

L'  valuation des comp  tences en milieu scolaire et en milieu professionnel

**EVALUER DES COMPETENCES EN MATHEMATIQUES DANS LE CADRE D'UNE
EPREUVE EXTERNE A LARGE ECHELLE AU DEPART DE TACHES COMPLEXES,
DE TACHES DECOMPOSEES ET DE TACHES ELEMENTAIRES :
QUEL POUVOIR INFORMATIF ?**

Christophe Dierendonck* et Annick Fagnant**

* *Universit   du Luxembourg – christophe.dierendonck@uni.lu*

** *Universit   de Li  ge – afagnant@ulg.ac.be*

***Mots-cl  s.** T  che complexe – Math  matiques – Evaluation    large   chelle – Evaluation diagnostique*

***R  sum  .** En vue d'aider    cerner les « forces » et les « faiblesses » en math  matiques d'  l  ves de l'enseignement secondaire, un dispositif exp  rimental int  grant des t  ches   valuant des comp  tences a   t   test   dans le cadre de deux   preuves d'  valuation externe (2009 et 2010) au Luxembourg. Les   preuves comportaient des t  ches complexes, des t  ches complexes d  compos  es et des t  ches n  cessitant l'application directe des ressources impliqu  es dans ces t  ches complexes (cf. mod  le d'  valuation en phases de Rey et al., 2003), mais aussi des probl  mes   l  mentaires ind  pendants   valuant la capacit      mobiliser ces m  mes ressources dans d'autres contextes. Elles comprenaient   galement d'autres items permettant de couvrir de mani  re plus large les domaines math  matiques investigu  s. Dans les deux   preuves, la pr  sence d'un nombre suffisant d'items (et leur ind  pendance) a autoris   l'utilisation d'un mod  le de r  ponse    l'item (mod  le de Rash) qui permet d'analyser, sur une   chelle commune, le degr   de difficult   des t  ches et le niveau de comp  tence des   l  ves. La combinaison des informations issues des divers types de t  ches permet d'  clairer les difficult  s des   l  ves face aux t  ches complexes.*

1. Introduction

La notion de « comp  tence » a pris une place consid  rable dans la plupart des r  f  rentiels europ  ens de formation. On attend aujourd'hui des   l  ves qu'ils soient capables de mobiliser, de fa  on int  gr  e, un ensemble de ressources internes et externes pour accomplir des t  ches complexes voire in  dites (Beckers, 2005 ; Carette, 2007 ; De Ketele & G  rard, 2005 ; Rey, Carette, Defrance & Kahn, 2003). Si l'int  r  t de l'approche par comp  tences est bien d'avoir replac   le concept de « mobilisation » au centre du d  bat   ducatif, le recours    des t  ches d'une grande complexit   pour   valuer la comp  tence des   l  ves a conduit    plusieurs paradoxes. Parmi ceux-ci, nous en d  velopperons trois en guise d'introduction de la pr  sente   tude.

Le premier paradoxe pose la question de la validit   et de la fiabilit   des   preuves d'  valuation compos  es d'un nombre forc  ment restreint de t  ches complexes,   tant donn   le temps n  cessaire    leur r  solution. Sur le plan de la validit  , on peut s'interroger sur le fait d'  valuer les   l  ves sur la base de t  ches qui n'ont peut-  tre pas fait l'objet d'un r  el apprentissage et qui sont sans doute trop cibl  es pour permettre de couvrir le curriculum enseign  . Sur le plan de la fiabilit   des informations r  colt  es, on rappellera que plus on dispose d'items qui   valuent ce que l'on souhaite mesurer, plus la mesure que l'on prend devrait   tre pr  cise (Laveault & Gr  goire, 1997) et ce, quel que soit le mod  le de mesure utilis   (th  orie classique des tests ou th  orie de r  ponse    l'item).

Le deuxi  me paradoxe se situe au niveau de l'impact que peuvent avoir ces   preuves d'  valuation. Il s'av  re en effet que tr  s peu d'  l  ves parviennent    fournir une r  ponse convaincante face    ce type de t  ches tr  s complexes (Rey et al., 2003). Le danger est donc de consid  rer, au d  part de ces quelques t  ches complexes, qu'un   l  ve n'est pas comp  tent et de sanctionner ainsi son

L'évaluation des compétences en milieu scolaire et en milieu professionnel

parcours scolaire, alors qu'il maîtrise certainement par ailleurs un certain nombre de connaissances et d'habiletés.

Le troisième paradoxe est lié à la difficulté d'établir un diagnostic des difficultés des élèves au départ de quelques tâches complexes. Non seulement il est probable qu'un nombre important d'élèves ne parviennent même pas à entrer dans les tâches étant donné leur complexité, mais, sans prise d'informations sur la démarche de résolution adoptée par les élèves, il devient impossible d'identifier précisément leurs lacunes face à de telles tâches. S'agit-il d'une mauvaise interprétation de la tâche à réaliser, d'une difficulté de mobilisation, d'un problème de coordination/intégration de plusieurs procédures ou de plusieurs étapes de résolution, de ressources élémentaires non maîtrisées, de simples erreurs de calcul, ... ?

Dans cet article, nous désirons apporter des éléments de réponse à ces trois paradoxes en rendant compte d'un dispositif d'évaluation qui a été expérimenté à large échelle, qui est fondé sur une définition moins exigeante de la complexité et qui tente d'apporter un éclairage relatif aux difficultés rencontrées par les élèves face aux tâches complexes.

2. Méthode de la recherche

2.1. Une définition moins exigeante de la complexité

Dans les modèles d'évaluation construits en référence à l'approche par compétences (voir notamment Rey et al., 2003 ; de Ketele et Gérard, 2005), les tâches ou les situations complexes sont définies à partir d'un niveau de complexité ultime : la résolution de ces tâches nécessite la mobilisation et l'intégration d'un grand nombre de ressources internes et/ou externes et donc beaucoup de temps. Pour que l'évaluation puisse couvrir plus largement le domaine de contenu investigué, nous avons suggéré d'adopter une définition plus raisonnable de la complexité des tâches d'évaluation qui s'inscrivent en référence à l'approche par compétences. Ainsi, une tâche/une situation serait considérée comme complexe à partir du moment où elle nécessite l'identification, la mobilisation et l'intégration de plus d'une procédure apprise et qu'elle nécessite dès lors une interprétation (ou un cadrage) de la situation et une organisation de la démarche de résolution (Dierendonck et Fagnant, 2010a, p. 13, définition actualisée). Cette définition moins ambitieuse de la complexité autorise selon nous l'élaboration d'évaluations constituées d'un nombre suffisant d'items pour s'assurer de la fiabilité des données récoltées et permettre une analyse statistique robuste.

Tenant compte des avancées de certains dispositifs d'évaluation des compétences, nous avons conçu un design expérimental qui s'appuie pour partie sur les travaux de Rey et al. (2003) et pour partie sur ceux de Crahay et Detheux (2005).

2.2. Les principes repris du modèle de Rey et al. (2003)

Rey et al. (2003) distinguent des compétences de 1^{er} degré qui consisteraient essentiellement à appliquer des procédures dans des situations fermées ; des compétences de 2^e degré qui feraient appel à la notion de mobilisation mais qui n'impliqueraient qu'un seul type de ressources et enfin, des compétences de 3^e degré qui nécessiteraient la mobilisation intégrée de différentes ressources. Sur la base de cette distinction, les auteurs développent un modèle d'évaluation des compétences en trois phases (Carette, 2007). En phase 1, les élèves sont confrontés à une tâche complexe exigeant le choix et la combinaison d'un nombre significatif de procédures. En phase 2, on leur propose la même tâche complexe, mais découpée cette fois en tâches élémentaires présentées dans l'ordre où elles doivent être accomplies. En phase 3, on leur soumet une série de tâches simples décontextualisées correspondant aux procédures élémentaires qui ont dû être mobilisées pour accomplir la tâche complexe.

L'  valuation des comp  tences en milieu scolaire et en milieu professionnel

2.3. Les principes repris du mod  le de Crahay et Detheux (2005)

Partant de l'hypoth  se que la ma  trise isol  e des proc  dures n'est pas suffisante pour r  soudre des probl  mes complexes impliquant l'int  gration (mobilisation et coordination) de celles-ci, Crahay et Detheux (2005) ont d  velopp   un dispositif exp  rimental permettant de mieux cerner les forces et les faiblesses des   l  ves face aux t  ches complexes. Plut  t que de proposer une d  composition de la t  che complexe comme dans le mod  le de Rey et al. (2003), ils proposent d'  valuer isol  ment les proc  dures dans des t  ches ind  pendantes qui sont propos  es aux   l  ves un autre jour que la t  che complexe elle-m  me.

Dans leur article, les auteurs d  crivent deux probl  mes complexes propos  s    1436   l  ves de grade 6. Chaque probl  me complexe implique la ma  trise de plusieurs proc  dures inscrites au programme de l'  cole primaire en Communaut   fran  aise de Belgique. La ma  trise de chacune des proc  dures impliqu  es dans la r  solution des deux probl  mes complexes a   t   test  e sous des formes de questionnement s'apparentant    des probl  mes   l  mentaires impliquant la mobilisation d'une seule proc  dure (*comp  tence de 2^e degr   selon Rey et al., 2003 – que nous nommerons « t  ches   l  mentaires en contexte » dans la suite du texte*) ou sous la forme de t  ches d  contextualis  es faisant directement appel    l'application de la proc  dure vis  e (*comp  tence de 1^{er} degr   ou proc  dure selon Rey et al., 2003 – que nous appellerons « t  ches   l  mentaires d  contextualis  es » dans la suite du texte*).

A titre illustratif de l'  preuve de Crahay et Detheux, un exemple de t  che   l  mentaire en contexte (SF1.3.) et un exemple de t  che   l  mentaire d  contextualis  e (SF 1.4.) sont pr  sent  es dans la figure 1.

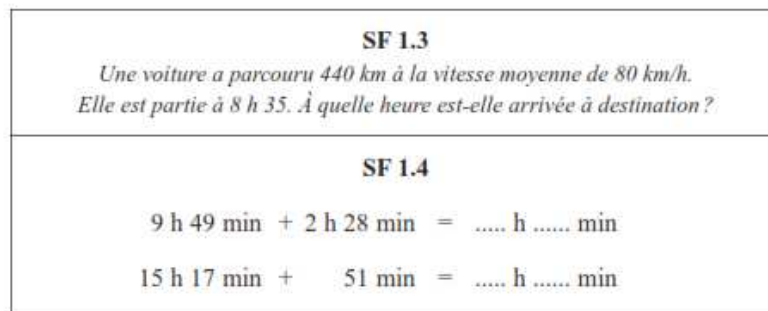


Figure 1 : Exemples de t  ches   l  mentaires propos  e par Crahay & Detheux (2005)

2.4. L'  laboration d'un dispositif d'  valuation incluant quatre types de t  ches

Pour r  soudre une t  che complexe et faire preuve de comp  tence, l'  l  ve doit   tre capable de mobiliser les ressources ad  quates, de les mettre en   uvre correctement et de les combiner si n  cessaire. Dans une perspective diagnostique, on peut donc   laborer des t  ches pr  sentant diff  rents niveaux de complexit  . De mani  re sp  cifique, nous avons voulu exp  rimer un dispositif d'  valuation de la comp  tence des   l  ves    r  soudre un probl  me de g  om  trie n  cessitant un « cadrage » de la situation, la mobilisation de plus d'une ressource ou proc  dure apprise et la coordination de plusieurs   tapes de r  solution. Concr  tement, nous avons distingu   quatre types de t  ches qui, combin  es au sein d'un m  me dispositif d'  valuation, pourraient apporter des informations diagnostiques int  ressantes : des t  ches complexes de r  f  rence, des t  ches complexes d  compos  es, des t  ches   l  mentaires d  contextualis  es et des t  ches   l  mentaires pr  sent  es dans un autre contexte probl  matique.

L'objectif de l'  tude est de rendre compte du potentiel diagnostique d'un tel dispositif d'  valuation en tentant de r  pondre aux questions suivantes :

L'  valuation des comp  tences en milieu scolaire et en milieu professionnel

- (1) Les diff  rents types de t  ches se positionnent-ils de mani  re hi  rarchique du point de vue de leur difficult   ?
 - a. Ce positionnement hi  rarchique est-il absolu ou des t  ches dites   l  mentaires peuvent-elles pr  senter un degr   de difficult   aussi   lev   qu'une t  che complexe ?
 - b. Les t  ches   l  mentaires pr  sent  es dans un autre contexte probl  matique apportent-elles des informations compl  mentaires aux informations fournies par les t  ches complexes d  compos  es ? Autrement dit, l'  valuation des comp  tences de 2^e degr   dans des situations ind  pendantes les unes des autres fournit-elle d'autres informations sur la ma  trise de ces comp  tences que leur   valuation dans des t  ches d  compos  es ?
- (2) L'int  gration d'un tel dispositif d'  valuation au sein d'une   preuve d'  valuation externe    large   chelle peut-elle d  boucher sur un diagnostic pr  cis et individualis   des lacunes et des forces des   l  ves ?

2.5. Le dispositif d'  valuation int  gr   aux   preuves standardis  es de grade 9 au Luxembourg

Au Luxembourg, des   preuves d'  valuation externes et standardis  es sont soumises chaque ann  e aux   l  ves de grade 9 en math  matiques et en langues (allemand et fran  ais), et ce dans trois types d'enseignement (ES, EST, PR). En 2009, une premi  re   tude exploratoire a   t   men  e au d  part d'une t  che complexe, d'une t  che d  compos  e et de quelques t  ches   l  mentaires propos  es en contexte ou non (quelques exemples sont propos  s en annexe). Les r  sultats ont   t   analys  s dans le type d'enseignement secondaire g  n  ral (ES, N= 1769) uniquement (Dierendonck & Fagnant, 2010b). En 2010, un design exp  rimental plus complexe a   t     labor  , avec ancrage commun entre les types d'enseignement (ES, Est et PR, N = 6399) et des parties sp  cifiques    chacune d'elles.

En plus des t  ches-cibles d  crites au point pr  c  dent, ces   preuves comprenaient une vari  t   de t  ches   valuant des comp  tences de 1^{er}, de 2^e ou de 3^e degr   de fa  on    couvrir le plus largement possible (en fonction des contraintes temporelles, deux heures de cours de 50 minutes au maximum) les domaines de contenu investigu  s (   savoir les « nombres et op  rations » et les « figures du plan et de l'espace »).

3. R  sultats

Les donn  es ont   t   analys  es    l'aide du mod  le de r  ponse    l'item (MRI)    un param  tre (mod  le de Rash)¹ au moyen du logiciel Conquest. Pour cette analyse, les t  ches complexes ont   t   recod  es 1 (r  ponse finale correcte) ou 0 (r  ponse finale incorrecte) tant dans leur format initial (t  che complexe de r  f  rence) que dans leur format d  compos   (dans ce cas pr  cis, seule la r  ponse    la derni  re sous-t  che a   t   consid  r  e pour l'analyse MRI).

Les r  sultats sont pr  sent  s en deux   tapes correspondant aux deux questions de recherche mentionn  es pr  c  demment : la premi  re traite de la hi  rarchisation des t  ches et sera analys  e sur l'ensemble de l'  preuve externe en confrontant les r  sultats obtenus en 2009 (en ES) et en 2010 (dans les trois types d'enseignement, gr  ce    la proc  dure d'ancrage) ; la deuxi  me traite du potentiel diagnostique des   preuves, en confrontant les r  sultats des diff  rentes t  ches   valuant la mise en   uvre des m  mes proc  dures dans des t  ches de complexit   diff  rente.

¹ Pour rappel, la particularit   des MRI est de placer sur une m  me   chelle le niveau de difficult   des items et le niveau de comp  tence des   l  ves. Concr  tement, la comp  tence d'un   l  ve est d  finie en fonction de l'item pour lequel il a une probabilit   de r  ussite de 50%. Sur le sch  ma, les   l  ves sont situ  s en vis-  -vis des items pour lesquels ils ont une probabilit   de r  ussite de 50% ; ils ont donc une probabilit   inf  rieure    50% de r  ussir les items plus difficiles (situ  s plus haut sur le graphique) et une probabilit   sup  rieure    50% de r  ussir les items plus faciles (situ  s plus bas sur le graphique). Les sujets situ  s en haut du graphique sont donc les sujets les plus performants ou qui ont le niveau de comp  tence le plus   lev  .

L'évaluation des compétences en milieu scolaire et en milieu professionnel

3.1. Comment les différents types de tâches se positionnent-ils du point de vue de leur difficulté ?

A titre illustratif, la figure 2 et synthétise les résultats de l'analyse MRI obtenus avec les données de l'épreuve standardisées 2009 (ES). La présentation des fichiers de sortie a été retravaillée pour faire ressortir l'appartenance des items aux quatre types de tâches précédemment décrits. La tâche complexe « judo », la tâche parallèle décomposée « menuisier »² et les tâches élémentaires évaluant isolément les procédures impliquées dans la tâche complexe (en contexte ou hors contexte) sont identifiées dans le schéma par leur nom respectif et par un encadré. Deux autres tâches complexes sont également identifiées (cercle et étoile), ainsi que les autres tâches conceptuellement liées.

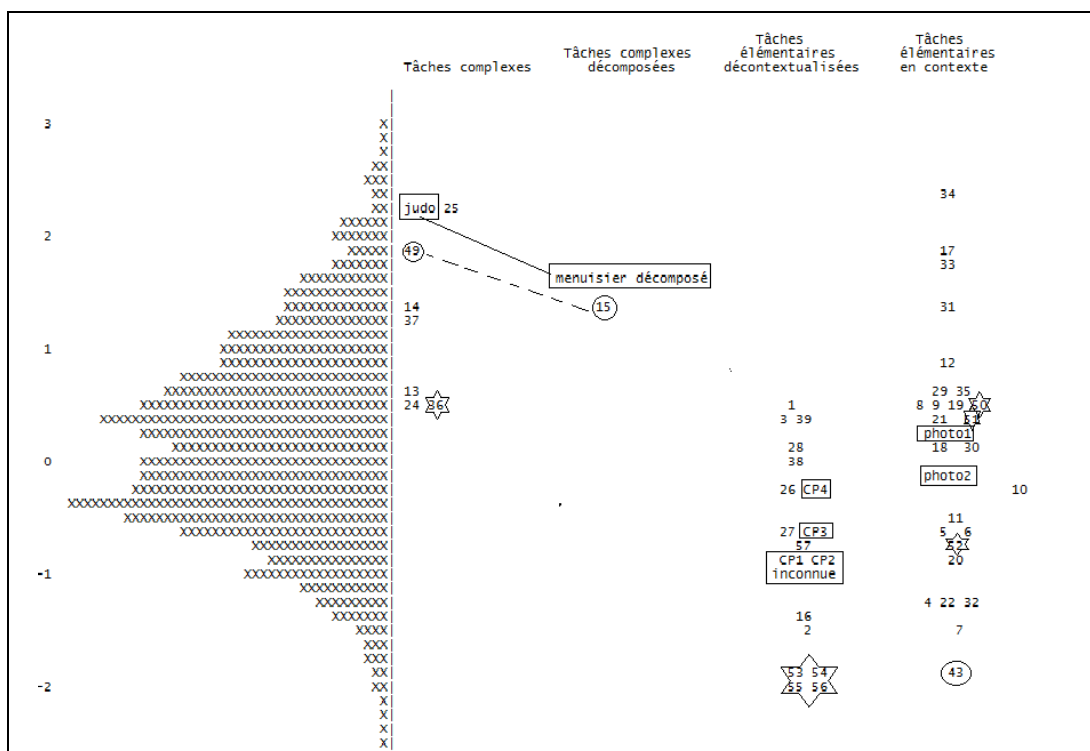


Figure 2 – Résultats en ES dans l'épreuve de 2009 (chaque 'X' représente 2.9 cas)

Les résultats de l'épreuve de 2010 (ES, EST et PR) présentent une distribution comparable à celle de l'épreuve de 2009 et permettent de tirer trois constats principaux :

- (1) On n'observe pas de hiérarchie stricte entre les quatre types de tâches. En effet, si les tâches les plus difficiles de l'épreuve sont des tâches complexes, quelques tâches élémentaires en contexte se situent sensiblement au même niveau de difficulté. Par ailleurs, toutes les tâches complexes ne présentent pas un niveau de difficulté extrêmement élevé et certaines apparaissent aussi difficiles que plusieurs tâches élémentaires (essentiellement en contexte) ;
- (2) Les tâches décomposées présentent un niveau de difficulté assez proche de celui des tâches complexes, ce qui laisse entendre que la décomposition aide finalement assez peu les élèves ;

² Les tâches « judo » et « menuisier » sont des tâches complexes qui impliquent la mobilisation intégrée de deux procédures mathématiques : (1) calculer la longueur d'un côté d'un carré au départ de son aire (judo) ou de son périmètre (menuisier) et (2) déduire la longueur d'un segment au départ de la longueur de deux autres segments (judo et menuisier).

L'évaluation des compétences en milieu scolaire et en milieu professionnel

- (3) De nombreuses tâches élémentaires en contexte présentent un niveau de difficulté moindre que celui des tâches décomposées. Plus particulièrement, les tâches élémentaires en contexte qui évaluent isolément les procédures impliquées dans la tâche complexe « judo » ou « menuisier » présentent un niveau de difficulté nettement inférieur à celui des tâches décomposées correspondantes.

3.2. D'un diagnostic global à un diagnostic spécifique

Les données synthétisées dans la figure 2 permettent d'établir un premier diagnostic qui porte sur la façon dont une population scolaire (en l'occurrence les élèves d'ES de grade 9 au Luxembourg) parvient à résoudre différents types de tâches plus ou moins complexes. Moyennant quelques adaptations de cette figure, ce diagnostic peut également être formulé à l'échelle d'une classe ou d'un élève en particulier (cf. Dierendonck & Fagnant, 2010b). Au départ de ce corpus de données, il semble intéressant de se concentrer sur les tâches les moins bien réussies par les élèves et d'identifier le type de tâches à travailler en priorité avec eux.

A partir de l'analyse MRI et d'un jugement opéré sur la nature des processus demandés par chaque item, on pourrait également définir différents niveaux de compétence (par exemple « niveau socle avancé », « niveau socle », « niveau en dessous du socle ») et situer ainsi les élèves par rapport à ces niveaux en procédant comme dans les analyses des épreuves PISA. Cette démarche de « standards setting » serait une autre façon de dresser un constat global sur le niveau de compétence des élèves en mathématiques et aurait l'avantage de concrétiser, au travers d'items, ce qui est réellement attendu des élèves (et des enseignants) en termes d'évaluation des socles de compétences.

Une autre possibilité est d'analyser plus finement les tâches complexes de référence pour lesquelles nous avons conçu le dispositif incluant quatre types de tâches en observant comment les élèves ont répondu à chacune d'elles. A titre d'illustration, nous avons conduit une telle analyse auprès des 2869 élèves de la filière d'enseignement technique (EST) en 2010 en croisant les résultats obtenus à la tâche complexe « judo », à sa tâche parallèle décomposée « menuisier » et aux 7 tâches élémentaires en contexte ou hors contexte censées évaluer de manière isolée les ressources et procédures sollicitées par la tâche complexe (tableau 1).

		Tâche complexe décomposée « Menuisier »	Tâches élémentaires décontextualisées (TED) et contextualisées (TEC)
Tâche complexe « Judo » 2869 élèves EST	Echec 2782 (97%)	Echec 2410 (86%)	→ 629 élèves (26%) réussissent 3 ou 4 TED et 2 ou 3 TEC
			→ 382 élèves (16%) réussissent 3 ou 4 TED et 0 ou 1 TEC
			→ 436 élèves (18%) réussissent 0, 1 ou 2 TED et 2 ou 3 TEC
			→ 963 élèves (40%) réussissent 0, 1 ou 2 TED et 0 ou 1 TEC
	Réussite 372 (14%)	Réussite 372 (14%)	→ 178 élèves (48%) réussissent 3 ou 4 TED et 2 ou 3 TEC
			→ 47 élèves (12%) réussissent 3 ou 4 TED et 0 ou 1 TEC
			→ 96 élèves (26%) réussissent 0, 1 ou 2 TED et 2 ou 3 TEC
			→ 51 élèves (14%) réussissent 0, 1 ou 2 TED et 0 ou 1 TEC
	Réussite 87 (3%)	Echec 52 (60%)	→ 39 élèves (75%) réussissent 3 ou 4 TED et 2 ou 3 TEC
			→ 7 élèves (13%) réussissent 3 ou 4 TED et 0 ou 1 TEC
→ 5 élèves (10%) réussissent 0, 1 ou 2 TED et 2 ou 3 TEC			
Réussite 35 (40%)		Réussite 35 (40%)	→ 1 élève (2%) réussit 0, 1 ou 2 TED et 0 ou 1 TEC
			→ 33 élèves (94%) réussissent 3 ou 4 TED et 2 ou 3 TEC
			→ 1 élève (3%) réussit 3 ou 4 TED et 0 ou 1 TEC
			→ 1 élève (3%) réussit 0, 1 ou 2 TED et 2 ou 3 TEC
			→ 0 élève (0%) réussit 0, 1 ou 2 TED et 0 ou 1 TEC

Tableau 1 - Croisement des résultats aux différents items liés à la tâche complexe « Judo » et à la tâche décomposée « Menuisier » pour les élèves de l'EST (données 2010)

L'évaluation des compétences en milieu scolaire et en milieu professionnel

Les résultats, détaillés dans le tableau 1, permettent de tirer quelques constats qui démontrent assez nettement que les élèves se comportent différemment de ce que l'on attend *a priori* d'un modèle diagnostique « en phases » supposées hiérarchiques :

- 1er constat : seuls 3% des élèves (87 sur 2869) réussissent la tâche complexe « Judo ». Même avec une définition simplifiée de la tâche complexe (la tâche nécessitait de mobiliser et d'intégrer deux procédures et d'organiser une démarche de résolution en trois étapes), celle-ci reste sensiblement « hors de portée » des élèves de la filière d'enseignement technique. Par ailleurs, ils sont moins de la moitié à réussir les tâches élémentaires connexes (33 sur 87 réussissent la tâche décomposée « Menuisier » et la majorité des tâches élémentaires tandis que 7 sur 87 échouent à la tâche décomposée « Menuisier », mais réussissent la majorité des tâches élémentaires) ;
- 2e constat : 14% des élèves qui échouent à la tâche complexe « Judo » réussissent la tâche complexe décomposée « Menuisier », mais 60% des élèves qui réussissent la tâche complexe « Judo » échouent à la tâche complexe décomposée « Menuisier ». Autrement dit, non seulement l'aide apportée par la décomposition semble relativement négligeable, mais cette décomposition imposée paraît même déstabiliser une part importante d'élèves. En outre, parmi les élèves qui réussissent la tâche grâce à la décomposition, seuls 48% réussissent également une majorité de tâches élémentaires ;
- 3e constat : la réussite de la majorité des tâches élémentaires (en contexte ou non) ne suffit pas à réussir la tâche complexe « Judo » (au total, 29% des élèves sont dans ce cas - 629+178+39) ou à réussir la tâche décomposée « Menuisier » (au total, 24% des élèves sont dans ce cas : 629+39). Ces résultats témoignent de l'importance de proposer des tâches évaluant des compétences de 2e degré en dehors de la décomposition de la tâche complexe ;
- 4e constat : au total, environ 15% des élèves (382+48+7+1) semblent maîtriser les tâches élémentaires décontextualisées (TED) alors qu'ils échouent aux tâches élémentaires contextualisées (TEC). A l'inverse, ils sont environ 19% (436+96+56+1) à réussir les tâches élémentaires en contexte (TEC), mais pas les tâches décontextualisées (TED), ce qui complique encore davantage le type de diagnostic à poser ;
- 5e constat : sur les 2410 élèves qui échouent à la tâche complexe « Judo » et à la tâche complexe décomposée « Menuisier », 40 % ne résolvent quasi aucune tâche élémentaire décontextualisée ou contextualisée, ce qui témoigne clairement de leur non maîtrise des procédures élémentaires. A ces élèves s'ajoutent encore quelques 52 cas particuliers (environ 2% des élèves) qui ne semblent pas maîtriser ces tâches élémentaires alors qu'ils ont réussi la tâche complexe décomposée (51 élèves) ou la tâche complexe non décomposée (1 seul élève).

Pour encore éclairer les résultats, on peut s'intéresser aux pourcentages de réussite observés aux différentes tâches élémentaires du dispositif d'évaluation (tableau 2).

Types de tâches	Tâches	% de réussite	Types de tâches	Tâches	% de réussite
Tâches élémentaires décontextualisées	Tableau 1	36	Tâches élémentaires en contexte	Cadre	31
	Tableau 2	37		Record	60
	Tableau 3	31		Distance	57
	Tableau 4	38			
	Tableau 5	23			
	Tableau 6	36			
	Inconnue	55			
	Formule	44			
	Représentation	55			
	Vocabulaire	73			

Tableau 2 - Pourcentages de réussite aux tâches élémentaires évaluant les ressources et procédures impliquées dans la tâche complexe « Judo/Menuisier » (épreuve 2010 – Résultats des élèves de l'ES)

Au niveau des tâches élémentaires, déterminer la longueur d'un côté d'un carré au départ de son aire (« Cadre ») est plus complexe qu'au départ de son périmètre (« Record »). Par contre, dans les tâches décontextualisées, le niveau de difficulté des questions est relativement équivalent, qu'il

L'  valuation des comp  tences en milieu scolaire et en milieu professionnel

s'agisse de calculer la longueur d'un c  t   d'un carr   au d  part de l'aire ou du p  rim  tre ou de rechercher l'une de ces deux mesures au d  part de la longueur d'un c  t   (l'ensemble des t  ches « Tableau »). Par ailleurs, rappelons que 60% des   l  ves qui ont r  ussi la t  che complexe « Judo » ont   chou      la t  che complexe d  compos  e « Menuisier », ce qui pourrait laisser supposer que, dans ce cas, c'est le calcul de la longueur au d  part du p  rim  tre qui pose le plus de difficult   aux   l  ves. Finalement, on peut s'interroger sur le r  el parall  lisme des t  ches « judo » et « menuisier », mais cela n  cessiterait des analyses compl  mentaires. Enfin, le probl  me « Distance » et la t  che   l  mentaire d  contextualis  e « Inconnue » sont les deux items qui   valuent la capacit   des   l  ves    d  duire une longueur au d  part de deux autres. Ces deux t  ches sont r  ussies respectivement par 57 et 55% des   l  ves alors que l'on aurait pu penser qu'il s'agirait des t  ches   l  mentaires les plus complexes puisque l'erreur la plus courante dans la t  che complexe « Judo » est d'aboutir    la r  ponse « 3 » qui correspond    une erreur se situant    ce niveau pr  cis³. Ces r  sultats rejoignent ainsi un des constats de l'  tude de Crahay et Detheux (2005) montrant que, dans les t  ches complexes, les   l  ves omettent souvent de mobiliser une proc  dure qu'ils ma  trisent par ailleurs. Ils montrent   galement la complexit   de cerner les difficult  s sp  cifiques des   l  ves et pointent les limites diagnostiques d'une   preuve qui ne couvrirait pas le champ de comp  tences investigu   avec suffisamment d'items et avec des items suffisamment vari  s.

Soulignons pour terminer – mais ce n'est pas l'objet du pr  sent article – que pour autoriser un r  el diagnostic pouvant aider les enseignants    cerner les forces et les faiblesses de leurs   l  ves, il conviendrait de d  passer les tendances globales et de conduire une analyse d'items au niveau individuel⁴. Con  ues dans cette optique, les   valuations externes permettraient sans doute de fournir des informations diagnostiques utiles aux enseignants, qui en semblent d'ailleurs demandeurs selon une   tude r  alis  e en 2010 au Luxembourg (Dierendonck & Fagnant, 2010c).

4. Discussion

Pour Carette et Dupriez (2009), l'  valuation de r  elles comp  tences n  cessiterait de « mettre l'  l  ve devant une feuille blanche » (p. 39) ; les   preuves « classiques » ne rempliraient pas cette exigence en proposant de nombreuses questions ouvertes    r  ponses courtes ou des questions ferm  es en vue de permettre une correction ais  e (et rapide) et de soumettre aux   l  ves un nombre important d'items. De notre point de vue, ce dernier argument (le format de la r  ponse) ne permet pas r  ellement de diff  rencier le niveau de complexit   ou non d'une t  che (ceci dans le domaine des math  matiques tout au moins) et ne constitue pas non plus un frein    une premi  re forme d'  valuation diagnostique de ces comp  tences. En effet, Loye (2005) pr  cise que si l'observation des sujets en activit   de r  solution de probl  mes ou l'analyse des traces de leurs d  marches de r  solution devraient permettre d'inf  rer leurs d  marches (et donc leurs processus cognitifs), d'autres approches plus   conomiques et plus objectives (comme des questions    choix    multiples ou des questions    r  ponse br  ve, cod  es de fa  on dichotomique) permettent   galement de r  aliser ce type d'inf  rence. En nous accordant sur cette position (possibilit   de proposer des questions    r  ponse br  ve    corriger de fa  on dichotomique) et sur notre d  finition « simplifi  e » des t  ches complexes, nous avons tent   d'  valuer le pouvoir diagnostique d'une   preuve visant      valuer plusieurs degr  s ou niveaux de comp  tence.

Les r  sultats observ  s montrent que la hi  rarchisation des t  ches ne d  pend pas uniquement du degr   de comp  tence consid  r   (selon la terminologie de Rey et al., 2003), mais que d'autres facteurs interviennent comme, sans doute, les contenus impliqu  s (la difficult   des proc  dures en jeu ou la familiarit   des   l  ves avec celles-ci) et la formulation-m  me des t  ches pouvant par

³ L'  l  ve calcule correctement la longueur d'un c  t   de la salle (12 m) et la longueur d'un c  t   du tatami (9 m) mais d  duit la largeur du chemin en soustrayant simplement les deux nombres (sa r  ponse correspond alors    deux largeurs de chemin).

⁴ A ce niveau, on pourrait compl  ter le feed-back par une analyse de la t  che complexe d  compos  e, pour voir o   se situent sp  cifiquement les erreurs des   l  ves. Cette analyse pourrait en effet   tre r  alis  e au d  part du dispositif mis en place dans la mesure o   les r  ponses    chacune des sous-questions ont   t   encod  es, avant d'  tre recod  es sur base de la r  ponse finale en termes de « 1 » ou « 0 » pour l'analyse MRI.

L'  valuation des comp  tences en milieu scolaire et en milieu professionnel

exemple rendre plus ou moins « opaque » le contenu impliqu   et les proc  dures    mobiliser. Autrement-dit, on peut supposer qu'il existe   galement diff  rents niveaux de complexit   du « cadrage » n  cessaire pour d  gager les proc  dures    mobiliser.

Un constat interpellant est que l'information fournie par l'  valuation des comp  tences de 2^e degr   dans des situations ind  pendantes les unes des autres ne conduit pas aux m  mes constats que l'  valuation de ces « m  mes » comp  tences de 2^e degr   dans des t  ches d  compos  es. En effet, sur l'  chelle issue de l'analyse de rash, les t  ches complexes d  compos  es se situent    un niveau de complexit   nettement sup  rieur    celui des t  ches ind  pendantes   valuant ces « m  mes » comp  tences de 2^e degr  . Autrement dit, si les seules t  ches   valuant des comp  tences de 2^e degr     taient des t  ches d  compos  es, on aurait sans doute sous-estim  , pour un nombre non n  gligeable d'  l  ves, le niveau de ma  trise de ces comp  tences⁵.

Par ailleurs, l'absence de hi  rarchie stricte entre les diff  rents types de t  ches et les r  sultats mettant en lumi  re le faible pouvoir « aidant » apport   par la d  composition invitent    interroger le pouvoir diagnostique pr  sum   du mod  le de d  composition en phases de Rey et al. (2003). Ce dernier constat (faible pouvoir « aidant » de la d  composition) nous a surpris dans la mesure o   les r  sultats d'une   tude pilote que nous avons soumise en 2009    176   l  ves (fili  re ES) en vue de tester une proc  dure inspir  e du mod  le d'  valuation en trois phases de Rey et al. (2003) et men  e au d  part de deux t  ches « complexes » laissait supposer un apport int  ressant de la d  composition propos  e en phase 2 (on passait par exemple de 20 %    48% de r  ussite pour le probl  me « Menuisier »). Les r  sultats obtenus aux   valuations externes sont nettement moins encourageants quant    l'aide apport  e par cette d  composition. Pour expliquer ce constat, on pr  cisera que, lors de l'  tude pilote, une information importante   tait fournie en classe aux   l  ves : les   l  ves ne recevaient la t  che d  compos  e que s'ils avaient   chou      la t  che complexe. Ces derniers   taient non seulement inform  s de leur   chec    la t  che complexe de r  f  rence, mais ils savaient par ailleurs que dans un second temps on leur proposerait    nouveau la t  che mais dans un format d  compos   en sous-questions guidant leur d  marche et ayant pour objectif affirm   de les aider    r  soudre la t  che. Ce constat m  riterait d'  tre approfondi pour voir dans quelle mesure il s'agit d'un « biais d'  chantillonnage » (l'  tude pilote n'  tait nullement repr  sentative) ou s'il s'agit d'un « biais de proc  dure de passation » qui soul  verait alors d'autres questionnements quant aux possibilit  s d'utiliser ou non des t  ches d  compos  es dans des   preuves externes    large   chelle, notamment en exploitant les potentialit  s du testing adaptatif par ordinateur.

Plus globalement, la diversit   des r  sultats observ  s plaide pour l'int  r  t d'  preuves constitu  es d'un nombre suffisants d'items et d'items suffisamment vari  s pour avoir le plus de chances de « saisir » les comp  tences r  elles des   l  ves dans le domaine de contenus investigu  . Bien que des analyses compl  mentaires soient n  cessaires pour mieux cerner le r  el pouvoir diagnostique des   preuves propos  es, les premiers r  sultats nous semblent t  moigner de l'int  r  t du mod  le propos   (combinant les apports de Rey et al. 2003 et de Crahay & Detheux, 2005) et, *a contrario*, du « danger »      valuer les comp  tences uniquement dans quelques t  ches complexes, m  me d  compos  es selon un mod  le « en phases ».

R  f  rences

- Beckers, J. (2005). Est-il possible de faire de la p  dagogique par comp  tences un alli   de l'  quit      l'  cole ? *Les cahiers du Service de P  dagogique exp  rimentale*, 21&22, 41-63.
- Carette, V. (2007). L'  valuation au service de la gestion des paradoxes li  s    la notion de comp  tence. *Mesure et   valuation*. 30(2), 49-71.

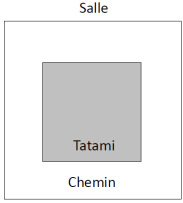
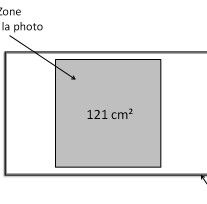
⁵ Pour nuancer ces propos, rappelons toutefois que nous avons recod   les t  ches complexes d  compos  es sur la base de la r  ponse finale uniquement et qu'une analyse fine des sous-questions impliqu  es dans ces t  ches complexes aurait probablement permis de nuancer « en partie » le constat d'  chec (« en partie » seulement, dans la mesure o   l'imbrication des sous-questions, conduit g  n  ralement une proportion non n  gligeable d'  l  ves    « abandonner » la t  che...).

L'évaluation des compétences en milieu scolaire et en milieu professionnel

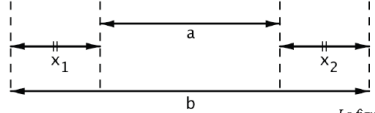
- Carette V., Dupriez V. (2009). La lente émergence d'une politique scolaire en matière d'évaluation des élèves. Quinze ans de transformations en Belgique francophone. *Mesure et Evaluation en Education*, 32 (3), p. 23-45
- Crahay, M. & Detheux, M. (2005). L'évaluation des compétences, une entreprise impossible ? (Résolution de problèmes complexes et maîtrise de procédures mathématiques). *Mesure et évaluation en éducation*, 28 (1), p. 57-76.
- De Ketele, J.-M. & Gérard, F.-M. (2005). La validation des épreuves d'évaluation selon l'approche par les compétences, *Mesure et évaluation en éducation*, 28(3), 1-26.
- Dierendonck, C. & Fagnant, A. (2010a). Quelques réflexions autour des épreuves d'évaluation développées dans le cadre de l'approche par compétences. *Le Bulletin de l'ADMEE-EUROPE*, 2010/1, 5-20.
- Dierendonck, C. & Fagnant, A. (2010b). *Monitoring du système scolaire et évaluation des compétences en mathématiques: une étude exploratoire en vue de donner un feedback diagnostique aux enseignants*. Communication orale au 22e colloque international de l'Admée-Europe. Evaluation des curriculums et des programmes d'éducation et de formation. Du 14 au 16 janvier 2010, Braga (Portugal).
- Dierendonck, C. & Fagnant, A. (2010c). *Comment les épreuves externes d'évaluation des acquis des élèves sont-elles perçues par les enseignants de l'enseignement secondaire au Luxembourg ?* Actes du congrès international de l'AREF (Actualité de la Recherche en Education et en Formation. Du 13 au 16 septembre 2010, Genève (Suisse) <http://www.unige.ch/aref2010/index.html>.
- Loye, N. (2005). Quelques modèles de mesure. *Mesure et Evaluation en Education*, 28(3), 51-68.
- Rey, B., Carette, V., Defrance, A., Kahn, S. (2003). *Les compétences à l'école : apprentissages et évaluation*. Bruxelles, De Boeck.

Annexes

Un exemple de tâche complexe (« Judo ») et de tâche élémentaire en contexte (« Cadre »)

<p>Le club de judo <i>Sudoku</i> organise un tournoi. La salle est de forme carrée et a une aire de 144 m². On veut placer un tapis (tatami) carré d'une superficie de 81 m². Le tatami est placé exactement au milieu de la salle. Il est donc entouré d'un chemin d'une certaine largeur.</p> <div style="text-align: center;">  </div> <p style="text-align: right; font-size: small;">Le dessin n'est pas à l'échelle.</p> <p>Quelle est la largeur de ce chemin? La largeur de ce chemin est de m.</p>	<p>Charles a construit un cadre pour la fête des pères. Au milieu du cadre, il a prévu une zone spéciale pour placer les photos. Cette zone de forme carrée a une aire de 121 cm².</p> <div style="text-align: center;">  </div> <p style="text-align: right; font-size: small;">La figure ne respecte pas les dimensions réelles.</p> <p>Quelles sont les dimensions de la plus grande photo que l'on peut placer dans la zone prévue à cet effet ? Les dimensions maximales sont : cm sur cm.</p>
---	---

Deux exemples de tâches élémentaires décontextualisées (« Tableau » et « Inconnue »)

<p>Complète le tableau suivant.</p> <table style="width: 100%; border-collapse: collapse; text-align: center;"> <thead> <tr> <th style="border-bottom: 1px solid black;">Longueur d'un côté du carré</th> <th style="border-bottom: 1px solid black;">Périmètre du carré</th> <th style="border-bottom: 1px solid black;">Aire du carré</th> </tr> </thead> <tbody> <tr> <td>.... cm</td> <td>.... cm</td> <td>121 cm²</td> </tr> <tr> <td>9 cm</td> <td>.... cm</td> <td>.... cm²</td> </tr> <tr> <td>.... cm</td> <td>.... cm</td> <td>64 cm²</td> </tr> <tr> <td>11 cm</td> <td>.... cm</td> <td>.... cm²</td> </tr> <tr> <td>.... cm</td> <td>64 cm</td> <td>.... cm²</td> </tr> <tr> <td>.... cm</td> <td>44 cm</td> <td>.... cm²</td> </tr> </tbody> </table>	Longueur d'un côté du carré	Périmètre du carré	Aire du carré cm cm	121 cm ²	9 cm cm cm ² cm cm	64 cm ²	11 cm cm cm ² cm	64 cm cm ² cm	44 cm cm ²	<p>a vaut 12 cm et b vaut 18 cm. x₁ et x₂ sont égales.</p> <div style="text-align: center;">  </div> <p style="text-align: right; font-size: small;">La figure ne respecte pas les dimensions réelles.</p> <p>Quelle est la longueur x₁ ? La longueur de x₁ est de cm.</p>
Longueur d'un côté du carré	Périmètre du carré	Aire du carré																				
.... cm cm	121 cm ²																				
9 cm cm cm ²																				
.... cm cm	64 cm ²																				
11 cm cm cm ²																				
.... cm	64 cm cm ²																				
.... cm	44 cm cm ²																				