plus l'écart est grand, plus l'item discrimine entre les sujets. Le groupe fort est constitué des

sujets qui ont obtenu un score total qui les situe dans la catégorie des 27% supérieurs et le groupe faible dans la catégorie des 27% inférieurs (Laveault & Grégoire, 2002). On voit clairement non seulement que cette méthode se fonde sur une distribution normale, mais aussi que l'épreuve elle-même est conçue de telle sorte à déboucher sur une distribution normale des résultats afin de pouvoir discriminer au mieux, de la manière la plus fine possible, les élèves.

Selon De Ketele et Dufays (2003), la validité des épreuves d'évaluation fondée sur les contenus ou sur les objectifs se caractérise ainsi par :

- le caractère représentatif de l'échantillon d'items par rapport à l'univers de référence ;
- la position de la moyenne (pas trop éloignée du centre de l'échelle de notation) ;
- la dispersion des résultats (variance la plus grande possible pour discriminer au mieux les élèves) ;
- la forme de la distribution (courbe gaussienne) ;
- l'unidimensionnalité du trait mesuré (en l'occurrence, une aptitude scolaire spécifique).

Les limites des épreuves selon l'approche par les compétences

La première de ces caractéristiques nous permet d'aller plus loin dans la problématique de la validation des épreuves d'évaluation fondées sur l'approche par les compétences. De la même manière que l'on souhaite, lors d'un sondage d'opinion, obtenir un échantillon représentatif de la population en procédant par tirage aléatoire, le caractère représentatif de l'échantillon des items d'une épreuve d'évaluation est atteint au mieux lorsqu'il est constitué de manière aléatoire. C'est par exemple le cas lors de l'examen théorique du permis de conduire géré par ordinateur. Celui-ci propose aléatoirement une série de questions issues de l'univers de toutes les questions possibles à propos du code de la route (uniquement composé de « contenus »). L'échantillon aléatoire est éventuellement stratifié en contenant, par exemple, *x* questions relatives à des infractions graves. Tous les candidats reçoivent des questionnaires différents, et on pourrait donc considérer qu'ils n'ont pas tous la même chance face à l'examen. Mais c'est justement cette dimension aléatoire de la construction du questionnaire qui en garantit la représentativité et en assure la validité, pour autant évidemment que le nombre de questions soit suffisamment important pour limiter l'erreur de mesure. Si on s'en tient aux règles relatives à la représentativité d'un échantillon, cette condition est rarement satisfaite. En effet, l'erreur de mesure sur un échantillon peut être

estimée par la formule $\frac{1}{\sqrt{n}}$ où n est le nombre d'éléments contenus dans l'échantillon. Pour une erreur de mesure limitée à 5 %, il faudrait ainsi que l'épreuve d'évaluation contienne 400 items, ce qui est pratiquement infaisable dans la plupart des cas. L'exemple de l'examen du permis de conduire illustre bien la situation D de notre schéma où l'effort se concentre sur la fiabilité et la validité de l'évaluation des ressources au détriment de la pertinence : c'est la raison pour laquelle une épreuve pratique en situation est organisée et nous savons très bien qu'une réussite brillante à l'épreuve écrite peut s'accompagner d'une grave incompétence lors de l'épreuve pratique.

Les épreuves élaborées selon l'approche par compétences risquent cependant de présenter des difficultés de validité et de fiabilité, étant donné la taille de l'échantillon issu de l'univers de référence. Par essence, ces épreuves vont consister à présenter à l'élève une, voire deux situations complexes, demandant de la part de l'élève une production elle-même complexe, nécessitant un certain temps de résolution[2]. Les tâches demandées aux élèves varieront selon les situations :

- parfois la production sera constituée d'une tâche unique complexe, par exemple produire un texte répondant à une situation de communication ou, dans l'épreuve pratique du permis de conduire, parcourir un circuit routier choisi pour exiger la mobilisation d'un ensemble de ressources jugées essentielles ;
- parfois la production pourra être décomposée en plusieurs tâches, comprenant elles-mêmes diverses étapes, sans qu'il soit nécessairement possible de préciser le nombre exact d'étapes, comme c'est le cas par exemple dans la résolution de situations-problèmes mathématiques complexes où plusieurs voies de résolution sont possibles.

Mais, en tous les cas, il ne sera pas possible de constituer un nombre d'« items », parce que ceux-ci reviendraient à décomposer la tâche complexe en sous-unités, ce qui ne permettrait plus d'évaluer la compétence qui consiste bien à gérer la complexité. Il n'est donc pas question, dans des épreuves d'évaluation construites selon l'approche par les compétences, de valider les outils comme on le fait avec des épreuves classiques (fondées sur les contenus et les objectifs) en prenant comme unité d'analyse les « items » constituant l'échantillon issu de l'univers de référence. Dans une épreuve selon les compétences, ces « items » n'existent pas : il n'y a qu'une production complexe, ne constituant certainement pas un échantillon statistiquement représentatif de toutes les productions complexes possibles relativement à la compétence.

Les techniques de validation des épreuves d'évaluation selon l'approche par compétences sont donc à inventer, avec d'autres postulats que ceux de la théorie classique des scores.

Les validations des épreuves selon l'approche par les compétences

Divers types de validation peuvent être envisagées de façon complémentaire :

- une validation *a priori*, par recours à des juges ;
- une validation empirique interne ;
- une validation empirique externe.

Validation a priori par recours à des juges

La validation *a priori* par recours à des juges est une technique également utilisée dans les épreuves classiques. Dans ce cas, elle consiste à demander aux juges de vérifier si les items font bien partie de l'univers de référence. Dans le cadre de l'approche par les compétences, le travail sera du même type, mais ce sont les « items » et l'univers de référence qui sont fondamentalement différents. Les « items » sont les situations-problèmes concrètes qui sont proposées aux élèves et l'univers de référence est la famille de situations qui correspond à la compétence. Les juges devront donc vérifier que la ou les situations appartiennent bien à la famille de situations. C'est la question de l'équivalence des situations. Cette question est très difficile, voire impossible à résoudre. En effet, ces situations étant par définition complexes, il est très difficile de certifier qu'elles sont bien équivalentes les unes avec les autres, et qu'elles appartiennent donc effectivement à la famille de situations. La condition indispensable pour réaliser ce travail est de disposer d'une description suffisamment précise de la famille de situations, c'est-à-dire d'une définition des paramètres qui la caractérisent.

Le **paramétrage** de la famille de situations détermine les caractéristiques auxquelles doivent répondre toutes les situations appartenant à la famille. Il peut porter sur différents aspects :

- l'univers de référence quant aux ressources à mobiliser, lorsque par exemple toutes les situations doivent nécessiter dans leur résolution une soustraction suivie d'une addition, portant chacune sur des nombres entiers compris entre 0 et 1 000 ;

- le type de situations, par exemple des situations de communication orale impliquant deux personnes l'une en présence de l'autre (excluant ainsi une communication téléphonique) ;
- le type et le nombre de supports, par exemple la présence d'au moins un schéma ou d'un texte d'une centaine de mots ;
- le type de tâche attendue, par exemple la production d'un texte argumentatif d'au moins dix phrases, accompagné d'un schéma explicatif ;
- les conditions de résolution, dont notamment les aides mises à disposition de l'élève (un dictionnaire, un formulaire, etc.) ;
- le type de critères utilisés pour évaluer la production, par exemple pour une résolution de problèmes mathématiques (1) la pertinence des ressources mobilisées face au problème posé, (2) le caractère correct des ressources mobilisées (qu'elles soient pertinentes ou non), (3) le sens donné à la réponse finale dans le cadre du problème posé ;
- etc.

Le paramétrage de la famille de situations ne doit pas être confondu avec l'**habillage** des situations. Celui-ci concerne chaque situation particulière, qui présente chaque fois des aspects spécifiques qui font qu'elle est une situation parmi d'autres. L'habillage concerne le contexte spécifique de la situation, les données concrètes qu'elle contient, les supports spécifiques, etc. Il est donc différent pour chaque situation, alors que le paramétrage, lui, est commun à toutes les situations issues de la même famille, sans quoi elles ne seraient pas équivalentes.

La validation par les juges consistera donc à vérifier que les paramètres de la famille de situations sont bien présents dans la situation concrète, et seulement eux. L'équivalence des situations porte sur le paramétrage, et non sur l'habillage. La difficulté essentielle est de disposer d'une définition suffisamment précise et complète de la famille de situations. Sans elle, on ne pourra jamais garantir l'équivalence. On pourrait donc dire qu'avant encore de confronter les situations à l'avis des juges, le travail de validation des épreuves élaborées selon l'approche par les compétences consiste à vérifier qu'on dispose de paramètres suffisamment précis et complets, et à les élaborer si on n'en dispose pas.

Validations empiriques a posteriori

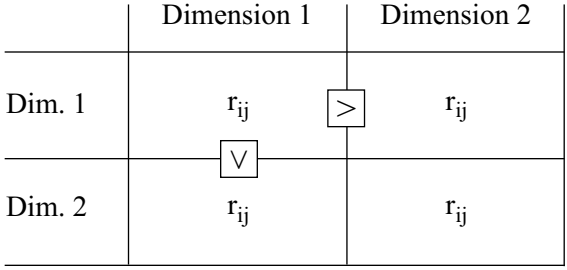
Une autre validation peut être effectuée *a posteriori*, sur la base des résultats obtenus. Classiquement, on distingue

- une validation empirique interne, si elle vise à vérifier l'unité conceptuelle de l'épreuve en elle-même,
- une validation empirique externe, si elle tend à montrer la convergence de l'épreuve avec d'autres épreuves similaires elles-mêmes validées ou encore à confronter les résultats avec d'autres épreuves qu'on sait avoir une relation inverse ou non avec la dimension évaluée.

Validation empirique interne

La validation empirique interne d'un test classique se fait essentiellement par le recours à deux types de techniques : l'analyse factorielle et le calcul de coefficients d'homogénéité (le plus souvent le coefficient alpha de Cronbach). Par l'analyse factorielle, on cherche à vérifier que, dans le cas où l'épreuve est censée mesurer une aptitude scolaire précisée, tous les items du test présentent des saturations fortes sur la première composante principale ou facteur extrait[3] et, en conséquence, contribuent à obtenir une variance expliquée très importante. Par contre, si l'épreuve est considérée comme composite (exemple : la maîtrise en mathématique est censée comporter deux dimensions, l'une algébrique, l'autre géométrique), l'analyse factorielle cherchera à vérifier si, au-delà d'une première composante principale forte représentant une aptitude mathématique générale, on trouvera deux facteurs après rotation, l'un représentant la dimension algébrique, l'autre la dimension géométrique. Le recours au coefficient alpha de Cronbach obéit aux mêmes préoccupations. Dans le premier cas de figure (tous les items sur une même dimension), on cherchera à vérifier que le coefficient obtenu est suffisamment élevé ; si tel n'est pas le cas, on cherchera à déterminer le ou les items qui ne mesurent pas l'aptitude à évaluer. Dans le second cas de figure (l'épreuve est censée mesurer plusieurs dimensions), on calculera autant de coefficients qu'il y a de dimensions (et donc de regroupements d'items).

Pour illustrer de manière simple ce type de raisonnement, et à des fins strictement heuristiques, nous nous référerons au calcul de corrélations, même si cette technique n'est pas parfaitement appropriée pour l'analyse d'items. Admettons qu'une épreuve évalue deux dimensions : la corrélation de tous les items relatifs à une même dimension devrait être supérieure à celle entre les items relatifs à une dimension et ceux relatifs à une autre dimension.



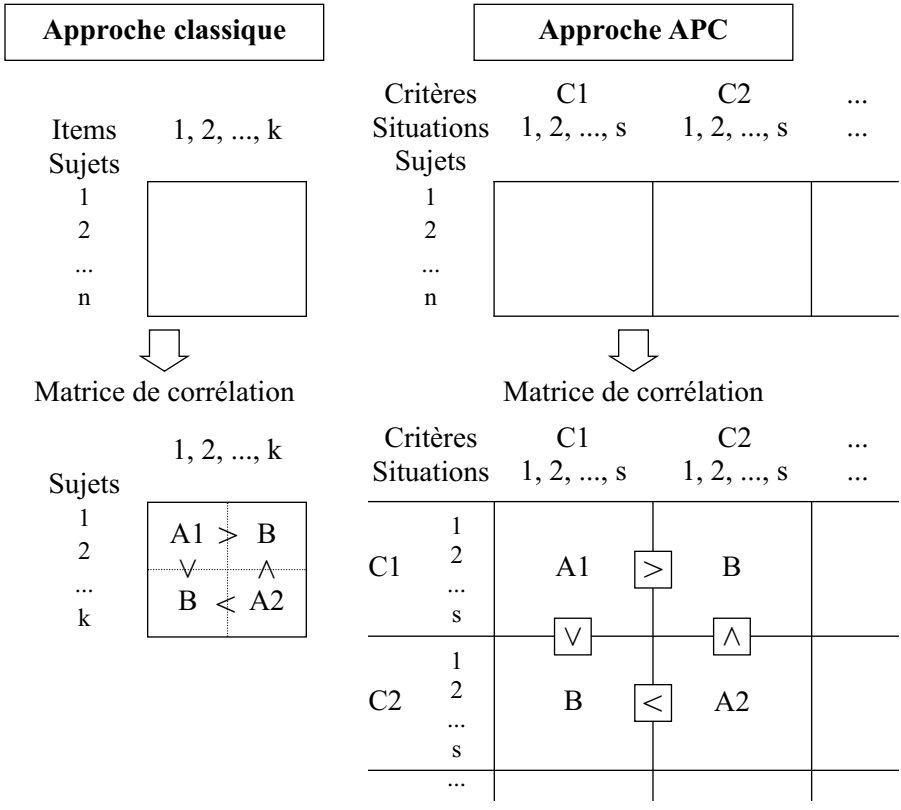
La difficulté de la validation empirique interne d'une épreuve élaborée selon l'approche par les compétences et constituée d'une ou de plusieurs situations complexes est qu'il n'existe ni de dimension unique, ni d'un ensemble dénombré d'items censés la mesurer puisqu'il s'agit d'une production complexe : il n'y a donc pas de stratégie de réponse à l'item (tâche unique complexe) permettant de fonder les analyses de corrélation.

La validation devrait pouvoir se fonder sur les critères d'évaluation, qui sont invariants aux situations d'une même famille, à l'inverse des indicateurs qui, pour la plupart d'entre eux, dépendent de chaque situation. Les critères sont les différents regards que l'évaluateur porte sur la production d'un élève ou les différentes qualités qui sont attendues de celle-ci. Dans l'approche classique des tests, il n'existe la plupart du temps qu'un seul critère : l'exac-titude de la réponse. Face à une situation complexe, il n'existe pas de réponse unique et donc de « bonne » réponse. Pour évaluer celle-ci, il faut se référer à des critères : La production est-elle pertinente par rapport à la situation ? Est-elle complète ? Les outils utilisés (algorithmes, formules, règles, etc.) sont-ils adéquats ? Sont-ils correctement utilisés ? La structure de la production est-elle cohérente ? La présentation de la production est-elle soignée ? etc. Chacun de ces critères pourra être opérationalisé dans une situation à travers des indicateurs qui sont des éléments directement observables, la plupart du temps spécifiques à la situation. Par exemple, le critère « Adéquation de la production à la situation de communication » sera vérifié si, face à une demande de renseignements relative à un itinéraire, l'élève indique effectivement un chemin à prendre, que celui-ci soit correct ou non, exprimé ou non dans le respect des règles grammaticales[4], etc.

Pour chaque compétence évaluée, il importe de déterminer les critères qui seront utilisés pour évaluer les productions des élèves (ceci fait partie du paramétrage). La validation des épreuves consistera à comparer diverses situations relatives à la même compétence : les élèves qui ont des scores élevés sur un

des critères dans une situation devraient avoir des scores élevés sur le même critère dans une autre situation appartenant à la même famille, et inversement. Un premier type de stratégies est donc de vérifier que ce principe est respecté pour chacun des critères.

Idéalement, il faudrait soumettre les mêmes élèves à un éventail de situations jugées équivalentes, c'est-à-dire paramétrées de la même façon. Mais on sait que les épreuves construites selon l'approche par compétences exigent du temps et que l'on ne peut guère, dans beaucoup de cas, imposer aux mêmes sujets plus d'une épreuve sans introduire des biais (fatigue, démotivation, effet d'entraînement, moment de la passation, etc.) difficilement dissociables. Cependant, il est des cas où l'on peut raisonnablement présenter aux mêmes élèves deux épreuves (voire plus) aux mêmes élèves. Dans ce cas, on peut recourir aux modèles corrélationnels, en sachant que l'on ne corrèle plus des items deux à deux, mais des situations deux à deux quant à la maîtrise du critère concerné. Dans le schéma suivant, on peut comparer la démarche de validation dans l'approche classique et dans l'approche par compétences.



Comme nous l'avons souligné antérieurement, on s'attend, dans l'approche classique, à obtenir des corrélations inter-items plus fortes dans les sous-matrices A1 et A2 (items appartenant à une même dimension) que dans la matrice B (items appartenant à des dimensions différentes). Dans l'approche par compétences, toutes les corrélations inter-situations à propos d'un même critère (A1 et A2) devraient être proches de 1 et supérieures à celles appartenant à des critères différents (B).

Comme il est souvent impossible de soumettre les mêmes sujets à des situations strictement équivalentes, on peut recourir à des stratégies détournées à partir de situations rencontrées dans les pratiques. En effet, l'apprentissage dans le cadre de l'approche par compétences implique de présenter aux élèves des situations progressivement plus complexes : l'élève mobilise dans une situation X+1 des ressources déjà mobilisées dans une situation X. Sous l'angle des critères d'évaluation, deux cas de figure peuvent se présenter selon que les critères communs aux situations X et X+1 sont de même niveau d'exigence ou non (exemple pour ce dernier cas : dans le critère « correction orthographique », on prend aussi en considération des règles plus complexes).

Imaginons deux situations d'évaluation S_x et S_{x+1} où C1(C2) de S_x est un critère de même niveau d'exigence que C'1 (C'2) de S_{x+1}, où C3 et C4 sont soit deux critères spécifiques à la situation S_{x+1} ou bien représentent des niveaux d'exigence supérieurs à C1 et C2. Le schéma suivant représente respectivement les données à recueillir, les indices à calculer et les résultats attendus.

Situations	S_x		S_{x+1}			
Critères	C1	C2	C'1	C'2	C3	C4
Sujets						
1						
2						
...						
n						
Moy.						
SD						

⇩

Matrice de corrélation

Critères	C1	C2	C'1	C'2	C3	C4
C1						
C2						
C'1						
C'2						
C3						
C4						

Résultats attendus

Moy. C'1 ≥ Moy. C1
SD C'1 ≤ SD C1

Moy. C'2 ≥ Moy. C2
SD C'2 ≤ SD C2

Moy. C3 ou C4 ≤ Moy. C'1 ou C'2
SD C3 ou C4 ≥ SD C'1 ou C'2

r (C1, C'1) et r (C2, C'2) très élevées et supérieures aux autres corrélations

Comme indiqué dans le schéma, on doit s'attendre à une maîtrise des critères communs en S_{x+1} au moins égale à celle de S_x compte tenu de leur consolidation dans le temps. De même, les déviations standards ne devraient pas être supérieures en S_{x+1}, car une bonne appropriation des critères à travers le temps homogénéise les résultats. L'exemple suivant illustre bien ce phénomène : si l'on donne à un an d'intervalle des problèmes qui nécessitent le recours à une démarche d'addition, on doit s'attendre en S_{x+1} à une performance meilleure (moyenne des élèves plus élevée) de tous (SD plus faible). C'est ce qui explique aussi que les corrélations sur un même critère entre les deux moments doivent être plus fortes que si les critères sont différents.

Validation empirique externe

Les stratégies que nous venons de présenter sont des tentatives de validation empirique interne. Mais il peut être intéressant de recourir à des stratégies complémentaires de validation empirique externe, c'est-à-dire en recourant à des critères externes. De nombreuses possibilités existent, qui peuvent être classées selon que les dispositifs impliquent des études comparatives ou corrélationnelles.

Les études comparatives consistent à comparer un groupe expérimental (approche par compétences) à un groupe contrôle équivalent. Deux grands cas de figure peuvent exister pour constituer le groupe contrôle selon que l'approche par compétences est ou n'est pas généralisée. Lorsqu'elle n'est pas généralisée, il s'agit de constituer un échantillon comparable à l'échantillon expérimental en puisant dans les écoles où l'approche n'est pas en vigueur et où les enseignants n'ont pas encore été formés à l'approche (attention aux phénomènes de contamination). Lorsqu'elle est généralisée, le groupe contrôle est constitué d'élèves qui ont bien appris les ressources nécessaires à la résolution de la situation, mais n'ont pas encore appris à les mobiliser sur ce type de situation[5]. Si les groupes expérimental et contrôle sont équivalents par ailleurs et qu'on leur présente une même situation, la maîtrise de chacun des critères évaluant la compétence concernée devrait être supérieure dans le groupe expérimental. Il peut être intéressant de construire une épreuve en deux parties : la première évalue les ressources nécessaires à la seconde ; cette dernière évalue la compétence. Au-delà d'une meilleure maîtrise de chaque critère de la deuxième partie, le groupe expérimental ne devrait idéalement pas être moins performant sur la première partie d'épreuve (Aden & Roegiers, 2003 ; O/ Didiye, El Hadj Amar, Gerard & Roegiers, 2005 ; Rajonhson, Ramilijaona, Randrianirina, Razafindralambo, Razafindranovona, Ranorovololona & Gerard, 2005).

Alors que les études comparatives opèrent à partir de groupes d'élèves différents, les études corrélationnelles reposent sur des mesures répétées : un même groupe d'élèves est évalué sur une épreuve compétence et sur une ou plusieurs autres variables critères. La ou les variables critères sont censées connues comme ayant des relations suffisamment démontrées avec la compétence à valider. Ainsi, puisque être compétent implique de pouvoir résoudre des problèmes ou effectuer des tâches complexes, un élève très (ou peu) performant pour une compétence X a plus de chances d'être plus (ou moins) performant pour une compétence ou une variable critère pertinente Y. Il est d'ailleurs possible de prendre en considération plusieurs variables critères dont certaines sont plus proches et d'autres moins proches de la compétence à valider ; on vérifie alors si les résultats observés (maîtrise de chacun des critères d'évaluation de la compétence) sont conformes aux hypothèses.

La certification par la *preuve* d'une note chiffrée ?

Toutes les stratégies proposées jusqu'à présent sont intéressantes dans le sens où elles constituent un faisceau d'indicateurs de validation (dans le vocabulaire des enquêtes policières, ce sont des présomptions), mais elle ne constituent pas des preuves (elles ne permettent pas d'avoir la preuve de la culpabilité ou de l'innocence). En effet, si nous avons démontré que la maîtrise des ressources (validation par une approche classique) n'est pas suffisante, la maîtrise de chacun des critères d'évaluation d'une épreuve construite sur une approche par compétences (validation selon les stratégies énoncées ci-dessus) n'est pas encore suffisante. En effet, il n'est pas sûr que la maîtrise de chacun de ces critères soit la preuve de la maîtrise de la compétence visée, non seulement parce que celle-ci est de l'ordre du complexe et implique donc des dimensions dont on ne connaît guère le degré d'interrelations, mais aussi parce que son évaluation est relative aux critères utilisés[6] (leur nombre, leur opérationalisation, leur statut).

Cette question est d'autant plus cruciale lorsque l'on s'efforce à tenter de traduire le degré de maîtrise d'une compétence, évaluée sur la base d'une situation précise, en lui attribuant une note unique, comme le veulent la plupart des textes ministériels ou le souhaitent, à des fins comparatives, beaucoup d'experts. Lorsqu'il s'agit de traduire sur une échelle mathématique unidimensionnelle le degré de maîtrise à un ensemble d'items simples (la réussite à un item est approximativement comparable à la réussite à un autre item, comme cela est souvent le cas dans des QCM bien construits), nous disposons

maintenant de modèles psychométriques performants (*cf. Response items theorie*). Dans ce cas, le résultat chiffré et la chose évaluée sont relativement isomorphes. Par contre, cette règle de l'isomorphisme n'est pas tenable lorsqu'on évalue une réalité multidimensionnelle, ce qui est le cas dans l'approche par compétences.

Faut-il renoncer dans tous les cas à traduire en une note unique l'évaluation d'une compétence, ou plus exactement l'évaluation d'une performance complexe censée être le signe d'une compétence ? Ne peut-on trouver dans le cadre de l'évaluation certificative une stratégie sommative[7] qui permette de fonder la décision de réussite ou d'échec ?

L'enjeu est ici de taille puisqu'il s'agit d'une évaluation certificative et que celle-ci aboutit à prendre deux types de décision. Particulièrement importante, la première est dichotomique et consiste à certifier socialement la réussite ou la non-réussit ; accessoire[8] à nos yeux, la seconde consiste à ranger les élèves par niveau de performance. La seconde nous intéresse peu pour de nombreuses raisons : l'impossibilité de vérifier que le rangement sur une échelle mathématique unidimensionnelle d'une performance multidimensionnelle est un rangement valide ; une telle opération correspond à un modèle sélectif inadéquat dans une pédagogie de la réussite, etc. Par contre, la première doit nous préoccuper au plus haut point et il faut donc s'interroger sur la signification de la décision réussite versus échec : c'est la question de la pertinence.

L'évaluation certificative pour décider la réussite ou l'échec

La décision de faire réussir un élève sera **pertinente** si on peut certifier que l'élève maîtrise à un seuil suffisant les compétences minimales requises pour commencer avec fruit les apprentissages de l'année suivante ou, en cas de qualification, d'accomplir les tâches professionnelles pour lesquelles il a été formé (moyennement éventuellement un complément de formation sur le terrain). La décision sera **valide** si on peut démontrer que l'élève maîtrise effectivement ce qu'il est censé maîtriser. La décision sera **fiable** si celle-ci ne dépend pas de l'évaluateur ou des circonstances de l'évaluation.

La validation d'une évaluation certificative passe donc d'abord par une évaluation *a priori* qui consistera à vérifier, à l'aide d'experts, si (1) l'évaluation certificative porte bien sur la ou les compétences minimales[9] ; (2) on a bien distingué les critères d'évaluation minimaux du ou des critères de perfectionnement[10] ; (3) on a déterminé adéquatement les seuils de maîtrise minimale. Dans le cadre scolaire, il faut ajouter une quatrième condition : (4) l'évaluation certificative d'une compétence ne peut être menée qu'une fois

l'apprentissage totalement achevé. Cette dernière condition est fréquemment enfreinte : les enseignants utilisent des évaluations censées être formatives pour prendre leurs décisions certificatives, cette confusion des fonctions avantageant les élèves qui ont le moins besoin de l'action de l'enseignant au détriment des autres.

Cette validation *a priori* de la **pertinence** peut faire l'objet d'une validation empirique externe. Dans ce cas, la variable critère est le fait que l'élève en question entame les apprentissages nouveaux ou la vie professionnelle sans ou avec problème. Les données d'une telle validation peuvent se mettre sous la forme du tableau suivant.

Variable-critère

	avec problème	sans problème
Réussite	a	b
Échec	c	d

Résultats attendus

a : devrait être nul ou proche de zéro quant à l'effectif
b : donc proche de a + b
c : proche de c + d
d : proche de 0

Il existe une difficulté de taille à mettre en place un tel dispositif de validation empirique externe, puisqu'en principe l'élève qui échoue ne passe pas dans l'année suivante. La difficulté peut être contournée en choisissant des élèves qui ont bénéficié d'une décision de réussite abusivement (car l'évaluation certificative n'a pas porté sur un dispositif pertinent[11]) ; ceci suppose de les identifier sur la base de l'épreuve standard que l'on cherche à valider.

La décision prise sur la base d'une situation X sera **valide** si la décision prise avec une situation Y ou Z, censée appartenir à la même famille de situations et évaluée dans les mêmes conditions, sera identique. Le tableau suivant illustre ce qui est attendu en cas de validité.

		Situation Y	
		Échec	Réussite
Situation X	Réussite	≈ 0	
	Échec		≈ 0

La décision prise sur la base d'une situation X sera **fiable** si des évaluateurs différents prennent la même décision. Les données peuvent être recueillies et traitées comme suit.

		Évaluateur B	
		Échec	Réussite
Évaluateur A	Réussite	a	b
	Échec	c	d

$$\text{PHI de Guilford} = \frac{(b + c) - (a - d)}{\sqrt{(a + b)\,(c + d)\,(a + c)\,(b + d)}} \quad \text{où } 0 \leq \text{PHI} \leq 1$$

$$\text{Coefficient K de Cohen} = \frac{p_o - p_e}{1 - p_e} \quad \text{où } -1 \leq K \leq +1$$

$$\text{où } p_o = \text{proportion d'accords observés} = \frac{b + c}{b + c + a + d}$$

$$p_e = \text{proportion d'accords attendus par le hasard} = \frac{(a + b)(b + d)(a + c)(c + d)}{N^2}$$

La fiabilité peut être vérifiée par un coefficient PHI de Guilford : celui-ci devrait tendre vers 1 (accord parfait pour l'ensemble des épreuves évaluées sur la base de la même situation X). Elle peut être vérifiée également par un coefficient d'accord de Cohen. Celui-ci varie de –1 (désaccord total : les évaluateurs prennent des décisions inverses pour tous les élèves) à +1 (accord total : les mêmes décisions pour tous les élèves) en passant par 0 (distribution aléatoire des accords et des désaccords). Le coefficient K de Cohen devrait également tendre vers +1.

L'évaluation certificative pour attribuer une note

Dans les paragraphes précédents, nous avons envisagé l'évaluation certificative sous forme de décision dichotomique. Peut-on tenter cependant d'établir une note qui, à la fois, respecte ce type de décision et tente de ranger les élèves les uns par rapport aux autres (en sachant que cette dernière opération n'est pas sans poser de nombreux problèmes) ?

Nous avons tenté de le faire en étant soucieux de nous fixer des règles qui respectent avant tout la validité de la décision réussite – échec. Ces règles, dont les quatre premières ont déjà été énoncées, sont les suivantes :

- l'évaluation certificative porte bien sur la ou les compétences minimales ;
- les critères d'évaluation minimaux ont été précisés et clairement distingués du ou des critères de perfectionnement ;
- les seuils de maîtrise minimale ont été adéquatement précisés ;
- l'évaluation certificative d'une compétence ne peut être menée qu'une fois l'apprentissage totalement achevé ;
- 75 % au moins de l'échelle des notes sont réservés aux critères minimaux et 25 % au plus aux critères de perfectionnement (règle des 3/4) ;
- le seuil de maîtrise minimal pour un critère minimal est fixé à 2/3 : cette règle est basée sur le fait que, dans la majorité des cas, commettre une erreur sur trois (exemple : se tromper dans un calcul sur trois) n'est pas nécessairement le fait d'une non-maîtrise, mais peut être le fruit d'une distraction ou de tout autre facteur ; en prenant cette règle des 2/3, on essaie de minimiser les échecs abusifs[12] ;
- le seuil de la réussite est fixé à 50 % de l'échelle des notes : en effet, 2/3 de 75 % égalent 50 % de l'échelle des notes ; par cette opération, on minimise les échecs et les réussites abusifs.

Le tableau suivant illustre bien ces conditions.

Niveaux de maîtrise	*Critères minimaux*			*Critères de perfectionnement*	
	C1	C2	C3	C4	C5
Absence totale	0	0	0	0	0
Maîtrise partielle	2	2	2	1	0
Maîtrise minimale	**4**	**4**	**4**	2	1
Maîtrise maximale	5	5	5	3	2
TOTAL	15/20			5/20	

Note : La maîtrise minimale des critères minimaux, indiquée en gras, entraîne une réussite (ici avec 12/20).

La validation *a priori* consistera à vérifier que ces conditions sont bien remplies. Quant à la validation empirique, elle consistera à développer le type de démarches suivantes :

- vérifier qu'on a bien une distribution en J et non une distribution gaussienne ; si tel n'est pas le cas, on peut douter du respect de certaines conditions (tout particulièrement les quatre premières) ; cela serait un manque de pertinence ;
- comparer les distributions des résultats des mêmes élèves (ou éventuellement de deux échantillons d'élèves appareillés) évalués sur une situation X et une autre situation Y appartenant à une même famille de situations (*cf.* modèles statistiques dits d'ajustement) et calculer la corrélation entre les deux distributions (elle devrait tendre vers +1) ;
- étudier la corrélation entre la distribution des résultats sur la situation X avec les résultats obtenus avec une ou plusieurs variables critères (validation empirique externe) ;
- calculer des coefficients de fiabilité (corrélations) entre les notes attribuées par des évaluateurs deux à deux ou un coefficient de concordance de Kendall.

D'autres procédures sont susceptibles d'être dérivées de celles-ci. L'une d'entre elles est particulièrement utile au praticien, dans le cas où l'évaluation ne serait pas définitive. Elle consiste à distinguer quatre catégories d'élèves : les élèves en très grande difficulté, voire à réorienter (maîtrise inférieure à 25 % par exemple) ; les élèves en difficulté et auxquels des remédiations s'imposent pour leur permettre de passer le seuil de réussite (entre 25 et 49 % par exemple) ; les élèves ayant réussi mais nécessitant une consolidation (entre 50 et 74 % par exemple) ; les élèves sans aucun problème (plus de 75 %). Si l'on soumet ses élèves à deux situations de la même famille, on obtient un tableau du type de suivant.

		Situation Y			
		(a)	(b)	(c)	(d)
Situation X	Élèves sans problème				
	Élèves ayant réussi mais consolidation à prévoir				
	Élèves en difficulté				
	Élèves en très grande difficulté				

Non seulement ce tableau permet une validation, mais il fournit une aide diagnostique à l'enseignant dans la mesure où l'évaluation certificative n'est pas encore définitive. En effet, une pédagogie de la réussite suppose que l'enseignant et l'élève fassent tout ce qui est en leur pouvoir pour acquérir un niveau de compétence au moins supérieur au seuil de réussite.

Conclusions

La validation d'épreuves construites en cohérence avec l'approche par compétences est loin d'être un problème résolu de façon satisfaisante. Les propositions faites dans cet article ne sont que des approches du problème à résoudre. Pour reprendre l'exemple de l'enquête policière, nous n'avons pas de preuve, car le coupable n'a pas été pris sur le fait et n'a pas avoué non plus. Nous n'avons qu'un faisceau de présomptions plus ou moins fortes. C'est bien comme cela qu'il faut comprendre nos propositions.

Peut-on espérer un jour disposer d'une édumétrie satisfaisante pour évaluer des compétences complexes ? Rien n'est moins sûr dans l'état actuel de nos connaissances. Faut-il pour autant se résigner à n'évaluer que les ressources apprises (savoir-reproduire et savoir-faire élémentaires), puisque celles-ci sont évaluables de façon valide et fiable ? La tentation est forte, car l'arsenal technique est bien au point. Et on connaît l'attrait, le sentiment de puissance et de sécurité, l'impression de scientificité que procure la manipulation de techniques sophistiquées.

Dans cet article, nous défendons avec force le principe que la pertinence est première par rapport à la validité et à la fiabilité. Ces dernières sont évidemment importantes à nos yeux, mais on ne peut s'en contenter : former un chasseur compétent implique qu'il soit apte à faire un tir groupé au centre de la cible mobile, et non simplement au centre de la cible fixe.

En attendant une édumétrie plus adéquate pour évaluer la résolution de problèmes complexes ou la réalisation de tâches complexes (fondement de la notion de compétence), il nous semble important, à l'image de l'enquêteur sans preuve flagrante, d'adopter une série de stratégies qui constituent un faisceau d'indices suffisants de la valeur de nos épreuves d'évaluation (à la fois pertinentes, valides et fiables).

NOTES

1. Ces ressources regroupent à la fois des ressources internes à la personne et des ressources externes (Gerard & Braibant, 2004) :
   - les ressources internes appartiennent en propre à un individu donné et guident son action pour résoudre la situation-problème. Il s'agit essentiellement de savoir-reproduire, de savoir-faire et de savoir-être qui peuvent tous trois s'exercer dans les domaines cognitif, psycho-sensori-moteur ou socio-affectif (Gerard, 2000, 2001) ;
   - les ressources externes sont, quant à elles, tout ce qui pourra être mobilisé en dehors de la personne. Elles concernent dès lors des ressources matérielles (p. ex. un outil, un texte), des ressources sociales (p. ex. une réunion, un réseau de relations), des ressources procédurales (p. ex. un algorithme, un règlement), etc.

   Si les définitions du concept de compétence présentent quelques différences (exemple : savoir agir pour les uns ; potentialité pour les autres), elles convergent cependant sur le fait que la personne compétente est en mesure de mobiliser les ressources pertinentes pour résoudre un certain type de problèmes ou effectuer un certain type de tâches complexes.
2. Ce qui explique qu'il est difficile de proposer plus qu'une ou deux situations.
3. Le terme d'analyse factorielle est un terme générique fréquemment utilisé qui masque la différence trop peu connue entre la signification d'une composante principale et d'un facteur. Dans le cas présent, c'est la première composante principale qui est la plus adéquate pour mesurer l'aptitude scolaire globale mesurée par l'épreuve.
4. Cela est l'illustration d'une règle importante : l'indépendance des critères d'évaluation.
5. Le principe méthodologique de l'approche par les compétences de base est d'apprendre, de manière relativement classique, les différentes ressources relatives à une compétence durant, par exemple, cinq semaines. Ensuite, une semaine sera consacrée à l'intégration des ressources : face à des situations complexes, les élèves apprendront à identifier quelles ressources sont nécessaires pour résoudre la situation et à les mobiliser de manière intégrée pour résoudre celle-ci. (Gerard & Braibant, 2004 ; Rey, Carette, Defrance & Kahn, 2003 ; Roegiers, 2000).
6. Ceci est en lien avec l'inévitable et nécessaire subjectivité de l'évaluation (Gerard, 2002).
7. Pour des raisons historiques (les travaux de Scriven et de Bloom notamment), une confusion continue à régner, même chez de nombreux experts actuels, entre les fonctions de l'évaluation et les démarches de l'évaluation. L'évaluation peut servir des fonctions d'orientation (décisions relatives à une nouvelle action à entreprendre), de régulation (décisions relatives à l'amélioration de l'action en cours comme dans l'évaluation dite formative) et de certification (décisions relatives à des objectifs de reconnaissance sociale et administrative). Chacune de ces fonctions peut être réalisée avec l'une ou l'autre des démarches suivantes : descriptive (je décris des faits, des performances, etc.) ou sommative (je somme mathématiquement des faits, des performances, etc.) ou herméneutique (je donne du sens intuitivement à un faisceau de signes). La confusion vient de l'assimilation abusive de l'évaluation certificative et de l'évaluation sommative, comme si la première ne pouvait se faire que par une méthode sommative.
8. Accessoire à nos yeux, mais pas aux yeux de beaucoup de parents.

9. À nos yeux, l'évaluation certificative sera d'autant meilleure que l'on aura reconnu une macrocompétence qui oblige l'élève à mobiliser l'ensemble des compétences minimales requises : nous appelons celle-ci un objectif terminal d'intégration (De Ketele, 1993a et b).
10. Un critère est minimal s'il doit absolument être maîtrisé pour certifier la réussite de la compétence, alors qu'un critère de perfectionnement ne doit pas absolument être maîtrisé : c'est mieux s'il l'est, mais ce n'est pas indispensable. Par exemple, dans le cas de l'évaluation d'une compétence de résolution de problèmes mathématiques, le critère « Utilisation correcte des outils de la discipline » serait un critère minimal, alors que la « Qualité de la présentation » serait un critère de perfectionnement. Il s'agit en réalité d'une décision des évaluateurs, préalable à l'évaluation, et qui fait partie du paramétrage de la famille de situations.
11. Cette situation est fréquente dans les pays où la variance interécoles est forte.
12. On parle d'échec abusif quand un élève échoue non pas en raison de sa non-maîtrise des compétences minimales requises, mais du fait du système d'évaluation lui-même : il échoue alors qu'il ne le devrait pas, car il aurait pu très bien suivre les apprentissages de l'année suivante. Si un élève réussit alors qu'il ne le devrait pas, on parle de réussite abusive : sa non-maîtrise des compétences de base requises ne lui permet pas d'entamer les apprentissages fondamentaux de l'année suivante.

RÉFÉRENCES

Aden, H.M., & Roegiers, X. (2003). À *quels élèves profite l'approche par les compétences de base ? Étude de cas à Djibouti*. Document inédit, accessible à l'adresse suivante : [http ://www.bief.be].

Bloom, B.S., Hastings, J.Th.H., & Madaus, C.F. (éds) (1971). *Handbook on formative and summative evaluation of student learning*. New York.

Cardinet, J. (1973). L'adaptation des tests aux finalités de l'évaluation. *Les sciences de l'Éducation – Pour l'ère nouvelle, 2-3*(1973), 148-182.

Cardinet, J., & Tourneur, Y. (1985). *Assurer la mesure*. Berne – Francfort-s.Main – New York : Peter Lang.

Carver, R.P. (1974). Two dimensions of tests : psychometric and edumetric. *American Psychologist, 29*, 512-518.

Crahay, M. (1996). *Peut-on lutter contre l'échec scolaire ?* Bruxelles : De Boeck Université.

De Ketele, J.-M. (1993a). L'évaluation conjuguée en paradigmes. *Revue française de pédagogie*, *103*, 59-80.

De Ketele, J.-M. (1993b). Objectifs terminaux d'intégration et transfert des connaissances. In R. Hivon (éd.), *L'évaluation des apprentissages. Réflexions, nouvelles tendances et formation*, (chap. 1, pp. 15-26). Sherbrooke (Canada) : CRP.

De Ketele, J.-M. (2000). En guise de synthèse : Convergences autour des compétences. In C. Bosman, F.-M. Gerard & X. Roegiers (éds), *Quel avenir pour les compétences ?* (pp. 187-191). Bruxelles : De Boeck Université.

De Ketele, J.-M. (2001a). Place de la notion de compétence dans l'évaluation des apprentissages. In G. Figari & M. Achouche (éds), *L'activité évaluative réinterrogée. Regards scolaires et socioprofessionnels* (pp. 39-43). Bruxelles : De Boeck Université.

De Ketele, J.-M. (2001b). Enseigner des compétences : repères. In J.-L. Jadoulle & M. Bouhon (éds), *Développer des compétences en histoire* (pp. 13-22). Louvain-la-Neuve-Bruxelles : Université catholique de Louvain et Ministère de l'éducation, de la recherche et de la formation.

De Ketele, J.-M., Chastrette, M., Cros, D., Mettelin, P., & Thomas, J. (1989). *Guide du formateur* (2e éd.). Bruxelles : De Boeck Université.

De Ketele, J.-M., & Dufays, J.-L. (2003). Vers de nouveaux modes d'évaluation des compétences. In L. Collès, J.-L. Dufays & C. Maeder, *Enseigner le français, l'espagnol et l'italien. Les langues romanes à l'heure des compétences* (pp. 171-182). Bruxelles-Paris : Éditions De Boeck Duculot.

De Ketele, J.-M., & Roegiers, X. (1991, 2e éd. 1993). *Méthodologie du recueil d'informations*. Bruxelles : De Boeck Université.

de Landsheere, G. (1980). *Examens et évaluation continue. Précis de docimologie*. Bruxelles-Paris : Labor-Nathan.

Dolz, J., & Ollagnier, E. (éds) (2002). *L'énigme de la compétence en éducation*. Bruxelles : De Boeck Université.

Findley, W. G. (1956). A rationale for evaluation of item discrimination statistics. *Educational and Psychological Measurement*, *16*, 175-180.

Fourez, G. (1999). Compétences, contenus, capacités et autres casse-têtes. *Forum des pédagogies,* mai, 26-31.

Gerard, F.-M. (2000). Savoir, oui... mais encore ! *Forum - pédagogies*, mai, 29-35.

Gerard, F.-M. (2001). L'évaluation de la qualité des systèmes de formation. *Mesure et évaluation en éducation*, *24*(2-3), 53-77.

Gerard, F.-M. (2002). L'indispensable subjectivité de l'évaluation, *Antipodes*, *156*, avril, 26-34.

Gerard, F.-M. (2005), Évaluer des compétences, ou ne pas se tromper de cible. *Liaisons*, *40*, février, Beyrouth, Liban, 7-9.

Gerard, F.-M., & Braibant, J.-M. (2003). Activités de structuration et activités fonctionnelles, même combat ? Le cas de l'apprentissage de la compétence en lecture à l'école primaire. *Français 2000*, *190-191*, avril 2004, 24-38.

Jonnaert, Ph. (2002). *Compétences et socioconstructivisme - Un cadre théorique.* Bruxelles : De Boeck.

Lasnier, F. (2000). *Réussir la formation par compétences*. Montréal : Guérin.

Laveault, D., & Grégoire, J. (1997, 2e éd. 2002). *Introduction aux théories des tests en psychologie et en sciences de l'éducation*. Bruxelles : De Boeck Université.

Le Boterf, G. (1994). *De la compétence. Essai sur un attracteur étrange.* Paris : Éditions de l'organisation.

Legendre, M.-F. (2001). Sens et portée de la notion de compétence dans le nouveau programme de formation. *Revue de l'AQEFLS*, *23*(1), 12-30.

O/ Didiye, D., El Hadj Amar, B., Gerard, F.-M., & Roegiers, X. (2005). *Étude relative à l'impact de l'introduction de l'APC sur les résultats des élèves mauritaniens*. Document inédit, accessible à l'adresse suivante : [http ://www.bief.be].

Perrenoud, Ph. (1997). *Construire des compétences dès l'école*. Paris : ESF.

Poirier Proulx, L. (1999). *La résolution de problèmes en enseignement. Cadre référentiel et outils de formation*. Bruxelles : De Boeck Université.

Rajonhson, L., Ramilijaona, F., Randrianirina, P., Razafindralambo, M.H., Razafindranovona, O., Ranorovololona, E., & Gerard, F.-M. (2005). *Premiers résultats de l'APC : invitation à continuer…* Document inédit, accessible à l'adresse suivante : [http ://www.bief.be].

Raynal, F., & Rieunier, A. (1997, 4e éd. 2003). *Pédagogie : dictionnaire des concepts clés*. Paris : ESF éditeur.

Rey, B. (1996). *Les compétences transversales en question.* Paris : ESF.

Rey, B., Carette, V., Defrance, A., & Kahn, S. (2003). *Les compétences à l'école – Apprentissage et évaluation*. Bruxelles : De Boeck.

Roegiers, X. (2000). *Une pédagogie de l'intégration.* Bruxelles : De Boeck.

Roegiers, X. (2003). *Des situations pour intégrer les acquis*. Bruxelles : De Boeck.

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par les compétences*. Bruxelles : De Boeck Université.

Scriven, M. (1967). The methodology of evaluation, *AERA Monograph Series on Curriculum Evaluation*, *1*. Chicago : Rand McNally.

Tardif, J., et al. (1995). Le développement des compétences, cadre conceptuel pour l'enseignement. In J.-P. Goulet, *Enseigner au collégial* (pp. 157-168). Montréal : Association québécoise de pédagogie collégiale.

Tilman, F. (2000). Qu'est-ce qu'une compétence ? *Exposant neuf, 2*(2000), 28-31.