Mesure et évaluation en éducation

Approches psychométrique et didactique de la validité d'une évaluation externe en mathématiques : quelles complémentarités et quelles divergences ?

Nadine Grapin and Brigitte Grugeon-Allys

Volume 41, Number 2, 2018

URI: https://id.erudit.org/iderudit/1059172ar
DOI: https://doi.org/10.7202/1059172ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (print)
2368-2000 (digital)

Explore this journal

Cite this article

Grapin, N. & Grugeon-Allys, B. (2018). Approches psychométrique et didactique de la validité d'une évaluation externe en mathématiques : quelles complémentarités et quelles divergences ? *Mesure et évaluation en éducation*, *41*(2), 37–66. https://doi.org/10.7202/1059172ar

Article abstract

Even if many researches deal with validity, especially in psychometrics, this notion hasn't been significantly discussed in the field of didactics of mathematics. This paper examines how didactics tools, especially a priori analysis of tasks, can be used to bring validity evidence in addition to those provided by psychometrics. A first part of the paper aims to expose a methodology, using a combination of didactic and psychometric approaches, to analyze and design external assessment tests. In a second part, we implement this methodology to study two national large scale assessments, in two mathematical domains at two levels of school (arithmetic at the end of primary school and algebra at the end of grade 9), in France.

Approches psychométrique et didactique de la validité d'une évaluation externe en mathématiques : quelles complémentarités et quelles divergences ?

Nadine Grapin
Brigitte Grugeon-Allys
Université Paris-Est Créteil, Laboratoire de didactique André Revuz

Mots clés : évaluation externe, validité, didactique des mathématiques, psychométrie, comparaison d'approches, arithmétique, algèbre

Si la notion de validité fait l'objet de multiples recherches, notamment dans le champ de la psychométrie, elle n'a pas encore été beaucoup étudiée dans le cadre de la didactique des mathématiques. Nous proposons d'expliciter la façon dont les outils de ce champ de recherche, en particulier l'analyse a priori des tâches, peuvent être exploités pour apporter des preuves de validité complémentaires à celles de la psychométrie. Une première partie vise à proposer une méthodologie d'analyse et de conception du contenu d'une évaluation externe dans laquelle l'approche didactique de la validité est pensée en complémentarité de celle psychométrique. Nous illustrons, dans une deuxième partie, la mise en œuvre de cette méthodologie sur l'étude de deux domaines mathématiques (l'arithmétique en fin d'école et l'algèbre en fin de collège) dans des évaluations menées nationalement en France à ces deux ordres d'enseignement.

Key words: external assessment, validity, didactics of mathematics, psychometrics, comparison of approaches, arithmetic, algebra

Even if many researches deal with validity, especially in psychometrics, this notion hasn't been significantly discussed in the field of didactics of mathematics. This paper examines how didactics tools, especially a priori analysis of tasks, can be used to bring validity evidence in addition to those provided by psychometrics. A first part of the paper aims to expose a methodology, using a combination of didactic and psychometric approaches, to analyze and design external assessment tests. In a second part, we implement this methodology to study two national large scale assessments, in two mathematical domains at two levels of school (arithmetic at the end of primary school and algebra at the end of grade 9), in France.

Palavras-chave: avaliação externa, validade, didática das matemáticas, psicometria, comparação de abordagens, aritmética, álgebra

Embora o conceito de validade seja objeto de muitas investigações, particularmente no campo da psicometria, não foi ainda muito estudado no quadro da didática das matemáticas. Ora, o nosso propósito é explicar a maneira como as ferramentas deste campo de investigação, em particular a análise prévia de tarefas, podem ser exploradas para fornecer provas de validade complementares às da psicometria. A primeira parte visa propor uma metodologia de análise e conceção do conteúdo de uma avaliação externa na qual a abordagem didática da validade é pensada em complementaridade à da psicometria. Numa segunda parte, mostramos a implementação desta metodologia no estudo de dois domínios matemáticos (aritmética no final do 2.º ciclo e de álgebra no final do 3.º ciclo do ensino básico) nas avaliações realizadas a nível nacional em França nestes dois níveis de ensino.

Note des auteures : La correspondance liée à cet article peut être envoyée à [nadine.grapin @u-pec.fr] et [brigitte.grugeon-allys@u-pec.fr].

Cette recherche a reçu le soutien de l'Agence nationale de la recherche (ANR) dans le cadre du projet NéoPRAEVAL. Le travail présenté dans cet article émane de celui des participants à la tâche 1 de ce projet. En plus des auteures ont collaboré à ces recherches Rémi Goasdoué et Marc Vantourout, chercheurs en sciences de l'éducation à l'Université Paris-Descartes.

Les données statistiques et les items illustrant cet article ont été communiqués par le bureau B2 de la Direction de l'évaluation, de la prospective et de la performance (DEPP) du ministère de l'Éducation nationale. Les auteures remercient en particulier Thierry Rocher, Philippe Arzoumanian et Jean-Marc Pastor pour leur disponibilité et leur aide.

Introduction

Les évaluations externes standardisées à grande échelle telles que le PISA, TIMSS[1] ou, nationalement en France, le Cycle des évaluations disciplinaires réalisées sur échantillon (CEDRE) exploitent des modèles statistiques (p. ex., modèles de réponse à l'item à un ou deux paramètres) qui leur garantissent une qualité certaine.

Si, pour toutes ces évaluations, ces modèles statistiques sont clairement définis et présentés publiquement, tout comme le cadre de l'évaluation, le processus pour concevoir et sélectionner les items est moins connu (Bodin, De Hosson, Decamp, Grapin et Vrignaud, 2016). Les différents domaines mathématiques et cognitifs sont précisément décrits et illustrés par des « items libérés » : ils permettent de préciser les objectifs d'évaluation relativement au cadre de l'évaluation et de montrer la répartition de l'ensemble des items selon ces différents domaines, mais sans donner accès aux énoncés de l'ensemble de ces items ni à la méthode pour les sélectionner avant la passation. Par exemple, en France, pour la réalisation des épreuves standardisées, les concepteurs ne disposent pas d'outils d'analyse des tâches mathématiques proposées.

Or, comme toute autre évaluation standardisée, ces évaluations se doivent d'être valides, c'est-à-dire qu'elles doivent évaluer ce qu'elles prétendent évaluer, et uniquement cela. Si le concept de validité a beaucoup évolué, Laveault et Grégoire (2014) soulignent cependant qu'en psychométrie « il reste le concept le plus fondamental et le plus important » (p. 163) et que différentes preuves de validité peuvent être apportées. Le premier enjeu de cet article est de montrer en quoi une analyse didactique du contenu du test et des processus de réponse des élèves peut se révéler complémentaire aux preuves de validité apportées par la psychométrie. Le second enjeu est de définir un cadre d'analyse et de conception des évaluations externes standardisées en mathématiques permettant de concevoir les questions du test et de les analyser, mais aussi d'interpréter les résultats. Comme l'analyse du contenu est considérée à différents niveaux de granularité (à une échelle locale, tâche par tâche, mais aussi globale sur l'ensemble des tâches du test), les éléments théoriques sous-tendant ce cadre sont issus de différentes théories de la didactique.

Une première partie fera donc état des éléments théoriques issus de la psychométrie et de la didactique des mathématiques fondant le cadre d'analyse et de conception développé. Comme notre travail se situe dans le champ de la didactique et que la prise en compte des questions relatives à la validité des évaluations dans ce champ de recherche est nouvelle, les développements consacrés à cette approche seront plus explicités. Nous présenterons sous l'angle de résultats, dans une deuxième partie, la méthodologie pour analyser ou concevoir une évaluation et un exemple de sa mise en œuvre sur deux évaluations standardisées passées en France. Nous montrerons ainsi, dans deux domaines mathématiques (les nombres entiers en fin d'école et l'algèbre en fin de collège), la façon dont des approches didactique et psychométrique peuvent s'articuler localement sur certains items et globalement sur l'ensemble du test pour apporter des preuves de validité à l'évaluation.

Cadre d'analyse et de conception : éléments théoriques

La notion de validité a fait et fait encore l'objet de nombreuses définitions et de débats (Goldstein, 2015) dans les différents champs scientifiques qui s'intéressent à l'évaluation. Si de multiples types de preuves de validité peuvent être apportés, Laveault et Grégoire (2014) retiennent entre autres des preuves basées sur le contenu du test, sur les processus de réponse et sur la structure interne du test. Ainsi, les modèles psychométriques apportent davantage des preuves quant à la structure interne du test, alors que l'approche didactique que nous développons concerne le contenu du test et les processus de réponse mis en jeu par les élèves. Quelles hypothèses fondent chacune de ces deux approches ? Quels indicateurs peuvent alors être pris en compte pour déterminer la validité de l'évaluation ? Enfin, comment penser ces deux approches en complémentarité l'une de l'autre, pour permettre une méthode mixte – à la fois quantitative et qualitative – recourant de façon conjointe et articulée à des analyses psychométrique et didactique ?

Approche psychométrique de la validité

Il ne s'agit pas pour nous de présenter les différentes méthodes psychométriques utilisées dans les évaluations à grande échelle, mais plutôt de revenir sur les caractéristiques principales de deux modèles : celles de l'approche classique et celles des modèles de réponse à l'item (MRI).

Dans l'approche classique, les items sont surtout caractérisés par deux indicateurs : leur difficulté et leur indice de discrimination (Vrignaud, 2006). Le premier est estimé (dans le cas d'un item dichotomique) par la proportion d'élèves ayant donné une réponse correcte ; le second renseigne sur la qualité de l'item à différencier les élèves selon leur score global, en particulier ceux qui ont un score élevé de ceux qui ont un score faible (Laveault et Grégoire, 2014). Les évaluations du CEDRE, que nous considérons dans cet article, visent à évaluer un niveau de performance et nécessitent donc que les items soient les plus discriminants possible ; l'indice de discrimination retenu est le Rbis (coefficient point-bisérial) (Rocher, 2015).

L'utilisation des modèles de réponse à l'item repose sur une hypothèse fondamentale, celle d'unidimensionnalité. Si Laveault et Grégoire (2014) considèrent ce critère comme primordial, Rocher explique que :

> « L'unidimensionnalité stricte n'existe probablement pas. Les processus mis en œuvre pour réussir un ensemble d'items sont complexes et varient selon les élèves et les contextes. Dès lors, il est difficilement concevable que ces processus se réduisent rigoureusement à une seule et même dimension (Goldstein, 1980). C'est pourquoi, en pratique, évaluer l'unidimensionnalité revient en fait à évaluer l'existence d'une dimension dominante (Blais et Laurier, 1997) » (Rocher, 2015, p. 52).

En particulier, la résolution d'une tâche mathématique met bien souvent en jeu des compétences et des connaissances qui relèvent de dimensions différentes, particulièrement selon le domaine duquel relève la tâche (nombres, géométrie, algèbre, statistiques, etc.). En proposant une approche didactique de la validité, nous visons à spécifier les savoirs mathématiques en jeu dans chacune des tâches et, par conséquent, pour un domaine mathématique donné, à décrire les compétences évaluées et à assurer que la dimension principale évaluée est bien celle voulue.

Approche didactique de la validité d'une évaluation

Dans ce cadre de la théorie anthropologique du didactique (TAD ; Chevallard, 1999), chaque élève est assujetti aux institutions dans lesquelles il apprend ou a appris, les savoirs étant relatifs aux institutions. Évaluer les connaissances de l'élève amène à prendre en compte les effets du contexte institutionnel, en particulier le processus de transposition didactique à travers les programmes scolaires et leur construction (savoir à enseigner), le savoir enseigné par l'enseignant, mais aussi ce qui se noue entre l'élève et le savoir dans la relation didactique, rapport personnel de

l'élève au savoir confronté au rapport institutionnel (Maury et Caillot, 2003). Évaluer les connaissances revient donc à évaluer le rapport personnel de l'élève au savoir dans un domaine mathématique donné, c'est-à-dire à étudier l'adéquation du rapport personnel au rapport institutionnel au savoir. C'est à travers l'activité mathématique mise en jeu qu'il est possible d'étudier le rapport personnel de l'élève au savoir. Or, une connaissance n'existe pas de façon isolée. Aussi, l'activité mathématique est modélisée selon une praxéologie mathématique (Chevallard, 1999), c'est-à-dire des types de tâches et de techniques les résolvant (savoir-faire). Une technique est justifiée par un discours rationnel appelé technologie, lui-même découlant d'une théorie (savoir). Plusieurs techniques peuvent permettre de résoudre une tâche de type donné. Dans une institution, seules quelques techniques sont reconnues institutionnellement selon l'année d'études. Étudier l'adéquation du rapport personnel au rapport institutionnel au savoir revient donc à évaluer les praxéologies développées par l'élève au regard des praxéologies à enseigner et enseignées. Ces dernières sont elles-mêmes codéterminées par une hiérarchie de niveaux institutionnels qui se conditionnent et se contraignent successivement, par le biais du système d'enseignement, lui-même assujetti au sein du système international par les évaluations internationales à grande échelle, par la société et par la civilisation (Artigue et Winslow, 2010). Cette modélisation permet de dégager, à travers les techniques correctes ou incorrectes mobilisées par l'élève, le bloc technologico-théorique (savoir) que les élèves mobilisent de façon prégnante. Dans ce cas, si l'on s'intéresse à l'activité d'un élève comme praxéologie mathématique, des preuves didactiques de la validité d'une évaluation peuvent être apportées « du côté du savoir ».

L'élève est aussi un sujet cognitif spécifique, développant une certaine activité lors d'une tâche d'évaluation. Ainsi, dans une approche ergonomique, les notions de tâche et d'activité au sens de Leplat et Hoc (1983) sont distinctes : la tâche est ce qui est à faire et l'activité, ce qui est mis en jeu par l'élève pour faire la tâche. Dans ce cas, si l'on s'intéresse à l'activité de l'élève par le biais des processus de réponse, d'autres types de preuves, « du côté de l'élève », peuvent être apportés.

En didactique des mathématiques, l'analyse *a priori* des tâches, permettant d'anticiper l'activité mathématique des élèves au regard de celle visée, est fréquemment exploitée pour analyser les situations et pour déterminer leurs potentialités d'apprentissage. Pour étudier le contenu d'un test

et les processus de réponse des élèves au regard d'un objectif d'évaluation, nous faisons l'hypothèse, comme Vantourout et Goasdoué (2014), que l'analyse *a priori* est aussi un outil adapté. Nous allons montrer ses potentialités pour définir des preuves didactiques de la validité d'une évaluation à la fois « du côté du savoir » et « du côté de l'élève ».

Preuves didactiques de la validité du côté du savoir

À partir de l'approche anthropologique, les tâches de l'évaluation sont considérées comme relevant d'un échantillon de l'ensemble des types de tâches constitutives des praxéologies mathématiques évaluées (Chevallard, 2007). Afin de déterminer leur représentativité dans un domaine mathématique donné, il est nécessaire au préalable de définir un référent, indépendant du système scolaire ou des programmes en vigueur, « tout en se situant dans son champ d'action […] qui permette une double analyse, à la fois du côté élève et du côté institutionnel » (Grugeon, 1997). La définition d'une praxéologie épistémologique de référence relative à un domaine mathématique obtenue à la suite d'une étude épistémologique du domaine mathématique (Bosch et Gascon, 2005) fonde la définition d'un tel référent. Elle conduit par la suite à décrire et à mettre en relation les praxéologies à enseigner, enseignées et développées par les élèves au cours du processus de transposition didactique.

Par conséquent, pour une tâche d'évaluation donnée relevant d'un certain type de tâche, il est possible de décrire a priori les techniques possibles ainsi que les éléments technologico-théoriques qui justifient ces techniques et de sélectionner les tâches qui nécessitent leur usage pour les résoudre, en fonction du choix des valeurs des variables didactiques les caractérisant. Nous distinguons ainsi une procédure ou démarche de résolution d'une tâche d'une technique qui se réfère à la résolution d'une tâche d'un type donné, avec catégorisation de techniques relevant de discours technologiques et de théories distincts. Une technique ne peut réussir que sur une partie des tâches de type donné correspondant à la portée de la technique.

Lors de l'interprétation des résultats, l'appui sur la praxéologie épistémologique de référence permet de les mettre en regard de ceux effectivement mis en œuvre par les élèves et de ceux attendus à une année d'études donnée. Si l'on étudie les réponses des élèves selon les techniques, la justification de ces techniques (technologie et théorie) conduit à identifier

les praxéologies mises en jeu par l'élève. L'interprétation des résultats demande ensuite de mettre en perspective ces praxéologies au regard de celles à enseigner et enseignées.

À partir de ces éléments théoriques, nous pouvons désormais décrire un premier type de preuve didactique de la validité, à l'échelle locale, se situant du côté du savoir mathématique. L'analyse praxéologique *a priori* de chaque tâche, telle que nous l'avons évoquée dans le cadre de la TAD, permet d'étudier sa pertinence et sa représentativité au regard de l'objectif d'évaluation qui lui est assigné, en particulier à travers la technique et les éléments technologico-théoriques visés.

L'analyse transversale de l'ensemble des tâches du test permet d'étudier la couverture du domaine mathématique selon les types de tâches définis dans la praxéologie épistémologique de référence et figurant dans le champ d'action des programmes. La couverture du domaine correspond à un deuxième type de preuve didactique de la validité, à l'échelle globale.

Un autre type de preuve de validité concerne la variété de la complexité des tâches sur l'ensemble du test. La complexité d'une tâche dépend du nombre de types de tâches convoqués pour la résoudre et de la variété des types de convocation mis en jeu, par la tâche elle-même ou à la charge de l'élève (Castela, 2008). Par exemple, les trois tâches suivantes « Factoriser l'expression $x^2 - 3^2$ », « Factoriser l'expression $4x^2 - 16y^2$ » et « a et b sont deux nombres tels que $a + b = 5$ et que $a - b = 3$. Quelle est la valeur de $a^2 - b^2$? » n'ont pas la même complexité. La première est une tâche d'application directe. La deuxième met en jeu la réécriture de l'expression $4x^2 - 16\,y^2$ comme $(2x)^2 - (4y)^2$, donc l'élève a à sa charge de *réécrire* l'expression. La troisième convoque la factorisation de $a^2 - b^2$, puis demande de réaliser le calcul du produit demandé ; cette tâche peu habituelle est la plus complexe.

Preuves didactiques de la validité du côté de l'élève

Il s'agit ici de nous centrer sur l'interaction entre l'élève et la tâche en étudiant l'écart éventuel entre la tâche prescrite par le concepteur de l'évaluation et l'activité effective de l'élève, un écart faible ou nul constituant une preuve didactique de la validité.

Dans ce cadre, l'analyse *a priori* des tâches permet de prendre en compte et d'interroger par exemple l'impact du format des questions sur le processus de réponse des élèves au regard du savoir évalué, en particulier celui

des distracteurs sélectionnés comme réponse (Maury, 1985 ; Sayac et Grapin, 2014). Ainsi, dans le cas d'un questionnaire à choix multiple (QCM), la présence de la bonne réponse peut apporter à l'élève des éléments de contrôle quant à la résolution demandée ou induire d'autres processus de réponse que ceux visés (par écartement des distracteurs). Par ailleurs, le format QCM peut aussi impliquer un changement de type de tâche puisqu'une tâche proposée sous un format ouvert peut se transformer en tâche de reconnaissance dans un format QCM.

D'autres éléments sont aussi à considérer lors de cette analyse : la nature du contexte de la tâche, l'impact du support (papier-crayon ou numérique), le niveau de langue, etc. Autant de variables didactiques qui seront prises en compte lors de l'analyse des tâches, mais qui demandent, pour en connaître l'impact sur l'activité effective de l'élève, des observations « cliniques » d'élèves en situation de résolution de la tâche (Vantourout et Goasdoué, 2014).

Critères d'analyse a priori des tâches pour une étude de la validité didactique de l'évaluation

Du côté du savoir, l'analyse praxéologique *a priori* des tâches est réalisée en appui sur la praxéologie épistémologique de référence du domaine mathématique évalué. Néanmoins, lorsque le test vise à évaluer des savoirs au regard des programmes scolaires, ces derniers sont aussi pris en compte et permettent de renseigner les praxéologies à enseigner. Chaque tâche est donc décrite par les éléments caractéristiques suivants :

- le type de tâche et la/les technique(s) attendue(s) en lien avec les éléments technologico-théoriques qui les sous-tendent ;
- les objets mathématiques en jeu, leur ancienneté et leur nombre;
- les variables didactiques associées aux objets mathématiques (ou à la tâche) en jeu et leurs valeurs, compte tenu de l'année d'études et des objectifs d'évaluation ;
- les registres de représentation sémiotiques en jeu, en entrée et en sortie, leur éventuelle congruence (Duval, 1996) ;
- le nombre de types de tâches convoqués dans la résolution et le type de convocation par la tâche ou par l'élève (Castela, 2008) définissant du côté épistémologique la complexité de la tâche.

En ce qui concerne les preuves didactiques se situant du côté de l'élève, nous sélectionnons des critères d'analyse nécessaires à l'étude du format, du contexte et du support de chaque tâche. Pour cela, compte tenu des techniques et des technologies évaluées pour chaque tâche, nous retenons les éléments caractéristiques suivants :

- la nature du format des questions (QCM, question ouverte) ;
- la nature du contexte dans le cas d'une situation extra-mathématique ;
- la nature de l'environnement du test (papier-crayon, environnement informatique, oral avec temps limité) ;
- la nature des instruments autorisés ;
- le niveau de langue de l'énoncé.

Une tâche représentative de l'objectif d'évaluation doit permettre d'associer à une réponse juste le savoir attendu par l'élève, à une année d'études donnée. La prise en compte des praxéologies permet de distinguer des élèves qui peuvent résoudre une tâche avec une technologie ancienne de ceux qui utilisent la technologie visée, ce qui permet de ne pas situer ces élèves dans le même groupe de performance.

Deux organisations mathématiques de référence

Du côté du savoir, l'analyse *a priori* est menée relativement à une praxéologie de référence. Pour notre étude, nous avons défini deux praxéologies selon deux domaines : l'arithmétique en fin d'école et l'algèbre en fin de collège[2].

Le domaine de l'arithmétique est structuré en trois praxéologies (Grapin, 2015) portant respectivement sur la résolution de problèmes arithmétiques, sur la numération et sur le calcul (posé et mental réfléchi). L'ensemble des types de tâches est défini en référence à différents travaux en didactique des mathématiques (p. ex., Collet, 2003 ; Chambris, 2008 ; Mounier, 2010 et Tempier, 2013). Nous n'abordons pas dans cet article la partie relative à la résolution de problèmes. La numération décimale est donc décrite par des types de tâches relevant : de transformations d'écriture d'un système de représentation à un autre en tenant compte du caractère canonique ou non de l'expression, de l'aspect ordinal du nombre (tel que la comparaison) et de l'aspect cardinal (tel que le dénombrement).

Pour le calcul sont distingués des types de tâches qui relèvent du calcul posé et d'autres, du calcul mental réfléchi. Puisque les propriétés additive, multiplicative, décimale et positionnelle de la numération écrite chiffrée justifient les techniques de calcul posé, les praxéologies relatives à la numération interviennent ainsi dans la justification de certaines techniques de calcul.

Nous caractérisons une praxéologie mathématique épistémologique de référence du domaine algébrique à partir d'une synthèse des travaux en didactique de l'algèbre (Chevallard, 1985, 1989 ; Drouhard, 1992 ; Grugeon, 1997 ; Kieran, 2007 ; Pilet, 2012). 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*, *substituer*, *développer*, *factoriser* pour les expressions algébriques ; *modéliser*, *traduire*, *reconnaître* pour les formules ; et *mettre en équation*, *traduire*, *reconnaître, tester*, *résoudre une équation* pour les équations. Ces praxéologies ponctuelles se regroupent en praxéologies locales, en praxéologie *modélisation* et en praxéologie *de calcul*. La praxéologie de référence du domaine algébrique caractérise les propriétés idoines pour un calcul « raisonné et contrôlé » (prise en compte des aspects procédural et structural des expressions algébriques, des équations, 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 sémiotiques du domaine et leur mise en relation.

Nous exploitons chacune de ces praxéologies de référence dans deux analyses *a priori* de tâches d'évaluation extraites du CEDRE en fin d'école et en fin de collège.

1re tâche

Elle a pour objectif d'évaluer la maîtrise de la numération écrite chiffrée et interroge les élèves sur le passage d'une écriture en unités de numération non canonique en une écriture en chiffres (voir Figure 1).

Le nombre composé de 1 centaine, 3 dizaines et 14 unités s'écrit :

- ☐ 10 031 014
- ☐ 1314
- ☐ 144
- ☐ 18

Figure 1. Exemple de tâche de numération extraite du CEDRE 2014, fin d'école

Cette tâche permet d'évaluer la maîtrise de la valeur positionnelle de la numération décimale avec une certaine complexité puisque c'est à l'élève de convoquer la praxéologie *convertir* pour écrire 14 unités sous la forme 1 dizaine et 4 unités. Au-delà de la réponse correcte (144), les distracteurs sont construits pour permettre d'interpréter les erreurs qui ont conduit à leur choix. Ainsi, le choix de réponse 10 031 014 est une transcription de l'écriture en unités de numération (Chambris, 2008) en chiffres (1 centaine s'écrit 100, 3 dizaines s'écrit 3 10 et 14 unités 14). Le choix de réponse 1314 correspond à la juxtaposition des chiffres correspondant aux différentes positions, plutôt qu'à la nécessité de convertir 14 unités en 1 dizaine et 4 unités. Le dernier choix de réponse (18) correspond à la somme des nombres engagés dans l'énoncé (1 + 3 + 14), privilégiant ainsi la propriété additive de la numération. Les élèves choisissant ce distracteur n'ont donc pas compris les propriétés à la fois additive, multiplicative, décimale et positionnelle de l'écriture chiffrée.

2e tâche

Elle a pour objectif d'évaluer la maîtrise de la substitution d'une lettre par un nombre (voir Figure 2).

On donne l'expression $A = 1 + 2x$
Cocher la valeur de A pour $x = 7$

- ☐ 15
- ☐ 21
- ☐ 28
- ☐ 37

Figure 2. Exemple de tâche d'algèbre extraite du CEDRE 2014, fin de collège

Une analyse *a priori* du QCM dans le domaine algébrique montre que la tâche n'est pas complexe et met en évidence que l'attribution des distracteurs permet l'analyse didactique des erreurs au regard des propriétés mises en jeu. Le choix de réponse 15 correspond à la bonne réponse privilégiant la priorité opératoire. Le choix de réponse 21 correspond à une interprétation séquentielle de l'expression numérique et du calcul présenté (1+ 2) ´ 7. Le choix de réponse 28 correspond à une interprétation qui juxtapose les valeurs 2 et 7 dans l'expression $2x$ (27) et au calcul de 1 + 27. Le choix de réponse 37 correspond à la conjonction des deux interprétations précédentes (une séquence suivie d'une juxtaposition), 1+ 2 représentant le nombre de dizaines.

Dans ces deux tâches, et plus généralement pour les QCM présents dans les évaluations du CEDRE, les distracteurs sont construits à partir de types d'erreurs établis en didactique des mathématiques. Les distracteurs peuvent émaner aussi d'observations cliniques d'élèves en train de résoudre la tâche, mais, dans ce cas, ils ne conduisent pas toujours à une interprétation aussi fiable en matière de techniques erronées et de technologies incomplètes. De telles analyses étant menées sur toutes les tâches du test, elles permettent de dégager localement des preuves de validité du « côté du savoir » et sont exploitées ensuite transversalement pour une étude globale de la couverture du domaine et de la variété de la complexité des tâches.

Complémentarité des approches psychométrique et didactique

L'étude didactique, a contrario de l'analyse psychométrique, ne repose pas sur des résultats obtenus et permet d'apporter a priori des preuves de validité du côté du savoir et de l'élève. De plus, elle s'appuie sur la définition d'une référence épistémologique, indépendante des systèmes éducatifs concernés et des programmes d'enseignement, tout en se situant dans leur champ d'action tant du point de vue institutionnel que cognitif. La définition de cette référence permet donc, théoriquement, de réaliser des comparaisons *a priori* sur le contenu des tests et de formuler des hypothèses pour interpréter les résultats des élèves. Une telle étude permet de caractériser les groupes de l'échelle par des tâches relevant des praxéologies visées.

Dans l'approche psychométrique, les items sont décrits par leurs caractéristiques statistiques. Les indices de difficulté ou de discrimination, par exemple, sont associés à chacun des items. Plus globalement, des indices permettant de calculer la fidélité du test (α de Cronbach pour mesurer la consistance interne) sont eux aussi calculés a posteriori à partir des résultats des élèves obtenus à la suite de la passation du test.

Dans chacune des deux approches, des descripteurs locaux sur chacun des items sont donc apportés en complément de descripteurs globaux. Nous pensons à la sélection des items pour la passation finale et à l'interprétation des résultats en prenant en compte les deux types de descripteurs. Ainsi, il ne serait guère judicieux de retenir un grand nombre de tâches « similaires » (même type de tâche, même technique, même complexité), aussi valides soient-elles d'un point de vue statistique, si des tâches représentant un type de tâche donné sont absentes de l'évaluation. Inversement, des items peuvent se montrer peu discriminants, tout en étant didactiquement valides.

C'est bien sous l'angle de la complémentarité de ces deux approches didactique et psychométrique que nous décrirons le cadre d'analyse et de conception des évaluations standardisées. Au-delà de la sélection des items, cette complémentarité amène le didacticien à étudier les caractéristiques statistiques de l'épreuve. Ainsi, localement, pour un item donné, sa difficulté, son caractère discriminant ou un fonctionnement différentiel (s'il y a lieu) peuvent être interprétés à partir d'études produites sur l'apprentissage et sur l'enseignement de contenus mathématiques précis (Roditi et Chesné, 2012). Globalement, la construction des échelles de scores (comme pour le PISA, TIMSS et le CEDRE) est réalisée à partir de modèles de réponse à l'item. Or, l'analyse *a priori* permet de qualifier les tâches réussies pour chaque groupe à partir de ses caractéristiques didactiques, des types de tâches et des éléments technologiques développés. L'interprétation d'une échelle de scores, soit à partir d'une analyse par groupe de l'échelle, soit à partir des éléments descripteurs des tâches (Grugeon-Allys et Grapin, 2015), permet alors d'interroger les pratiques d'enseignement et, selon la nature des évaluations externes standardisées à grande échelle, le contenu des programmes mis en œuvre, en particulier les types de tâches manquants ou ceux surreprésentés. Cela explique ainsi l'intérêt de penser ces deux approches en complémentarité.

Résultats

De la complémentarité entre ces approches résulte tout d'abord une méthodologie d'analyse et de conception des évaluations standardisées, que nous présentons comme premier résultat. Nous la mettons par la suite en œuvre pour analyser la validité de deux évaluations standardisées.

Conception d'une évaluation externe : une méthodologie d'expertise

La méthodologie retenue se décompose en deux étapes principales (séparées par la passation de l'évaluation) et se situe à des échelles locale (tâche par tâche ou item par item) et globale sur l'ensemble du test. Nous précisons ici que, dans les évaluations standardisées, le terme « item » est fréquemment employé pour désigner un élément de questionnaire. Toutefois, dans notre analyse, nous distinguons un item (en lien avec l'approche psychométrique) d'une tâche, celle-ci relevant d'un certain type de tâche mis en jeu dans un item donné.

1re étape : avant la passation

Localement

- Du côté du savoir : étude du contenu du test pour déterminer a priori dans quelle mesure les tâches sont représentatives du domaine évalué et mettent en jeu les savoirs, les techniques et les éléments technologico-théoriques à évaluer. Cette étude est menée à partir de la praxéologie de référence du domaine mathématique évalué ;
- Du côté de l'élève : analyse des processus de réponse pour décider a priori si les tâches proposées permettent d'évaluer ce qui est visé, d'engager des processus de réponse effectifs en adéquation avec les tâches prescrites et de recueillir les techniques et technologies sous-jacentes d'élève.

Globalement

- Du côté du savoir : étude de l'ensemble des tâches d'évaluation pour vérifier a priori s'il permet de recouvrir le domaine des savoirs évalués au regard de la praxéologie de référence et d'avoir une variété de la complexité des tâches.

2e étape : après la passation, localement et globalement

- Mise en relation entre les analyses *a priori* et les résultats des élèves ;
- Interprétation et mise en relation entre les résultats des élèves, les programmes (s'il y a lieu) et les pratiques d'enseignement ;
- Exploitation des résultats, observation de leurs conséquences sur le système d'enseignement (en particulier à partir des échelles des scores calculés statistiquement après la passation) et pistes pour le pilotage au regard des différents niveaux de codétermination institutionnels.

La méthodologie que nous développons vise ainsi à rendre complémentaires les preuves de validité apportées par la psychométrie et celles émanant d'un point de vue didactique sur le contenu. Si chacune des approches (psychométrique ou didactique) présente des spécificités, une première étape d'analyse, se basant uniquement sur l'analyse *a priori*, concerne la conception et la sélection des tâches, avant la passation du test.

Celle-ci permet de déterminer les caractéristiques statistiques des items. Comme nous l'avons signalé précédemment, elle est elle-même déclinée en plusieurs stades et conduit non seulement à la sélection finale des items retenus pour l'évaluation, mais aussi, *in fine*, à la production des résultats au test et en particulier, pour les évaluations considérées dans cet article, à une échelle de scores.

Une seconde étape consiste à interpréter les résultats. C'est à cette étape que l'articulation entre les deux approches prend son sens. L'analyse didactique menée lors de la conception des tâches peut alors être exploitée pour interpréter les résultats obtenus, qu'ils soient globaux, mais aussi plus locaux.

L'analyse *a priori* de chacune des tâches joue donc un rôle central dans la méthodologie, aussi bien pour concevoir le test que pour en interpréter les résultats. De plus, à partir des résultats produits statistiquement, l'approche didactique permettra, au-delà de la description des connaissances des élèves, de formuler des hypothèses relatives à l'enseignement (besoins d'apprentissage ignorés dans les programmes, pratiques enseignantes à faire évoluer) et d'expliciter certains des résultats obtenus. Ce sont autant de pistes en direction des politiques de pilotage.

Étude de la validité du CEDRE : le cas de l'arithmétique en fin d'école et de l'algèbre en fin de collège dans les évaluations du CEDRE

Le Cycle des évaluations disciplinaires réalisées sur échantillon (CEDRE) a pour objectif, en France, d'évaluer tous les six ans les connaissances et les compétences des élèves de fin d'école et de fin de collège, dans une discipline donnée, relativement aux programmes scolaires en vigueur. Quelle que soit la discipline évaluée, la méthodologie de conception de l'évaluation est similaire et conduit, par l'utilisation d'un MRI, à la définition d'une échelle de scores répartissant, de façon arbitraire, les élèves en six groupes : un premier seuil est défini pour que 15 % des élèves soient en dessous (groupes 0 et 1) et un second, pour que 10 % des élèves soient au-dessus (groupe 5). Trois groupes intermédiaires sont ensuite définis par partage d'égale amplitude entre ces deux scores seuils. Les connaissances et compétences associées à chacun de ces groupes sont ensuite décrites par les items correspondants (Rocher, 2015).

Pour les mathématiques, ces évaluations ont eu lieu en 2008 et 2014, et leurs résultats ont fait l'objet de différentes publications (pour le CEDRE 2014 : Dalibard et Arzoumanian, 2015 ; Dalibard et Pastor, 2015). Le cadre de l'évaluation est défini à partir des programmes scolaires. Les items du test, conçus essentiellement par des enseignants et des formateurs, sont répartis selon les différents domaines identifiés dans les programmes.

Les praxéologies de référence sur deux domaines ayant été présentées dans la partie théorique, nous allons illustrer la façon dont nous avons mis en œuvre la méthodologie d'analyse et de conception, en décrivant les deux grandes étapes de cette dernière : avant la passation et après.

Conception du test, avant la passation

Niveau local d'analyse : étude de l'adéquation entre la tâche proposée et l'objectif d'évaluation par une analyse a priori

Étant donné une année d'études et un objectif d'évaluation, l'analyse *a priori* d'une tâche permet de lister les savoirs nécessaires pour la résoudre ainsi que les techniques possibles de résolution et les éléments technologiques les sous-tendant, en distinguant celles qui sont attendues à cette année d'études de celles qui relèvent d'une année d'études inférieure.

Dans la partie relative aux éléments théoriques, nous avons réalisé l'analyse *a priori* de deux tâches en lien avec leur objectif d'évaluation (voir Figures 1 et 2). Ces deux tâches apparaissaient alors comme étant adéquates avec leur objectif d'évaluation, mais nous montrons maintenant, avec l'analyse des deux tâches suivantes, que ce n'est pas toujours le cas.

Dans le CEDRE 2008 en fin d'école, l'objectif de la tâche « poser et effectuer 3 257 + 6 431 » était d'évaluer le calcul posé d'une somme. Or, elle paraît non adaptée puisque l'élève ne peut pas se tromper sur l'alignement des chiffres, les deux nombres étant de même taille, et puisqu'aucune retenue n'est mise en jeu dans le calcul. Il aurait été préférable de proposer un calcul du type 456 + 7 895, comme cela a été proposé en 2014 pour évaluer, dans la technique opératoire, non seulement la maîtrise de l'aspect positionnel de la numération écrite chiffrée, mais aussi celle de l'aspect décimal.

Dans le cadre de l'algèbre en fin de collège, la tâche « Chantal » (voir Figure 3) vise à étudier la compétence d'un élève à mettre en équation un problème du premier degré, puis à le résoudre par le biais d'une équation du premier degré à une inconnue.

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. Exemple de problème extrait du CEDRE 2008, fin de collège

L'analyse *a priori* permet d'envisager trois techniques possibles de résolution : une technique par essai (en calculant les âges de Chantal et de Sophie au cours des cinq années suivantes), 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 du premier degré (l'inconnue x représentant le nombre d'années et l'équation à résoudre étant $53 + x = 2(24 + x)$ [technologie algébrique]). Or, l'équation du premier degré ayant des coefficients entiers et une solution entière, une étude didactique (Combier, Guillaume et Pressiat, 1995) indique que la résolution de cette tâche ne nécessite pas la mobilisation d'une technologie algébrique. De plus, les deux autres démarches sont de

fait beaucoup plus faciles. Cette tâche n'est donc pas adaptée pour tester cette compétence et n'est pas représentative du type de tâche *mettre en équation*. En effet, la réussite de cet item peut être obtenue par des techniques qui ne relèvent pas de raisonnement algébrique.

Après l'analyse *a priori* de chacune des tâches, nous proposons une synthèse sur l'ensemble du domaine.

Niveau global d'analyse du contenu du test

Pour le domaine de l'arithmétique en fin d'école, Grapin (2015) a observé, au regard de la praxéologie de référence, une redondance de certains types de tâches, alors que d'autres ne sont pas représentés. Ainsi, la plupart des tâches de numération mettant en jeu des transformations d'écriture portent sur le passage du nom du nombre vers l'écriture chiffrée, tandis que très peu d'entre elles, comme celle de la figure 1, mobilisent des décompositions non canoniques nécessitant de maîtriser les propriétés de la numération écrite chiffrée. Plus globalement, nous observons une surreprésentation des tâches de calcul (voir Annexe 1) et, pour le calcul posé, les soustractions et les divisions sont beaucoup plus présentes que les additions et les multiplications.

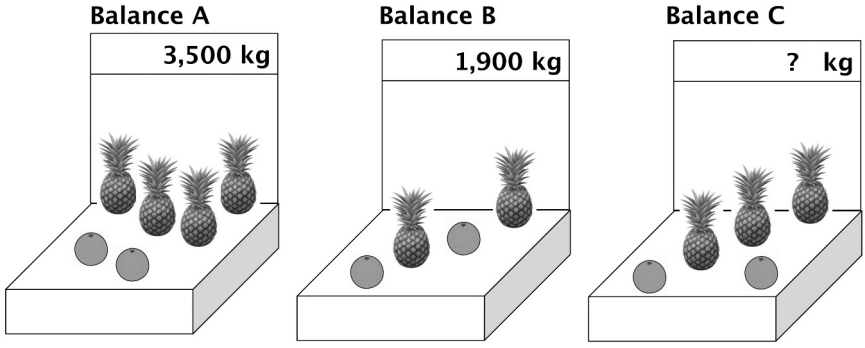
En ce qui concerne le domaine de l'algèbre, nous avons montré, en appui sur la praxéologie épistémologique relative aux objets de l'algèbre, qu'il y avait en 2008 une absence des tâches du côté *outil* de type *généraliser*, *modéliser*, *mettre en équation*, *prouver* (voir Annexe 2). En effet, les tâches de type *produire* indiquent une lettre à utiliser et ne permettent pas de repérer si les élèves mobilisent d'eux-mêmes l'outil algébrique. En 2014, trois tâches de traduction sont proposées, les représentations étant congruentes d'un registre à un autre. De même, les trois tâches relevant du type de tâche *mettre en équation* données en 2014 (voir Tableau 1) peuvent être résolues par des techniques arithmétiques, et non exclusivement par des techniques algébriques fondées sur la modélisation algébrique.

Tableau 1
Trois tâches de mise en équation extraites du CEDRE 2014, fin de collège

Tâche « livres et magazines »	**Tâche « roses et iris »**
Brigitte va dans une librairie. Elle y achète autant de livres que de magazines. Les magazines coûtent 2 € chacun et les livres coûtent 6€ chacun. Elle dépense en tout 40 €. Combien de livres a t-elle achetés ?	Un fleuriste compose des bouquets de roses et d'iris. Toutes les roses sont au même prix. Tous les iris sont au même prix. Un bouquet composé de quatre roses et de quatre iris revient à 34 euros. Un bouquet composé de six roses et de deux iris revient à 38 euros. Quel est le prix d'un bouquet composé de cinq roses et de trois iris.

Tâche « ananas et oranges »

Dans cet exercice, tous les ananas ont la même masse
et toutes les oranges ont la même masse.



Quelle est la masse affichée sur la balance C?

Les tâches de type calcul sont plus nombreuses (80 % en 2008) que les autres types de tâches, ce qui provoque un déséquilibre entre les types de tâches *outil/objet* (calcul littéral). Pour le CEDRE 2014, nous constatons une augmentation du nombre de tâches de type *outil* (35 % par rapport à 20 %).

De façon générale, nous constatons, en 2008, aussi bien pour l'évaluation de fin d'école que pour celle de fin de collège, que les tâches proposées sont de faible complexité (Grugeon-Allys et Grapin, 2015). En effet, pour l'épreuve du CEDRE 2008 en fin de collège, seules 8 tâches sur 30 convoquent, lors de la résolution, plus d'un type de tâche, les autres étant des tâches d'application directe. Deux tâches sur trois nécessitent la reformulation d'un énoncé, les deux registres de représentation sémiotiques

n'étant pas congruents, mais l'une des tâches, « Chantal », ne nécessite pas la mise en équation. Pour l'évaluation de fin d'école, nous constatons cependant une plus grande variété de complexité de tâches en 2014 pour les deux domaines. Toutefois, en algèbre, les items relevant du type de tâche *mettre en équation* ne mettent pas en jeu la reformulation de la première représentation.

Après la passation du test

Niveau local d'analyse

Nous étudions d'abord localement les tâches qui ont des caractéristiques statistiques spécifiques, en particulier celles pour lesquelles nous observons un décalage important entre la complexité (déterminée a priori) et la difficulté (calculée après la passation) ou celles qui présentent un fonctionnement différentiel.

En fin d'école, il apparaît que les tâches de transformation d'écriture mettant en jeu des décompositions non canoniques (comme celle présente dans la figure 1) ne sont réussies qu'à partir du groupe 5 de l'échelle de scores, soit par environ 10 % des élèves, alors que ce sont des tâches qui apparaissent comme peu complexes. Ce décalage important entre la difficulté mesurée à partir des réponses effectives des élèves et la complexité déterminée a priori peut s'expliquer, dans ce cas, par le fait que peu d'exercices de ce type sont proposés aux élèves de l'école élémentaire en France, ces constats étant effectués à partir d'étude de manuels et d'observations de pratiques (Chambris, 2008 ; Tempier, 2013). Plus précisément, différents travaux sur l'apprentissage de la numération, dont ceux de Deblois (1996) et de Collet (2003), montrent que les activités de dénombrement participent à la conceptualisation du nombre ; il serait donc pertinent d'étudier la pratique des enseignants quant à ce type de tâche spécifique.

En fin de collège, les tâches de mise en équation correspondent au groupe 5 de l'échelle de scores, ce qui apparaît en adéquation avec la complexité d'un tel type de tâche. Or, la tâche « Chantal », analysée précédemment, a été réussie, toutes techniques prises en compte, par environ 35 % des élèves, avec 23 % d'élèves n'ayant donné aucune réponse. Il y a donc un décalage entre la complexité étudiée a priori et la difficulté calculée à partir des réponses des élèves. Une hypothèse peut être formulée quant à ce décalage : des pratiques enseignantes visent à développer prioritairement une stratégie de résolution, ici algébrique, même si d'autres stratégies sont envisageables.

Nous pouvons aussi remarquer que les trois items relevant du type de tâche *mettre en équation* (voir Tableau 1) et qui peuvent être résolus soit par des techniques arithmétiques, soit par des techniques algébriques ont des difficultés différentes et qualifient trois groupes différents : G3 pour « livres et magazines » avec 59 % de réussite et 14 % de non-réponse ; G4 pour « ananas et oranges » avec 46 % de réussite et 27 % de non-réponse ; et G5 pour « roses et iris » avec 36 % de réussite et 26 % de non-réponse. Ces items ne sont pas représentatifs d'items relevant du type de tâche *mettre en équation*, dont la résolution nécessite des techniques fondées sur la modélisation algébrique.

Localement, il est aussi possible d'étudier les items présentant des fonctionnements différentiels, en particulier dans la comparaison entre 2008 et 2014. De tels résultats peuvent être la conséquence de changements de programme, mais cela demande une étude que nous ne présentons pas dans cet article.

Niveau global d'analyse

Nous ne revenons pas ici sur les résultats de chacune des deux évaluations ni sur la caractérisation des groupes de l'échelle de scores dans chacun des domaines, mais montrons plutôt la façon dont l'analyse *a priori* du contenu peut permettre de qualifier différemment les résultats obtenus, notamment en recherchant, dans l'échelle de scores, une hiérarchie des techniques ou des technologies.

Pour l'arithmétique en fin d'école, nous observons que les tâches qui mettent uniquement en jeu l'aspect positionnel de la numération sont réussies par environ 84 % des élèves de fin d'école (tâches caractérisant principalement les groupes 2 et inférieurs à 2 de l'échelle de scores), alors que les tâches mettant en jeu l'aspect décimal (comme celle de la figure 1) ne sont réussies qu'à partir du groupe 5. Comme nous l'avons souligné dans l'étude globale du contenu, il figure très peu de tâches de transformation d'écriture mettant en jeu l'aspect décimal de la numération. Par conséquent, il est difficile d'évaluer avec précision la maîtrise de la numération décimale en fin d'école. L'ajout de tâches évaluant cet aspect de la numération et la prise en compte des erreurs réalisées par les élèves permettraient alors de mieux caractériser chacun des groupes. Par exemple, les élèves qui ont répondu 1314 à la question de la figure 1 juxtaposent les chiffres correspondant aux différentes positions, plutôt que de convertir 14 unités en 1 dizaine et 4 unités. Il serait alors intéressant de déterminer le

groupe de l'échelle qui est caractérisé par de telles réponses pour pouvoir ensuite rechercher une certaine hiérarchie liée à la maîtrise des différentes propriétés de la numération écrite chiffrée, dans cette échelle.

En ce qui concerne l'algèbre en fin de collège, la comparaison de l'échelle de scores entre 2008 et 2014 met en évidence des invariants dans la difficulté des types de tâches du groupe Hors échelle (*prouver par contre-exemple*, *traduire une situation par une expression algébrique*, *ordre de grandeur*), travaillés en fin de collège. Toutefois, l'absence de types de tâches *outil* (*mettre en équation*, *produire une expression algébrique*) pour décrire les groupes ne permet pas complètement de caractériser les connaissances et compétences des élèves de ces groupes. Des tâches sur les programmes de calcul (*calculer*, *trouver l'antécédent*, *associer une expression au programme*) ont été rajoutées dans le CEDRE 2014. Leur prise en compte, avec variation des valeurs de variables didactiques (degré et complexité de l'expression algébrique, nature des nombres [variables, images et antécédents]), permet d'affiner la description des groupes (ici, niveau <1, 1 et 2) au regard des connaissances et compétences évaluées. Comme nous l'avons déjà indiqué, la prise en compte des techniques mobilisées pour résoudre un problème du premier degré permettrait de mieux caractériser les différents groupes, la technique de modélisation permettant de traduire une relation verbale entre données par une équation ciblant les groupes de plus haut niveau. Or, pour cela, il faudrait sélectionner une tâche se ramenant à une équation du type $ax + b = cx + d$, les valeurs choisies pour a, b, c et d conduisant à un nombre fractionnaire non décimal comme solution, qui ne peut être résolue que par la technique algébrique de modélisation. La présence de la variable didactique « registres de représentation sémiotiques en jeu, leur éventuelle congruence » permettrait d'affiner la hiérarchie des groupes.

Discussion

Nous avons montré, au fil de notre propos, l'intérêt de considérer d'un point de vue didactique le contenu d'une évaluation pour favoriser la conception d'une évaluation valide, et pour permettre d'interpréter plus précisément les résultats des élèves et de les penser en lien avec les programmes et l'enseignement offert.

Le cadre d'analyse présenté dans cet article vise non seulement à analyser le contenu d'une évaluation, mais aussi à en concevoir. En effet, l'étude *a priori* que nous avons menée localement et globalement peut être développée, comme nous l'avons expliqué, avant la passation du test. La sélection finale des items pour l'interprétation des résultats résulte donc à la fois d'indicateurs didactiques et psychométriques qui se montrent complémentaires, mais qui ne conduisent pas toujours aux mêmes choix. Par exemple, faut-il conserver ou écarter un item qui ne se montre pas suffisamment discriminant alors qu'il est didactiquement pertinent ? Inversement, faut-il conserver un item statistiquement pertinent (p. ex., fortement discriminant) alors qu'il n'est pas didactiquement pertinent ? Ce second dilemme ne devrait pas se poser si l'analyse *a priori* de la tâche a été réalisée au préalable. En revanche, si le contenu du test est didactiquement pertinent (localement et globalement) lors de l'expérimentation, écarter certains items de la passation finale parce qu'ils ne présentent pas les caractéristiques statistiques attendues ne signifie pas pour autant qu'ils ne peuvent donner aucune information sur les connaissances des élèves ou sur les pratiques des enseignants. Ce type de résultat demande une démarche différente et une étude plus importante pour pouvoir être interprété (observation clinique d'élèves, analyse de manuels ou de pratiques enseignantes). Les deux approches se révèlent donc être encore complémentaires dans ce cas.

L'expérience que nous menons actuellement pour la conception du CEDRE 2019 en ayant proposé aux concepteurs un tel cadre montre une évolution dans le contenu des items. Nous faisons l'hypothèse qu'il sera possible de qualifier plus précisément les connaissances et les compétences des élèves *in fine.* Si la définition des praxéologies de référence relève d'un travail spécifique de chercheur en didactique, leur exploitation semble pouvoir être confiée à des concepteurs d'évaluation, à condition qu'ils connaissent suffisamment l'enseignement des mathématiques et qu'ils soient for-

més à l'usage d'un tel type d'analyse. En effet, pour un domaine donné, une fois établie la liste de types de tâches, des techniques et des propriétés les justifiant, il est assez aisé de caractériser une tâche relative à ces éléments et d'en faire une analyse préalable pour repérer différentes techniques et différents éléments technologico-théoriques sous-jacents et erreurs envisageables.

Dans le modèle d'évaluation diagnostique développé par Grugeon (1997) et menant à des parcours d'enseignement différencié (Pilet, 2012), des cohérences de fonctionnement dans les apprentissages sont recherchées sur les technologies sous-tendant les techniques à partir des réponses des élèves. Dans ce modèle, les réponses sont analysées non seulement en termes « correct/incorrect », mais prennent aussi en compte les technologies sous-tendant les techniques mobilisées par les élèves. Le profil de l'élève ou la géographie de la classe (Grugeon-Allys, Pilet, Chenevotot et Delozanne, 2012 ; Grugeon-Allys, 2017) sont alors décrits à partir d'une analyse transversale sur les technologies. Les élèves sont répartis par groupe de profil proche.

Pour une évaluation diagnostique mise en œuvre en classe par l'enseignant, le nombre de tâches se doit d'être limité par contrainte de temps. Il est donc nécessaire de proposer des tâches nécessitant la technique la plus experte pour être réussies, la mise en œuvre des autres techniques étant prise en compte par le codage des réponses. Dans le cas d'une évaluation-bilan telle que le CEDRE, dans laquelle le nombre de tâches peut être important grâce à la passation sur cahiers tournants, ce sont les tâches elles-mêmes qui doivent être conçues pour être résolues selon une hiérarchie des technologies. Les élèves pourront alors être répartis par groupe selon leur réussite à des tâches, réussite liée exclusivement à l'usage des technologies visées par l'institution. Les groupes seront caractérisés par des tâches, de type donné, réussies par plus de 50 % des élèves.

Les critères de choix des tâches sont donc nécessairement distincts, mais visent à caractériser les praxéologies apprises par les élèves et à les mettre en relation avec l'enseignement reçu. Par conséquent, le cadre de conception d'une évaluation externe tel que nous l'avons décrit précédemment se doit d'être pensé aussi selon les objectifs de l'évaluation et selon ses contraintes.

Conclusion

Nous avons montré les apports d'une approche mixte (didactique et psychométrique) pour développer la qualité d'une évaluation et l'exploitation qui pourrait être faite des données à l'échelle institutionnelle. L'analyse didactique des résultats des élèves à certains items amène ainsi à dégager des besoins d'apprentissage ignorés par les programmes (Castela, 2008) ainsi qu'à engager une réflexion sur les contenus de ces derniers et sur les pratiques des enseignants. Il est aussi envisageable d'utiliser les praxéologies de référence, moyennant éventuellement quelques transpositions, pour que les enseignants puissent analyser les tâches qu'ils proposent à leurs élèves au filtre d'une référence, que ce soit en formation initiale ou continue. Il ne s'agit plus, dans ce cadre, de penser uniquement les pratiques d'évaluation, mais aussi les pratiques des enseignants en général.

Enfin, dans le cadre de la théorie anthropologique, Artigue et Winslow (2010) ont conçu un cadre conceptuel afin de comparer les études menées internationalement sur les performances des élèves. L'exploitation des niveaux de codétermination (Chevallard, 1999) permet ainsi de considérer les conditions et les contraintes qui pèsent sur l'enseignement et sur l'apprentissage sur différents plans. Les moyens méthodologiques développés dans ce cadre se révèlent être une perspective de travail dans la suite de cette étude sur la validité.

Réception : 6 janvier 2017

Version finale : 16 février 2018

Acceptation : 2 mai 2018

NOTES

1. PISA : Programme international pour le suivi des acquis des élèves ; TIMSS : *Trends in International Mathematics and Science Study*.
2. La fin d'école en France équivaut à la 5e année du primaire au Québec, tandis que la fin de collège équivaut à la 3e année du secondaire.

RÉFÉRENCES

Artigue, M., & Winslow, C. (2010). International comparative studies on mathematics education: A viewpoint from the anthropological theory of didactics. *Recherches en didactique des mathématiques*, *30*(1), 47-82. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.470.193&rep=rep1&type=pdf

Bodin, A., De Hosson, C., Decamp, N., Grapin, N. et Vrignaud, P. (2016). *Acquis des élèves: comprendre les évaluations internationales PISA et TIMSS* (vol. 1 et 2). Rapport scientifique. Paris : Conseil national d'évaluation du système scolaire.

Bosch, M. et Gascon, J. (2005). La praxéologie comme unité d'analyse des processus didactiques. Dans A. Mercier et C. Margolinas (dir.), *Balises pour la didactique des mathématiques* (pp. 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. Repéré à http://rdm.penseesauvage.com/Travailler-avec-travailler-sur-la.html

Chambris, C. (2008). *Relations entre les grandeurs et les nombres dans les mathématiques de l'école primaire : évolution de l'enseignement au cours du 20e siècle – Connaissances des élèves actuels* (Thèse de doctorat). Université Paris-Diderot, Paris.

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. Repéré à www-irem.ujf-grenoble.fr/revues/revue_x/fic/5/5x3.pdf

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. Repéré à www-irem.ujf-grenoble.fr/revues/revue_x/fic/19/19x5.pdf

Chevallard, Y. (1999). L'analyse des pratiques enseignantes en théorie anthropologique du didactique. *Recherches en didactique des mathématiques*, *19*(2), 221-266. Repéré à http://rdm.penseesauvage.com/L-analyse-des-pratiques.html

Chevallard, Y. (2007). *Une épreuve expérimentale de mathématiques?* Réponse à une question de Michèle Artigue dans le cadre d'un forum du site EducMath. Repéré à http://yves.chevallard.free.fr/spip/spip/IMG/pdf/Une_epreuve_experimentale_de_mathematiques.pdf

Collet, M. (2003). Le développement du système en base 10 chez des élèves de 2ème et 3ème année primaire : une étude exploratoire. *Éducation et francophonie*, *31*(2), 218-241. Repéré à www.acelf.ca/c/revue/pdf/XXXI_2_218.pdf#page=10&zoom=auto,-169,508

Combier, G., Guillaume, J.-C. et Pressiat, A. (1995). *Calcul littéral : savoirs des élèves de collège* (J. Colomb, dir.). France : INRP.

Dalibard, E. et Arzoumanian, P. (2015). CEDRE 2014 – Mathématiques en fin de collège : une augmentation importante du pourcentage d'élèves de faible niveau. *Note d'information*, *19*, Paris : MEN-DEPP.

Dalibard, E. et Pastor, J.-M. (2015). CEDRE 2014 – Mathématiques en fin d'école primaire : les élèves qui arrivent au collège ont des niveaux très hétérogènes. *Note d'information*, *18*, Paris : MEN-DEPP.

Deblois, L. (1996). Une analyse conceptuelle de la numération de position au primaire. *Recherches en didactique des mathématiques*, *16*(1), 71-128. Repéré à http://rdm.penseesauvage.com/Une-analyse-conceptuelle-de-la.html

Drouhard, J.-P. (1992). *Les écritures symboliques de l'algèbre élémentaire* (Thèse de doctorat). Université Paris 7, Paris.

Duval, R. (1996). Quel cognitif retenir en didactique des mathématiques ? *Recherches en didactique des mathématiques*, *16*(3), 349-380. Repéré à http://rdm.penseesauvage.com/Quel-cognitif-retenir-en.html

Goldstein, H. (2015). Validity, science and educational measurement. *Assessment in Education : Principles, Policy & Practice*, *22*(2), 193-201. doi: 10.1080/0969594X.2015.1015402

Grapin, N. (2015). *Étude de la validité de dispositifs d'évaluation et conception d'un modèle d'analyse multidimensionnelle des connaissances numériques des élèves de fin d'école* (Thèse de doctorat). Université Paris-Diderot, Paris.

Grugeon, B. (1997). Conception et exploitation d'une structure d'analyse multidimensionnelle en algèbre élémentaire. *Recherches en didactique des mathématiques*, *17*(2), 167-210. Repéré à http://rdm.penseesauvage.com/Conception-et-exploitation-d-une.html

Grugeon-Allys, B. (2017). Modéliser le profil diagnostique des élèves dans un domaine mathématique et l'exploiter pour gérer l'hétérogénéité des apprentissages en classe : une approche didactique multidimensionnelle. *Évaluer : Journal international de recherche en éducation et formation*, *2*(2), 63-88. Repéré à admee.ulg.ac.be/journal/index.php/ejiref/article/download/104/58

Grugeon-Allys, B. et Grapin, N. (2015). Validité d'une évaluation externe : complémentarité des approches. Dans A.-C. Mathé et É. Mounier (dir.), *Actes du séminaire national de didactique des mathématiques 2015* (pp. 13-26). Paris : IREM Paris 7. Repéré à https://hal.archives-ouvertes.fr/hal-01317134/document

Grugeon-Allys, B., Pilet, J., Chenevotot, F. et Delozanne, E. (2012). Diagnostic et parcours différenciés d'enseignement en algèbre élémentaire. *Recherches en didactique des mathématiques*, Enseignement de l'algèbre, bilan et perspectives (hors série), 137-162. Repéré à http://lutes.upmc.fr/delozanne/Publi/Publi2012/RDM154b-Lingot2012.pdf

Kieran, C. (2007). Learning and teaching algebra at the middle school through college levels. In F. K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 707-762). Charlotte, NC: Information Age.

Laveault, D. et Grégoire, J. (2014). *Introduction aux théories des tests en sciences humaines.* Bruxelles : De Boeck.

Leplat, J. et Hoc, J.-M. (1983). Tâche et activité dans l'analyse psychologique des situations. *Cahiers de psychologie cognitive*, *3*(1), 49-63. Repéré à http://jeanmichel-hoc.free.fr/pdf/LepHoc%201983.pdf

Maury, S. (1985). Influence de la question dans une épreuve relative à la notion d'indépendance. *Educational Studies in Mathematics*, *16*, 283-301. doi: 10.1007/BF00776734

Maury, S. et Caillot, M. (2003) *Rapport au savoir et didactique*. Paris : Faber.

Mounier, É. (2010). *Une analyse de l'enseignement de la numération au CP : vers de nouvelles pistes* (Thèse de doctorat). Université Paris-Diderot, Paris.

Pilet, J. (2012). *Parcours d'enseignement différencié en algèbre élémentaire* (Thèse de doctorat). Université Paris 7, Paris.

Rocher, T. (2015). Mesure des compétences : méthodes psychométriques utilisées dans le cadre des évaluations des élèves. *Éducation et formation*, *86-87*, 37-60. Repéré à http://cache.media.education.gouv.fr/file/revue_86-87/59/3/depp-2015-EF-86-87-mesure-competences-methodes-psychometriques-utilisees_424593.pdf

Roditi, E. et Chesné, J.-F. (2012). Un point de vue didactique sur les questions d'évaluation en éducation. Dans M. Lattuati, J. Penninckx et A. Robert (dir.), *Une caméra au fond de la classe* (pp. 279-292). Besançon : Presses universitaires de Franche-Comté.

Sayac, N. et Grapin, N. (2014). Évaluer les capacités des élèves à résoudre des problèmes dans le cas d'une évaluation externe en France : les spécificités de la forme QCM. *Éducation et francophonie*, *42*(2), 64-83. doi: 10.7202/1027906ar

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.

Vantourout, M. et Goasdoué, R. (2014). Approches et validité psycho-didactiques des évaluations. *Éducation et formation*, e-302. Repéré à https://halshs.archives-ouvertes.fr/halshs-01239551/document

Vrignaud, P. (2006). La mesure de la littéracie dans PISA : la méthodologie est la réponse, mais quelle était la question ? *Revue française de pédagogie*, *157*, 27-41. doi: 10.4000/rfp.409

Annexe 1 : Répartition des items du CEDRE 2014, fin d'école, en arithmétique

Types d'items		N^bre d'items	%	% par domaine
Numération	Gestion de la numération	10	10	21
	Nombre « cardinal »	3	3	
	Nombre « ordinal »	8	8	
Calcul	Calcul posé	22	22	64
	Calcul réfléchi	37	37	
	Savoirs déclaratifs	4	4	
Résolution de problèmes		15	15	15
Total		99	100	100

Annexe 2 : Répartition des items du CEDRE 2008 et 2014, fin de collège, en algèbre

Types de tâches	Année 2008	Année 2014
Généraliser	0	0
Modéliser	0	0
Produire	2 (lettre donnée)	3 (lettre donnée)
Traduire	0	3
Mettre en équation	1 (équation non nécessaire)	3 (équation non nécessaire)[1]
Prouver	0	0
Associer	0	1
Développer	2	0
Factoriser	2	2
Réduire	0	0
Calculer	3	4
Substituer	5	4
Tester	6	0
Résoudre une équation	2 (produit, système)	2 (premier degré, produit)
Reconnaître la structure	4	4
Choisir la forme la plus adaptée	0	0
Conjecturer	0	0
Trouver un contre-exemple	3	2
Total	**30**	**28**

1. Mais résolution de problèmes du premier degré permettant plusieurs techniques, dont une mise en équation non nécessaire.