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**

* *Université du Luxembourg – christophe.dierendonck@uni.lu*

** *Université de Liège – afagnant@ulg.ac.be*

Mots-clés. *Tâche complexe – Mathématiques – Evaluation à large échelle – Evaluation diagnostique*

Résumé. *En vue d'aider à cerner les « forces » et les « faiblesses » en mathématiques d'élèves de l'enseignement secondaire, un dispositif expérimental intégrant des tâches évaluant des compétences a été testé dans le cadre de deux épreuves d'évaluation externe (2009 et 2010) au Luxembourg. Les épreuves comportaient des tâches complexes, des tâches complexes décomposées et des tâches nécessitant l'application directe des ressources impliquées dans ces tâches complexes (cf. modèle d'évaluation en phases de Rey et al., 2003), mais aussi des problèmes élémentaires indépendants évaluant la capacité à mobiliser ces mêmes ressources dans d'autres contextes. Elles comprenaient également d'autres items permettant de couvrir de manière plus large les domaines mathématiques investigués. Dans les deux épreuves, la présence d'un nombre suffisant d'items (et leur indépendance) a autorisé l'utilisation d'un modèle de réponse à l'item (modèle de Rash) qui permet d'analyser, sur une échelle commune, le degré de difficulté des tâches et le niveau de compétence des élèves. La combinaison des informations issues des divers types de tâches permet d'éclairer les difficultés des élèves face aux tâches complexes.*

1. Introduction

La notion de « compétence » a pris une place considérable dans la plupart des référentiels européens de formation. On attend aujourd'hui des élèves qu'ils soient capables de mobiliser, de façon intégrée, un ensemble de ressources internes et externes pour accomplir des tâches complexes voire inédites (Beckers, 2005 ; Carette, 2007 ; De Ketele & Gérard, 2005 ; Rey, Carette, Defrance & Kahn, 2003). Si l'intérêt de l'approche par compétences est bien d'avoir replacé le concept de « mobilisation » au centre du débat éducatif, le recours à des tâches d'une grande complexité pour évaluer la compétence des élèves a conduit à plusieurs paradoxes. Parmi ceux-ci, nous en développerons trois en guise d'introduction de la présente étude.

Le premier paradoxe pose la question de la validité et de la fiabilité des épreuves d'évaluation composées d'un nombre forcément restreint de tâches complexes, étant donné le temps nécessaire à leur résolution. Sur le plan de la validité, on peut s'interroger sur le fait d'évaluer les élèves sur la base de tâches qui n'ont peut-être pas fait l'objet d'un réel apprentissage et qui sont sans doute trop ciblées pour permettre de couvrir le curriculum enseigné. Sur le plan de la fiabilité des informations récoltées, on rappellera que plus on dispose d'items qui évaluent ce que l'on souhaite mesurer, plus la mesure que l'on prend devrait être précise (Laveault & Grégoire, 1997) et ce, quel que soit le modèle de mesure utilisé (théorie classique des tests ou théorie de réponse à l'item).

Le deuxième paradoxe se situe au niveau de l'impact que peuvent avoir ces épreuves d'évaluation. Il s'avère en effet que très peu d'élèves parviennent à fournir une réponse convaincante face à ce type de tâches très complexes (Rey et al. , 2003). Le danger est donc de considérer, au départ de ces quelques tâches complexes, qu'un élève n'est pas compétent et de sanctionner ainsi son

parcours scolaire, alors qu'il maitrise 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 maitrisées, 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 1er degré qui consisteraient essentiellement à appliquer des procédures dans des situations fermées ; des compétences de 2e degré qui feraient appel à la notion de mobilisation mais qui n'impliqueraient qu'un seul type de ressources et enfin, des compétences de 3e 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 ou 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.

2.3. Les principes repris du modèle de Crahay et Detheux (2005)

Partant de l'hypothèse que la maîtrise isolée des procédures n'est pas suffisante pour résoudre des problèmes complexes impliquant l'intégration (mobilisation et coordination) de celles-ci, Crahay et Detheux (2005) ont développé un dispositif expérimental permettant de mieux cerner les forces et les faiblesses des élèves face aux tâches complexes. Plutôt que de proposer une décomposition de la tâche complexe comme dans le modèle de Rey et al. (2003), ils proposent d'évaluer isolément les procédures dans des tâches indépendantes qui sont proposées aux élèves un autre jour que la tâche complexe elle-même.

Dans leur article, les auteurs décrivent deux problèmes complexes proposés à 1436 élèves de grade 6. Chaque problème complexe implique la maîtrise de plusieurs procédures inscrites au programme de l'école primaire en Communauté française de Belgique. La maîtrise de chacune des procédures impliquées dans la résolution des deux problèmes complexes a été testée sous des formes de questionnement s'apparentant à des problèmes élémentaires impliquant la mobilisation d'une seule procédure (*compétence de 2e degré selon Rey et al., 2003 – que nous nommerons « tâches élémentaires en contexte » dans la suite du texte*) ou sous la forme de tâches décontextualisées faisant directement appel à l'application de la procédure visée (*compétence de 1er degré ou procédure selon Rey et al., 2003 – que nous appellerons « tâches élémentaires décontextualisées » dans la suite du texte).*

A titre illustratif de l'épreuve de Crahay et Detheux, un exemple de tâche élémentaire en contexte (SF1.3.) et un exemple de tâche élémentaire décontextualisée (SF 1.4.) sont présentées dans la figure 1.

SF 1.3

Une voiture a parcouru 440 km à la vitesse moyenne de 80 km/h.
Elle est partie à 8 h 35. À quelle heure est-elle arrivée à destination ?

SF 1.4

9 h 49 min + 2 h 28 min = ... h ..... min

15 h 17 min + 51 min = ... h ..... min

Figure 1 : Exemples de tâches élémentaires proposée par Crahay & Detheux (2005)

2.4. L'élaboration d'un dispositif d'évaluation incluant quatre types de tâches

Pour résoudre une tâche complexe et faire preuve de compétence, l'élève doit être capable de mobiliser les ressources adéquates, de les mettre en œuvre correctement et de les combiner si nécessaire. Dans une perspective diagnostique, on peut donc élaborer des tâches présentant différents niveaux de complexité. De manière spécifique, nous avons voulu expérimenter un dispositif d'évaluation de la compétence des élèves à résoudre un problème de géométrie nécessitant un « cadrage » de la situation, la mobilisation de plus d'une ressource ou procédure apprise et la coordination de plusieurs étapes de résolution. Concrètement, nous avons distingué quatre types de tâches qui, combinées au sein d'un même dispositif d'évaluation, pourraient apporter des informations diagnostiques intéressantes : des tâches complexes de référence, des tâches complexes décomposées, des tâches élémentaires décontextualisées et des tâches élémentaires présentées dans un autre contexte problématique.

L'objectif de l'étude est de rendre compte du potentiel diagnostique d'un tel dispositif d'évaluation en tentant de répondre aux questions suivantes :

(1) Les différents types de tâches se positionnent-ils de manière hiérarchique du point de vue de leur difficulté ?

   a. Ce positionnement hiérarchique est-il absolu ou des tâches dites élémentaires peuvent-elles présenter un degré de difficulté aussi élevé qu'une tâche complexe ?
   b. Les tâches élémentaires présentées dans un autre contexte problématique apportent-elles des informations complémentaires aux informations fournies par les tâches complexes décomposées ? Autrement dit, l'évaluation des compétences de 2e degré dans des situations indépendantes les unes des autres fournit-elle d'autres informations sur la maîtrise de ces compétences que leur évaluation dans des tâches décomposées ?

(2) L'intégration d'un tel dispositif d'évaluation au sein d'une épreuve d'évaluation externe à large échelle peut-elle déboucher sur un diagnostic précis et individualisé des lacunes et des forces des élèves ?

2.5. Le dispositif d'évaluation intégré aux épreuves standardisées de grade 9 au Luxembourg

Au Luxembourg, des épreuves d'évaluation externes et standardisées sont soumises chaque année aux élèves de grade 9 en mathématiques et en langues (allemand et français), et ce dans trois types d'enseignement (ES, EST, PR). En 2009, une première étude exploratoire a été menée au départ d'une tâche complexe, d'une tâche décomposée et de quelques tâches élémentaires proposées en contexte ou non (quelques exemples sont proposés en annexe). Les résultats ont été analysés dans le type d'enseignement secondaire général (ES, N= 1769) uniquement (Dierendonck & Fagnant, 2010b). En 2010, un design expérimental plus complexe a été élaboré, avec ancrage commun entre les types d'enseignement (ES, Est et PR, N = 6399) et des parties spécifiques à chacune d'elles.

En plus des tâches-cibles décrites au point précédent, ces épreuves comprenaient une variété de tâches évaluant des compétences de 1er, de 2e ou de 3e degré de façon à couvrir le plus largement possible (en fonction des contraintes temporelles, deux heures de cours de 50 minutes au maximum) les domaines de contenu investigués (à savoir les « nombres et opérations » et les « figures du plan et de l'espace »).

3. Résultats

Les données ont été analysés à l'aide du modèle de réponse à l'item (MRI) à un paramètre (modèle de Rash)[1] au moyen du logiciel Conquest. Pour cette analyse, les tâches complexes ont été recodées 1 (réponse finale correcte) ou 0 (réponse finale incorrecte) tant dans leur format initial (tâche complexe de référence) que dans leur format décomposé (dans ce cas précis, seule la réponse à la dernière sous-tâche a été considérée pour l'analyse MRI).

Les résultats sont présentés en deux étapes correspondant aux deux questions de recherche mentionnées précédemment : la première traite de la hiérarchisation des tâches et sera analysée sur l'ensemble de l'épreuve externe en confrontant les résultats obtenus en 2009 (en ES) et en 2010 (dans les trois types d'enseignement, grâce à la procédure d'ancrage) ; la deuxième traite du potentiel diagnostique des épreuves, en confrontant les résultats des différentes tâches évaluant la mise en œuvre des mêmes procédures dans des tâches de complexité différente.

[1] Pour rappel, la particularité des MRI est de placer sur une même échelle le niveau de difficulté des items et le niveau de compétence des élèves. Concrètement, la compétence d'un élève est définie en fonction de l'item pour lequel il a une probabilité de réussite de 50%. Sur le schéma, les élèves sont situés en vis-à-vis des items pour lesquels ils ont une probabilité de réussite de 50% ; ils ont donc une probabilité inférieure à 50% de réussir les items plus difficiles (situés plus haut sur le graphique) et une probabilité supérieure à 50% de réussir les items plus faciles (situés plus bas sur le graphique). Les sujets situés en haut du graphique sont donc les sujets les plus performants ou qui ont le niveau de compétence le plus élevé.

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.

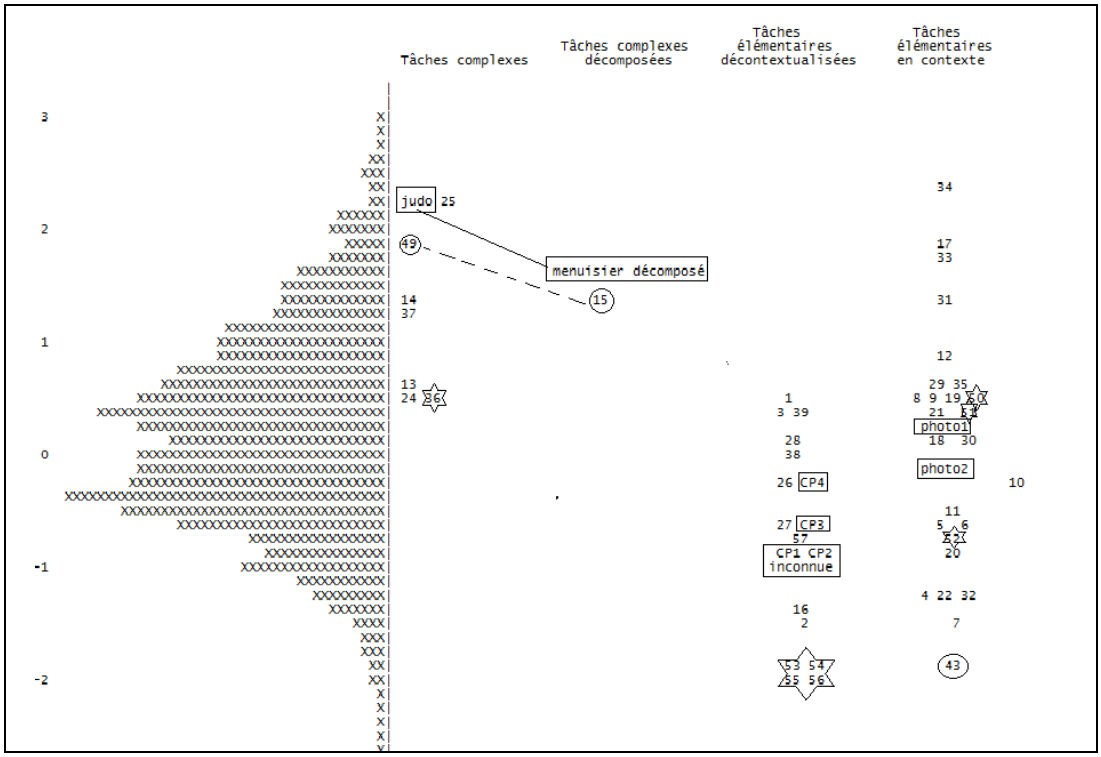


Figure 2 – Résultats en ES dans l'épreuve de 2009 (chaque 'X' représente 2.9 cas)

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).

(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%)	→ 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
			→ 1 élève (2%) réussit 0, 1 ou 2 TED et 0 ou 1 TEC
		Réussite 35 (40%)	→ 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)

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 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

s'agisse de calculer la longueur d'un côté d'un carré au départ de l'aire ou du périmètre ou de rechercher l'une de ces deux mesures au départ de la longueur d'un côté (l'ensemble des tâches « Tableau »). Par ailleurs, rappelons que 60% des élèves qui ont réussi la tâche complexe « Judo » ont échoué à la tâche complexe décomposée « Menuisier », ce qui pourrait laisser supposer que, dans ce cas, c'est le calcul de la longueur au départ du périmètre qui pose le plus de difficulté aux élèves. Finalement, on peut s'interroger sur le réel parallélisme des tâches « judo » et « menuisier », mais cela nécessiterait des analyses complémentaires. Enfin, le problème « Distance » et la tâche élémentaire décontextualisée « Inconnue » sont les deux items qui évaluent la capacité des élèves à déduire une longueur au départ de deux autres. Ces deux tâches sont réussies respectivement par 57 et 55% des élèves alors que l'on aurait pu penser qu'il s'agirait des tâches élémentaires les plus complexes puisque l'erreur la plus courante dans la tâche complexe « Judo » est d'aboutir à la réponse « 3 » qui correspond à une erreur se situant à ce niveau précis[3]. Ces résultats rejoignent ainsi un des constats de l'étude de Crahay et Detheux (2005) montrant que, dans les tâches complexes, les élèves omettent souvent de mobiliser une procédure qu'ils maîtrisent par ailleurs. Ils montrent également la complexité de cerner les difficultés spécifiques des élèves et pointent les limites diagnostiques d'une épreuve qui ne couvrirait pas le champ de compétences investigué avec suffisamment d'items et avec des items suffisamment variés.

Soulignons pour terminer – mais ce n'est pas l'objet du présent article – que pour autoriser un réel diagnostic pouvant aider les enseignants à cerner les forces et les faiblesses de leurs élèves, il conviendrait de dépasser les tendances globales et de conduire une analyse d'items au niveau individuel[4]. Conçues dans cette optique, les évaluations externes permettraient sans doute de fournir des informations diagnostiques utiles aux enseignants, qui en semblent d'ailleurs demandeurs selon une étude réalisée en 2010 au Luxembourg (Dierendonck & Fagnant, 2010c).

4. Discussion

Pour Carette et Dupriez (2009), l'évaluation de réelles compétences nécessiterait de « mettre l'élève devant une feuille blanche » (p. 39) ; les épreuves « classiques » ne rempliraient pas cette exigence en proposant de nombreuses questions ouvertes à réponses courtes ou des questions fermées en vue de permettre une correction aisée (et rapide) et de soumettre aux élèves un nombre important d'items. De notre point de vue, ce dernier argument (le format de la réponse) ne permet pas réellement de différencier le niveau de complexité ou non d'une tâche (ceci dans le domaine des mathématiques tout au moins) et ne constitue pas non plus un frein à une première forme d'évaluation diagnostique de ces compétences. En effet, Loye (2005) précise que si l'observation des sujets en activité de résolution de problèmes ou l'analyse des traces de leurs démarches de résolution devraient permettre d'inférer leurs démarches (et donc leurs processus cognitifs), d'autres approches plus économiques et plus objectives (comme des questions à choix à multiples ou des questions à réponse brève, codées de façon dichotomique) permettent également de réaliser ce type d'inférence. En nous accordant sur cette position (possibilité de proposer des questions à réponse brève à corriger de façon dichotomique) et sur notre définition « simplifiée » des tâches complexes, nous avons tenté d'évaluer le pouvoir diagnostique d'une épreuve visant à évaluer plusieurs degrés ou niveaux de compétence.

Les résultats observés montrent que la hiérarchisation des tâches ne dépend pas uniquement du degré de compétence considéré (selon la terminologie de Rey et al., 2003), mais que d'autres facteurs interviennent comme, sans doute, les contenus impliqués (la difficulté des procédures en jeu ou la familiarité des élèves avec celles-ci) et la formulation-même des tâches pouvant par

[3] L'élève calcule correctement la longueur d'un côté de la salle (12 m) et la longueur d'un côté du tatami (9 m) mais déduit la largeur du chemin en soustrayant simplement les deux nombres (sa réponse correspond alors à deux largeurs de chemin).

[4] A ce niveau, on pourrait compléter le feed-back par une analyse de la tâche complexe décomposée, pour voir où se situent spécifiquement les erreurs des élèves. Cette analyse pourrait en effet être réalisée au départ du dispositif mis en place dans la mesure où les réponses à chacune des sous-questions ont été encodées, avant d'être recodées sur base de la réponse finale en termes de « 1 » ou « 0 » pour l'analyse MRI.

exemple rendre plus ou moins « opaque » le contenu impliqué et les procédures à mobiliser. Autrement-dit, on peut supposer qu'il existe également différents niveaux de complexité du « cadrage » nécessaire pour dégager les procédures à mobiliser.

Un constat interpellant est que l'information fournie par l'évaluation des compétences de 2e degré dans des situations indépendantes les unes des autres ne conduit pas aux mêmes constats que l'évaluation de ces « mêmes » compétences de 2e degré dans des tâches décomposées. En effet, sur l'échelle issue de l'analyse de rash, les tâches complexes décomposées se situent à un niveau de complexité nettement supérieur à celui des tâches indépendantes évaluant ces « mêmes » compétences de 2e degré. Autrement dit, si les seules tâches évaluant des compétences de 2e degré étaient des tâches décomposées, on aurait sans doute sous-estimé, pour un nombre non négligeable d'élèves, le niveau de maîtrise de ces compétences[5].

Par ailleurs, l'absence de hiérarchie stricte entre les différents types de tâches et les résultats mettant en lumière le faible pouvoir « aidant » apporté par la décomposition invitent à interroger le pouvoir diagnostique présumé du modèle de décomposition en phases de Rey et al. (2003). Ce dernier constat (faible pouvoir « aidant » de la décomposition) nous a surpris dans la mesure où les résultats d'une étude pilote que nous avons soumise en 2009 à 176 élèves (filière ES) en vue de tester une procédure inspirée du modèle d'évaluation en trois phases de Rey et al. (2003) et menée au départ de deux tâches « complexes » laissait supposer un apport intéressant de la décomposition proposée en phase 2 (on passait par exemple de 20 % à 48% de réussite pour le problème « Menuisier »). Les résultats obtenus aux évaluations externes sont nettement moins encourageants quant à l'aide apportée par cette décomposition. Pour expliquer ce constat, on précisera que, lors de l'étude pilote, une information importante était fournie en classe aux élèves : les élèves ne recevaient la tâche décomposée que s'ils avaient échoué à la tâche complexe. Ces derniers étaient non seulement informés de leur échec à la tâche complexe de référence, mais ils savaient par ailleurs que dans un second temps on leur proposerait à nouveau la tâche mais dans un format décomposé en sous-questions guidant leur démarche et ayant pour objectif affirmé de les aider à résoudre la tâche. Ce constat mériterait d'être approfondi pour voir dans quelle mesure il s'agit d'un « biais d'échantillonnage » (l'étude pilote n'était nullement représentative) ou s'il s'agit d'un « biais de procédure de passation » qui soulèverait alors d'autres questionnements quant aux possibilités d'utiliser ou non des tâches décomposées dans des épreuves externes à large échelle, notamment en exploitant les potentialités du testing adaptatif par ordinateur.

Plus globalement, la diversité des résultats observés plaide pour l'intérêt d'épreuves constituées d'un nombre suffisants d'items et d'items suffisamment variés pour avoir le plus de chances de « saisir » les compétences réelles des élèves dans le domaine de contenus investigué. Bien que des analyses complémentaires soient nécessaires pour mieux cerner le réel pouvoir diagnostique des épreuves proposées, les premiers résultats nous semblent témoigner de l'intérêt du modèle proposé (combinant les apports de Rey et al. 2003 et de Crahay & Detheux, 2005) et, *a contrario*, du « danger » à évaluer les compétences uniquement dans quelques tâches complexes, même décomposées selon un modèle « en phases ».

Références

Beckers, J. (2005). Est-il possible de faire de la pédagogie par compétences un allié de l'équité à l'école ? *Les cahiers du Service de Pédagogie expérimentale, 21&22,* 41-63.

Carette, V. (2007). L'évaluation au service de la gestion des paradoxes liés à la notion de compétence. *Mesure et évaluation. 30(2),* 49-71.

[5] Pour nuancer ces propos, rappelons toutefois que nous avions recodé les tâches complexes décomposées sur la base de la réponse finale uniquement et qu'une analyse fine des sous-questions impliquées dans ces tâches complexes aurait probablement permis de nuancer « en partie » le constat d'échec (« en partie » seulement, dans la mesure où l'imbrication des sous-questions, conduit généralement une proportion non négligeable d'élèves à « abandonner » la tâche…).

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.

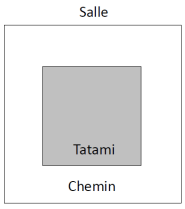
Loye, N. (2005). Quelques modèles de mesure. *Mesure et Evaluation en Education, 28(3)*, 51-68.

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 »)

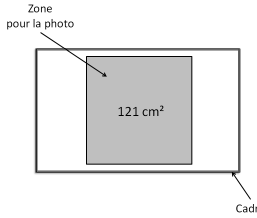
Le club de judo *Sudoku* organise un tournoi. La salle est de forme carrée et a une aire de 144 m². On veut placer un tapis (tatami) carré d'une superficie de 81 m². Le tatami est placé exactement au milieu de la salle. Il est donc entouré d'un chemin d'une certaine largeur.



Le dessin n'est pas à l'échelle.

Quelle est la largeur de ce chemin?
La largeur de ce chemin est de …..... m.

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 cm^2.



La figure ne respecte pas les dimensions réelles.

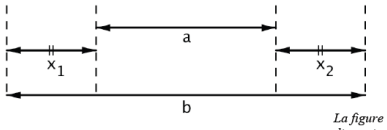
Quelles sont les dimensions de <u>la plus grande</u> photo que l'on peut placer dans la zone prévue à cet effet ?
Les dimensions maximales sont : …… cm sur ……. cm.

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

Complète le tableau suivant.

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

a vaut 12 cm et b vaut 18 cm.
x_1 et x_2 sont égales.



La figure ne respecte pas les dimensions réelles.

Quelle est la longueur x1 ?
La longueur de x_1 est de …… cm.