

Carulla, Cristina, Kassam, Shanoor, Corti, Diego & Ruiz-Primo, Maria Araceli. (2015). Outil pour la caractérisation d'items mathématiques pour des épreuves à grande échelle. In Detroz, P. & Borsu, O. (éds), *L'évaluation à la lumière des contextes et des disciplines : actes du 27ème colloque de l'ADMEE-Europe, 28-30 janvier 2015, Liège* (pp. 219-222). Liège : Université de Liège.

Outil pour la caractérisation d'items mathématiques pour des épreuves à grande échelle

Cristina Carulla, Shanoor Kassam, Diego Corti & Maria Araceli Ruiz-Primo

Traditionnellement décentralisé, le système scolaire suisse se fédère avec l'entrée en vigueur des standards nationaux (CDIP, 2011) et de plans d'études pour chaque région linguistique. Il est prévu également, que des tests de référence soient réalisés pour vérifier si les standards nationaux sont atteints. Dans cet élan, les cantons de la Suisse romande (région francophone) se sont ralliés sous l'égide de la Convention scolaire romande (CSR). Cette Convention vise principalement à harmoniser les pratiques éducatives et à créer un espace de formation homogène. C'est ainsi que le Plan d'études romand (PER), un référentiel d'enseignement commun, a vu le jour en 2010 (CIIP, 2010). L'article 15 de la CSR prévoit la mise en place d'épreuves romandes communes (EpRoCom) afin de "vérifier l'atteinte des objectifs du Plan d'études". Cet article de la Convention précise que le test national pourrait remplacer une partie de l'épreuve romande dans le cas où la discipline du test de référence coïnciderait avec l'une des disciplines visées par les EpRoCom.

La réalisation d'une évaluation commune à tous les cantons de Suisse romande est un projet complexe pour plusieurs raisons. D'abord, l'évaluation du système de formation est un domaine réservé des cantons qui présente de grandes diversités tant dans les pratiques que dans les objectifs visés par les dispositifs existants. Ainsi, chaque canton définit son propre dispositif d'évaluation, qui peut varier suivant le choix des objets à évaluer, le nombre d'épreuves, les visées, la structure, les types d'items et la périodicité. Ensuite, la mise en application du PER se déroule avec des rythmes différents et est interprétée selon les habitudes cantonales. De plus, les moyens d'enseignement utilisés par les cantons ne sont pas encore tous en accord au PER. Actuellement, seuls les moyens d'enseignement des mathématiques pour la fin de la scolarité (cycle 3) ont été construits en se basant sur la structure et les contenus du PER.

1. La problématique

Dans un tel contexte, plusieurs questions de validité doivent être examinées concernant la réalisation de ces EpRoCom. Bodin (2006) parle de la validité interne d'épreuves en mathématiques quand les contenus mathématiques des items sont alignés sur une manière de choisir, d'organiser et de concevoir les contenus visés par l'évaluation (ex. alignement des items de l'épreuve PISA sur les sous-échelles de contenu et de processus selon le cadre théorique de l'épreuve OCDE (2013)). Dans la situation des EpRoCom, la validité interne de l'épreuve doit être garantie par le recouvrement des contenus des items de l'épreuve sur des contenus mathématiques choisis et organisés selon la structure du PER. Bodin (2006) parle également de la validité externe d'une épreuve, quand on étudie le degré de recouvrement du contenu mathématique des items sur le contenu mathématique d'autres référentiels ou manières de choisir et d'organiser des connaissances mathématiques. Prenant en compte d'une part le texte de la CSR qui considère que les tests de référence dans une discipline, dont le contenu des items est aligné au contenu mathématique déterminé et organisé selon les standards nationaux, pourraient aussi servir à évaluer l'atteinte des objectifs du PER et, d'autre part, l'existence de nouveaux dispositifs cantonaux d'évaluation dont le contenu mathématique des items est explicitement aligné sur le contenu mathématique selon la structure du PER, nous voulons mettre en place un dispositif de validation externe des futures EpRoCom qui examine le degré de recouvrement du contenu mathématique des items des épreuves selon le PER et selon les standards nationaux.

Dans ce but, nous avons développé une stratégie de recherche pour aborder la validité des EpRoCom dans l'étape de conception, en attendant que des décisions politiques soient prises quant à la finalité, la visée et les caractéristiques des EpRoCom. Nous avons prévu trois études de validité : l'alignement* des items d'une épreuve aux standards nationaux et au PER ; le *mapping* (carte) des caractéristiques des épreuves cantonales ; et l'analyse des biais et inégalités dans la conception des épreuves.

Par rapport au thème de l'ADMEE 2015, notre communication contribue à construire des liens entre contextes (pratiques d'évaluation régionales et épreuves cantonales), la didactique des mathématiques (analyse du contenu mathématique des items selon le contenu mathématique du PER et le contenu des standards) et l'évaluation (EpRoCom). Notre but est de présenter l'outil développé pour la réalisation d'études de validité. Il est à la fois méthodologique (*mapping*), analytique (caractérisation des épreuves) et informatique (plateforme web et base de données).

* Ici, le sens donné au mot alignement est réduit au degré de recouvrement du contenu mathématique —connaissances, démarches, stratégies, notions, concepts et raisonnements— du corpus d'items d'une épreuve sur le contenu d'un référentiel des connaissances mathématiques.

À partir de la technique du *mapping* inspirée de Ruiz-Primo et Li (2015), l'outil permet la visualisation du contenu mathématique des évaluations cantonales et des caractéristiques des items ; permet l'étude du recouvrement du contenu des items des évaluations cantonales sur le contenu du PER et des standards nationaux ; et contribue au développement d'une interprétation inter-cantonale du PER par rapport aux items des épreuves. L'outil est constitué d'un corpus d'items et d'épreuves, organisé suivant divers critères taxonomiques liés aux épreuves, aux items et aussi aux structures de référentiels (p.ex. PER et standards nationaux) ou modèles d'évaluation et d'apprentissage. Avec cet outil, nous aimerions répondre aux questions suivantes : quel alignement peut être établi entre le PER, les standards nationaux et les tâches évaluatives (items) des épreuves cantonales ? Quels contenus du PER ont été évalués par les épreuves cantonales ? Comment ont-ils été évalués ? Quels objets du PER n'ont pas été évalués ?

2. L'outil, sa méthodologie et quelques perspectives de résultats

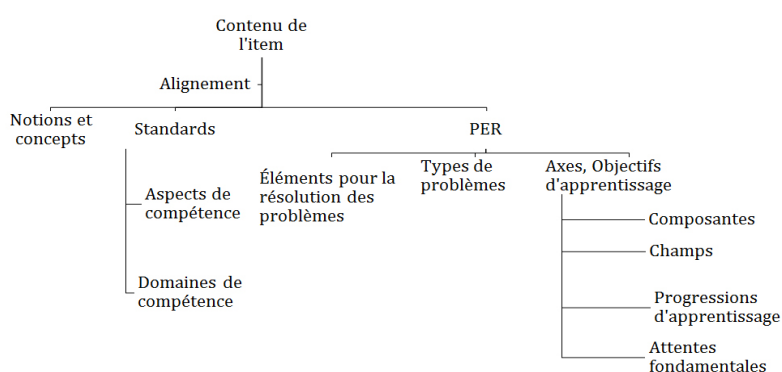
Le *mapping* est un outil qui s'inspire de l'approche de «profiling» utilisée par Ruiz-Primo et Li (2015) dans le cadre d'une recherche visant à caractériser la relation entre le contexte des items PISA en sciences et les résultats des élèves. Le *mapping* permet de créer une cartographie des connaissances et compétences mathématiques visées par les évaluations cantonales et des caractéristiques des items utilisées dans les épreuves. Notre corpus de données est constitué des évaluations cantonales en mathématiques de fin de second cycle (fin du primaire, élèves de 12 ans) pour la période 2011-2014.

Le dispositif de codage est une plateforme Web dont le nom de code est «Projet d'outillage et de suivi du PER» (ci-après ProSPER). La création de cette plateforme s'inscrit dans le projet de création d'une banque d'items intercantonale et de mutualisation des compétences des différents services cantonaux. Les classifications des items devraient servir d'abord à caractériser les items pour les travaux des chercheurs dans le domaine de l'évaluation, mais également aux concepteurs d'épreuves cantonales et EpRoCom et aux enseignant-e-s (sélection de « bons » items à leur disposition). Le système de codage de ProSPER offre une organisation hiérarchique en arborescence. Par exemple, si un codeur sélectionne la dimension de l'axe mathématique du PER « Espace » pour un item, seul les objets liés à cet axe seront visibles pour la sélection suivante de codes. La validité de l'analyse des items par le *mapping* sera évaluée selon deux aspects : 1) la consistance avec laquelle les items des épreuves seront codés par différents codeurs (validité inter-juges) et 2) le jugement d'experts en didactique des mathématiques sur les dimensions du *mapping* (validité interne de Bodin). La consistance sera analysée entre plusieurs codeurs et en utilisant un échantillon stratifié d'items par épreuve et par canton. Un panel d'experts en conception d'épreuves mathématiques des cantons sera invité à juger de la qualité du cadre théorique du *mapping*.

Notre corpus de données est constitué d'épreuves cantonales et nos unités d'analyse sont essentiellement les items de ces épreuves. L'item est défini comme un élément individuel de l'évaluation, noté pour lui-même et incluant une consigne et/ou une question qui guide l'élève vers la réalisation des tâches mathématiques. Il est l'unité par laquelle des informations relatives aux compétences et connaissances de l'élève vont être inférées. C'est une tâche d'évaluation créée avec l'intention de recueillir de l'information sur les savoirs et savoir-faire des élèves en mathématiques.

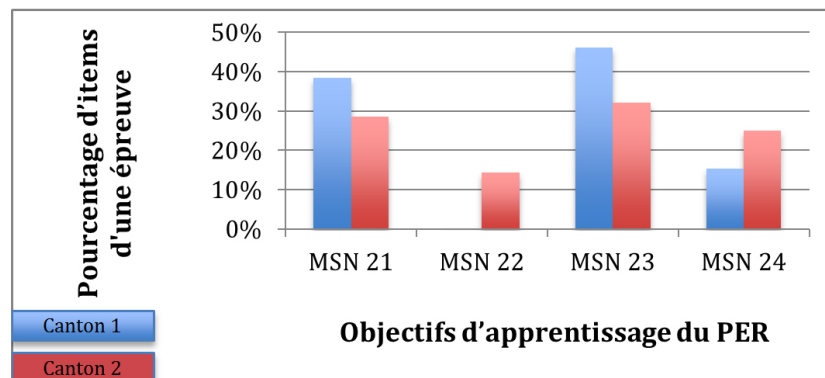
L'analyse est faite au niveau des contenus mathématiques de l'item, en examinant l'alignement avec différentes dimensions : les dimensions structurant le PER et son contenu mathématique (démarches, connaissances, savoirs et savoir-faire) les dimensions structurant les standards nationaux et son contenu, et les notions et concepts mathématiques nommés dans les deux référentiels. Par ailleurs, la forme de l'item a été décrite à partir des catégories émergentes de l'observation : taxonomie des demandes cognitives, habillage et contexte donnant du sens aux connaissances mathématiques, longueur du texte de la consigne, utilisation des graphiques ou autre type de représentations, etc. Chaque analyse permet de caractériser et définir les différences et similitudes par rapport à ce qui a été évalué ou pas et comment cela a été évalué.

Nous avons défini le contenu d'un item selon les objets de la structure du PER (CIIP, 2010) et les standards nationaux des compétences en mathématiques (CDIP, 2011). Les objets du PER se déclinent en objectifs d'apprentissage (correspondant également à des axes mathématiques : espace, nombre, opération, longueurs et mesures et modélisation), composantes d'objectifs, progressions d'apprentissage, attentes fondamentales, éléments pour la résolution de problèmes et en une typologie des problèmes mathématiques. Les éléments des standards nationaux (HarmoS) se déclinent quant à eux en aspects de compétences (savoir reconnaître et décrire, mathématiser et modéliser...) et en domaines de compétences (correspondant à la géométrie, nombres et opération...).



Graphique 1. Mapping au niveau du contenu de l'item

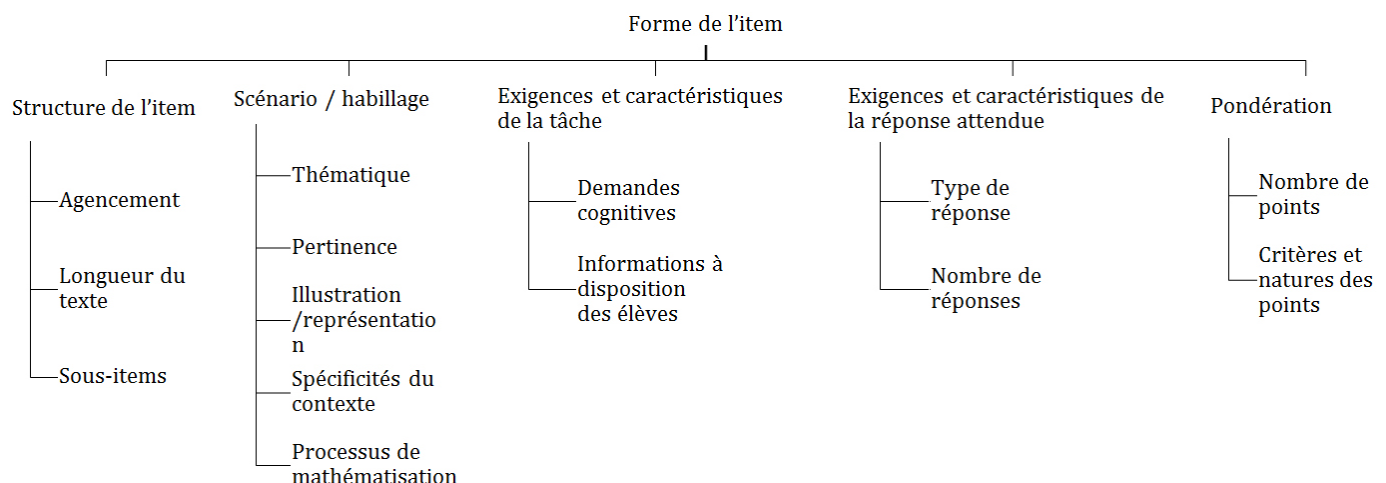
Chaque item d'une épreuve est codé selon cette structure. Un processus d'accord inter-juges sur le codage permet d'explicitier l'interprétation du PER selon les tâches analysées dans les épreuves. L'analyse permet de repérer le nombre d'items qui a été mis en lien avec un objectif du PER ou précisément par rapport aux divers objets qui constituent cet objectif. Par exemple, après codage des items de l'épreuve de 2013 du canton 1 et du canton 2, on a établi un bilan par rapport à la quantité d'items qui ont été alignés à 4 objectifs du PER : MSN 21 — Poser et résoudre des problèmes pour structurer le plan et l'espace... ; MSN 22 — Poser et résoudre des problèmes pour construire et structurer des représentations des nombres rationnels... ; MSN 23 — Résoudre des problèmes additifs et multiplicatifs... ; et MSN 24 — Utiliser la mesure pour comparer des grandeurs...



Graphique 2. Analyse de deux épreuves cantonales au niveau du contenu

On pourra dire par exemple, que l'épreuve de 2013 du canton 1 n'a pas évalué l'objectif MSN 22. En effet, les codeurs n'ont pas identifié d'items ayant des contenus de l'objectif MSN 22, par contre, dans le canton 2, 14.29% des items de l'épreuve ont été codés en lien aux connaissances faisant partie de cet objectif.

La caractérisation de la forme de l'item considère la structure de l'item en termes d'agencement des éléments (titre, énoncé, espace pour la réponse des élèves, espace pour les recherches de l'élève, illustration...), de longueur du texte (nombre de mots, de phrases, de paragraphes) et de présence éventuelle de sous-items ; les caractéristiques de l'habillage de l'item incluent notamment la nature de la thématique (aucune, réelle ou pseudo-réelle), sa déclinaison en sous-thématique (environnement, social, etc.), sa pertinence pour la tâche demandée à l'élève ainsi que le type d'illustration présent (graphique, croquis...) ; les demandes cognitives selon la nature de la tâche et les informations qui lui sont fournies ; le type de réponses (choix multiple, vrai-faux ou oui-non, lacunaire, champ ouvert pour des dessins, des justifications, des calculs...) ; les critères de pondération et ce qui est concrètement mesuré par l'item.



Graphique 3. Mapping au niveau de la forme de l'item

D'autres types d'analyses peuvent être réalisées par rapport aux items selon leur forme. Par exemple, on pourra comparer le type d'habillage (contexte) des items. Dans le cas des cantons 1 et 2 pour l'épreuve de 2013, 40% des items de l'épreuve du canton 1 ont été codés comme « possédant un habillage pseudo-réel », c'est-à-dire que la situation non mathématique qui donne sens aux objets mathématiques de l'item n'est pas réelle, car elle n'utilise pas de données d'une réalité particulière (ex. Jean a acheté une pomme au supermarché avec 2 francs. Il avait 7 francs, combien de francs lui reste-t-il après avoir acheté la pomme ?). La

plupart de ces items ont utilisé des situations sociales, non liées à l'environnement, aux sciences naturelles ou au vivant visés par le PER. On constate notamment que l'épreuve du canton 1 utilise uniquement des habillages faisant référence à des situations cataloguées comme sociales, alors que celle du canton 2 est plus diversifiée (technique et naturelle). En allant encore plus en profondeur dans l'analyse, le champ social se divise lui-même en thématiques (jeux, emploi du temps, sport, achats, légendes et voyages) et alors que le canton 1 a privilégié les thématiques du jeu, de l'emploi du temps et des légendes, les items du canton 2 portent quant à eux sur le sport et les achats.

3. Conclusions

La démarche du *mapping* des épreuves des cantons romands a trois buts : 1) définir quelles connaissances et compétences mathématiques ont été mesurées dans les épreuves cantonales ; 2) identifier comment ces connaissances et compétences ont été mesurées ; et 3) construire une manière romande d'interpréter les objets du PER pour la construction des items des EpRoCom.

Le *mapping* permet de caractériser chaque item d'une épreuve cantonale. Cette démarche nous aidera au développement des EpRoCom au moins de trois manières : 1) pour développer des épreuves qui soient sensibles aux différences cantonales que nous aurons « cartographiées » ; 2) pour déterminer les types d'items qui sont utilisés par tous les cantons (par exemple, il ne serait pas juste pour certains élèves de répondre seulement à des questions de type QCM s'ils n'y ont jamais été confrontés auparavant dans les épreuves de leur canton) ; et 3) pour identifier les connaissances mathématiques et les compétences qui ont été visées dans les épreuves de chaque canton.

L'information issue de l'approche du *mapping* nous donnera des indicateurs de validité pour la construction d'items. En outre, les données provenant du *mapping* des épreuves de différents cantons pourraient contribuer à concevoir une épreuve commune qui soit « plus juste », en considérant dans sa conception ce qui a été appris par rapport aux épreuves de chaque canton.

Le travail proposé contribue au domaine de la didactique des mathématiques : 1) en proposant une approche pour caractériser des items mathématiques dans des évaluations ; 2) en proposant un modèle d'analyse d'items pour estimer la validité interne (Bodin) d'une épreuve (ex. cet item va-t-il visibiliser le raisonnement mathématique ?) ; et 3) en contribuant à la connaissance des propriétés techniques et des caractéristiques d'items mathématiques (p.ex. partant des performances des élèves à une épreuve, quelles sont les caractéristiques des items qui présentent un indice de difficulté et de discrimination approprié ? Quelles sont les caractéristiques des contextes mathématiques qui conduisent à une discrimination basse ou à une difficulté haute ?).

4. Références bibliographiques

Bodin, A. (2006). *Ce qui est vraiment évalué par PISA en mathématiques. Ce qui ne l'est pas. Un point de vue français*. Besançon : Université de Franche-Comté. Institut de recherche sur l'enseignement des mathématiques, [En ligne]. http://educmath.ens-lyon.fr/Educmath/ressources/etudes/experimentation-math/com_pisa_ff_matheduc.pdf (Page consultée le 13 mars 2015)

Conférence intercantonale de l'instruction publique de la Suisse romande et du Tessin (CIIP) (2010). *Plan d'études romand (PER)*. Neuchâtel : CIIP, [En ligne]. <http://www.plandetudes.ch/> (Page consultée le 13 mars 2015).

Conférence suisse des directeurs cantonaux de l'instruction publique (CDIP) (2011). *Compétences fondamentales pour les mathématiques : standards nationaux de formation adoptés par l'Assemblée plénière de la CDIP le 16 juin 2011*. Berne : CDIP, [En ligne]. http://edudoc.ch/record/96783/files/grundkomp_math_f.pdf (Page consultée le 13 mars 2015)

Organisation de coopération et de développement économiques (OCDE) (2013). *Cadre d'évaluation et d'analyse du cycle PISA 2012 : compétences en mathématiques, en compréhension de l'écrit, en sciences, en résolution de problèmes et en matières financières*. Paris : OCDE, [En ligne]. http://www.oecd.org/pisa/pisaproducts/Cadre%20d'%C3%A9valuation%20PISA%202012%20e-book_FR.pdf (Page consultée le 13 mars 2015)

Marc, V. & Wirthner, M. (2013). *Développement d'un modèle d'évaluation adapté au PER : rapport scientifique du projet d'épreuves romandes communes*. Neuchâtel : IRDP, [En ligne]. http://publications.irdp.relation.ch/ftp/1368438968epreuves_romandes_communes_bouquin.pdf (Page consultée le 13 mars 2015)

Ruiz-Primo, M.A. & Li, M. (2015). The Relationship Between Item Context Characteristics and Student Performance : The Case of the 2006 and 2009 PISA Science Items. *Teachers College Record*, vol. 117, no 1