Article

« Approche par compétences et évaluation à large échelle : deux logiques incompatibles ? »

Christophe Dierendonck et Annick Fagnant

Mesure et évaluation en éducation, vol. 37, n° 1, 2014, p. 43-82.

Approche par compétences et évaluation à large échelle : deux logiques incompatibles ?

Christophe Dierendonck
Université du Luxembourg

Annick Fagnant
Université de Liège

Mots clés : approche par compétences, évaluation à large échelle, tâche complexe, dispositif d'évaluation

Depuis les années 1990, un changement de paradigme s'opère dans la plupart des systèmes d'éducation : le pilotage centré sur les ressources investies (inputs) se transforme progressivement en un pilotage centré sur les résultats (outputs). C'est dans ce contexte que se sont développés les dispositifs d'évaluation externe à large échelle des acquis scolaires des élèves, tant au niveau international (comme PIRLS et PISA) qu'au niveau national ou même au niveau local. En parallèle, l'approche par compétences a été instaurée progressivement dans les référentiels scolaires et les classes. Basée sur un travail au départ de tâches complexes, l'approche par compétences n'est cependant pas sans poser certains problèmes sur le plan de l'évaluation. Certains auteurs affirment même que les évaluations externes à large échelle et l'approche par compétences présentent des logiques contradictoires et que, dès lors, les deux éléments sont inconciliables. L'objectif de cet article est de nuancer ce genre d'affirmations en rendant compte d'un dispositif d'évaluation exploratoire à large échelle combinant une évaluation « classique » (constitué d'un nombre important d'items à corriger de façon standardisée) et une évaluation de compétences à visée diagnostique (tâches complexes, tâches complexes décomposées et tâches élémentaires associées).

Key words: competency-based approach, large scale assessment, complex task, assessment design

Since the 1990's, there is a paradigm shift in most educational systems: the initial input-oriented school monitoring shift to an output-oriented monitoring system. In this context, large scale external assessments of students achieve-

ments were developed both at the international level (see PIRLS or PISA), at the national level or even at the local level. In parallel, there was the progressive introduction of the competency-based approach in curricula and classrooms. Based on complex tasks, competency-based approach is however not without problems in terms of evaluation. Some authors even claim that large scale assessments and competency-based approach have conflicting logics and that both elements are irreconcilable. The aim of this paper is to qualify such statements by reporting an exploratory large-scale assessment combining a "classical" evaluation (with an important number of items to be corrected in a standardized way) and a competency-based evaluation with a diagnostic aim (complex tasks, decomposed complex tasks and associated elementary tasks).

Palavras-chave: abordagem por competências, avaliação em larga escala, tarefa complexa, dispositivo de avaliação

Depois dos anos 90, operou-se uma mudança de paradigma na maior parte dos sistemas educativos: de uma pilotagem centrada sobre os recursos investidos (inputs), passou-se para uma pilotagem centrada nos resultados (outputs). É neste contexto que se desenvolvem os dispositivos de avaliação externa em larga escala das aprendizagens escolares dos alunos, quer ao nível internacional (como PIRLS e PISA), quer ao nível nacional e até ao nível local. Em paralelo, assistiu-se à instauração progressiva da abordagem por competências nos referenciais escolares e nas turmas. Baseada num trabalho de tarefas complexas, a abordagem por competências, no entanto, não deixa de colocar certos problemas no plano da avaliação. Alguns autores afirmam mesmo que as avaliações externas em larga escala e a abordagem por competências apresentam lógicas contraditórias e que, portanto, são dois elementos inconciliáveis. O objetivo deste artigo é matizar este género de afirmações, dando conta de um dispositivo de avaliação exploratório em larga escala que combina uma avaliação "clássica" (constituída por um número importante de itens a corrigir de modo estandardizado) e uma avaliação de competências para fins de diagnóstico (tarefas complexas, tarefas complexas decompostas e tarefas básicas associadas).

Note des auteurs – Cet article constitue une synthèse des travaux exploratoires menés au Luxembourg. La correspondance liée à cet article peut être adressée à Christophe Dierendonck, Adjoint de recherche, Université du Luxembourg, Faculté des Lettres, des Sciences humaines, des Arts et des Sciences de l'éducation, Campus Walferdange, Bât. XI, 2.01, Route de Diekirch, L-7220, Walferdange, téléphone : +352 46 66 44 9485, ou par courriel à l'adresse suivante : [christophe.dierendonck@uni.lu]

Introduction

Ces dernières années ont vu fleurir, un peu partout dans le monde, d'importantes réformes des systèmes d'éducation et de formation professionnelle (Gauthier, 2006 ; UNESCO, 2007). À la base de ce mouvement se trouve une notion centrale – la compétence – dont il semble difficile de retracer les origines multiples. Certes, la notion a le mérite d'avoir renouvelé la question du transfert des apprentissages et celle de l'apprentissage en situation, mais elle demeure un concept flou, régulièrement remis en cause pour sa polysémie ou son manque d'assise théorique (Crahay, 2006 ; Dierendonck, Loarer, & Rey, 2014 ; Romainville, 2006). Pourtant, en dépit de ses interprétations multiples et de ses fondements apparemment fragiles, la notion de compétence a donné naissance à un courant devenu dominant – l'approche par compétences (APC) – qui a progressivement investi les référentiels de la plupart des systèmes d'éducation des pays européens (Eurydice, 2012) avec l'apparition de finalités formulées en termes de niveaux de compétence ou de socles à atteindre aux différentes étapes de la scolarité obligatoire et qui s'est imposé aux enseignants et aux écoles.

Cette révolution curriculaire et pédagogique, qui a pris des formes différentes et connu des succès divers selon le pays considéré, s'accompagne dans la plupart des cas d'une mise en place de dispositifs d'assurance-qualité et de pilotage des systèmes scolaires (Lafontaine, Soussi, & Nidegger, 2009). Pour Crahay, Audigier et Dolz (2006), depuis les années 1990, un changement de paradigme qui consiste en une transition progressive d'un pilotage centré sur les *inputs* (les ressources investies dans le système) vers un pilotage et une gestion centrés sur les *outputs* (les résultats obtenus par le système) se manifesterait. C'est dans ce contexte que se sont fortement développés les dispositifs d'évaluation externe à large échelle des acquis des élèves, qui sont connus aujourd'hui tant sur le plan international, avec principalement les enquêtes de l'IEA (PIRLS, TIMMS) et de l'OCDE (PISA), que sur le plan national (évaluations-bilans en France, épreuves standardisées au Luxembourg, évaluations externes non certificatives en Communauté française de Belgique, etc.) ou même sur le plan local (épreuves cantonales en Suisse, épreuves communales en Communauté

française de Belgique, etc.). Le postulat qui sous-tend ces évaluations externes des acquis des élèves consiste à dire qu'elles peuvent jouer un double rôle : d'un part, aider au pilotage des systèmes scolaires en fournissant des indicateurs transversaux et longitudinaux et, d'autre part, réguler les pratiques en apportant un *feedback* aux différents niveaux du système (école, classe, élève).

Parallèlement (ou en réaction) à cette déferlante d'évaluations externes à large échelle, de plus en plus de chercheurs francophones[1] (Carette, 2007 ; De Ketele & Gérard, 2005 ; Gérard, 2008 ; Rey, Carette, Defrance, & Kahn, 2003 ; Scallon, 2004) se sont lancés de façon évidente dans une quête d'outils d'évaluation alternatifs qui seraient, selon eux, davantage adaptés aux principes de l'approche par compétences.

C'est cette opposition couramment dénoncée entre les outils alternatifs d'évaluation des compétences (Carette, 2007 ; Carette & Dupriez, 2009 ; Rey et al., 2003 ; De Ketele & Gérard, 2005) et les évaluations standardisées dites classiques (comme les enquêtes PIRLS, PISA ou les évaluations externes nationales) qui constitue le point de départ de cet article. En caricaturant, on dira que le débat oppose ceux qui affirment que les évaluations à large échelle dites classiques n'évaluent pas de réelles compétences et ceux qui affirment que les outils d'évaluation alternatifs développés en référence à l'approche par compétences ne sont pas fiables du point de vue de la mesure.

Un premier point de divergence entre ces deux positionnements tient notamment au fait que les objectifs prioritaires des deux types d'épreuves sont fondamentalement différents : outils de pilotage du système dans un cas, outils généralement à visée diagnostique dans l'autre. Si l'on s'accorde à reconnaître que les enquêtes internationales (PIRLS ou PISA, par exemple) n'ont pas une visée diagnostique ciblée sur le plan individuel (notamment puisque qu'elles testent des échantillons d'élèves et qu'elles n'organisent pas un retour d'information sur le plan des classes), il n'en va pas de même pour certaines épreuves externes régionales ou nationales qui déclarent de telles intentions diagnostiques. Par exemple, les évaluations externes non certificatives proposées en Belgique francophone offrent aux enseignants l'occasion de disposer des résultats individualisés de leurs élèves en vue de cerner leurs forces et leurs faiblesses face aux différentes dimensions évaluées dans l'épreuve. D'autres dispositifs nationaux d'évaluation externe des acquis des élèves (par exemple, les épreuves standar-

disées au Luxembourg) s'attachent, avant tout, aux résultats globaux du système et aux résultats spécifiques des écoles et des classes, mais ils ne donnent que très peu d'occasions aux enseignants de se pencher sur les résultats spécifiques de leurs élèves, ce qui rend leur portée diagnostique clairement limitée (au plus grand regret des acteurs concernés, comme l'ont montré Dierendonck & Fagnant, 2010b, 2010c).

Un autre point de divergence tient au fait que les deux types d'épreuves n'évaluent pas « la même chose » : les premières s'inscriraient dans une approche parcellisée des savoirs et des savoir-faire tandis que les secondes permettraient l'expression de réelles compétences en situation. Ainsi, les enseignants se trouveraient « perdus » face à certaines contradictions quant à la façon dont les attendus du système sont définis, d'une part, dans les référentiels de compétences et dans les divers documents qui les accompagnent (en ce inclus, certains « outils d'évaluation » proposés à titre illustratif) et, d'autre part, dans les différentes épreuves externes d'évaluation (nationales ou internationales) auxquelles leurs élèves sont soumis (Carette & Dupriez, 2009 ; Lafontaine, 2012 pour une discussion de cette question en Belgique francophone) [2].

Pour tenter d'éclairer ce débat, l'article se propose

1) de questionner plusieurs dispositifs d'évaluation existants (leurs apports respectifs et leurs complémentarités éventuelles) en les confrontant à la définition de la notion de compétence qu'ils véhiculent,
2) de présenter et discuter les résultats des travaux exploratoires des auteurs, qui ont tenté une conciliation partielle entre les approches à large échelle « classiques » et certains apports des nouveaux outils d'évaluation des compétences.

Concrètement, l'étude teste la possibilité d'intégrer, au sein d'une épreuve d'évaluation externe « classique » (au sens de constituée d'un nombre important d'items et pouvant être analysée à l'aide des modèles statistiques couramment utilisés dans les dispositifs d'évaluation à large échelle), quelques tâches complexes visant à évaluer des compétences, ainsi que des tâches décomposées ou apparentées, telles que proposées dans certains dispositifs d'évaluation à visée diagnostique. Cette possible « conciliation » sera étudiée dans le cadre particulier des mathématiques, domaine où il est possible de proposer des tâches complexes, nécessitant la mobilisation et l'intégration d'un ensemble de ressources préalablement

développées en classe pour faire face à un problème inédit, tout en aboutissant à une réponse numérique précise (Loye, 2005 ; Loye, Caron, Pineault, Tessier-Baillargeaon et Burney-Vincent, 2011). Ce choix facilite et rend fiables les corrections, mais il ne résout évidemment pas les problèmes qui pourraient être soulevés dans le cadre d'épreuves nécessitant des productions complexes face à des tâches ouvertes, qui pourraient par exemple être éprouvées lors de démonstrations ou d'argumentations mathématiques ou lors de tâches de production écrite ou orale en langues.

L'analyse de quatre approches évaluatives comme point de départ de la réflexion

Il sera tout d'abord présenté deux approches d'évaluation dites alternatives développées dans le cadre de l'approche par compétences : les épreuves d'évaluation par situations complexes de De Ketele et Gérard (2005), et le modèle d'évaluation en trois phases de Rey et al. (2003). Une approche (Crahay & Detheux, 2005), qui ne se pose pas en tant que modèle d'évaluation mais qui propose de considérer plusieurs types de tâches complémentaires pour cerner les forces et faiblesses des élèves et mieux comprendre de la sorte leurs difficultés face aux tâches complexes, sera ensuite décrite. Pour traiter des évaluations à large échelle dites classiques, le texte analysera enfin le modèle d'évaluation sous-jacent à l'épreuve de mathématiques développée dans le cadre du PISA 2003.

Trois approches pour évaluer les compétences scolaires

Les épreuves d'évaluation par situations complexes de De Ketele et Gérard (2005)

Pour De Ketele et Gérard (2005), l'approche par compétences cherche à développer la possibilité pour les apprenants de mobiliser un ensemble intégré de ressources afin de résoudre une situation-problème appartenant à une famille de situations[3]. Selon eux, les épreuves élaborées selon l'approche par compétences consistent, par essence, « 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 » (*ibid*, p. 8). Gérard (2008) affirme que l'élément le plus important dans la notion de compétence est l'aspect « intégration » : « la compétence se mani-

feste dans une situation complexe qui intègre un certain nombre d'éléments et qui nécessite de mobiliser – c'est-à-dire d'identifier et d'utiliser conjointement – un ensemble intégré de ressources » (p. 51).

Pour correspondre à leur définition de la compétence, De Ketele et Gérard (2005) estiment que les tâches demandées aux élèves peuvent être de deux ordres : soit on propose une tâche unique complexe (par exemple, produire un texte répondant à une situation de communication), soit on décompose une production en plusieurs tâches comprenant elles-mêmes plusieurs étapes. Mais, précisent les chercheurs, « en tous les cas, il ne sera pas possible de constituer un nombre d'« items », parce que ceci reviendrait à décomposer la tâche complexe en sous-unités, ce qui ne permettrait plus d'évaluer la compétence qui consiste bien à gérer la complexité » (*ibid*, p. 8).

Pratiquement, De Ketele et Gérard (2005) et Gérard (2008) fondent leurs évaluations par situations complexes sur l'idée qu'on peut évaluer une compétence à partir de deux ou trois situations complexes, de critères d'évaluation (les qualités minimales que doit respecter la production de l'élève) et de la « règle des 2/3 » (De Ketele, 1996). Cette règle, « validée empiriquement » selon Roegiers (2005, p. 7), postule d'offrir à l'élève au moins trois occasions indépendantes de démontrer sa compétence. Selon Gérard (2008), « on considérera qu'un élève maîtrise un critère lorsqu'il réussit au moins deux des trois occasions qui lui sont offertes » (p. 80). Il semblerait donc à la fois nécessaire et suffisant que l'élève réussisse deux tâches complexes sur trois pour être déclaré compétent.

Le modèle d'évaluation de Rey et al. (2003)

La définition de la notion de compétence proposée par Rey et al. (2003) va dans le même sens que celle donnée par De Ketele et Gérard (2005) puisqu'ils définissent « l'authentique compétence » comme « la capacité à répondre à des situations complexes et inédites par une combinaison nouvelle de procédures connues ; et non pas seulement à répondre par une procédure stéréotypée à un signal préétabli » (p. 26). À partir de cette définition, Rey et al. (2003) distinguent trois degrés de compétences qu'ils définissent comme suit :

1) Les *compétences de premier degré* (ou « procédures ») consistent à « savoir exécuter une opération (ou une suite prédéterminée d'opérations) en réponse à un signal (qui peut être en classe une question, une consigne, ou une situation connue et identifiable sans difficulté, ni ambiguïté) » (p. 26).
2) Les *compétences de deuxième degré* impliquent de « posséder toute une gamme de ces compétences élémentaires et savoir, dans une situation inédite, choisir celle qui convient » (p. 26). Face à ce type de tâche, une interprétation de la situation est nécessaire et l'élève doit mobiliser (et non simplement appliquer) la procédure adéquate.
3) Les *compétences de troisième degré* consistent à « savoir choisir et combiner plusieurs compétences élémentaires pour traiter une situation nouvelle et complexe » (p. 26). Dans ce type de situations, les élèves doivent « choisir et combiner, parmi les procédures qu'ils connaissent, plusieurs d'entre elles afin de résoudre adéquatement un problème nouveau pour eux » (*ibid*, p. 44).

En s'appuyant *stricto sensu* sur ces définitions, c'est le concept de « mobilisation » qui permettrait de distinguer le premier degré (savoir exécuter) du deuxième degré (savoir choisir la ressource appropriée) et c'est le concept d'« intégration » qui permettrait de distinguer le deuxième degré (choisir *la* ressource, *celle* qui est appropriée) du troisième degré (savoir choisir et combiner *des* ressources)[4].

Sur la base de cette distinction, les auteurs développent un modèle d'évaluation des compétences en trois phases décrites comme suit par Carette (2007) :

> Phase 1 : On demande aux élèves d'accomplir une tâche complexe, exigeant le choix et la combinaison d'un nombre significatif de procédures qu'ils sont censés posséder à la fin d'un cycle[5].
>
> Phase 2 : On propose à nouveau aux élèves la même tâche. Mais cette fois, la tâche complexe est découpée en tâches élémentaires dont les consignes sont explicites et qui sont présentées dans l'ordre ou elles doivent être accomplies pour parvenir à la réalisation de la tâche complexe globale. Mais il appartient à l'élève, pour chacune de ces tâches élémentaires, de déterminer la procédure à mettre en œuvre parmi celles qu'il est censé posséder.
>
> Phase 3 : On propose aux élèves une série de tâches simples décontextualisées, dont les consignes sont celles qui sont utilisées ordinairement dans l'apprentissage des procédures élémentaires qu'on propose à l'école : effectuer

> une soustraction ; écrire des mots ; accorder un verbe avec un sujet ; etc. Ces tâches correspondent aux procédures élémentaires qui ont dû être mobilisées pour accomplir la tâche complexe de la phase 1 (pp. 62-63).

Un des objectifs centraux du phasage est de permettre de poser un diagnostic sur ce qui est maîtrisé ou non par les élèves. Pour Carette (2007), il est en effet essentiel « d'aider les enseignants à mieux cerner le « savoir mobiliser » de leurs élèves, qui représente […] un enjeu essentiel de l'introduction de la notion de compétence dans le monde scolaire », ce qui, selon lui, nécessite « des outils qui leur permettent de recueillir une information pertinente sur cette *mystérieuse capacité* » (p. 60). Dans le dispositif en phases, la phase 1 observe les élèves qui gèrent la complexité du choix et de la combinaison des procédures nécessaires pour réaliser la tâche complexe. La phase 3 apporte une information sur la maîtrise des procédures nécessaires à la résolution de cette tâche complexe. Entre ces deux extrêmes, la phase 2 guide l'élève dans la résolution de la tâche complexe en lui proposant un découpage de celle-ci en autant d'étapes nécessaires à sa résolution. Si l'objectif des auteurs du modèle est ici de cerner les compétences de deuxième degré (le découpage permettant de proposer des sous-tâches nécessitant la mobilisation d'une seule ressource à la fois), force est de constater qu'implicitement surgit l'idée selon laquelle le découpage ainsi organisé devrait aussi aider l'élève dans la résolution de la tâche complexe[6], et ceci d'autant plus que le modèle s'appuie sur une hiérarchie présumée entre les trois degrés de compétence.

Les données présentées par Rey et al. (2003) fournissent d'ailleurs la preuve empirique qu'il existe une certaine hiérarchie entre les trois phases de leur modèle d'évaluation : en moyenne, les items les mieux réussis sont ceux de la phase 3 (compétences de premier degré), puis ceux de la phase 2 (compétences de deuxième degré) et enfin ceux de la phase 1 (compétences de troisième degré). Rien ne permet cependant d'affirmer que la hiérarchie découverte est une hiérarchie stricte. Tout d'abord, les items proposés en phase 3 par les auteurs dépassent (assez largement parfois) le cadre de l'application stricte des procédures mobilisées lors des phases 1 et 2. Par exemple, si une multiplication intervient lors de la phase 1, la phase 3 évalue plus largement cette procédure en proposant des items impliquant différents niveaux de maîtrise (par exemple[7] : $38 \times 33 = 209 \times \ldots$). Si l'intérêt de proposer une évaluation plus large des procédures est reconnu (voir Dierendonck & Fagnant, 2010a, pour une argumentation plus développée), force est de constater que cet exemple illustre assez bien la nuance

qu'il convient de faire entre les notions de difficulté et de complexité. Il est en effet possible qu'une tâche complexe (au sens de nécessitant la mobilisation et l'intégration de plusieurs ressources/procédures) présente un niveau de difficulté relativement faible et qu'à l'inverse, une tâche élémentaire s'avère extrêmement difficile pour les élèves. Tâche complexe et tâche compliquée (pour reprendre la terminologie de Roegiers, 2007), tout comme tâche élémentaire et tâche facile, ne sont en effet nullement synonymes[8].

Le dispositif expérimental d'évaluation 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 1 436 élèves de fin d'enseignement primaire. Chaque problème complexe implique la maîtrise de plusieurs procédures inscrites au programme de l'école primaire en Belgique francophone. 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 deuxième degré selon Rey et al., 2003) ou sous la forme de tâches décontextualisées faisant directement appel à l'application de la procédure visée (compétence de premier degré ou procédure selon Rey et al., 2003). Un exemple de tâche élémentaire en contexte (SF1.3.) et un exemple de tâche élémentaire décontextualisée (SF 1.4.) sont présentés 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)***

Que retenir de ces trois approches d'évaluation?

Les modalités d'évaluation proposées par De Ketele et Gérard (2005), par Rey et al. (2003) et par Crahay et Detheux (2005) semblent intéressantes d'un point de vue diagnostique. Toutefois, dès l'instant où une épreuve d'évaluation s'appuie uniquement sur deux ou trois tâches complexes (mêmes décomposées) pour évaluer les compétences des élèves, la fiabilité de la mesure prise peut être questionnée puisque les interprétations ou les décisions (d'ordre pédagogique ou autres) qui pourraient en découler ne sont, en réalité, basées que sur un nombre restreint de questions ou d'items. Or, quel que soit le modèle de mesure utilisé (théorie classique des tests ou théorie de réponse à l'item), 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). Des études réalisées sur les modèles de réponse à l'item montrent par ailleurs l'influence du nombre d'items sur la précision des estimations de la compétence des sujets (Burton, 2004; Hulin, Lissak, & Drasgow, 1982). Conscients de cette difficulté, inhérente selon eux aux nouveaux outils d'évaluation proposés, certains des auteurs précités font appel à « de nouvelles approches statistiques qui restent à ce jour à inventer » (Carette, 2009, p. 155) ou qu'ils ont tenté d'opérationnaliser, comme avec la règle des deux tiers précitée (De Ketele & Gérard, 2005).

L'évaluation des compétences dans l'épreuve de mathématique PISA 2003

Dans PISA 2003, la compétence mathématique (ou la culture mathématique pour reprendre la terminologie de l'OCDE) est définie comme « l'aptitude d'un individu à identifier et à comprendre le rôle joué par les mathématiques dans le monde, à porter des jugements fondés à leur propos, et à s'engager dans des activités mathématiques, en fonction des exigences de la vie en tant que citoyen constructif, impliqué et réfléchi » (OCDE, 2004, p. 27). Elle implique la capacité des élèves à analyser, raisonner et communiquer de manière efficace lorsqu'ils posent, résolvent et interprètent des problèmes mathématiques dans une variété de situations impliquant des quantités, des concepts spatiaux, probabilistes ou autres. Dans son optique de « culture mathématique », PISA (2003) confronte principalement les élèves à des problèmes ancrés dans le monde « réel ». L'objectif est de voir dans quelle mesure ils peuvent se servir d'un bagage mathématique qu'ils ont acquis au cours de leur scolarité pour résoudre des problèmes variés.

PISA (2003) propose une épreuve d'évaluation « classique » (au sens de constituée d'une série d'items indépendants[9]) qui peut dès lors être validée et analysée à l'aide de divers modèles statistiques. Le modèle statistique utilisé par PISA est un modèle de réponse à l'item (MRI) à un paramètre (le modèle de Rasch) qui permet de positionner les items et les élèves sur une même échelle (de difficulté pour les items, de compétence pour les élèves). Sur la base des résultats empiriques, l'échelle est ensuite décomposée en plusieurs niveaux de compétence (six niveaux en mathématiques) qui correspondent à des ensembles de tâches de difficulté croissante. Chaque élève participant à l'évaluation est alors positionné dans un niveau de l'échelle en fonction de sa performance au test. Un élève situé à un niveau de compétence donné a une probabilité de réussir 50 % des questions se situant à ce niveau de l'échelle. Autrement dit, si un élève est placé dans le niveau 2, il est capable de réussir au minimum 50 % des items qui composent ce niveau ; il a en outre une probabilité supérieure à 50 % de réussir les items situés au niveau 1 et une probabilité inférieure à 50 % de réussir les items des niveaux 3, 4, 5 et 6.

Il n'est pas question d'entrer dans l'analyse des résultats eux-mêmes. Ce qui intéresse ici, c'est la façon dont PISA traduit le principe d'évaluation des compétences. Les élèves sont soumis à une série de tâches et l'analyse MRI permet de les situer dans un niveau de compétence. Il ne s'agit donc pas ici de réfléchir en termes dichotomiques (l'élève est compétent ou il ne l'est pas) mais plutôt en termes de *niveaux de compétence* qui peuvent être décrits à partir des tâches qui les constituent, ce qui permet de cerner ce que les élèves sont capables ou non de faire ; tâches pour lesquelles ils ont un niveau de maîtrise acceptable, tâches qu'ils maîtrisent pleinement et tâches qu'ils ne maîtrisent pas.

De Ketele et Gérard (2005) critiquent la pertinence des épreuves d'évaluation telles que TIMSS ou PISA qui proposeraient, selon eux, « un ensemble d'items selon une structure très élaborée issue principalement d'une approche par les contenus ou par les objectifs » (p. 5). Ils estiment que si ces épreuves « 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 » (p. 5). De Ketele et Gérard (2005) définissent la pertinence comme étant « le caractère plus ou moins approprié de l'épreuve, selon qu'elle s'inscrit dans la ligne des objectifs visés » (p. 3). Pour être pertinentes dans une approche par les compétences, les épreuves d'évaluation devraient donc mesurer de réelles compétences. C'est là que les distinctions sémantiques opérées par la suite par De Ketele (2010, 2011)[10] s'avèrent intéressantes. L'auteur distingue ainsi les évaluations portant sur des « ressources » (connaissances, applications et applications habillées) de celles portant sur des « compétences » ; les premières impliquant un savoir-restituer ou des savoir-faire de base alors que les secondes impliqueraient des savoir-faire complexes ou un savoir-transférer. Pour De Ketele (2011), la différence qui existe entre « applications habillées » (savoir-faire de base) et « compétences » (savoir-faire complexe ou savoir-transférer) est que, pour les premières, les ressources à mobiliser sont indiquées dans la consigne, tandis que pour les secondes, les ressources à mobiliser sont à déduire de l'analyse de la tâche. De Ketele (2011) considère alors que les tâches dans PISA (2003) sont essentiellement des applications habillées et n'évaluent donc pas de réelles compétences.

À titre illustratif, on pourrait considérer que l'énoncé de la tâche « Vol spatial » (figure 2) fait explicitement appel aux ressources à mobiliser (calcul de la circonférence d'un cercle) puisqu'il contient le terme « circonférence » et la référence à la formule de calcul de celle-ci ($\pi \times 12\,700$).

VOL SPATIAL

La station MIR tournait autour de la Terre à une altitude d'à peu près 400 km. Le diamètre de la Terre est d'environ 12 700 km et sa circonférence d'environ 40 000 km ($\pi \times 12\,700$).

Donnez une estimation de la distance totale parcourue par la station Mir pendant les 86 500 révolutions qu'elle a accomplies lorsqu'elle était sur orbite. Arrondissez votre réponse à la dizaine de millions la plus proche.

Figure 2. ***Exemple de question issue de l'épreuve PISA 2003 (OCDE, 2004)***

Encore faut-il que les élèves comprennent que le calcul de la distance effectuée par la station Mir implique, dès lors, de s'appuyer sur ces ressources ! Pour ce faire, il faut qu'ils parviennent à se représenter le problème sous la forme d'un disque de diamètre équivalent à 13 500 km (12 700+400+400=13 500) ou de rayon équivalent à 6 750 km (voir figure 3).

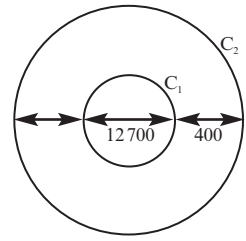
Première étape Transformation du problème en un problème mathématique	Un schéma annoté permet de mieux comprendre les relations évoquées dans le problème :  Quelle est la longueur correspondant à 86 500 fois la circonférence du cercle C_2 ?

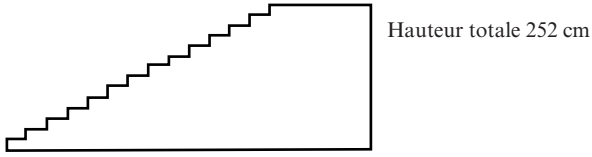
Figure 3. ***Représentation du problème que les élèves doivent construire pour résoudre le problème « vol spatial » (Demonty & Fagnant, 2004)***

Autrement dit, même si les ressources à mettre en œuvre sont plus ou moins explicitement évoquées dans l'énoncé, encore faut-il que les élèves se représentent correctement le problème pour le « transformer » en un problème mathématique pouvant être résolu au travers d'une mobilisation et d'une intégration adéquates des ressources impliquées.

A contrario, dans le problème « escalier » (figure 4), rien ne fait mention dans l'énoncé qu'il s'agit de faire appel à une seule ressource spécifique, en l'occurrence une simple division. Convient-il de dire que ce deuxième exemple est plus propice que le premier exemple à l'évaluation de compétences ?

ESCALIER

Le schéma ci-dessous représente un escalier de 14 marches, qui a une hauteur totale de 252 cm.



Profondeur totale 400 cm

Question 1 :

Quelle est la hauteur de chacune des 14 contremarches ?

Hauteur : = …… cm.

Figure 4. ***Exemple de question issue de l'épreuve PISA 2003 (OCDE, 2004)***

Dans d'autres problèmes, il est fait mention de plates-bandes et l'élève doit en déduire qu'il s'agit de faire appel à des calculs de périmètres. Les figures proposées empêchent toutefois le recours aux formules classiques et nécessitent un raisonnement plus complexe. Mesure-t-on dans ce genre de tâches la compétence ou simplement l'application d'une procédure apprise ? Est-il réellement possible d'opérationnaliser la typologie proposée par De Ketele (2010, 2011) pour analyser chaque question de l'épreuve dans le but de voir s'il s'agit d'applications habillées ou de tâches impliquant de réelles compétences ? Est-ce même pertinent de le faire dans la mesure où finalement, tout va dépendre en partie des élèves auxquels la question s'adresse et du contexte dans lequel elle est présentée ? Un problème aussi complexe et peu transparent soit-il quant aux ressources à

mobiliser ne sera qu'une simple tâche d'application pour des élèves qui viennent d'apprendre une procédure donnée et qui savent donc implicitement que le problème est un prétexte pour l'appliquer ; ce sera une tâche familière pour des élèves qui en ont déjà rencontré de semblables ou qui ont un niveau d'abstraction suffisant pour faire le parallélisme avec d'autres tâches connues ; cela pourra constituer une tâche totalement inédite pour d'autres.

Dans le prolongement de la position défendue par De Ketele et Gérard (2005), Carette et Dupriez (2009) pointent les « discordances » entre les différentes épreuves externes d'évaluation des acquis scolaires organisées en Belgique francophone. Ils avancent le même type d'arguments que De Ketele et Gérard (2005) pour conclure que seules les épreuves produites par les commissions d'outils d'évaluation évaluent de réelles compétences en proposant « une ou plusieurs tâches complexes et inédites qui demandent aux élèves de mobiliser leurs ressources » (p. 38) ; les autres épreuves (PISA, les évaluations externes certificatives et non certificatives) conduisant selon eux à proposer des épreuves où « l'évaluation des compétences ne semble pas fondamentalement se démarquer de la pédagogie par objectifs » (p. 38).

Mais pourquoi certains items de l'épreuve PISA ne pourraient-ils pas être considérés comme évaluant des compétences ? En effet, dans plusieurs questions de l'épreuve PISA, les élèves sont placés face à des situations (des problèmes à résoudre) qui nécessitent la mobilisation d'une ou de plusieurs ressources ou procédures élémentaires et qui dépassent le niveau de la simple application. Par exemple, dans la question « Vol spatial » (figure 2), les élèves doivent d'abord calculer le diamètre (ou le rayon) du cercle parcouru par la station Mir. Ils doivent ensuite en calculer la circonférence, puis calculer le trajet global parcouru pour enfin arrondir leur solution à la dizaine de million la plus proche.

Carette et Dupriez (2009) critiquent aussi le format des réponses de ce type d'épreuves, correspondant dans la majorité des cas à des questions fermées à choix multiples ou à réponses brèves (on rappellera toutefois que certaines questions PISA demandent une argumentation aux élèves). En mathématiques tout au moins, le format de la réponse ne traduit pas la complexité des processus cognitifs mis en œuvre (une tâche complexe, impliquant la mobilisation et l'intégration de plusieurs procédures, peut

conduire à une réponse numérique précise) et, à croire certaines auteurs, questions à choix multiples et tâches complexes ne devraient pas être considérées comme antinomiques (Loye, 2005 ; Loye et al., 2011).

Dans le cadre d'un article comparant différentes épreuves internationales, Leduc, Riopel, Raîche, et Blais (2011) s'appuient sur une analyse du cadre d'évaluation et des questions proposées dans PISA et estiment que « c'est la maîtrise fonctionnelle et l'acquisition de compétences d'une grande portée, combinant les connaissances, les habiletés ainsi que les attitudes […] qui dominent » (p. 108). Plus loin, ils ajoutent : « PISA comporte des items qui évaluent la mise en œuvre créative des compétences mathématiques dans des situations présentant toutes sortes de problèmes, des plus quotidiens et simples jusqu'aux plus inhabituels et complexes, et qui n'ont au départ aucune structure mathématique apparente » (*ibid*, pp. 111-112). Au final, au travers de l'analyse des items accessibles, les auteurs concluent que « PISA vise principalement à évaluer les compétences des élèves » (*ibid*, p. 127).

Le présent texte ne pourra pas trancher le débat, mais il paraissait important de pointer les divergences d'opinions des auteurs qui se sont penchés sur l'analyse des épreuves externes internationales, comme PISA qui est souvent pris en exemple. Finalement, il paraît impossible que chacun s'accorde quant à la question de savoir si telle ou telle question précise permet ou non d'évaluer des compétences (voire si telle ou telle question rencontre ou non le critère de « mobilisation » ou celui d'« intégration »). Il semble que c'est essentiellement une course à la complexité (relativement mal définie par ailleurs, voir Fagnant & Dierendonck, 2012) qui conduit à ce rejet systématique des épreuves « classiques ». Si l'on tente de sortir de ce débat, en partie stérile, on pourrait convenir que l'essentiel n'est pas de savoir dans quelle mesure chaque tâche isolée de PISA évalue ou non telle ou telle compétence, mais plutôt de dresser un bilan permettant de caractériser le niveau de compétence atteint par les élèves. Autrement dit, en s'éloignant du débat consistant à savoir quel type de tâches est suffisamment complexe ou non pour évaluer réellement des compétences, PISA envisage la problématique sous un autre angle en définissant ce qui constitue la compétence mathématique (ou la culture mathématique), en analysant le chemin à parcourir pour s'en approcher (les différents niveaux de compétences « à franchir ») et en pointant à quel endroit du chemin (à quel niveau de compétence) se situent les élèves au moment de l'évaluation.

Cette façon de considérer les choses est certes en rupture avec la l'idée selon laquelle une compétence précise serait évaluée au travers de sa mise en œuvre dans une tâche bien déterminée (ou dans une série de tâches de la même famille), mais elle s'accorde davantage avec l'idée de concevoir la compétence mathématique comme une « disposition mathématique » (De Corte & Verschaffel, 2005; Schonfeld, 1992) composée de cinq catégories d'habiletés qui rejoignent assez largement les ressources à mobiliser pour démontrer sa compétence :

1) une base de connaissances spécifiques au domaine,
2) des stratégies de recherche qui aident à appréhender plus efficacement la tâche à résoudre,
3) des connaissances et des stratégies métacognitives portant sur son fonctionnement cognitif,
4) des stratégies visant à réguler sa motivation et son engagement face à la tâche et, enfin,
5) des croyances relatives aux mathématiques, à leur apprentissage et à la résolution de problèmes (voir Fagnant & Dierendonck, 2012, et Fagnant, Demonty, Dierendonck, Dupont, & Marcoux, 2014, pour une présentation plus détaillée de cette perspective).

Une complémentarité des différentes approches est-elle possible ?

La conciliation entre les outils alternatifs d'évaluation des compétences et les épreuves d'évaluation à large échelle dites classiques pourrait trouver un terrain d'entente moyennant deux conditions essentielles : s'accorder sur une définition moins ambitieuse de la tâche complexe et s'assurer que les épreuves comportent un nombre suffisant d'items pour chaque dimension évaluée.

Dans les modèles d'évaluation construits en référence à l'approche par compétences, les tâches ou les situations complexes sont définies à partir d'un niveau de complexité ultime : seraient « complexes » les tâches nécessitant la mobilisation et l'intégration d'un grand nombre de ressources internes ou externes, les tâches pluridisciplinaires étant considérées comme les mieux à même de remplir ces critères. Théoriquement, on peut toujours imaginer soumettre aux élèves un nombre suffisant de tâches complexes pour s'assurer d'une mesure fiable des compétences visées. Le pro-

blème se pose davantage en termes pratiques puisqu'il n'est pas envisageable d'organiser des séances d'évaluation de plusieurs heures qui seraient nécessaires à la résolution d'un nombre significatif de tâches complexes.

Comme déjà suggéré ailleurs (Dierendonck & Fagnant, 2010a, 2012), il est dès lors proposé d'adopter une définition « minimaliste » des tâches complexes : 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 ressource (ou procédure) et qu'elle nécessite dès lors une interprétation (ou un cadrage) de la situation et une organisation de la démarche de résolution[11]. Soulignons que cette définition n'entre pas en contradiction avec la définition des compétences de troisième degré de Rey et al. (2003) : savoir choisir et combiner correctement plusieurs procédures de base pour traiter une situation nouvelle et complexe[12]. Par ailleurs, elle autorise l'élaboration d'épreuves constituées d'un plus grand nombre d'items et permet dès lors de s'approcher des épreuves standardisées habituellement construites sur le plan national ou international. La possibilité de décomposer ces tâches complexes en tâches élémentaires permet aussi d'introduire un phasage comme dans le modèle de Rey et al. (2003) et d'évaluer, dans des contextes différents ou hors contexte, les ressources impliquées dans celles-ci, comme dans l'étude de Crahay et Detheux (2005). La présence d'un nombre suffisant d'items (et leur indépendance) offre alors la possibilité d'envisager l'utilisation de modèles statistiques tels que les MRI utilisés dans PISA et dans les épreuves externes standardisées au Luxembourg notamment. C'est une telle tentative de conciliation qui a été développée dans les travaux exploratoires présentés dans les prochaines lignes.

Travaux exploratoires menés au Luxembourg

Les travaux exploratoires effectués dans l'enseignement secondaire luxembourgeois se composent de deux études menées respectivement en 2009 et 2010.

La première (étude A) s'est déroulée en deux temps : la première étape visait essentiellement à prétester le potentiel informatif des tâches complexes (telles qu'elles ont été définies) ainsi que de leur décomposition ; la deuxième étape visait à prétester la possibilité d'un modèle « mixte » combinant évaluation à large échelle « classique » et évaluation de compétences.

Lors de la première étape, menée en juin 2009, quelques tâches complexes ont été expérimentées auprès de 176 élèves de troisième secondaire générale. Dans une première phase, deux tâches complexes (au sens où plusieurs ressources doivent être mobilisées et intégrées dans une démarche caractérisée par plusieurs étapes de résolution) ont été soumises aux élèves. Lors d'une deuxième phase, ces deux tâches ont été proposées sous format décomposé. En phase 1 (troisième degré de compétence), les tâches complexes ont été réussies par respectivement 20 % (problème « Menuisier ») et 39 % (problème « Recette ») des élèves[13]. Le fait de présenter une seconde fois ces tâches complexes mais en les découpant en plusieurs tâches plus ciblées et imposant par la même occasion une démarche de résolution (décomposition de la tâche complexe, deuxième degré de compétence) engendre un gain, exprimé en pourcentage d'élèves, de 28 % pour le problème « Menuisier » et de 9 % pour le problème « Recette ». L'ampleur du gain observé entre la phase 1 et la phase 2 semble varier selon la tâche considérée et semble lié au degré de difficulté de la tâche proposée : plus la tâche initiale est difficile, plus le nombre d'élèves qui réussissent lors de la seconde phase est important et inversement. À l'issue de la phase 2, les pourcentages de réussite globaux semblent s'uniformiser pour les deux tâches (48 et 49 % de réussite), ce qui laisse penser qu'environ 50 % des élèves testés dans la présente étude se trouvent en grande difficulté face à ce genre de tâches complexes en mathématiques, même lorsque celles-ci sont proposées sous un format découpé. La phase 3, dont les résultats ne sont pas détaillés ici, permet de mieux cibler les procédures non maîtrisées par les élèves. Autrement dit, même avec une définition « minimaliste » des tâches complexes, les tâches proposées semblent engendrer des difficultés importantes et pouvoir faire l'objet d'une évaluation diagnostique visant à mieux cerner les difficultés éprouvées par les élèves. Cette première étude exploratoire permet de montrer qu'il est possible de mettre en œuvre une épreuve en phases, comme dans le dispositif à visée diagnostique de Rey et al. (2003), même avec des tâches « moins complexes ».

L'étape suivante consiste à voir dans quelle mesure il est possible d'insérer un tel dispositif au sein d'épreuves externes classiques. Ce type de tâches complexes, leur décomposition et des tâches élémentaires évaluant les procédures impliquées dans les tâches complexes initiales ont été intégrées au sein d'une épreuve d'évaluation à large échelle « classique », à savoir l'épreuve externe standardisée proposée dans toutes les classes de

grade 9 (troisième année de l'enseignement secondaire) au Luxembourg. Cette épreuve, proposée en version informatique, permet un encodage direct des réponses numériques fournies par les élèves.

En vue d'évaluer la faisabilité du dispositif, les résultats des élèves de l'enseignement général (N = 1 769) ont été analysés à l'aide du modèle de Rasch. Les résultats, qui ne seront pas développés ici (voir Dierendonck & Fagnant, 2012), ont permis d'attester de la faisabilité d'un tel dispositif, ainsi que du pouvoir informatif complémentaire apporté par le découpage des tâches complexes et par les tâches élémentaires évaluant les mêmes ressources isolément.

Dans la deuxième étude (étude B), un design expérimental plus complet que celui testé en 2009 a été développé. En s'appuyant sur la possibilité d'ancrage offerte par le modèle de Rasch, un dispositif d'évaluation (tableau 1) combinant un nombre important d'items, quelques tâches complexes et des tâches complémentaires à visée diagnostique inspirées du phasage proposé par Rey et al. (2003), combinées à l'évaluation des procédures dans des problèmes simples (tâches élémentaires en contexte) comme l'ont proposé Crahay et Detheux (2005), a été mis au point.

Tableau 1

Démarche expérimentale d'évaluation inspirée du « phasage » de Rey et al. (2003) et de l'approche de Crahay et Detheux (2005) testée pour quelques tâches complexes dans l'épreuve standardisée de mathématiques au grade 9

Modèle d'évaluation des compétences de Rey et al. (2003)			Proposition pour les épreuves standardisées au Luxembourg	
	Phases d'évaluation	Degrés de compétence	Phases d'évaluation	
Phase 1	« *On propose aux élèves une* ***tâche complexe exigeant le choix et la combinaison d'un nombre significatif de procédures*** *que les élèves doivent posséder. Il est préférable que cette tâche soit pluridisciplinaire et fonctionnelle*» (p. 92).	**Compétence de 3e degré** ou compétence complexe	*On soumet aux élèves un* ***problème complexe*** *(au sens de nécessitant la mobilisation et l'intégration d'au moins deux ressources).*	Phase 1
Phase 2	« *On propose la* ***même tâche complexe mais découpée en tâches élémentaires,*** *présentées dans l'ordre où elles doivent être accomplies pour réaliser la tâche globale. L'élève doit choisir, parmi les procédures qu'il connaît, celle qui convient à chacune des tâches élémentaires*» (p. 92).	**Compétence de 2e degré** ou compétence élémentaire avec cadrage	*On soumet à nouveau le* ***même problème complexe*** *mais, cette fois, celui-ci est* ***décomposé en sous-tâches*** *à réaliser successivement.*	Phase 2
			On évalue, à l'aide d' ***autres problèmes simples*** *(autres contextes), la maîtrise isolée des* ***ressources qui interviennent dans le problème complexe initial.***	Phase 3
Phase 3	« *On demande aux élèves d'accomplir les* ***procédures exigées dans les phases précédentes, mais sous une forme décontextualisée*** […]» (p. 92).	**Compétence de 1er degré** ou compétence élémentaire ou procédure	*On évalue,* ***hors contexte,*** *la maîtrise isolée des* ***ressources qui interviennent dans le problème complexe initial.***	

L'épreuve standardisée de mathématiques, soumise en octobre 2010 à l'ensemble des élèves de grade 9 (n = 6 399) au Luxembourg, permet de confirmer et de nuancer les constats dressés à l'issue de l'étude A. L'épreuve d'évaluation proprement dite a été construite en référence aux socles de compétences luxembourgeois avec l'objectif d'évaluer quatre dimensions au départ du pool d'items suivant (tableau 2).

Tableau 2
Caractéristiques de l'épreuve standardisée de mathématique

	Résolution de tâches complexes	**Résolution de tâches élémentaires contextualisées ou décontextualisées**	**Total**
Compétence en nombres et opérations	5 items	19 items	24 items
Compétence en figures du plan et de l'espace	3 items	20 items	23 items
Total	8 items	39 items	47 items

Ces 47 items n'ont pas été soumis à tous les élèves étant donné le temps imparti à l'épreuve de mathématique (50 minutes) et l'existence de filières d'enseignement (ES, EST, PR)[14] fortement hiérarchisées quant aux acquis scolaires moyens. Il a donc fallu organiser une sélection et un ancrage d'items au travers des trois versions de l'épreuve. Dans l'esprit de la première étude exploratoire (deuxième étape de l'étude A), trois des huit tâches complexes ont été décomposées en trois ensembles de sous-tâches appelés tâches complexes décomposées. La tâche complexe « Menuisier » et sa version décomposée sont reprises à titre d'illustration en annexes 1 et 2. Parmi les tâches complexes proposées, les tâches « Menuisier » et « Judo » constituent des versions parallèles au sens où elles sont présentées dans des contextes différents mais elles impliquent toutes les deux la mobilisation et l'intégration de deux procédures mathématiques :

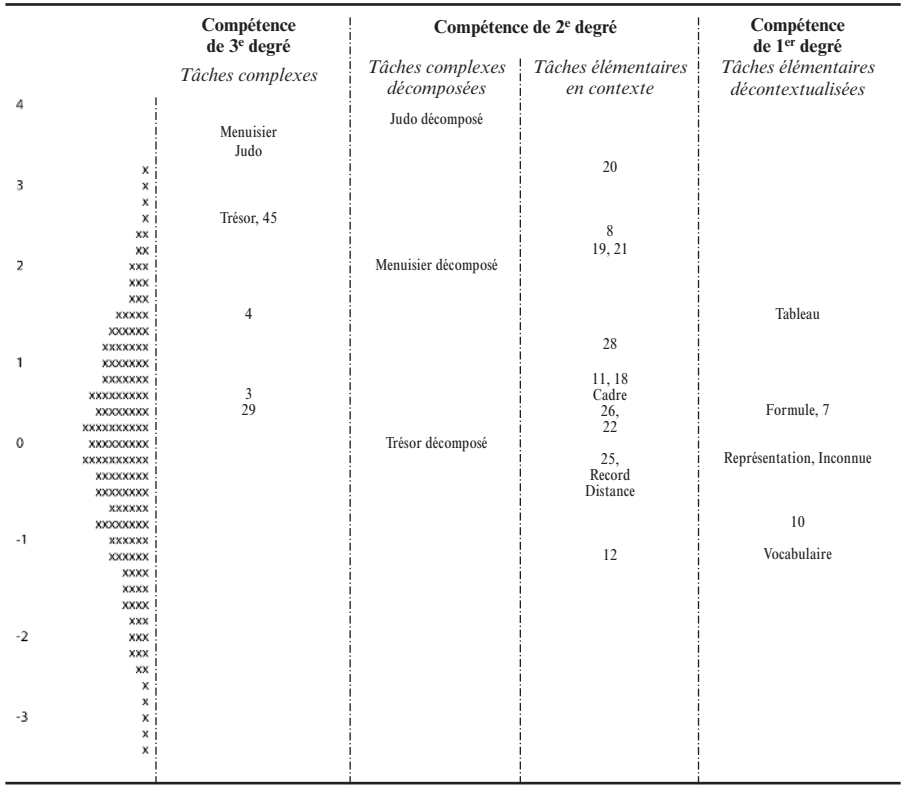
1) calculer la longueur d'un côté d'un carré au départ de son aire (Judo) ou de son périmètre (Menuisier),
2) déduire la longueur d'un segment au départ de la longueur de deux autres segments (Judo et Menuisier).

Plusieurs tâches ont été proposées pour évaluer autrement les ressources mobilisées dans ces tâches complexes. Un exemple de tâche élémentaire en contexte et un exemple de tâche élémentaire décontextualisée sont proposés en annexes 3 et 4.

Il n'a pas été possible, comme cela avait été le cas lors de la première étape de l'étude A (prétest dans quelques classes), de soumettre aux mêmes élèves à la fois les tâches complexes et les tâches complexes décomposées puisqu'ils auraient pu se servir des tâches décomposées pour modifier ensuite leurs réponses aux tâches complexes initiales (le *testing* informatique permettant de « voyager » dans l'épreuve). Durant la seconde étape de l'étude A (épreuve externe de 2009), cette difficulté a été contournée en optant pour des tâches parallèles (« Judo » pour la tâche complexe et « Menuisier » pour la tâche décomposée). Dans l'étude B (épreuve externe de 2010), le même principe a été employé de façon à ce que les élèves reçoivent une tâche complexe et la version parallèle de la tâche décomposée. Par ailleurs, la procédure d'ancrage offerte par le modèle de Rasch permet de comparer l'ensemble des tâches sur une échelle commune, et ceci même si elles ont été soumises à des élèves différents. Si, sur le plan du diagnostic individuel, il aurait été plus intéressant de proposer la tâche complexe et sa version décomposée aux mêmes élèves (ce qu'il serait possible de mettre en œuvre en améliorant les modalités du *testing* informatique, le texte y reviendra par la suite), l'apport du modèle de Rasch offre une autre information intéressante, sur un plan plus général, en permettant de situer sur une même échelle l'ensemble des tâches de l'épreuve et le niveau de compétence des élèves par rapport à l'ensemble de ces tâches[15].

Pour mettre en œuvre l'analyse de Rasch, les tâches complexes ont été codées 1 (réponse finale correcte) ou 0 (réponse finale incorrecte) tant dans leur format initial (tâche complexe non décomposée) que dans leur format décomposé (tâche complexe décomposée en sous-tâches conduisant pas à pas vers la réponse finale). Cette manière de procéder occasionne une perte d'information diagnostique importante, mais elle était nécessaire à ce niveau pour permettre d'appliquer la modélisation statistique au départ d'items indépendants. Comme il sera développé par la suite, cette perte d'information n'est que temporaire puisqu'il reste possible de retourner ultérieurement aux réponses données par les élèves à chacune des étapes de la tâche décomposée, ainsi que de retrouver leur réponse numérique chiffrée pour chacune des tâches ou sous-tâches de l'épreuve.

La figure 5 présente les résultats de la modélisation statistique obtenue avec le logiciel d'analyse Conquest, mais illustrée sous une forme plus lisible qui présente les items en fonction de trois degrés de compétence inspirés de Rey et al. (2003) et des quatre types de tâches définis précédemment dans le tableau 2.



Chaque « x » représente 38.7 individus.

Figure 5. ***Résultats de l'analyse MRI présentés en fonction des trois degrés de compétence de Rey et al. (2003) et des quatre types de tâches***

La figure 5[16] permet de dresser quatre constats principaux :

1) Une certaine hiérarchie est observée entre les différents types de tâches du dispositif d'évaluation : en général, les tâches les plus difficiles de l'épreuve sont des tâches complexes (à l'exception de « Judo décomposé ») et les tâches les plus faciles sont des tâches élémentaires présentées avec contexte ou hors contexte.

2) Il n'y a pas de hiérarchie stricte entre les quatre types de tâches. Tout d'abord, si les tâches les plus difficiles de l'épreuve sont essentiellement des tâches complexes, il apparaît que certaines tâches décomposées et certaines tâches élémentaires en contexte présentent des degrés de difficulté très élevés également. Par ailleurs, certaines tâches complexes présentent un degré de difficulté modéré. Enfin, si certaines tâches élémentaires en contexte présentent un degré de difficulté nettement plus élevé que les tâches élémentaires hors contexte, ceci ne se vérifie pas non plus face à tous les items.
3) L'effet de la décomposition est très variable d'une tâche complexe à l'autre : le découpage semble simplifier fortement le problème « Trésor », modérément le problème « Menuisier » et nullement le problème « Judo ».
4) De nombreuses tâches élémentaires en contexte présentent un degré de difficulté moindre que celui des tâches décomposées. En particulier, les tâches élémentaires en contexte qui évaluent isolément les procédures impliquées dans les tâches complexes « Judo » et « Menuisier » présentent un niveau de difficulté sensiblement inférieur à celui des tâches décomposées correspondantes. Autrement dit, l'information fournie par l'évaluation des compétences de deuxième degré dans des situations indépendantes les unes des autres ne conduit pas aux mêmes constats que l'évaluation de ces « mêmes » compétences de deuxième degré dans des tâches décomposées : le niveau de maîtrise de ce degré de compétence[17] aurait sans doute été sous-estimé, pour un nombre non négligeable d'élèves, si les seules tâches évaluant des compétences de deuxième degré étaient des tâches décomposées.

Il est à noter que les résultats de l'épreuve de 2009, analysés uniquement pour l'enseignement général et face à un *design* expérimental moins complet (Dierendonck & Fagnant, 2012) présentaient une distribution comparable et permettaient de dresser des constats similaires (excepté la nuance apportée ici par l'effet différentiel des trois tâches décomposées), ce qui semble pouvoir en renforcer la robustesse.

Il importe de rappeler, au risque de se répéter, que le modèle de Rasch situe sur une même échelle le degré de difficulté des items et le niveau de performance des élèves (ces derniers sont représentés par des croix dans la figure 5). Ainsi, les élèves en vis-à-vis de l'item « Trésor » ont une probabilité de 50 % de réussir cette tâche complexe, une probabilité inférieure

de réussir les tâches complexes « Menuisier », « Judo » et « Judo décomposé », ainsi que la tâche élémentaire numéro 20, et une probabilité supérieure de réussir les items situés plus bas sur l'échelle. *A contrario*, les élèves qui se situent en bas de l'échelle ont une faible probabilité de réussir la majorité des items. En s'appuyant sur le même type d'analyses que celles réalisées par l'OCDE pour l'épreuve PISA, il serait possible de déterminer des niveaux de compétence et de situer les élèves dans un niveau représentatif de leurs performances (par exemple, le niveau pour lequel ils ont une probabilité minimale de 50% de réussir l'ensemble des tâches de ce niveau). Si l'on parvient à qualifier chaque niveau de l'échelle par un descriptif des processus cognitifs impliqués dans les tâches qui le constituent, il devient alors possible de décrire le niveau de compétence atteint par chaque élève, tout en donnant une première indication, forcément relativement grossière, sur le chemin restant à parcourir (c'est-à-dire quels processus cognitifs demeurent non maîtrisés à ce stade) pour atteindre le niveau supérieur.

En proposant des épreuves d'évaluation constituées d'un nombre suffisant d'items (et d'items suffisamment variés) et en cherchant à définir des niveaux de compétence hiérarchisés au sein desquels pourraient se situer les élèves, il semble possible d'établir un premier diagnostic fiable quant à la mesure et permettant de renseigner le type de tâches qu'un élève donné maîtrise (les tâches situées dans les niveaux de compétence inférieurs), celles qu'il maîtrise partiellement (les tâches correspondant au niveau de compétence dans lequel il a été placé) et celles qui lui posent problème (les tâches situées dans les niveaux de compétence supérieurs). En dressant de tels constats, cette fois au niveau de la classe, un enseignant pourrait cerner les types de tâches qu'il doit impérativement travailler en classe avec une majorité d'élèves ou au sein de groupes à besoins spécifiques.

Comme mentionné précédemment, avec les niveaux de compétence qui pourraient être tirés d'une analyse MRI, la logique passe d'une « évaluation diagnostique d'une compétence spécifique face à une tâche donnée » à la « détermination d'un niveau de compétence en mathématiques ». Mais, dès lors, toute ambition diagnostique face aux tâches complexes proprement dite a-t-elle été perdue ? Sans doute que non, dans la mesure où les épreuves externes devraient aussi offrir la possibilité de retourner aux résultats bruts des élèves et, dès lors, fournir des renseignements plus

précis en s'appuyant sur une analyse des tâches résolues par les élèves. À ce niveau, il est alors possible de dépasser la dichotomie « réponse correcte ou incorrecte » pour analyser non seulement les réponses numériques fournies par les élèves face aux différents types de tâches, complexes ou non (ce qui a déjà en soit un pouvoir informatif non négligeable quant aux démarches développées par les élèves, Dierendonck & Fagnant, 2012), mais également les réponses données aux différentes étapes de décomposition des tâches complexes. Il serait dès lors possible de confronter les réponses fournies par les élèves aux tâches évaluant les mêmes ressources dans des contextes différents ou hors contexte pour mieux cerner où se situent précisément les difficultés éprouvées.

Discussion et conclusion

Les travaux exploratoires relatés dans cet article avaient pour objectif de dégager une voie de conciliation possible entre, d'une part, les principes que l'approche par compétences semble poser en termes d'évaluation et, d'autre part, les pratiques d'évaluation à large échelle des acquis scolaires des élèves. Cette conciliation repose sur une définition plus modeste du concept de compétence qui rompt avec l'idée que les tâches d'évaluation doivent nécessairement être d'une grande complexité, donner lieu à des productions ouvertes (difficilement évaluables de façon standardisée) ou présenter un caractère interdisciplinaire pour pouvoir évaluer correctement les apprentissages réalisés dans le cadre de l'approche par compétences.

En posant que pour être complexes (et ainsi rendre compte de la notion de compétence et permettre une décomposition à visée diagnostique, comme dans le modèle de Rey et al., 2003), les tâches d'évaluation devraient nécessiter l'identification, la mobilisation et l'intégration d'au minimum deux ressources ou procédures apprises, il devient possible de proposer aux élèves un nombre suffisant de tâches d'évaluation qui permet de respecter à la fois les critères de qualité de la mesure et certains principes liés aux dispositifs d'évaluation diagnostique des compétences.

L'article a également montré la faisabilité d'intégrer, au sein d'un dispositif d'évaluation externe classique, des tâches à visée diagnostique s'inspirant du dispositif développé par Rey et al. (2003) et de la démarche proposée par Crahay et Detheux (2005). L'intérêt de combiner ces approches

réside, tel que mentionné, dans la complémentarité de l'information apportée par les quatre types de tâches, comme l'ont montré les premiers résultats présentés ici. En s'appuyant sur l'analyse de Rasch, en définissant des niveaux de compétences et en situant les élèves dans ces niveaux, il est possible de préciser des ensemble de tâches que les élèves maîtrisent plus ou moins, ce qui fournit déjà un premier diagnostic général. En s'appuyant sur les réponses fournies par les élèves aux différentes phases et aux différents types de tâches, il est possible d'établir précisément ce qui pose problème aux élèves, en termes de maîtrise des ressources isolées, de leur mobilisation dans des tâches élémentaires ou encore de leur intégration dans les tâches complexes, ce qui permet un diagnostic plus fin. Évidemment, il y a lieu de regretter de ne pas avoir accès aux « traces » laissées par les élèves sur leurs feuilles de brouillon dans la mesure où ces éléments pourraient en effet avoir un pouvoir informatif important pour saisir les démarches des élèves et ainsi mieux comprendre leurs difficultés. Il s'agit cependant là d'une limite inhérente à la plupart des évaluations externes nationales ou internationales.

En ce qui concerne la décomposition de la tâche complexe, au-delà de l'information diagnostique apportée par les réponses données aux sous-tâches, il faut remarquer que le « pouvoir aidant » de la décomposition est variable selon la tâche complexe considérée. La tâche complexe « Trésor », par exemple, est très différente, quant aux contenus mathématiques impliqués, des deux autres problèmes complexes discutés jusqu'ici puisqu'elle impliquait notamment la maîtrise des règles de divisibilité. Il serait donc pensable que l'effet différentiel de la décomposition sur le pourcentage de réussite est à rechercher dans une analyse fine du contenu des tâches proposées. Pour rappel, un effet différentiel était également observé face aux deux tâches complexes prétestées lors de la première phase de l'étude A et il serait donc important de creuser davantage cette question pour mieux cerner l'effet potentiel de cette étape de décomposition[18]. Une hypothèse alternative serait que le calcul des paramètres de difficulté des items dans l'analyse MRI soit influencé par la sous-population d'élèves à qui les items ont été soumis, et ceci malgré la procédure d'ancrage utilisée qui devrait théoriquement permettre de comparer sur une même échelle des problèmes résolus par des élèves différents. Dans le cas présent, la tâche « Trésor décomposé » (pour laquelle l'effet de la décomposition semble le plus important) a été soumise aux élèves de l'enseignement général ; la tâche

« Menuisier décomposé » (pour laquelle l'effet de la décomposition semble nettement moindre) a été proposée aux élèves de l'enseignement technique et la tâche « Judo décomposé » (pour laquelle l'effet de la décomposition semble contre-productif) n'a été soumis qu'aux élèves de la filière académiquement la plus faible. En comparaison, dans PISA, la procédure d'ancrage est également basée sur des carnets de tests différents distribués aux élèves, mais ces différents carnets sont distribués de manière aléatoire, ce qui n'engendre pas la création de sous-populations de performances moyennes très différentes (il n'y a aucune raison que le groupe d'élèves qui reçoit le carnet A soit plus faible que celui qui reçoit le carnet B). En ce sens, des analyses complémentaires devraient être menées pour vérifier l'adéquation de l'analyse de Rasch au *design* expérimental testé dans le cadre des épreuves externes organisées au Luxembourg. Enfin, si on s'intéresse spécifiquement au problème « Menuisier », on notera également que le gain apporté par la décomposition est nettement plus faible que celui observée dans l'étude pilote menée dans quelques classes (phase 1 de l'étude A). Dans 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, la tâche leur serait proposée à nouveau 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.

Dans l'optique d'un *testing* adaptatif informatisé, il y aurait lieu de ne présenter à l'élève que les items qui sont nécessaires pour déterminer son niveau de compétence et préciser où se situent ses lacunes[19]. Concrètement, dans un premier temps, l'ordinateur soumettrait à l'élève une première tâche complexe (phase 1). En cas de réussite à cette tâche, l'ordinateur proposerait une deuxième tâche complexe. En cas d'échec, l'ordinateur présenterait à nouveau la première tâche complexe (phase 2) mais, cette

fois, accompagnée d'une décomposition précisant les étapes à suivre pour obtenir la réponse correcte (comme dans le modèle de Rey et al., 2003) ou de quelques indices visant à faciliter ou à aiguiller la construction de la représentation du problème ou sa résolution. Si, au terme de la phase 2, l'élève trouve la bonne réponse, l'ordinateur présente la deuxième tâche complexe. Si l'élève échoue en phase 2, l'ordinateur soumet les items évaluant la maîtrise isolée des ressources/procédures qui sont nécessaires à la résolution de la tâche complexe de départ (phase 3) et ainsi de suite jusqu'à couverture raisonnable (du point de vue de la mesure) du domaine évalué.

Rappelons, avant de conclure, que la présente étude n'est qu'une étude pilote et qu'au-delà de certains constats potentiellement intéressants, elle soulève de nombreux questionnements, souvent fondamentaux. En plus de ceux déjà mentionnés ci-dessus, se pose également la question de la coexistence d'une évaluation en phases (qui, par définition, est composée d'items dépendants) avec des modèles d'analyse statistique (comme la théorie de réponse à l'item) qui sont fondés sur des postulats d'indépendance des items et d'unidimensionnalité. En effet, la théorie de réponse à l'item postule que les items doivent être strictement indépendants les uns des autres. Or, le simple fait de proposer un phasage ou un *testing* adaptatif sur le plan de l'évaluation conduit à proposer des items liés, de près ou de loin, à une même tâche complexe initiale et, dès lors, qui peuvent être considérés comme des items dépendants. Face à ce problème, recourir à un modèle à crédit partiel pourrait peut-être s'avérer une piste intéressante puisque chaque tâche complexe et son éventuel découpage pourraient être recodés en un seul item à crédit partiel.

Bien que des analyses complémentaires soient nécessaires pour mieux cerner le réel pouvoir diagnostique du type d'épreuve présentée dans cet article, les premiers résultats semblent témoigner de l'intérêt potentiel du modèle proposé (combinant les apports de Rey et al., 2003 et de Crahay & Detheux, 2005) et permettent certains espoirs de possibles complémentarités entre ces deux approches. *A contrario*, il semble important de pointer, une fois encore, le « danger », non seulement en termes de fiabilité de la mesure mais aussi sur le plan pédagogique, d'évaluer la compétence mathématique des élèves (ou leur « disposition mathématique », pour reprendre la terminologie précédemment mentionnée) uniquement au départ de quelques tâches complexes de troisième degré.

Au final, toutes les interrogations soulevées dans cet article et les limites importantes posées par le cadre spécifique des études exploratoires ne doivent pas conduire à l'abandon d'une tentative de conciliation entre différents modèles d'évaluation, au prétexte que celle-ci poserait d'insurmontables défis sur le plan de la fiabilité de la mesure ou ne permettrait pas d'atteindre un idéal sur le plan diagnostique. Au contraire même, il semble primordial et urgent que les efforts de recherche se concentrent sur ces questions d'évaluation, qui constituent sans doute l'un des talons d'Achille de l'approche par les compétences.

NOTES

1. Avec Scallon (2004), on soulignera que cette volonté de disposer de nouvelles pratiques d'évaluation n'est pas propre à la francophonie puisqu'elle est apparue également aux États-Unis au milieu des années 1990 sous l'influence du courant *Competency-Based Éducation* caractérisé, notamment, par le rejet des tests standardisés classiques et par le développement de la notion de *performance assessment* (appréciation de la performance) qui suppose d'évaluer les compétences des élèves dans le cadre de situations complexes et authentiques. Selon Hébert, Valois et Frenette (2008), « malgré des rapprochements certains entre les approches anglo-saxonne (la situation de performance) et francophone (la situation de compétence), la conception francophone actuelle du concept de situation pour apprécier des compétences semble avoir progressé indépendamment du courant anglo-saxon, qui est de toute façon très peu cité dans les écrits francophones ».
2. Par exemple, en Belgique francophone, Carette et Dupriez (2009) dénoncent une contradiction entre, d'une part, les épreuves externes internationales (PISA par exemple) ou nationales (les évaluations externes certificatives – CEB par exemple – et non certificatives) et, d'autre part, les outils d'évaluation construits en s'appuyant (pour l'enseignement primaire et secondaire inférieur) sur le dispositif en phases développé par l'équipe de l'Université de Bruxelles (Rey et al., 2003) et proposés à titre illustratif d'outils diagnostiques d'évaluation de compétences par le Ministère de l'enseignement.
3. À ce propos, Gérard (2008) signale que « tout le monde ne s'accorde pas sur la notion de "familles de situations" ». Certains chercheurs en contestent même la validité (Crahay, 2006 ; Rey et al., Carette, Defrance et Kahn, 2003) » (p. 61).
4. En accord avec les auteurs, on reconnaîtra que le terme de compétence est peu approprié pour qualifier le 1er degré puisqu'une compétence nécessite de pouvoir mobiliser « à bon escient » des ressources préalablement apprises. C'est par ailleurs en nous appuyant sur la distinction entre le deuxième et le troisième degré que nous proposerons par la suite une définition « minimaliste » de la tâche complexe.

5. Pour la phase 1, Carette (2007) fait référence à « un nombre significatif de procédures ». Cette imprécision, sans doute voulue pour laisser une certaine marge de liberté aux concepteurs de tâches complexes, nous semble être un des éléments pouvant conduire à une « surenchère » débouchant sur des tâches très complexes et très longues à résoudre. *A contrario*, nous retiendrons que la tâche complexe doit impliquer, *a minima*, la mise en œuvre intégrée de deux ressources (ou procédures élémentaires) de façon à pouvoir être décomposée en phase 2 en « tâches élémentaires […] présentées dans l'ordre où elles doivent être accomplies pour parvenir à la résolution de la tâche complexe » (p. 62), c'est-à-dire en étapes successives de résolution.
6. C'est en ce sens que, dans la suite du texte, nous évoquerons l'idée d'un potentiel « pouvoir aidant » de l'étape de décomposition.
7. Exemple tiré d'une des épreuves d'évaluation développée par Rey et al. (2003, p. 84).
8. Selon Roegiers (2007), une tâche compliquée pour l'élève est une tâche « qui mobilise des acquis nouveaux, peu ou pas connus par lui, insuffisamment maîtrisés, ou qui lui sont peu familiers » (p. 16). La complexité mettrait en jeu quelque chose d'assez différent dans la mesure où elle ne dépendrait pas tellement du type de savoir, de savoir-faire ou de savoir être à mobiliser, mais surtout de leur quantité : « la difficulté vient non pas de chaque opération à exécuter, mais de l'articulation de ces opérations entre elles » (p. 121).
9. On peut toutefois s'interroger sur l'indépendance totale de certains items, notamment ceux qui sont liés au même stimulus.
10. La distinction entre les différents types de tâches proposée par De Ketele (2011) présente une certaine proximité avec la distinction opérée par Scallon (2004) entre « connaissances », « habiletés – de bas ou de haut niveau » et « compétences ».
11. Nous qualifions la définition de « minimaliste » au sens où elle constitue le minimum requis pour offrir la possibilité de décomposer la tâche complexe en tâches élémentaires évaluant des ressources isolées (cf. distinction entre les phases 1 et 2 du modèle de Rey et al., 2003). Voir note 5 ci-dessus pour une explication plus détaillée.
12. Cette façon de concevoir les tâches complexes s'accorde d'ailleurs assez bien avec certains exemples proposés dans le cadre des outils d'évaluation diagnostiques des compétences proposés en Belgique francophone sur la base du modèle de Rey et al. (2003). Si sur le plan de l'enseignement primaire, la plupart des tâches nécessitent « un nombre significatif de ressources » (Carette, 2007) (plusieurs sont même pluridisciplinaires et nécessitent de traiter des ressources externes, telles que des horaires, des plans, etc.), il n'en va pas de même au niveau de l'enseignement secondaire où les tâches de mathématiques sont davantage ciblées sur quelques ressources spécifiques. À titre illustratif, la tâche « cadeau collectif » (voir ci-après) se présente sous la forme d'un problème aboutissant à une réponse numérique précise et nécessitant deux types de ressources mathématiques (d'une part, un système de deux équations à deux inconnues pour trouver le nombre d'élèves et, d'autre part, une opération arithmétique simple – une multiplication pour trouver le prix du cadeau). L'énoncé est le suivant : *Claude raconte à son copain : « La semaine dernière, les élèves des classes de deuxième année ont décidé de se cotiser pour offrir un cadeau à Kevin hospitalisé. Chacun a donné 0,50 €. Chargé de l'achat du cadeau, j'ai constaté qu'il manquait 15 €. J'ai réclamé 0,20 € supplémentaires par élève et j'ai alors eu 11 € de trop. » Calcule le prix du cadeau. Écris les étapes deton raisonnement*

et tous tes calculs. Si une telle tâche complexe semblait *a priori* pouvoir être intégrée dans l'épreuve exploratoire réalisée, le contenu mathématique impliqué ne l'a pas autorisé puisque l'algèbre n'est pas à certifier en début d'enseignement secondaire au Luxembourg.

13. Le problème « Menuisier » est proposé en annexe 1. Le problème « Recette » est un problème de proportionnalité dans lequel les élèves doivent retrouver une erreur (une proportionnalité non respectée) dans une recette de crêpes (problème inspiré de celui proposé par Julo, 1995).
14. Les sigles ES, EST, PR désignent respectivement l'enseignement secondaire général, l'enseignement secondaire technique (orientations technique et polyvalente) et l'enseignement secondaire technique (orientations pratique et préparatoire).
15. Le modèle de Rasch permet en effet d'inférer pour chaque élève, une probabilité de réussite à chacune des tâches de l'épreuve, qu'il ait ou non résolu une tâche donnée. Autrement dit, on peut estimer qu'un élève situé à un niveau de compétence donné a une probabilité donnée de réussir l'ensemble des tâches de ce niveau, et ceci même s'il n'a pas réellement été confronté à chacune des questions constitutives de ce niveau.
16. Les items ne répondant pas aux contraintes statistiques (r.bis < 0,25 et/ou infit < 0,80 ou > 1,20) ont été retirés de l'analyse (Bond & Fox, 201307 ; Wright, Linacre, Gustafson, & Martin-Löf, 1994). Il s'agit là des standards classiquement fixés dans PISA et dans les épreuves externes au Luxembourg.
17. 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…).
18. L'analyse fine des contenus mathématiques impliqués dans les tâches est en effet essentielle. Dans une étude réalisée dans le domaine de l'algèbre élémentaire, des chercheurs ont ainsi pu constater que certains élèves parvenaient à réussir l'étape 1 de l'épreuve (en développant une démarche arithmétique) alors qu'ils échouaient lors de l'étape 2 (qui leur imposait une démarche algébrique) (voir Demonty, Fagnant, & Dupont, soumis, pour une présentation détaillée de cette étude).
19. La procédure adaptative qui est proposée ici est assez éloignée de ce qui se fait classiquement avec les modèles de réponse à l'item (voir Laveault & Grégoire, 1997, p. 306). Ici, à un moment ou à un autre, l'ordinateur propose à tous les élèves plusieurs tâches complexes sans cadrage. Les autres items sont présentés uniquement en cas d'échec aux tâches de phase 1 ou de phase 2.

RÉFÉRENCES

Bond, T., G., & Fox, C., M. (2013). *Applying the Rasch Model. Fundamental Measurement in the Human Sciences*, 2nd ed. New York, NY: Routledge.

Burton, R. (2004). Influence des distributions du trait latent et de la difficulté des items sur les estimations du modèle de Birnbaum : une étude du type Monte-Carlo. *Mesure et évaluation en éducation*, *27*(3), 41–62.

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

Carette, V. (2009). Et si on évaluait des compétences en classe ? À la recherche du cadrage instruit. In L. Mottier Lopez & M. Crahay (Eds.), *Évaluations en tension. Entre la régulation des apprentissages et le pilotage des systèmes* (pp. 149-163). Bruxelles, Belgique : De Boeck.

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 évaluation en éducation, 32*(3), 23-45.

Crahay, M. (2006). Dangers, incertitudes et incomplétude de la logique de compétence. *Revue française de pédagogie, 154*, 97-110.

Crahay, M., Audigier, F., & Dolz, J. (2006). En quoi les curriculums peuvent-ils être objets d investigation scientifique ? In F. Audigier, M. Crahay, & J. Dolz (Eds.), *Curriculum, enseignement et pilotage* (pp. 7-37). Bruxelles, Belgique : De Boeck.

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), 57-76.

De Corte, E., & Verschaffel. L. (2005). Apprendre et enseigner les mathématiques : un cadre conceptuel pour concevoir des environnements d'enseignement-apprentissage stimulants. In M. Crahay, L. Verschaffel, E. De Corte, & J. Grégoire (Eds.). *Enseignement et apprentissage des mathématiques : que disent les recherches psychopédagogiques ?* (pp. 25-54). Bruxelles : De Boeck Université.

De Ketele, J.-M. (1996). L'évaluation des acquis scolaires : Quoi ? Pourquoi ? Pour quoi ? *Revue tunisienne des sciences de l'éducation, 23*, 17-36.

De Ketele, J-M. (2010). Ne pas se tromper d'évaluation. *Revue française de linguistique appliquée, XV*(1), 25-37.

De Ketele, J.-M. (2011, 16 mars). *Scores multidimensionnels : une nécessité pour évaluer les compétences.* Communication orale lors du deuxième congrès français de psychométrie, Sèvres, France.

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.

Demonty, I., & Fagnant, A. (2004). *PISA 2003. Évaluation de la culture mathématique des jeunes de 15 ans.* Document à l'attention des professeurs de mathématiques des 1er et 2e degrés de l'enseignement secondaire. Ministère de la Communauté française. Service général du pilotage du système éducatif.

Demonty, I., Fagnant, A., & Dupont, V. (soumis à la revue *Mesure et évaluation en éducation*, en révision). *La mesure des compétences des élèves face à la résolution d'une tâche complexe en algèbre élémentaire : l'apport d'indicateurs issus de recherches centrées sur les difficultés spécifiques des élèves en algèbre*.

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'ADMÉÉ-Europe, 2010*(1), 5-20.

Dierendonck, C., & Fagnant, A. (2010b, janvier). *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'ADMÉÉ-Europe, 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,* Genève, Suisse. Abstract retrieved from: http://www.unige.ch/aref2010/index.html

Dierendonck, C. & Fagnant, A. (2012). Évaluer des compétences en mathématiques dans le cadre d'une épreuve externe à large échelle au départ de tâches complexes, de tâches décomposées et de tâches élémentaires : quel pouvoir diagnostique ? *Actes du 24e colloque international de l'ADMÉÉ-Europe*, Luxembourg, Luxembourg. Abstract retrieved from: http://admee2012.uni.lu/index.php/actes

Dierendonck, C, Loarer, E. & Rey, B. (sous presse). *L'évaluation des compétences en milieu scolaire et en milieu professionnel*. Bruxelles, Belgique : De Boeck.

Eurydice (2012). *Developing key competences at school in Europe: Challenges and opportunities for policy*. Eurydice Report. Luxembourg: Publications Office of the European Union.

Fagnant, A., Demonty, I., Dierendonck, C., Dupont, V. & Marcoux, G. (sous presse). Résolution de tâches complexes, évaluation "en phases" et compétence en mathématiques. In C. Dierendonck, E. Loarer, & B. Rey (Eds.). *L'évaluation des compétences en milieu scolaire et en milieu professionnel* (pp. 179-190). Bruxelles, Belgique : De Boeck.

Fagnant, A., & Dierendonck, C. (2012). Table ronde : Comment assurer l'évaluation diagnostique des compétences scolaires ? *Actes du 24e colloque international de l'ADMÉÉ-Europe,* Luxembourg, Luxembourg. Abstract retrieved from: http://admee2012.uni.lu/index.php/actes

Gauthier, R.-F. (2006). *Les contenus de l'enseignement secondaire dans le monde : état des lieux et choix stratégiques*. UNESCO.

Gérard, F.-M. (2008). *Évaluer des compétences - Guide pratique.* Bruxelles, Belgique : De Boeck.

Hébert, M.-H., Valois, P., & Frenette, E. (2008). La validation d'outils alternatifs d'évaluation des apprentissages : progressiste ou rétrograde ? *Actes du 20e colloque de l'ADMÉÉ-Europe*, Genève, Suisse. Abstract retrieved from https://plone. unige.ch/sites/admee08/communications-individuelles/v-a1/v-a1-1

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two-and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement*, *6*(3), 249-260. doi: http://dx.doi.org/10.1177/014662168200600301

Julo, J. (1995). *Représentation des problèmes et réussite en mathématiques. Un apport de la psychologie cognitive à l'enseignement.* Rennes, France : Presses de l'Université de Rennes.

Lafontaine, D. (2012). Des politiques aux pratiques d'évaluation en Belgique francophone. *Actes du 24e colloque international de l'ADMÉÉ-Europe,* Luxembourg, Luxembourg. Abstract retrieved from : http://admee2012.uni.lu

Lafontaine, L. Soussi, A., & Nidegger, C. (2009). Évaluations internationales et/ou épreuves nationales : tensions et changement de pratiques. In L. Mottier Lopez & M. Crahay (Eds.), *Évaluations en tension* (pp. 61-80). Bruxelles, Belgique : De Boeck.

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

Leduc, D., Riopel, M., Raîche, G., & Blais, J.-G. (2011). L'influence des définitions des habiletés disciplinaires sur la création et le choix d'items dans le PISA et le TEIMS. *Mesure et évaluation en éducation*, *34*(1), 97-130.

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

Loye, N., Caron, F., Pineault, J., Tessier-Baillargeaon, M., Burney-Vincent, C., & Gagnon, M. (2011). La validité du diagnostic issu d'un mariage entre didactique et mesure sur un test existant. In G. Raîche, K. Paquette-Côté, & D. Magis (Eds), *Des mécanismes pour assurer la validité de l'interprétation de la mesure en éducation, volume 2* (pp. 11-30). Ste-Foy, Québec : Presses de l'Université du Québec.

OCDE (2004). *Cadre d'évaluation de PISA 2003 : connaissances et compétences en mathématiques, lecture, science et résolution de problèmes.* Paris : OCDE.

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

Roegiers, X. (2005). *L'évaluation selon la pédagogie de l'intégration : est-il possible d'évaluer les compétences des élèves ?* In K. Toualbi-Thaâlibi & S. Tawil (Eds.), *La Refonte de la pédagogie en Algérie - Défis et enjeux d'une société en mutation* (pp. 108-124). Alger : UNESCO-ONPS.

Roegiers, X. (2007). *Des situations pour intégrer les acquis scolaires*. Bruxelles, Belgique : De Boeck.

Romainville, M. (1996). L'irrésistible ascension du terme "compétence" en éducation. *Enjeux, 37-38*, 132-142.

Romainville, M. (2006). L'approche par compétences en Belgique francophone : où en est-on ? *Les Cahiers pédagogiques, Quel socle commun ?*, 439, 24-25.

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

Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition and sense-making in mathematics. In D. A. Grouws (Ed.), *Handbook of research on mathematics learning and teaching* (pp. 334-370). New York: Macmillan.

UNESCO (2007). Curriculum Change and Competency-based Approaches: A Worldwide Perspective, Prospects, 37 (2). http://www.ibe.unesco.org/en/services/online-materials/publications/recent-publications/single-view/news/curriculum-change-and-competency-based-approaches-a-worldwide-perspective-prospects-n-142/2842/next/4.html

Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Löf, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Date de réception : 3 mai 2012

Date de réception de la version finale : 10 mai 2013

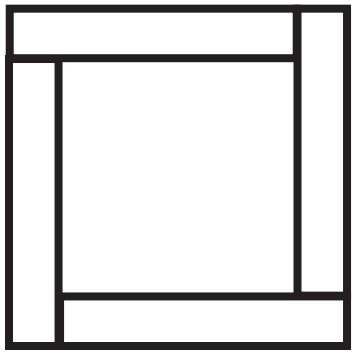
Date d'acceptation : 14 juin 2013

Annexe 1

Un exemple de tâche complexe de référence

Menuisier

Un menuisier assemble un cadre en bois avec quatre planches rectangulaires identiques. Le périmètre extérieur mesure 64 cm, le périmètre intérieur mesure 44 cm.



La figure ne respecte pas les dimensions réelles.

Quelle est la largeur d'une planche ?

La largeur d'une planche est de …… cm.

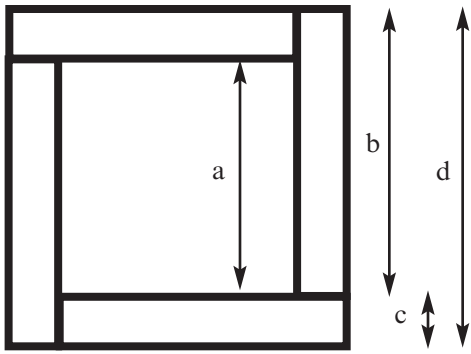
Annexe 2

Un exemple de tâche complexe décomposée

Menuisier (tâche décomposée)

Un menuisier assemble un cadre en bois avec quatre planches rectangulaires identiques. Le périmètre extérieur mesure 64 cm, le périmètre intérieur mesure 44 cm.

Quelle est la largeur d'une planche ?



La figure ne respecte pas les dimensions réelles.

On a représenté différentes mesures (a, b, c et d) sur le dessin du cadre.

1. Dans ce problème, il faut trouver …

… combien vaut a.

… combien vaut b.

… combien vaut c.

… combien vaut d.

2. Complète les phrases suivantes en choisissant la lettre adéquate (a, b, c ou d).

La longueur d'une planche est représentée par la lettre …

La longueur d'un côté extérieur du cadre est représentée par la lettre …

La longueur d'un côté intérieur du cadre est représentée par la lettre …

La largeur d'une planche est représentée par la lettre …

3. Si on connaît le périmètre d'un carré, comment peut-on trouver la longueur d'un côté ?

On multiplie le périmètre par 4.

On divise le périmètre par 2.

On divise le périmètre par 4.

On additionne la longueur de tous les côtés.

On calcule la racine carrée du périmètre.

On fait « côté fois côté ».

4. Sachant que le périmètre extérieur du cadre mesure 64 cm, quelle est la longueur d'un côté extérieur du cadre?

La longueur d'un côté extérieur du cadre est de …… cm.

5. Sachant que le périmètre intérieur mesure 44 cm, quelle est la longueur d'un côté intérieur du cadre?

La longueur d'un côté intérieur du cadre est de …… cm.

6. Tu connais à présent la longueur d'un côté extérieur du cadre et la longueur d'un côté intérieur du cadre, comment vas-tu procéder pour calculer la largeur d'une planche?

C'est simplement la longueur a.

C'est simplement la longueur d.

On doit faire d – a.

On doit faire (d – a) : 2.

On doit faire [(d – a) : 2] + a.

7. Calcule la largeur d'une planche.

La largeur d'une planche est de …… cm.