Pluralités culturelles et universalité des mathématiques :
enjeux et perspectives pour leur enseignement
et leur apprentissage

espace mathématique francophone
Alger : 10-14 Octobre 2015

MÉTHODOLOGIE D'ANALYSE D'ÉVALUATIONS EXTERNES

Brigitte GRUGEON-ALLYS[*] – Nadine GRAPIN[**]

Résumé – En France, des évaluations bilans réalisées sur échantillon sont menées tous les six ans en mathématiques, en fin d'école et de collège, pour déterminer les acquis des élèves. Après avoir présenté ce dispositif spécifique, nous présentons une méthodologie d'analyse des évaluations s'appuyant sur trois approches (épistémo-didactique, psycho-didactique et psychométrique) et nous montrons comment ces dernières se révèlent complémentaires pour garantir la validité des résultats produits. Enfin, nous utilisons cette méthodologie pour étudier les évaluations bilans 2008 et 2014 sur deux domaines : l'arithmétique des nombres entiers en fin d'école et l'algèbre en fin de collège.

Mots-clefs : évaluation standardisée, validité, épistémo-didactique, psycho-didactique, psychométrie.

Abstract – In France, assessments realised on sample are conducted every six years in mathematics, at the end of school and college, to determine students' knowledge. After presenting this specific device, we present an assessment analysis methodology founded on three approaches (epistemo-didactic, psycho-didactic and psychometric) and show how these prove complementary to ensure the validity of the results. Finally, we use this methodology to study the results of assessments 2008 and 2014 on two areas: arithmetic of whole numbers at the the end of school and algebra at the end of college.

Keywords: standardized assessment, validity, epistemo-educational, psycho-educational, psychometrics

En France, en parallèle des enquêtes internationales PISA et TIMSS, des évaluations standardisées sont menées régulièrement nationalement, sur échantillon ou sur la totalité des élèves d'un niveau scolaire donné, par l'intermédiaire de la DEPP (Direction de l'évaluation, de la performance et de la prospective) (Roditi & Chesne 2012). Alors que les résultats des élèves à ces évaluations sont utilisés pour réguler les programmes d'enseignement, voire pour orienter les politiques éducatives (Mons 2009, p.7) la didactique, en particulier la didactique des mathématiques, s'empare davantage des questions d'évaluation telles que – l'évaluation diagnostique et la régulation (Grugeon 1997, Grugeon-Allys 2012, Pilet 2012), l'évaluation de compétences (Winslow 2005 ou Schneider 2006), l'évaluation formative (El Hage & al. 2014), la comparaison des évaluations nationales ou internationales (Artigue & Winslow 2010 ou Ruminot-Vergara 2014), les perspectives à partir de données récoltées sur ces épreuves réalisées à grande échelle (Deblois, Freiman & Rousseau 2007), l'impact des évaluations standardisées sur les pratiques des enseignants (Ruminot-Vergara 2014[340]).

Dans la continuité des travaux de Bodin (1997), nous cherchons à étudier la validité d'un test, ce qui revient à déterminer si un test mesure effectivement ce pour quoi il a été construit et uniquement cela. Cette question apparaît comme essentielle pour les concepteurs de test et

[*] Université Paris Est Créteil – ESPE de Créteil – France – brigitte.grugeon-allys@u-pec.fr
[**] Université Paris Est Créteil – ESPE de Créteil – France – nadine.grapin@u-pec.fr
[340] Étude menée à partir de l'évaluation SIMCE au Chili

Grugeon-Allys B., Grapin N. (2015) Méthodologie d'analyse d'évaluations externes. In Theis L. (Ed.) *Pluralités culturelles et universalité des mathématiques : enjeux et perspectives pour leur enseignement et leur apprentissage* – Actes du colloque EMF2015 – Spé4, pp. 1017-1027.

de nombreux travaux en psychométrie ont été menés autour de ce concept (Grégoire & Laveault 2014 p. 163), mais ici nous nous centrons principalement sur la validité de contenu qui « se réfère à la qualité des questions de l'épreuve » (De Landsheere, 1998) c'est-à-dire la représentativité des questions relativement au domaine mathématique évalué. Nous fondons notre travail sur des approches didactique, prenant en compte des aspects épistémologique et, psychologique, et psychométrique en nous situant à deux niveaux de granularité - global sur l'ensemble des items d'un domaine puis local pour chaque item. Afin d'éclairer pourquoi ces différentes approches sont nécessaires et complémentaires, nous illustrons notre questionnement en présentant d'abord un dispositif d'évaluation nationale, le bilan CEDRE (Cycle des évaluations disciplinaires réalisées sur échantillon) puis nous explicitons la méthodologie d'analyse de la validité que nous avons établie et enfin, nous l'appliquons à deux domaines distincts : celui de l'arithmétique des entiers en fin d'école et celui de l'algèbre en fin du collège.

II. PRESENTATION DU DISPOSITIF D'EVALUATION ET QUESTIONNEMENT

1. Dispositif de conception des épreuves

En France, les acquis des élèves en mathématiques à la fin de l'école primaire (élèves âgés de 10 ans) et du collège (élèves âgés de 15 ans) sont en partie évalués par le bilan CEDRE. Cette évaluation externe est renouvelée tous les 6 ans (2008 - 2014 pour les mathématiques) et permet par conséquent une comparaison temporelle des apprentissages des élèves. Il ne s'agit pas d'établir un diagnostic des difficultés des élèves, ni de comparer des groupes d'élèves selon les établissements, mais d'établir un bilan des connaissances et des compétences des élèves au regard des programmes scolaires en vigueur (Lescure & Pastor 2012).

La conception des items et l'analyse des résultats sont menées par des personnels de l'enseignement qui définissent des items, la forme des questions et le codage des réponses puis les sélectionnent et par des statisticiens qui calculent différents indicateurs psychométriques. Selon les enjeux assignés à l'évaluation, les items du test sont répartis selon les différents domaines définis par les programmes : 385 items répartis sur 6 domaines pour le primaire (organisation des données numériques, résolution de problèmes, nombres et calcul, espace et géométrie, grandeurs et mesures) et 172 items répartis sur 5 domaines pour le collège (géométrie, nombres et calculs, organisation de données & fonctions, grandeurs & mesures). Même s'il a été demandé aux concepteurs de produire des items de difficulté variée (Brun & Huguet 2012), aucun outil d'analyse didactique ne leur a été fourni pour ce faire.

Enfin, trois formes de questions sont proposées : des questions à choix multiples (QCM) avec le plus souvent quatre choix de réponse, des Vrai-Faux et des questions ouvertes. La correction des QCM et des Vrai-Faux est faite automatiquement alors que la plupart des questions ouvertes sont corrigées par des enseignants.

Dans le dispositif CEDRE, et comme pour les évaluations internationales PISA et TIMSS, une pré-expérimentation a lieu l'année précédant l'expérimentation définitive, qui permettra aux concepteurs de choisir les items retenus définitivement pour l'évaluation. Enfin, comme il est impossible que chaque élève passe la totalité du test, un principe de cahiers tournants est mis en place (Brun & Huguet 2008) ; l'étude psychométrique rend possible d'estimer, pour tous les items du test, la probabilité de réussite de chaque élève, même s'il n'a pas été confronté à toutes les questions.

2. *Éléments de méthodologie statistique*

La méthodologie employée pour recueillir les résultats des élèves dans CEDRE est similaire à celle des évaluations internationales, telles que PISA. Sans revenir sur les différents modèles de mesure (théorie classique des scores et modèles de réponse à l'item) détaillés dans différents ouvrages de synthèse, tels que Grégoire et Laveault (2014, pp. 281-304) ou de façon plus synthétique dans Bottani & Vrignaud (2005, pp. 93-105), nous précisons que le modèle statistique utilisé est le modèle de réponse à l'item (MRI). Il permet la réalisation d'échelles de scores et de comparaisons temporelles (Lescure & Pastor 2012, Brun & Huguet 2008), mais repose sur deux contraintes fortes : l'unidimensionnalité du test et l'indépendance locale des items (Bottani & Vrignaud 2005 p. 100).

En psychométrie, un item est principalement caractérisé par son indice de difficulté (calculé *a posteriori* en fonction du nombre d'élèves qui l'ont réussi) et par son indice de discrimination. Les items de l'évaluation CEDRE étant codés 0 ou 1 selon l'échec ou la réussite, l'indice de difficulté correspond à la proportion des élèves qui réussissent l'item et par conséquent est compris entre 0 et 1 (Grégoire & Laveault 2014, p. 204) ; le score d'un élève correspond au nombre d'items réussis par cet élève. La discrimination de l'item est déterminée par la valeur du coefficient bisérial (r_{bis}), défini comme « le coefficient de corrélation linéaire entre le score et une variable latente (ici, la compétence mathématique), régie par une loi normale, conditionnant la réussite à l'item » (Megherbi & al. 2009). Cet indice renseigne sur le « pouvoir discriminant » de l'item, c'est à dire « dans quelle mesure l'item s'inscrit dans la dimension générale (corrélation item-test). Il indique également la différence de performance constatée entre les individus qui réussissent l'item et ceux qui l'échouent. » (ibid, p.72). Dans des évaluations telles que CEDRE, où les résultats conduisent à la construction d'échelles, il est important de différencier entre eux les scores, et par conséquent d'avoir des items qui ont un fort pouvoir de discrimination. Par conséquent, à la suite de la pré-expérimentation, certains items sont écartés en vue de l'évaluation définitive puisqu'ils ne sont pas suffisamment discriminants (par rapport à la compétence mesurée) et n'apportent pas d'informations supplémentaires par rapport aux autres.

3. *Résultats produits*

Notons d'abord que les réponses des élèves ne sont pas analysées pour révéler des difficultés spécifiques (ce n'est pas l'objectif de l'évaluation CEDRE), mais uniquement traitées en tant que réponse correcte ou incorrecte ; à la différence d'évaluation telles que PISA, aucun crédit partiel n'est accordé pour des réponses à des questions ouvertes.

La mesure de la performance des élèves produite à la suite du test par le MRI permet de déterminer les compétences maîtrisées par les élèves ; ce qui se traduit pour le bilan CEDRE, par une échelle de performance qui répartit les élèves en six groupes (de 0 à 5, du plus faible au plus performant). Chacun des groupes est ensuite caractérisé par un niveau de compétence (globalement sur l'ensemble du test, mais aussi par domaine), les élèves d'un groupe maîtrisant aussi les compétences définies dans les groupes de niveaux inférieurs (Brun & Huguet 2008, Lescure & Pastor 2012). A titre d'exemple, nous présentons ci-dessous un premier graphique (Figure 1) qui représente, pour le collège, les taux de réussite par domaine suivant les différents groupes.

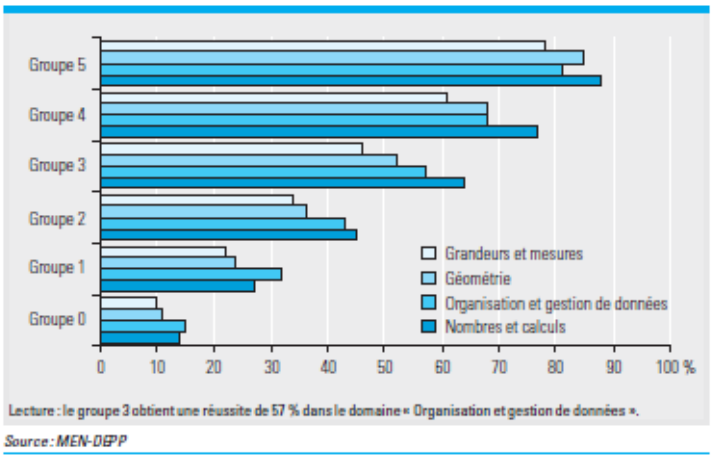

Figure 1 *- Groupes et taux de réussite à chacun des quatre champs (extrait de Brun & Huguet 2008)*

Si la lecture directe de ce graphique permet de comparer selon les domaines, les pourcentages de réussite et de constater par exemple que pour les groupes 3-4-5, la réussite est la plus importante dans le domaine « nombres et calcul » alors que pour les groupes de plus bas niveaux, ils réussissent davantage dans le domaine « organisation et gestion de données », la question fondamentale reste posée : quelles sont les connaissances que possèdent les élèves de ces groupes ? à partir de quels items l'évaluation de ces connaissances a-t-elle été faite ?

4. Questionnement

Comme nous l'avons illustré précédemment à partir de la conception des évaluations CEDRE, la détermination des compétences des élèves dans de telles évaluations repose majoritairement sur une approche psychométrique (Bottani & Vrignaud 2005, p. 101), mais comment garantir que le test est valide ? Comment s'assurer que l'ensemble des tâches permet effectivement d'évaluer les connaissances des élèves au regard des programmes ?

Il s'agit ainsi de rechercher des preuves de la validité du test en interrogeant le contenu (les items sont-ils représentatifs du domaine évalué ? Couvrent-ils le domaine évalué au regard des programmes ?) ou en étudiant les processus de réponse afin de s'assurer que les élèves qui répondent mobilisent bien les connaissances qui sont supposées être évaluées et que chaque item possède les caractéristiques psychométriques demandées. Ces trois conditions correspondent aux trois approches qui fondent notre travail :

1. préalablement à la conception des items
   - approche épistémo-didactique : étude de la représentativité des items et de la couverture du domaine mathématique concerné au filtre de l'épistémologie et de la didactique des mathématiques relativement au domaine (types de tâche proposés, modalités de réponse, codage des réponses de l'élève en fonction du type de raisonnement,...),
   - approche psycho-didactique : analyse des processus de réponses mis en jeu par les élèves pour produire une réponse en lien avec la forme des énoncés et des variables extra-mathématiques (contexte, découpage de l'énoncé, ..),

2. a *posteriori* à partir des réponses des élèves
   - approche psychométrique : calcul des caractéristiques statistiques des items et production de l'échelle des scores et des groupes selon leur niveau de compétence.

La question du choix des tâches dans une évaluation a déjà été soulevée par Bodin (2006), mais peut difficilement être traitée de façon effective à partir de l'analyse d'évaluations externes existantes. En effet, même si les évaluations internationales telles que PISA et TIMSS sont conçues à partir d'un cadre reposant sur des travaux en didactique, et que nous supposons que les groupes d'experts de ces évaluations s'assurent de la représentativité des tâches choisies, les chercheurs n'ont pas accès à l'ensemble des items des évaluations et ne peuvent par conséquent établir une analyse didactique de l'ensemble des tâches proposées. Deux recherches ont néanmoins été menées en ce sens : celle de Roditi & Salles (2015) sur l'évaluation PISA 2012 ou celle de Sayac & Grapin (2014) sur l'analyse du bilan CEDRE 2008 fin d'école à partir de « facteurs de complexité et de compétences ».

Pour notre recherche, nous avons eu l'opportunité d'avoir accès à l'ensemble du dispositif des évaluations CEDRE fin d'école et fin de collège de 2008 et de 2014, et nous présentons dans le paragraphe suivant la méthodologie que nous avons adoptée, puis quelques résultats sur l'analyse de ces évaluations.

III. METHODOLOGIE D'ANALYSE

1. Complémentarité des trois approches : un point de vue global et local sur le test

L'enjeu de CEDRE étant d'évaluer les acquis des élèves au regard des savoirs visés dans les différents domaines mathématiques dans le cadre des programmes scolaires, l'ensemble des items du test doit « couvrir » les différents aspects épistémologiques des savoirs décrits dans les programmes ; par conséquent, un premier niveau d'analyse concerne la pertinence des tâches proposées dans chaque domaine mathématique et leur couverture du domaine (validité épistémo-didactique). Pour ce faire, nous avons situé notre travail dans le cadre de la Théorie Anthropologique du Didactique (Chevallard 1999). A partir des tâches proposées dans les évaluations, il s'agit de caractériser les connaissances apprises des élèves en relation avec les savoirs à enseigner et enseignés mis en jeu dans les différentes étapes de la transposition didactique. En effet, quelle validité accorder aux inférences faites à partir des scores si les tâches présentes dans l'évaluation ne sont pas pertinentes ou « représentatives » de celles relevant du domaine évalué, au regard des programmes et de ce qui est enseigné en classe ?

De ce fait, au delà d'une étude cognitive, nous fondons l'analyse des connaissances apprises par les élèves, dans un domaine donné, sur la caractérisation d'une référence épistémologique (Bosch & Gascon 2005). La définition de cette dernière repose sur une analyse épistémologique des « savoirs savants », relativement au domaine mathématique concerné, ici, respectivement en algèbre en fin de collège (Grugeon 1997) et aux nombres entiers en fin d'école (en particulier Tempier 2013); elle conduit, relativement au domaine étudié, à une description des types de tâches, des techniques permettant de les traiter et des éléments technologiques et théoriques justifiant les techniques. Nous prenons aussi en compte, dans l'analyse des tâches, leur complexité relativement au niveau d'enseignement auquel l'évaluation a lieu ainsi que la congruence sémantique des registres de représentation sémiotique (Duval 1996).

Par conséquent, au regard de la référence, relativement au domaine considéré, nous pouvons analyser la validité d'un test à partir de preuves basées sur le contenu d'une évaluation par une étude « globale » du contenu à savoir : la représentativité des types de tâches et la couverture du domaine selon les tâches proposées dans l'évaluation, la complexité des tâches à partir de l'étude des valeurs des variables didactiques en jeu et du niveau d'intervention du type de tâche et des savoirs en jeu (Castela 2008), les différents registres sémiotiques mis en jeu.

Ce point de vue épistémo-didactique, sur la totalité des items relevant d'un domaine donné, permet d'apporter *a priori* des éléments relatifs à la validité de contenu. Mais au-delà, quelque soit la tâche proposée, il faut s'assurer, en situation, qu'elle permet bien d'évaluer ce qu'elle est sensée évaluer, lorsque les élèves vont la résoudre, et en particulier la nature du raisonnement mobilisé au regard du niveau scolaire visé. Pour ceci, il est nécessaire de s'assurer que l'élève mobilisera bien les savoirs attendus, ce qui dépend aussi d'autres paramètres liés en particulier à l'énoncé.

Développée dans le cadre des approches psycho-didactiques en évaluation, la validité d'un test, qualifiée de « psycho-didactique » Vantourout et Goasdoué (2015) s'appuie sur des preuves basées sur le « processus de réponse de l'évalué, avec le souci de comprendre avant tout son fonctionnement cognitif ». Un deuxième niveau d'analyse se pose donc au niveau local, item par item, et se révèle d'autant plus complémentaire à l'analyse épistémo-didactique développée précédemment, que bon nombre des items sont sous la forme de QCM ou de Vrai-Faux. Les études menées par Sayac et Grapin (2014) sur les stratégies employées par les élèves de fin d'école pour répondre à des questions sous forme de QCM montrent que ce ne sont pas toujours des stratégies de savoir qui sont employées et que les élèves changent de stratégies au cours d'un même test, selon leur niveau de connaissance. On peut alors raisonnablement questionner l'impact de la forme de la question sur les connaissances que mobilisera l'élève, et étudier le décalage entre ce qui est attendu et ce qui est effectif. A ce stade, nous entrevoyons comment une approche psycho-didactique, fondée sur les processus de réponse de l'élève vient en complément d'une analyse épistémo-didactique. Comment l'approche psychométrique s'articule-t-elle par la suite ?

Le troisième niveau d'analyse vise à vérifier les caractéristiques psychométriques des items comme par exemple, leur pouvoir de discrimination, leur indice de difficulté ou encore, dans le cadre de comparaison temporelle entre 2008 et 2014, les items présentant des fonctionnements différentiels (items biaisés). Pour ces items, une analyse didactique relativement aux savoirs en jeu et à leur enseignement ou une observation d'élèves en train de résoudre l'exercice peut permettre d'expliquer leurs caractéristiques statistiques ; et réciproquement, les indicateurs statistiques correspondant à ces items peuvent nous conduire à interroger les pratiques enseignantes (praxéologies enseignées) et les savoirs évalués.

Par ailleurs, en psychométrie, la difficulté d'un item se détermine *a posteriori*, selon la réussite des élèves au test. En didactique des mathématiques, la complexité d'une tâche peut être définie à partir de différents descripteurs (que nous listons dans le paragraphe suivant) lors de l'analyse *a priori* mais aussi en considérant les difficultés des élèves recensées par différentes recherches relativement à l'apprentissage d'un savoir donné. Par conséquent, il est intéressant d'étudier la corrélation entre la difficulté de l'item (*a posteriori*) et ces éléments théoriques, déterminés *a priori* : nous faisons l'hypothèse que ce double éclairage (épistémo-didactique et psychométrique) permettra ainsi de préciser l'état des connaissances des élèves en fin d'école ou en fin de collège, et réciproquement, d'apporter des éléments supplémentaires pour garantir la validité des inférences faites à partir des résultats.

2. *Descripteurs de tâches retenus pour l'analyse*

Présentons désormais plus spécifiquement la méthodologie d'analyse que nous avons retenue.

Une première étape consiste à faire, pour le domaine étudié, une analyse de toutes les tâches de l'évaluation en précisant : les types de tâches et les objets mis en jeu, les valeurs des variables didactiques associées, les techniques possibles et celles mettant en jeu le raisonnement attendu, les registres de représentations sémiotiques en jeu et leur congruence (Duval 1996). Ensuite, comme nous l'avons expliqué précédemment, nous étudions la

couverture du domaine par les tâches proposées relativement à une organisation mathématique de référence : ce qui nous permet de repérer les manques ou les redondances en termes de types de tâches, mais aussi la variété des représentations sémiotiques convoqués dans le dispositif d'évaluation.

Pour la deuxième étape, nous nous centrons plus localement sur chaque item et au delà de l'analyse *a priori*, par des observations d'élèves, étudions le processus de réponse développé par l'élève, notamment en fonction du format de question.

Ces deux analyses devraient être menées avant passation du test et permettre à la fois l'échantillonnage des tâches et le choix du format de question, pour garantir *a minima* que les distracteurs des QCM correspondent à des erreurs d'élèves et que le processus de réponse engagé pour répondre est bien celui attendu.

La troisième étape d'analyse correspond à la mise en perspective des caractéristiques statistiques des items (difficulté de l'item, indice de corrélation) avec les analyses épistémo-didactique et psycho-didactique : y a t-il une cohérence localement, globalement ? Si oui, quel éclairage apporte l'approche didactique ? si non, quelles modifications apporter pour enrichir l'évaluation ? Quelles conclusions apporter sur l'état des connaissances des élèves selon les groupes de l'échelle de performance ? Quelles hypothèses formuler alors sur la raison de ces difficultés en lien avec les programmes d'enseignement existants et les pratiques des enseignants ?

IV. ANALYSE DES EVALUATIONS EXISTANTES

1. Observations générales et évolution 2008-2014

Sur les évaluation CEDRE fin d'école et fin de collège, nous observons que les questions relatives au contenu, que ce soit dans sa globalité ou sur un domaine donné, mais aussi plus localement, item par item (choix des distracteurs pour les QCM ou modalités de correction pour les questions à réponse ouverte) restent insuffisamment prises en compte, notamment parce qu'elles ne font pas l'objet d'une analyse *a priori* (avant expérimentation) ou *a posteriori* (après expérimentation). Parallèlement, les questions ouvertes notamment en résolution de problèmes ont des consignes de correction qui ne permettent pas de prendre en compte différents types d'erreurs, la réponse étant considérée comme correcte ou incorrecte au regard du résultat attendu.

La question du choix des distracteurs dans les QCM est importante puisqu'elle impacte sur la réussite de l'élève. Même s'il est recommandé (Leclercq 2000) de choisir des distracteurs qui correspondent à des erreurs d'élèves correspondants à des conceptions ou des démarches erronées, nous avons pu constater, notamment dans l'évaluation 2008, que ce n'était pas forcément toujours le cas, et que par conséquent ces items ne se révélaient pas toujours valides d'un point de vue épistémo-didactique et psycho-didactique.

Nous présentons ci-après, pour les deux domaines, des résultats synthétiques de l'analyse que nous avons menée selon les deux premières étapes : une analyse épistémo-didactique et psycho-didactique.

2. L'arithmétique des entiers en fin d'école

Nous entendons par « arithmétique des nombres entiers » le domaine couvert par la numération (écriture et lecture des nombres dans différents systèmes sémiotiques), le calcul

posé et mental (addition, soustraction, multiplication et division euclidienne) ainsi que la résolution de problèmes additifs et multiplicatifs.

En ce qui concerne les registres sémiotiques mis en jeu, une évolution est constatée entre 2008 et 2014 : alors qu'ils étaient absents en 2008, des items mobilisant des écritures avec des unités de numération et des représentations symboliques avec du matériel de numération ont été insérés dans l'évaluation 2014. Les autres registres (écriture chiffrée, droite graduée, décomposition additive ou en puissances de dix) figurent parmi les exercices proposés, mais les écritures en lettres (et plus largement la numération parlée) sont très peu présentes. Par ailleurs, la plupart des nombres mobilisés sont de l'ordre du millier ou de la centaine, mais « les grands nombres » sont peu présents (une quinzaine d'items pour chacune des évaluations).

Certains manques constatés dans l'évaluation 2008, en termes de types de tâches ont pu être comblés en 2014, en particulier celles portant sur les conversions de registres (par exemple, le passage d'une écriture en chiffres à une écriture en lettre n'était pas évalué en 2008). Enfin, aussi bien en 2008 qu'en 2014, les quatre opérations (maîtrise des techniques opératoires) ont fait l'objet de nombreuses tâches similaires (ce qui interroge leur redondance et l'équilibre global), mais on observe une répartition équilibrée des types de problèmes proposés (additifs et multiplicatifs).

Dans un deuxième temps, si on s'intéresse plus localement à certains items, on peut constater par exemple un effet du format de la question sur la réussite de l'élève. L'item suivant (Figure 2) a été proposé sous forme ouverte et sous forme de QCM :

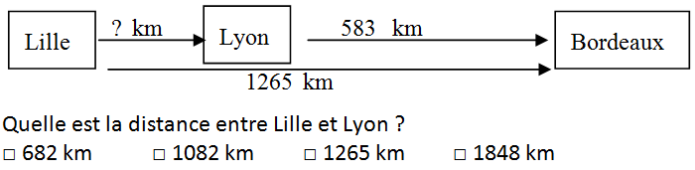

Figure 2 - *Item extrait de CEDRE fin d'école 2008*

Dans le cas du QCM, les choix proposés étaient : 682 km (la bonne réponse), 1082 km (pas d'explication reposant sur un type d'erreur reconnu), 1265 km (recopie du nombre le plus grand) et 1848 km (la somme des deux nombres). Alors que l'item est réussi en ouvert à 62 %, il est réussi en QCM à 73 % ; le constat de cet écart de réussite (11 %) cible bien la question de ce qui est évalué en fonction de la forme de la question et l'impact de cette dernière sur l'activité de l'élève. Dans ce type de cas, une première analyse épistémo-didactique permettra de mettre en avant les procédures de résolution possibles, celles attendues, les erreurs possibles et conduira à trouver des distracteurs « pertinents » pour le QCM ; dans cet item, nous interrogeons justement le choix du distracteur « 1082 » qui aurait pu avantageusement être remplacé par 1322 correspondant à une mauvais maitrise de la technique opératoire de la soustraction (soustraire le plus grand du plus petit systématiquement sur chaque colonne). Par ailleurs, l'écart de réussite avec la question sous forme ouverte peut s'expliquer par l'effet rétroactif du QCM (si l'élève, après calcul, ne trouve pas une des réponses proposées, il peut vérifier ou modifier son raisonnement), mais aussi par la possibilité d'utiliser les choix de réponse et les tester un par un (calculer 682 + 583 et constater que la somme est égale 1265). Mais, c'est en observant l'activité de l'élève lors de la résolution effective du problème que l'on pourra savoir s'il mobilise effectivement les procédures attendues ou s'il met en place des stratégies autres, en particulier dans le cas de QCM.

3. L'algèbre en fin de collège

Nous entendons par algèbre, en fin de collège, le domaine couvert par la résolution de différents types de problèmes intra ou extra mathématiques et dans différents cadres - problèmes de généralisation (expressions littérales), problèmes de modélisation (formules), des problèmes de mise en équation (équations ou inéquations), problèmes de preuve - ainsi que le calcul sur les expressions littérales, les équations, mobilisant des propriétés d'ordre syntaxique et sémantique (propriété de distributivité, conservation de l'égalité, mais aussi équivalence des expressions) et des modes de représentation associés à différents registres sémiotiques de représentation.

Pour la première étape d'analyse, les tâches mises en jeu dans l'évaluation CEDRE en 2008 et en 2014 relèvent bien du domaine algébrique mais ne couvrent tous les types de tâches du domaine.

Dans CEDRE 2008, au-delà des tâches de calcul habituelles bien représentées - substituer, développer ou factoriser une expression algébrique, tester si un nombre est solution d'une équation ou inéquation du premier degré, résoudre une équation produit, un système de deux équations du premier degré à deux inconnues – certaines, plus complexes, permettent d'étudier la capacité des élèves à articuler calcul algébrique et numérique (figure 3). Cette tâche n'a pas été reprise en 2014.

a et b sont deux nombres tels que $a + b = 5$ et que $a - b = 3$. Quelle est la valeur de $a^2 - b^2$?

Figure 3 - *Item extrait CEDRE fin de collège 2008*

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 4 - *Item extrait CEDRE fin de collège 2008*

En 2014, la résolution d'une équation du premier à une inconnue « $8x - 3 = 5x + 30$ », sous forme ouverte, permet de repérer les techniques utilisées par les élèves. Une tâche « développer » de format V/F, repris en 2014, permet d'aborder la problématique de l'équivalence des expressions algébriques. Du côté objet, globalement les items en 2008 sont représentatifs du domaine, l'aspect calcul intelligent étant même présent via l'item de la figure 3 et les exercices V/F.

En 2008, au delà des tâches de production de formules, les problèmes numériques proposés, au vu de leur structure et des valeurs numériques retenues, peuvent être résolus par des démarches numériques et ne nécessitent pas la mise en équation, la résolution algébrique du problème s'avérant en plus difficile (réponse décimale) (cf. Figure 4). On peut donc remettre en cause la validité de cet item, au vu de l'objectif visé « mettre en équation ». Cet item a été repris en 2014, mais l'objectif visé est devenu « résoudre un problème par diverses démarches ». Aucun problème de généralisation, de modélisation fonctionnelle, de mise en équation ou de preuve ne permet d'étudier si les élèves rentreraient d'eux mêmes dans une démarche algébrique pour les résoudre. Des tâches de généralisation ont été ajoutées en 2014, sans prise en charge de la mobilisation des lettres, mais des types de tâches restent absents.

En ce qui concerne la deuxième étape d'analyse, on peut s'interroger sur le codage des réponses, limité à correct / incorrect et le choix des distracteurs pour certains problèmes. Dans le cas des questions ouvertes, ce codage ne permet pas de distinguer les types de technique utilisés par les élèves conduisant à une réponse correcte, ni les types d'erreurs. Ces techniques relèvent-elles de l'algébrique (Cf. figure 4) ? En 2014, davantage d'énoncés à « question ouverte » ont été proposés mais quelle exploitation en sera faite ?

En bilan, l'analyse des items du test, au niveau global (couverture des items relativement au domaine) et local (choix des distracteurs ou modalités de correction pour les réponses ouvertes), nous amène à interroger partiellement la répartition de la population sur l'échelle de performance, réalisée *a posteriori* à partir de la réussite aux exercices.

V. CONCLUSION ET PERSPECTIVES

La méthodologie d'analyse proposée ici et illustrée à partir de quelques exemples montre la richesse et l'intérêt de croiser différents champs de recherche (épistémo-didactique, psycho-didactique et psychométrique). Si une analyse épistémologique et didactique permet de contrôler le contenu de l'évaluation au niveau local (par tâche) et global (sur le recouvrement et la variété), il s'agit aussi de mettre en perspective la complexité de la tâche déterminée par l'analyse *a priori* avec l'indice de difficulté de l'item pour repérer d'éventuelles incohérences. Des questions similaires peuvent aussi se poser sur des items qui pourraient sembler pertinents d'un point de vue didactique, mais qui ne se révèleraient pas discriminants après calcul du r_{bis}. Dans les deux cas, une analyse de la validité psycho-didactique des items, ou la formulation d'hypothèses relatives aux pratiques enseignantes permettraient d'interpréter ces écarts. On comprend ainsi qu'il ne s'agit pas de trois approches successives, mais bien de points de vue complémentaires, parfois sur un même item, qui renseignent différemment sur le contenu des évaluations et sur leurs résultats.

Au-delà de l'analyse des évaluations existantes, les résultats obtenus devraient ainsi contribuer à perfectionner la conception des évaluations externes existantes. Signalons pour conclure que les évaluations CEDRE en mathématiques entre 2008 et 2014 ont vu leurs dispositifs de conception évoluer grâce à deux types de travaux en didactique : l'intégration des exercices extraits du dispositif Pépite (Grugeon 1997) pour évaluer les compétences des élèves en algèbre (fin de collège) et l'exploitation de l'outil d'analyse « facteurs de complexité et de compétence » développé par Sayac & Grapin (2014) pour équilibrer les items de l'évaluation CEDRE fin d'école.

REFERENCES

Artigue M., Winslow C. (2010) International comparative studies on mathematics education: a view from the anthropological theory of didactics. *Recherches en didactique des mathématiques* 30(1/3), 47–82.

Bodin A. (2006) L'évaluation du savoir mathématique. *Recherches en didactique des mathématiques* 17(1/3), 49–96.

Bosch M., Gascon J. (2005) La praxéologie comme unité d'analyse des processus didactiques. In Mercier A., Margolinas C. (Eds.) *Balises pour la didactique des mathématiques* (pp. 197 – 122). Grenoble : La Pensée Sauvage.

Bottani N., Vrignaud P. (2005) La France et les évaluations internationales. Rapport établi à la demande du Haut Conseil de l'Évaluation de l'École, 16. Paris : DEP/Bureau de l'édition.

Brun A., Huguet T. (2008) Les compétences des élèves en mathématiques en fin de collège, *Note d'information*, 10-18. Direction de l'Évaluation de la Prospective et de la Performance.

Castela C. (2008) Travailler avec, travailler sur la notion de praxéologie mathématique pour décrire les besoins d'apprentissage ignorés par les institutions d'enseignement. *Recherches en didactique des mathématiques* 28(2/3), 135–182.

Chevallard Y., Feldmann S. (1986) *Pour une analyse didactique de l'évaluation*. IREM d'Aix-Marseille.

Chevallard Y. (1999) L'analyse des pratiques enseignantes en théorie anthropologique du didactique. *Recherches en didactique des mathématiques* 19 (2/3), 221 – 266.

De Landsheere (1988) *Faire réussir, faire échouer.* Paris : PUF

Duval R. (1996) Quel cognitif retenir en didactique des mathématiques ? *Recherches en didactique des mathématiques* 16(3/3), 349–380.

El Hage S., Le Hebel F., Coppé S., Tiberghien A. (2014) Identifier l'évaluation formative en classe. In *Cultures et politiques de l'évaluation en éducation et en formation*, *Actes du 26ème colloque ADMEE-Europe.* Marrakech.

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

Grugeon B. (1997) Conception et exploitation d'une structure d'analyse multidimensionnelle en algèbre élémentaire. *Recherches en didactique des mathématiques*17 (2/3), 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 (pp. 137-162). Paris : La pensée sauvage.

Leclercq D. (1986) *La conception des questions à choix multiple*. Bruxelles : Labor.

Lescure S. & Pastor J-M. (2012) Mathématiques en fin d'école primaire. Le bilan des compétences. Paris : Scéren.

Megherbi H., Rocher T., Gyselink V., Troseille B., Tardieu H. (2009) Évaluation de la compréhension de l'écrit chez l'adulte. *Économie et statistique*, 424-425, 63-86.

Mons N. (2009) Les effets théoriques et réels de l'évaluation standardisée. Compléments à l'étude Eurydice. Réseau Eurydice.

Pilet J. (2012) *Parcours d'enseignement différencié appuyés sur un diagnostic en algèbre élémentaire à la fin de la scolarité obligatoire : modélisation, implémentation dans une plateforme en ligne et évaluation*. Thèse de doctorat Paris : Université Paris-Diderot.

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

Roditi E., Salles F. (2015) Nouvelles analyses de l'enquête PISA 2012 en mathématiques, un autre regard sur les résultats. *Revue Education et Formations* 86 – 87, 235-255.

Ruminot-Vergara C. (2014) Effets d'un système d'évaluation sur l'enseignement des mathématiques : le cas de SIMCE au Chili. Thèse de doctorat. Paris : Université Paris Diderot.

Sayac N., & Grapin N. (2014) Evaluer les capacités des élèves à résoudre des problèmes dans le cadre d'une évaluation externe en France ; la spécificité des QCM. *Education et francophonie* XLII :2.

Schneider M. (2006) Comment des théories didactiques permettent-elles de penser le transfert en mathématiques ou dans d'autres disciplines ? *Recherches en didactique des mathématiques* 26(1), 9 – 38.

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

Vantourout M., Goasdoué R. (2015) Approches et validité psycho-didactique des évaluations. *Revue Éducation et Formations* e302, 139-156.

Winslow C. (2005) Définir les objectifs de l'enseignement mathématique : la dialectique matières - compétences. *Annales de didactique et de sciences cognitives* 10, 131 – 155.