VALIDITÉ D'UNE ÉVALUATION EXTERNE. COMPLÉMENTARITÉ DES APPROCHES DIDACTIQUE ET PSYCHOMÉTRIQUE

Brigitte **GRUGEON-ALLYS**

Université Paris Est-Créteil - ESPE
Laboratoire de didactique André Revuz
brigitte.grugeon-allys@u-pec.fr

Nadine **GRAPIN**

Université Paris Est-Créteil - ESPE
Laboratoire de didactique André Revuz
nadine.grapin@u-pec.fr

Résumé

Comment étudier la validité d'un dispositif d'évaluation ? Dans la continuité des travaux de Bodin (1997, 2006), nous présentons une recherche portant sur l'étude des validités d'une évaluation. Cette recherche conduit à mettre en place une méthodologie d'analyse multidimensionnelle articulant l'analyse *a priori* des tâches à différents niveaux de granularité. L'analyse concerne d'une part, au niveau global de l'ensemble du test une étude épistémologique du savoir évalué, d'autre part, au niveau local une étude du processus de résolution de chacune des tâches. Ces deux approches épistémo-didactique (Grugeon 1997) et psycho-didactique (Vantourout & Goasdoué 2014) sont mises en relation avec des caractéristiques psychométriques du test. Nous montrons comment ces deux recherches ont permis de préciser les résultats de l'évaluation nationale CEDRE (Cycle des évaluations disciplinaires réalisées sur échantillon) 2008 et 2014, sur deux domaines (numération et arithmétique des entiers en fin d'école, algèbre en fin de collège) et d'interroger la méthodologie de conception d'évaluations à grande échelle.

Mots clés

Évaluation externe ; épistémologie ; validité ; approche anthropologique ; psychométrie ; psycho-didactique.

INTRODUCTION

Nous abordons ici des questions relatives à la validité d'évaluations externes. Nous entendons par « externe », des évaluations à grande échelle (type évaluation nationale ou internationale),

ce qui explique que nous prenions en compte des questions psychométriques, leur conception leur accordant une place essentielle. Qu'entend-on par validité d'évaluations ? Une évaluation valide doit évaluer ou mesurer effectivement ce qu'elle est sensée évaluer et rien d'autre (Grégoire et Laveault, 2014). Nous développons une approche multidimensionnelle pour aborder cette question très complexe, à partir de plusieurs champs de recherche, éducatifs, sciences de l'éducation et didactique des mathématiques. Cette approche permet d'utiliser les filtres conceptuels de chaque champ pour découper le phénomène étudié et mobiliser leurs objets d'étude afin d'aborder un phénomène complexe selon différentes entrées.

L'approche didactique permet d'interroger à différents niveaux de granularité des évaluations standardisées sur les acquis des élèves, en mathématiques, conçues essentiellement dans une approche psychométrique. En France, le dispositif CEDRE à la fin de l'école primaire (élèves âgés de 9-10 ans) et à la fin du collège (élèves âgés de 14-15 ans) a pour objectifs d'évaluer les acquis des élèves en mathématiques au regard des programmes scolaires. La conception des items, l'analyse des réponses et l'interprétation des résultats sont menées par des enseignants et inspecteurs et par des statisticiens qui calculent différents indicateurs psychométriques et sélectionnent les items retenus. Le modèle statistique utilisé dans ces évaluations est celui du modèle de réponse à l'item (MRI¹) (Lescure & Pastor, 2012). La validité de CEDRE repose de ce fait majoritairement sur une approche psychométrique.

Nous interrogeons la validité de CEDRE à partir d'une approche didactique à deux niveaux de granularité : aux niveaux local et global sur la représentativité des items du test et leur couverture du savoir évalué à partir d'une étude épistémologique, et au niveau local sur les processus de résolution de chacune des tâches. Nous précisons ces deux approches épistémologiques (Grugeon 1997) et psycho-didactique (Vantourout & Goasdoué 2014) et les illustrons dans deux domaines : l'arithmétique des entiers en fin de CM2 et l'algèbre élémentaire en fin de 3e. Nous mettons en relation les résultats de l'analyse didactique avec des caractéristiques psychométriques du test. En retour, nous interrogeons d'un point de vue didactique les résultats des évaluations en relation avec les praxéologies mathématiques et didactiques des programmes et des pratiques enseignantes. Nous montrons comment cette recherche a permis d'analyser les résultats de l'évaluation nationale CEDRE 2008, 2014, sur les deux domaines, et d'interroger la méthodologie de conception d'évaluations à grande échelle.

DEUX APPROCHES COMPLEMENTAIRES DES VALIDITES D'UNE EVALUATION

CEDRE : UNE EVALUATION STANDARDISEE DES ACQUIS DES ELEVES

Cette évaluation a lieu en fin de CM2 ou de 3e et concerne différentes disciplines, qui sont évaluées avec un cycle de six ans ; les mathématiques ont été ainsi concernées en 2008 et 2014 pour les deux dernières versions. CEDRE a pour finalités d'évaluer les connaissances et les compétences des élèves au regard des programmes scolaires en vigueur en France. Les items appartiennent aux domaines mathématiques des programmes évalués. Le cadre de l'évaluation

¹ « Les MRI sont une classe de modèles probabilistes. Ils modélisent la probabilité qu'un élève donne une certaine réponse à un item, en fonction de paramètres concernant l'élève et l'item [...]. Les MRI ont un intérêt pratique pour la construction de tests [...] : si le modèle est bien spécifié sur un échantillon donné, les paramètres des items – en particulier leurs difficultés – peuvent être considérés comme fixes et applicables à d'autres échantillons dont il sera alors possible de déduire les paramètres relatifs aux élèves – en particulier, leur niveau de compétence. » (Rocher 2015, p. 44)

impose une certaine répartition des items entre ces différents domaines ; les concepteurs doivent donc concevoir des items répondant à cette exigence, tout en veillant à ce que les items soient plus ou moins complexes afin de pouvoir caractériser les connaissances de tous élèves, des plus faibles comme des plus forts. Les questions recouvrent plusieurs formats : QCM, question Vrai/Faux, questions ouvertes à réponse courte, contrainte ou longue (pour la version 2014 en collège). Les QCM sont très majoritaires.

Fondée sur une approche psychométrique et sur le MRI, l'évaluation CEDRE est conçue sur deux ans. Une fois les items conçus par des enseignants experts, un groupe de pilotage s'assure de leur conformité au cadre de l'évaluation et du programme. Une expérimentation du dispositif est réalisée l'année $n-1$ sur un échantillon représentatif d'élèves du niveau scolaire concerné. Seuls les items discriminants² sont sélectionnés pour le test final. De plus, les items retenus doivent relever de niveaux de difficulté variés³. Ces caractéristiques psychométriques des items, calculées après l'étude statistique des réponses aux items, déterminent les conditions de validité du dispositif d'évaluation (Rocher, 2015). L'étude statistique fournit une échelle de performance en six groupes⁴, hiérarchisés selon les connaissances mathématiques qu'ils maîtrisent. Cette échelle est définie *a posteriori* à partir des scores sur l'ensemble des items.

Validité d'une évaluation

Dans la plupart des pays de l'OCDE, l'étude des effets des évaluations standardisées sur les performances des élèves en termes d'efficacité ou d'inégalité scolaire reste faible (Mons, 2009). Et pourtant, le poids accordé aux évaluations standardisées à grande échelle, à l'international ou en France devient de plus en plus important. On peut se demander si ce type d'évaluation externe permet de donner des informations fiables, exploitables sur les acquis des élèves, surtout qu'il s'agit d'une mesure, ici un score. L'élaboration de ces dispositifs d'évaluation s'appuie-t-elle sur une représentativité des tâches au regard du programme pour évaluer les différents aspects des compétences évaluées ? Les items permettent-ils effectivement aux élèves de développer des processus de réponses envisageables et correspondant à ceux visés ?

Ces questions concernent la qualité de l'évaluation qui peut être caractérisée par la pertinence, la validité et la fiabilité du dispositif d'évaluation. La pertinence concerne le caractère plus ou moins approprié de l'épreuve relativement aux objectifs visés, la validité définit le degré d'adéquation entre ce que l'on déclare évaluer et ce que l'on fait réellement, la fiabilité établit le degré de confiance que l'on peut accorder aux résultats observés (De Ketele et Gérard, 2005). Nous centrons notre étude uniquement sur la validité d'une évaluation et distinguons en appui sur cinq types de preuves de validité définies par Grégoire & Laveault (2014, p.165) :

- la validité du contenu pour étudier dans quelle mesure, les items sont représentatifs du domaine évalué,
- la validité des processus de réponse pour décider si les items proposés permettent d'évaluer ce qui est visé et de recueillir les processus de réponse et les démarches d'élèves,
- la validité liée à la structure interne du test pour vérifier l'unidimensionnalité du test

² Les items doivent avoir un bon indice de discrimination entre les élèves qui réussissent et ceux qui échouent. Les items peu discriminants ($R\text{-bis} < 0,2$) apportent peu d'information et sont écartés après la phase d'expérimentation du test et avant la mise en place de l'évaluation finale.

³ La difficulté d'un item est déterminée à partir de la proportion d'élèves qui a donné une bonne réponse. Plus la proportion d'élèves ayant réussi est élevée plus l'item est facile.

⁴ « Les scores estimés sont alors standardisés de sorte que les élèves de 2007 [pour les sciences physiques] aient une moyenne de 250 et un écart-type de 50. Puis, la distribution des scores est « découpée » en six groupes de la manière suivante : nous déterminons le score-seuil en deçà duquel se situent 15 % des élèves (groupes 0 et 1), nous déterminons le score seuil au-delà duquel se situent 10 % des élèves (groupe 5). Entre ces deux niveaux, l'échelle a été scindée en trois parties d'amplitudes de scores égales correspondant à trois groupes intermédiaires » (Rocher, 2015 p56).

pour évaluer les connaissances et compétences visées.

Approche multidimensionnelle des validités d'une évaluation

Approche psychométrique

La conception de CEDRE à partir d'une approche psychométrique est fortement ancrée sur l'unidimensionnalité du test - chacun des items étant sensé mesurer une seule dimension, que l'on pourrait nommer de façon générale, « la compétence mathématique » - et l'indépendance des items. La validité des résultats de l'évaluation est déterminée à partir de deux caractéristiques psychométriques : les niveaux de difficulté des items doivent être variés et les items discriminants. Les items peu discriminants, c'est-à-dire qui apportent peu d'informations sur les acquis des élèves sont donc écartés après la phase d'expérimentation et avant l'évaluation finale. Par ailleurs, lors de la comparaison temporelle des résultats des évaluations 2008 et 2014, des items peuvent aussi présenter un fonctionnement différentiel⁵ et être écartés de l'échelle des performances, après la passation de l'évaluation.

D'un point de vue psychométrique, les preuves de validité portent peu sur le contenu effectif des items ; la qualité de ce dernier est prise en compte et interrogée lors de la conception de l'évaluation, mais l'analyse de tâche pour chaque item d'un point de vue mathématique (global et local) n'est pas réalisée de façon systématique à partir d'outils d'analyse - démarches ou erreurs envisageables au vu de l'énoncé et du niveau scolaire. En effet, souvent seul le résultat en termes correct ou incorrect est pris en compte. Aussi, l'étude didactique du contenu mathématique en tant que tel, est de ce fait peu prise en compte dans l'approche psychométrique, comme celle des processus de réponse.

Approche didactique

On comprend bien alors le rôle de l'analyse *a priori* des tâches, à des niveaux local ou global. Ce type d'analyse permet d'éclairer différents aspects de validité didactique pour caractériser la pertinence des items : d'une part, au regard des objectifs d'apprentissage visés et des programmes scolaires, et d'autre part, au regard des processus de réponse potentiels. En premier lieu, il s'agit d'étudier leur représentativité en fonction des programmes évalués - types de tâche, variables didactiques associées et couverture du domaine mathématique concerné : c'est la facette de la validité épistémologique-didactique du contenu. La deuxième facette correspond à la validité psycho-didactique définie par Vantourout et Goasdoué (2014). Ces caractéristiques didactiques peuvent être définies lors de la conception et avant la passation de l'épreuve (*a priori*) et non post passation (*a posteriori* comme c'est le cas pour les caractéristiques psychométriques de l'item qui sont calculées à partir des réponses des élèves). L'approche didactique permet donc d'interroger les choix réalisés au regard de la validité psychométrique. L'analyse *a priori* est bien au centre de l'étude des deux facettes de la validité didactique mais conduit à des exploitations différentes pour l'analyse du test. Nous nous centrons, ici, sur la facette qualifiée d'« épistémologique-didactique ».

⁵ « Un fonctionnement différentiel d'item (FDI) apparaît entre des groupes d'individus dès lors qu'à niveau égal sur la variable latente mesurée, la probabilité de réussir un item donné n'est pas la même selon le groupe considéré. » (Rocher 2015, p.53)

REGARD THEORIQUE SUR LA VALIDITE EPISTEMO-DIDACTIQUE

L'apport de la théorie anthropologique du didactique (TAD) sur la validité épistémologique-didactique

Éléments de la TAD

Dans l'approche épistémologique et didactique, nous nous intéressons d'abord à la représentativité des items d'un point de vue local pour chaque item mais aussi d'un point de vue global sur l'ensemble du test : ce point de vue est un préalable à l'étude de la validité psycho-didactique. Il permet d'étudier si l'item proposé est adapté et en cohérence avec les programmes du niveau auquel il est proposé ou des niveaux précédents. L'étude de la représentativité des items doit se faire au regard des savoirs à enseigner et enseignés : les tâches choisies doivent produire des résultats qui permettent d'analyser de façon opérationnelle les connaissances et les compétences mathématiques des élèves en vue d'une exploitation par l'évaluateur (que ce soit l'enseignant ou l'institution).

Nous avons choisi le cadre de la TAD pour cette analyse. Chevallard (1999) modélise l'activité mathématique en termes de praxéologies mathématiques, c'est-à-dire, de types de tâches et de techniques les résolvant (savoir-faire), une technique étant justifiée par un discours technologique, lui-même étant justifié par une théorie (savoir). Evaluer l'activité mathématique d'un élève sur un domaine mathématique donné, revient à évaluer les praxéologies apprises au regard des praxéologies à enseigner (programmes) à partir d'un échantillon de l'ensemble des types de tâches constitutifs des praxéologies visées (Chevallard 2007, Rey et Feyfant, 2014). Comment construire de tels échantillons ?

Pour un domaine donné de savoir, la définition d'une praxéologie épistémologique de référence à partir d'une étude épistémologique du domaine mathématique (Bosch et Gascon, 2005) vise à décrire et à mettre en relation au cours du processus de transposition didactique à la fois les praxéologies à enseigner, enseignées et apprises. Une praxéologie de référence peut servir de fondement à la sélection d'échantillons de l'ensemble de types de tâches constitutifs des praxéologies visées pour évaluer des praxéologies apprises pour un domaine mathématique, à un niveau scolaire (ici fin CM2 ou fin 3e)⁶.

Éléments d'analyse a priori

Nous organisons l'étude de la validité épistémologique-didactique à partir d'une analyse praxéologique des tâches. Pour un domaine mathématique donné, nous sommes amenées à croiser la praxéologie à enseigner et la praxéologie épistémologique de référence, ce qui conduit à lister les genres de tâches qui relèvent des praxéologies visées : chaque item relèvera d'un type de tâches. Nous étudions ainsi pour chaque item :

- le type de tâche ;
- la nature du cadre en jeu ;
- les objets mathématiques en jeu et leur nombre. Nous distinguons s'ils sont des données de l'item, des objets à mobiliser pendant la résolution, des résultats ;
- les variables didactiques associées aux objets mathématiques (ou à la situation) en jeu et leurs valeurs compte tenu du niveau scolaire et des enjeux d'évaluation ;
- les registres de représentation sémiotiques en jeu, en entrée et en sortie, s'ils sont congruents ou non (Duval, 1996),

⁶ La conception de ce type d'évaluation doit aussi prendre en compte les praxéologies à enseigner et enseignées, développées à différents niveaux scolaires, préalablement au niveau scolaire visé.

- le nombre de types de tâches convoqués dans la résolution, et le type de convocation (t ou r convoqués) (Castela, 2008). Le nombre de types de tâches convoqués dans la résolution et les types de convocation définiront la complexité de l'item. La complexité est ainsi liée au fait que l'énoncé contient des indices qui peuvent orienter le résolveur vers la technique à utiliser ou l'identification du type de tâche.⁷

Différents aspects de la validité épistémo-didactique

Nous spécifions les aspects de la validité épistémo-didactique en distinguant la conception des items du test, celle de l'analyse des réponses, de l'interprétation des résultats et de leur exploitation.

- *Conception du test*

Au niveau local

Pour étudier l'adéquation de chaque item au regard des objectifs d'apprentissage visés et des programmes, nous distinguons les items de format QCM, de ceux à question ouverte.

Dans les deux cas, nous analysons le choix des objectifs d'évaluation visés et des codages au regard de l'analyse *a priori*. En croisant praxéologies de référence et à enseigner, nous comparons la technique et technologie attendues et idoines au regard du niveau scolaire et celles potentiellement réalisables selon les valeurs des variables didactiques définissant la tâche. Pour les items de format QCM, nous analysons le choix des distracteurs au regard de l'analyse didactique praxéologique, techniques envisageables ou erreurs récurrentes.

Au niveau global

Nous cherchons si les items recouvrent le domaine évalué et les types de tâches sont absents ou sur représentés en comparant les types de tâches représentés et ceux des programmes via la comparaison avec la praxéologie épistémologique de référence.

- *Analyse des réponses*

Au cours de l'analyse *a priori* nous dépassons le codage des réponses tel qu'il existe dans l'évaluation et le considérons :

- non seulement en termes de réussite ou d'échec ;
- mais aussi en termes d'éléments technologiques et théoriques mobilisés dans les techniques : nous étudions si elles soient idoines ou non, les types d'erreurs associées qui vivent dans le cas de praxéologies apprises non idoines, s'il y a lieu. Les erreurs peuvent être liées à la mobilisation d'une technique en dehors de son domaine de validité. Le croisement entre approches anthropologique et cognitive permet de mettre en relation praxéologies non idoines à un niveau scolaire donné et types d'erreurs.

Nous pouvons ainsi interroger le pourcentage de réussite des items en lien avec les objectifs d'évaluation visés et la mobilisation des techniques envisageables ou visées.

- *Interprétation des résultats et exploitation*

L'analyse didactique permet d'étudier l'impact sur le contenu des choix réalisés pour écarter

⁷ L'OM intervient au niveau "t-convoquée" (convoquée par la tâche) lorsque l'énoncé et/ou le contexte contiennent des indices qui orientent directement le résolveur vers la technique à utiliser.

L'OM intervient au niveau "r-convoquée" (convoquée par le résolveur) dans le cas contraire qui peut se produire si l'identification du type de tâches T est à la charge du résolveur ou (non exclusif) si celui-ci doit choisir la technique parmi plusieurs techniques connues pour T.

certain items compte tenu de leurs caractéristiques psychométriques calculées *a posteriori*. Elle permet aussi d'interroger les résultats statistiques, en particulier les échelles de performance, au regard des programmes à enseigner et des pratiques des enseignants, à partir de la comparaison entre praxéologies à enseigner et enseignées et praxéologies épistémologiques de référence associées aux domaines mathématiques en jeu dans les programmes. Cette comparaison permet d'inférer des besoins d'apprentissages ignorés dans les programmes (Castela 2008) ou des pistes de régulation en ce qui concerne les manques praxéologiques mathématiques et l'organisation didactique.

Nous opérationnalisons cette démarche sur deux domaines : celui « d'arithmétique sur les entiers » en fin d'école et celui de l'algèbre en fin de collège.

Deux domaines mathématiques : Numération et arithmétique des entiers (fin de CM2) et Algèbre (fin du collège)

Le domaine de l'arithmétique des entiers (fin d'école)

Le domaine de la numération et de l'arithmétique des entiers est structuré autour de trois praxéologies relatives à la production d'expressions arithmétiques (en situation de résolution de problèmes), à la gestion de la numération décimale et au calcul. Chacune de ces praxéologies est définie par des organisations mathématiques locales caractérisée par des types de tâches, des techniques et des technologies définies à partir de travaux menés sur le domaine (en particulier Fuson et al. 1997, Chambris 2008, Mounier 2010, Tempier 2013). Les types de tâches définis pour le domaine prennent en compte les dimensions outil/objet, les différents registres de représentation du nombre et les modes de calcul possibles : modéliser, convertir pour la production d'expressions arithmétiques, traduire, gérer le nombre sous un aspect ordinal, et cardinal pour la numération et calculer mentalement, de façon posée pour le calcul. Les technologies (de modélisation, de la numération et du calcul) sont définies par l'OM de référence.

Le domaine de l'algèbre (fin du collège)

Nous caractérisons la praxéologie mathématique épistémologique du domaine algébrique à partir d'une synthèse des travaux en didactique de l'algèbre (Chevallard, 1985, 1989b ; Kieran, 2007). Elle est structurée à partir des praxéologies relatives aux expressions algébriques, aux formules et aux équations. Les types de tâches concernent ces différents objets : généraliser, traduire, prouver, reconnaître, calculer - développer, factoriser pour les expressions, modéliser, traduire, reconnaître pour les formules, mettre en équation, traduire, reconnaître, résoudre une équation pour les équations. Ces praxéologies ponctuelles se regroupent en praxéologies locales, praxéologie de calcul et praxéologie d'usage de l'outil algébrique. La praxéologie de référence caractérise les propriétés idoines pour un calcul « intelligent et contrôlé » (prise en compte des aspects procédural et structural des expressions algébriques, de l'équivalence des expressions algébriques, des équations, de la dialectique numérique / algébrique), les discours et modes de raisonnement associés ainsi que les modes de représentation dans les registres de représentation sémiotique du domaine et leur mise en relation.

QUELQUES RESULTATS D'ANALYSE SUR CEDRE

Nous présentons pour chacun des deux domaines des exemples de résultats obtenus après analyse de la validité d'items à un niveau local et à un niveau global en montrant comment cette

analyse à deux niveaux s'articule avec la psychométrie.

CEDRE CM2 – Arithmétique des entiers

Au niveau local

Nous retenons deux exemples d'items qui ne sont pas représentatifs du type de tâche mais qui illustrent les questions de validité que nous avons soulevées précédemment et la façon dont l'analyse didactique permet de les appréhender.

Le premier item (Figure 1) représente un type de tâche de traduction ; il est formulé sous la forme d'une succession de trois vrai-faux, pour lesquels un seuil de nombre de bonnes réponses est fixé à trois sur trois) pour déclarer l'item réussi :

Indiquer vrai ou faux pour chaque ligne :

$5\ 751 = 5\ 000 + 700 + 50 + 1$	☐ Vrai	☐ Faux
$9\ 054 = 9\ 000 + 500 + 4$	☐ Vrai	☐ Faux
$27\ 303 = 20\ 000 + 7\ 000 + 300 + 3$	☐ Vrai	☐ Faux

Extrait de CEDRE 2008

Figure 1 : tâche de traduction

Dans cet item, les trois décompositions des nombres sont des décompositions additives réduites (lorsque les égalités sont vraies). Les trois questions de cet item relèvent d'un type de tâche de traduction canonique et par conséquent peuvent être résolues avec une technique qui ne met pas en jeu de technologies reposant sur l'aspect décimal du nombre. Il est pertinent dans les successions de Vrai-Faux, d'avoir des questions similaires qui conduisent à évaluer le même savoir et qui mobilisent la même technique et les mêmes éléments technologiques, afin que le regroupement des questions pour déterminer un seuil ait un sens. Il n'aurait pas été pertinent de proposer, dans ce même item, une égalité comme « $5\ 751 = 5\ 200 + 500 + 51$ » qui ne peut pas être validée avec cette même technique. En revanche, si à l'intérieur d'un même item il est pertinent de ne pas proposer des tâches qui relèvent de types de tâches différents, il est complémentaire d'avoir des items similaires formulés dans un autre registre d'écriture (EUN ou EPD⁸) et/ou de façon non canonique pour repérer les compétences d'un élève à reconnaître différentes décompositions d'un nombre.

Nous avons aussi repéré des items moins usuels pour lesquels nous avons dû formuler un objectif d'évaluation, comme par exemple celui ci-dessous (figure 2) :

Marc effectue des opérations. Pour chaque opération, quel chiffre des unités obtient-il ?

1891×9 455×25 2077×57 400×12 919×6

Item posé sous une forme ouverte contenant les 5 opérations - Extrait de CEDRE 2008

Figure 2 : tâche de calcul réfléchi

Des tâches semblables sont proposées avec l'addition et avec la soustraction ; nous les avons considérées comme relevant du calcul réfléchi (même s'il ne s'agit pas de donner le résultat) puisque la technique qui permet le plus efficacement de donner la réponse consiste à appliquer mentalement la technique de calcul posé pour retrouver que le chiffre des unités du produit est le chiffre du produit des unités des deux nombres. De telles tâches permettent de repérer aussi la maîtrise des répertoires multiplicatif ou soustractif en lien avec la connaissance des unités

⁸ EUN : écriture en unités de numération ; EPD : Écriture selon les puissances de 10.

d'ordre 0 dans un exercice de calcul réfléchi.

Au niveau de l'échelle des scores, nous interrogeons le choix de cet item pour caractériser le groupe 3 : en effet, l'item identique posé avec des soustractions caractérise le groupe 4, alors qu'en calcul posé comme en calcul réfléchi, les soustractions sont mieux réussies que les multiplications. Plus généralement, nous avons pu observer que lorsqu'item ne se révélait pas ou peu pertinent localement d'un point de vue didactique, sa position dans l'échelle n'était pas cohérente avec celle des autres items qui représentent le même type de tâches.

Au niveau global

Après avoir analysé chacun des items du domaine pour les deux évaluations 2008 et 2014, nous observons que différents critères assurant la validité du contenu ne sont pas complètement respectés ; en particulier en 2008 :

- des types de tâches manquants pour la praxéologie de numération et un certain déséquilibre dans la répartition de ceux existants qui remettent en cause la représentativité des tâches sélectionnées : en particulier pas de types de tâches de conversion (Tempier, 2013), pas de types de tâches mettant en jeu l'aspect cardinal du nombre, alors que certains types de tâches sont surreprésentés (comme celui de « compléter une suite numérique de raison 10, 100... en avançant ou en reculant » qui est représenté par 17 tâches sur 44 en numération),
- les items proposés ne permettent pas souvent de voir une progressivité sur les techniques et les technologies,
- les praxéologies sont convoquées principalement à un niveau technique, et rarement à un niveau supérieur,
- les formats de question ne sont pas toujours adaptés à la tâche, en particulier pour le calcul réfléchi. Ce que nous avons montré localement dans le paragraphe précédent.

Évolution entre 2008 et 2014

Si de nouveaux items apparus en 2014 ont permis de combler certaines absences dans les types de tâches, d'autres se sont créés (disparition de types de tâches de comparaison ou de traduction entre écriture en unités de numération et écriture chiffrée) maintenant une répartition déséquilibrée. Nous observons néanmoins que les tâches où il s'agit de « compléter une suite numérique de raison 10, 100... en avançant ou en reculant » ne sont plus autant présentes en 2014 qu'en 2008 (4 sur 21 au lieu de 17 sur 44) amenant un meilleur équilibre, qui mériterait cependant d'évoluer encore pour qu'au moins une tâche relevant de chacun des genres de tâches soit proposée.

Il serait néanmoins hâtif de remettre en cause la validité du contenu des évaluations CEDRE 2008 et 2014 à partir de ces constats ; en effet, nous nous limitons ici à une étude sur les entiers et il serait judicieux et complémentaire de la mettre en perspective d'une étude similaire sur les nombres décimaux ; en effet, si certains types de tâches (de comparaison, de traduction, de calcul...) sont peu représentés avec des nombres entiers, ils peuvent l'être avec des nombres décimaux et donc rééquilibrer le contenu de l'évaluation. Par contre, la méthodologie que nous avons développée est performante pour permettre aux concepteurs de d'équilibrer le contenu dès la conception (si elle est utilisée en amont) puis de l'adapter (si elle est utilisée en aval).

CEDRE 3e – Calcul littéral

Nous opérationnalisons l'analyse dans le domaine du calcul littéral sur certains items illustrant les questions abordées.

Au niveau local

Les items sont-ils pertinents pour évaluer des objectifs d'apprentissage visés au regard des programmes ? Nous l'étudions pour deux types d'items : à énoncé ouvert et de format QCM.

Problème à énoncé ouvert

En 2008, Chantal fête ses 53 ans et sa fille Sophie ses 24 ans.
En quelle année, l'âge de Chantal sera-t-il le double de celui de sa fille Sophie ?

Figure 3

Cet item (figure 3) vise à étudier la compétence d'un élève à mettre en équation un problème du premier degré puis à le résoudre via une équation à une inconnue. Trois techniques possibles de résolution sont envisageables : une technique par essai, en calculant les âges de Chantal et de Sophie au cours des cinq années suivantes ou une technique à partir d'un schéma et d'un calcul arithmétique (technologie arithmétique), et enfin une technique algébrique de mise en équation et de résolution d'équation, l'inconnue x représentant le nombre d'années et l'équation à résoudre étant $53 + x = 2(24 + x)$ (technologie algébrique). L'équation du premier degré ayant des coefficients entiers et une solution entière, la résolution de cet item ne nécessite pas la mobilisation d'une technologie algébrique. Les deux autres démarches sont de fait beaucoup plus faciles.

25 % des élèves n'ont pas répondu et 31 % des réponses sont correctes. Or, le calcul du pourcentage de réussite prend en compte le résultat du problème et non la technique. Les 31 % de réussite recouvrent donc les techniques algébriques et non algébriques, ce qui est en contradiction avec l'objectif d'évaluation visé par l'item. Cet item n'est donc pas adapté pour l'objectif d'évaluation visé et aucun item de CEDRE 2008 ne permet de tester à quel groupe de l'échelle est associée la capacité « mettre en équation ». En 2014, le type de tâche mettre en équation n'est toujours pas évalué, mais des problèmes analogues au problème précédent ont changé d'objectif d'évaluation. Ils visent à évaluer la capacité des élèves à mobiliser une technique pour résoudre ce type de problème du premier degré « arithmétique » qui peut être résolu par plusieurs techniques.

Item de format QCM

Considérons les deux items de format QCM de CEDRE 2008 et 2014 (figures 4 et 5) proposés ci-dessous.

<p>Quelle est la forme développée et réduite de $(2x+3)(x-2)$?</p> <p>Cocher la bonne réponse.</p> <p>1 ☐ $2x^2 + 3x - 2$</p> <p>2 ☐ $2x^2 - x - 6$</p> <p>3 ☐ $2x^2 + 3x - 5$</p> <p>4 ☐ $2x^2 - x - 2$</p>	<p>On donne l'expression $A = 1+2x$</p> <p>Cocher la valeur de A pour $x = 7$</p> <table border="1" style="margin-left: auto; margin-right: auto; border-collapse: collapse;"><tr><td style="padding: 2px 10px;">1</td><td style="padding: 2px 10px;">☐</td><td style="padding: 2px 10px;">15</td></tr><tr><td style="padding: 2px 10px;">2</td><td style="padding: 2px 10px;">☐</td><td style="padding: 2px 10px;">21</td></tr><tr><td style="padding: 2px 10px;">3</td><td style="padding: 2px 10px;">☐</td><td style="padding: 2px 10px;">28</td></tr><tr><td style="padding: 2px 10px;">4</td><td style="padding: 2px 10px;">☐</td><td style="padding: 2px 10px;">37</td></tr></table>	1	☐	15	2	☐	21	3	☐	28	4	☐	37
1	☐	15											
2	☐	21											
3	☐	28											
4	☐	37											

Taux de réussite : 73,5%

Figure 4 : Développer une expression

Taux de réussite : 84,7%

Figure 5 : Calculer une expression

Comment ont été choisis les distracteurs ? Quel est leur impact possible sur le processus de réponse de l'élève ?

Le premier item est un QCM à choix multiple portant sur le développement d'une expression

algébrique via la distributivité de la multiplication par rapport à l'addition ou la soustraction. 73% des réponses proposées sont correctes. Les distracteurs sont des polynômes du second degré choisis parmi un panel de polynômes du second degré de formes proches, correct ou erronés, après l'usage de cette propriété. Une analyse didactique met en évidence d'autres distracteurs, des polynômes correspondant à plusieurs techniques erronées, les technologies impliquées et les contrôles mis en jeu laissant vivre différents types d'erreurs récurrentes, par exemple, un polynôme du second degré pour une application erronée de la distributivité, $3x+1$ pour une reconnaissance erronée de la structure de l'expression ou bien, 10×2 pour la forme réduite obtenue suite à une confusion entre processus et structure dans une technologie arithmétique.

Le choix des distracteurs pour le premier item ne permet pas d'interpréter son pourcentage 73,5% de réussite, étant donné que les distracteurs mobilisés ne prennent en compte que des usages erronés de la distributivité de la multiplication par rapport à l'addition et non d'autres démarches envisageables.

L'attribution des distracteurs pour l'item 2 s'appuie sur une analyse didactique : 21 pour l'interprétation incorrecte de la priorité opératoire en jeu dans l'expression numérique $(1+2) \times x$, 28 pour l'interprétation incorrecte de $2x$ comme une écriture chiffrée en base 10, 37 pour la conjonction des deux interprétations incorrectes précédentes.

Au niveau global

Au-delà de la pertinence de chaque item, nous interrogeons la représentativité des items et leur couverture du domaine algébrique. Pour CEDRE 2008, nous constatons l'absence de types de tâches *outil* - « produire une expression littérale », « prouver une propriété », « mettre en équation un problème du premier degré » – et un déséquilibre entre les types de tâches *outil* (6) et ceux de *calcul* (24), à l'avantage de ceux de calcul comme « substituer » ou « tester », « calculer », « reconnaître la structure », « développer » ou « factoriser ». Les items concernant la mise en équation ne sont pas présents dans l'un des six groupes les plus hauts de l'échelle construite dans CEDRE 2008 *a posteriori*.

Évolution entre 2008 et 2014

On constate une évolution entre 2008 et 2014. Les types de tâche « résoudre un système d'équations linéaires à deux inconnues » (groupe 5), « traduire une situation par une expression algébrique » (hors échelle) étaient évalués en 2008 et ne le sont plus en 2014. Mais, des types de tâches pré algébriques, comme « calculer un programme de calcul » (4), ou « remonter un programme de calcul » (4), apparaissent dans CÈDRE 2014, en particulier, pour le premier type afin de remplir l'échelle de compétences pour les groupes de bas niveau. Il en est de même pour « associer une expression générale à un programme de calcul d'un pattern », mais c'est une tâche très complexe, lié au décalage du numéro d'indice avec le numéro de l'étape du pattern. Il y a plus de types de tâches *outil* (11) en 2014, leur nombre restant inférieur à ceux de calcul (29), mais leur poids reste identique le nombre total de tâches augmentant.

De plus, la complexité des items est faible, les items visant principalement une application assez rapide de technique, sans laisser aux élèves beaucoup d'initiative. Dans aucun des items, les élèves n'ont à leur charge de mobiliser une lettre pour généraliser, modéliser ou mettre en équation.

D'un point de vue didactique, nous interrogeons donc l'échantillon des items retenus : ils ne sont pas complètement représentatifs du domaine algébrique, ici le domaine algébrique, tant en ce qui concerne les praxéologies d'usage de l'outil algébrique que la complexité des types de tâches.

CONCLUSION ET PERSPECTIVES

Nous poursuivons la recherche engagée dans le cadre du projet ANR Néopraéval (2014-2017) et en particulier, l'analyse de l'évaluation CEDRE. L'articulation entre les preuves de validité psychométrique et didactique a déjà montré son opérationnalité et la complémentarité des deux approches sur les deux domaines.

Comme nous l'avons montré, l'approche didactique fournit en amont de la conception des outils de description et d'analyse praxéologique *a priori* des tâches pour déterminer la pertinence et la représentativité d'une tâche, la couverture du domaine mathématique. Nous avons déjà montré les évolutions entre CEDRE 2008 et 2014. Elle permet aussi d'analyser des items écartés de l'évaluation pour des raisons psychométriques, en particulier les items peu discriminants avec un R-bis inférieur à 0,20, de mettre en relation l'indice de difficulté d'item avec des descripteurs de tâche (congruence ou non des représentations sémiotiques, niveaux de complexité, format de la question, ...) et de questionner des « incohérences » dans le niveau de difficulté d'un item en termes psycho-didactique. Dans le cadre de la comparaison temporelle entre deux évaluations CEDRE, il s'agit d'interpréter le fonctionnement différentiel d'items à partir de l'évolution éventuelle des praxéologies à enseigner.

En retour, l'analyse didactique permet d'interroger des résultats statistiques : quelles sont les origines de pourcentages de réussite faibles sur certains types de tâche ? Des technologies implicites dans les programmes ? Des pratiques d'enseignants à faire évoluer ?

Les évaluations CEDRE, comme les autres évaluations nationales et internationales, évoluent vers des passations sur support numériques, et se développent aussi avec des volets adaptatifs (par exemple pour le calcul mental). L'approche épistémologique et didactique donne aussi des outils pour développer des évaluations sur support numérique, comme l'a déjà montré la recherche autour de logiciel d'évaluation diagnostique *Pépîte* (Grugeon-Allys, Pilet, Chenevotot et Delozanne, 2012). Le fondement sur un modèle multidimensionnel de la compétence algébrique permet un codage qualitatif s'appuyant sur différentes dimensions liées à la caractérisation de la praxéologie épistémologique de référence en praxéologies locales agrégées. Cette approche permettrait d'interpréter les groupes des échelles de performance (géographie de la classe à partir de profils d'élèves) et de proposer des pistes de régulation adaptées aux besoins d'apprentissage repérés des élèves (Grugeon-Allys, Pilet, Chenevotot et Delozanne, 2012). Nous poursuivons ainsi les recherches en ce qui concerne les évaluations adaptatives.

Pour conclure, nous cherchons à déterminer des conditions pour "renforcer" les preuves de validité des dispositifs d'évaluation externe : une méthodologie de conception basée sur une analyse multidimensionnelle et la complémentarité des approches didactique et psychométrique, au cours des différentes étapes de la conception, en est une, tant en ce qui concerne le contenu (validité épistémo-didactique) que les processus de réponses (validité psycho-didactique). Nous visons ainsi un enrichissement des évaluations au service des apprentissages et de la formation. Cette méthodologie permet un double éclairage sur les programmes et les pratiques enseignantes, d'une part, pour comprendre les résultats des évaluations et leur évolution et, d'autre part, pour questionner leur impact sur les apprentissages à partir de l'exploitation des résultats des évaluations. Pour ceci, nous continuons à mettre en perspective les praxéologies à enseigner, enseignées et apprises, l'échantillon des types de tâches de l'évaluation via la caractérisation d'une praxéologie épistémologique de référence relativement aux domaines mathématiques en jeu.

REFERENCES BIBLIOGRAPHIQUES

- BODIN A. (1997) L'évaluation du savoir mathématique. *Recherches en didactique des mathématiques* 17(1) 49–96.
- BODIN A. (2006). Les mathématiques face aux évaluations nationales et internationales. *Repères IREM*, 65, 55-89.
- BOSCH M., GASCON J. (2005) La praxéologie comme unité d'analyse des processus didactiques. Dans Mercier A. & Margolinas C. (Dir) *Balises pour la didactique des mathématiques* (197 – 122). Grenoble : La Pensée Sauvage.
- CASTELA, C. (2008). Travailler avec, travailler sur la notion de praxéologie mathématique pour décrire les besoins d'apprentissage ignorés par les institutions d'apprentissage. *Recherches en Didactique des Mathématiques* 28(2) 135–182.
- CHAMBRIS C. (2008) Relations entre les grandeurs et les nombres dans les mathématiques de l'école primaire. Évolution de l'enseignement au cours du 20^e siècle. Connaissances des élèves actuels. Thèse de doctorat. Université Paris Diderot - Paris 7.
- CHEVALLARD Y. (1985) Le passage de l'arithmétique à l'algèbre dans l'enseignement des mathématiques au collège - Première partie. L'évolution de la transposition didactique *Petit x* 5 51–94.
- CHEVALLARD Y. (1989) Le passage de l'arithmétique à l'algèbre dans l'enseignement des mathématiques au collège - Deuxième partie. Perspectives curriculaires : la notion de modélisation. *Petit x* 19, 43–75.
- CHEVALLARD Y. (1999) L'analyse des pratiques enseignantes en théorie anthropologique du didactique. *Recherches en didactique des mathématiques* 19(2) 221 – 266.
- CHEVALLARD Y. (2007) Une épreuve expérimentale de mathématiques ? http://yves.chevallard.free.fr/spip/spip/IMG/pdf/Une_epreuve_experimentale_de_mathematiques.pdf
- DE KETELE, J.-M. & GERARD, F.-M. (2005). La validation des épreuves d'évaluation selon l'approche par les compétences, *Mesure et évaluation en éducation*. 28.3, 1-26.
- DUVAL R. (1996) Quel cognitif retenir en didactique des mathématiques ? *Recherches en didactique des mathématiques* 16(3) 349–380.
- FUSON, K. C., WEARNE, D., HIEBERT, J. C., MURRAY, H. G., HUMAN, P. G., OLIVIER, A. I., FENNEMA, E. (1997). Children's Conceptual Structures for Multidigit Numbers and Methods of Multidigit Addition and Subtraction. *Journal for Research in Mathematics Education*, 28(2), 130-162.
- GREGOIRE J., LAVEAULT D. (2014) *Introduction aux théories des tests en sciences humaines*. De Boeck : Bruxelles.
- GRUGEON B. (1997) Conception et exploitation d'une structure d'analyse multidimensionnelle en algèbre élémentaire. *Recherches en didactique des mathématiques* 17 (2) 167–210.
- GRUGEON-ALLYS B., PILET J., CHENEVOTOT-QUENTIN F., DELOZANNE E. (2012) Diagnostic et parcours différenciés d'enseignement en algèbre élémentaire. In Coulange L., Drouhard J.P., Dorier J.L. & Robert A. (Eds.), *Recherche en Didactique des Mathématiques* Enseignement de l'algèbre élémentaire, Bilan et perspectives Hors-série 137–162. Grenoble : La pensée sauvage.
- LESCURE S., PASTOR J-M. (2012) *Mathématiques en fin d'école primaire. Le bilan des compétences*. Paris : Scéren.
- MOUNIER E. (2010) Une analyse de l'enseignement de la numération au CP. Vers de nouvelles

pistes. Thèse de doctorat. Université Paris Diderot - Paris 7.

PILET J. (2012) Parcours d'enseignement différencié en algèbre élémentaire. Thèse de doctorat. Université Paris Diderot-Paris 7.

REY O., FEYFANT A. (2014) Évaluer pour (mieux) faire apprendre. Dossier de veille de l'Ifé, 94.

ROCHER T. (2015) Mesure des compétences. Méthodes psychométriques utilisées dans le cadre des évaluations des élèves. *Éducation et Formations* 86/87 37–60.

TEMPIER F. (2013) La numération décimale de position à l'école primaire. Une ingénierie didactique pour le développement d'une ressource. Thèse de doctorat. Université Paris Diderot-Paris 7.

VANTOUROUT M., GOASDOUE R. (2014) Approches et validités psycho-didactique des évaluations. *Éducation et Formation* N°e-302. <http://ute3.umh.ac.be/revues/>