Approche didactique de l'évaluation et de ses pratiques en mathématiques : enjeux d'apprentissages et de formation.

Nathalie Sayac

UNIVERSITÉ PARIS DIDEROT
Paris7

Note de synthèse pour d'Habilitation à Diriger des Recherches de :

Nathalie Sayac

Approche didactique de l'évaluation et de ses pratiques en mathématiques : enjeux d'apprentissages et de formation

Présentée le 29 Novembre 2017

Jury

Michèle Artigue, Professeure émérite, Université Paris Diderot (France)

Alain Kuzniak, Professeur, Université Paris Diderot (France)

Lucie Mottier Lopez, Professeure, Université de Genève (Suisse)

Éric Roditi, Professeur, Université Paris Descartes (France)

Christine Suurtamm, Professeur, Université d'Ottawa (Canada)

Luc Trouche, Professeur, Institut français de l'Éducation - École Normale Supérieure de Lyon (France)

LDAR
LABORATOIRE DE DIDACTIQUE
ANDRÉ REVUZ

à F.

SOMMAIRE

NOTA BENE
La loi du 4 août 2014 pour l'égalité réelle entre les femmes et les hommes encourage les établissements d'enseignement supérieur à prendre leur place dans la lutte contre les stéréotypes de genre. La langue reflète la société et sa façon de penser le monde. Une langue qui rend les femmes visibles est la marque d'une société où elles peuvent se projeter dans des rôles premiers à égalité des hommes, ce qui est particulièrement important en mathématiques. Soucieuse de m'inscrire dans cette perspective, j'ai rédigé mon dossier d'HDR avec une écriture inclusive.

INTRODUCTION

Faire le bilan de son parcours scientifique à un moment donné de sa carrière permet à la fois de rendre compte à une communauté de chercheur·e·s d'une somme de travaux scientifiques d'un·e de ses membres, mais aussi de lui permettre de relire ses travaux avec un regard distancié des contingences qui les ont produits, tourné vers des perspectives professionnelles et scientifiques nouvelles.

En ce qui me concerne, c'est à la communauté des didacticien·ne·s des mathématiques que j'adresse cette note de synthèse qui témoigne de mon activité scientifique depuis mon inscription en DEA en 1998, soit il y a près de vingt ans. Ce travail singulier m'a permis de réaliser à quel point le travail d'un·e chercheur·e était empreint de convictions personnelles et scientifiques qui l'animent et qui, sans que cela soit toujours explicite, orientent et déterminent ses travaux. Dans la présente note de synthèse, je propose donc de rendre compte des travaux que j'ai réalisés autour de l'évaluation en mathématiques, en précisant les principes qui les ont structurés et en les inscrivant dans une histoire nationale et internationale, à la croisée de plusieurs champs scientifiques.

Un des premiers principes que je revendique est celui d'être une chercheure engagée, animée par des problématiques de formation et de réussite des élèves. J'entends par là que, dans mes travaux ou dans mes engagements, les perspectives de formation des enseignant·e·s et d'amélioration des apprentissages mathématiques des élèves sont toujours présentes et m'amènent à faire des choix spécifiques. Quand j'étudie les pratiques de professeur·e·s ou de formateur·rice·s en mathématiques, c'est avec l'objectif de mieux les connaître pour mieux les comprendre et les faire évoluer. Quand je m'intéresse à l'évaluation, c'est pour permettre de comprendre ce qui se joue du point de vue des apprentissages des élèves en mathématiques et les améliorer. Je préciserai donc, dans la suite de mon exposé, comment ce principe a influencé mes travaux ou quel rôle il a pu jouer dans mes choix scientifiques. La dimension d'engagement est également une dimension forte de mon identité de chercheure. L'écriture épicène adoptée pour la rédaction de cette note de synthèse s'inscrit pleinement dans cette perspective. Cette dimension d'engagement me permet d'éprouver, dans d'autres cadres, mes questionnements et orientations scientifiques en les confrontant à d'autres réalités. Dans le cadre de l'association d'aide aux devoirs "Fibonacci" que j'ai créée il y a six ans, je suis confrontée à une réalité scolaire, aussi bien du point de vue des pratiques enseignantes que de celui des difficultés d'apprentissages des élèves, que ma fonction d'enseignante-chercheure me fait appréhender autrement. Dans le cadre du comité de suivi de la réforme de la formation des enseignant·e·s auquel je participe en tant qu'experte scientifique depuis novembre 2013, je suis confrontée à des visions de la formation et de la recherche qui sont celles d'autres acteurs et actrices de la formation (syndicalistes, cadres de l'éducation, membres de directions administratives des ministères de l'enseignement supérieur et de l'éducation nationale, etc.) ce qui m'oblige à clarifier, à justifier, à défendre mon positionnement scientifique m'amenant ainsi à questionner mes travaux et leurs résultats. Cette confrontation à d'autres cadres

(institutionnels, citoyens) s'inscrit pleinement dans le deuxième principe qui anime mes travaux.

Le deuxième principe concerne mon positionnement de chercheure qui aime éprouver les frontières des autres champs scientifiques pour enrichir mon approche de didacticienne des mathématiques. Dans le cadre de ma thèse, j'ai été amenée à m'initier, très modestement, à la sociologie à travers la lecture des travaux de Weber (1965) notamment sur la notion d'idéal-type et ceux de Huberman (1989) sur les différentes phases de la carrière d'un·e enseignant·e. Ces travaux m'ont ainsi amenée à définir les trois déterminants m'ayant permis, en partie, d'étudier les pratiques des professeur·e·s de mathématiques enseignant en lycée : le sexe, l'âge et le cursus. Le déterminant "sexe" m'a permis d'explorer une dimension qui m'a toujours intéressée pour penser et comprendre les rapports humains dans la société et dans le monde scolaire. Le déterminant "âge" a été retenu parce qu'il permettait de différencier, au moment de ma thèse (2000-2003), les professeur·e·s en fonction de leur expérience professionnelle et de l'influence potentielle de la réforme des "Maths Modernes" sur leurs pratiques. Le déterminant "cursus" a été pris en compte à travers le concours passé par les professeur·e·s, en distinguant Agrégation/Capes, mais aussi les voies interne/externe puisque ces différents concours pouvaient témoigner de différents niveaux d'étude (Maitrise et Licence, voire moins) et de performances (nombre de places au concours plus ou moins limité). Ces trois déterminants m'ont donc permis d'explorer, plus spécifiquement, une des composantes des pratiques enseignantes (la composante personnelle), mais c'est bien dans le cadre d'une approche didactique (celle de la "double approche" d'Aline Robert et Janine Rogalski, 2002) que j'ai réalisé mon étude et produit des résultats que je préciserai par la suite.
Dans le cadre de mes travaux sur l'évaluation, j'ai également été amenée à me confronter à d'autres cadres théoriques (psychométrie, sociologie, édumétrie, psychologie sociale, etc.) qui se sont, depuis longtemps, emparés des problématiques d'évaluation et qui ont produit de nombreux résultats. Il a donc fallu que je me penche sur ces différents travaux pour étudier en quoi ils pouvaient être utiles aux didacticien·ne·s, mais aussi en quoi les travaux en didactique pouvaient également les enrichir et les compléter. Cette confrontation m'a permis de réaliser que l'approche scientifique développée en sciences de l'éducation pour étudier les questions d'évaluation (externe, de classe, de pratiques, de formation) n'est pas suffisante quand des contenus mathématiques sont en jeu, mais que l'approche didactique ne l'est pas non plus quand elle ne prend pas en compte des résultats pourtant cruciaux développés en sciences de l'éducation. Cette confrontation a donc été indispensable à l'élaboration du cadre didactique que j'ai développé pour appréhender les questions d'évaluation en mathématiques et que je présenterai dans la deuxième partie de cette note de synthèse.

Le troisième principe auquel j'ai très tôt adhéré est celui de la nécessaire prise en compte de la composante personnelle des individus, acteurs et actrices de l'enseignement et de la formation en mathématiques, dans leur singularité et dans leur identité professionnelle, pour étudier les pratiques d'enseignement et de formation en mathématiques. Le sujet initial de ma thèse était "l'étude de l'impact d'un stage de formation continue sur les pratiques des professeur·e·s de mathématiques l'ayant suivi". J'avais donc suivi toutes les séances du stage qui ont abouti à l'élaboration d'un problème à proposer à des élèves de 1re S, pendant une séance de 2h. Ce problème élaboré par les professeur·e·s du stage intégrait différents concepts didactiques ayant été travaillés durant les premières séances et a fait l'objet de nombreuses discussions et explicitations entre les participant·e·s. Néanmoins lorsque j'ai commencé à observer les différentes passations de ce problème, je me suis rapidement rendue compte à quel point les différentes mises en œuvre proposées par les professeur·e·s orientaient *in fine* le problème vers des apprentissages différents, malgré le strict respect de son énoncé. Les différences étaient telles que j'ai choisi de réorienter mon sujet de thèse vers celui de l'étude des pratiques

des professeur·e·s de mathématiques de lycée à partir de l'influence de déterminants que j'ai souhaité objectivement définis.
Dans la recherche portant sur les pratiques des formateurs et formatrices en mathématiques réalisée entre 2008 et 2010 (Sayac, 2008, 2010, 2012, 2013), j'ai également pu constater à quel point les différentes biographies personnelles et professionnelles des formateurs et formatrices pouvaient avoir une incidence sur l'offre et les stratégies de formation proposées en mathématiques par les six formateurs et formatrices ayant participé à cette étude. Il s'est en effet avéré que suivant leur parcours (instituteur avec une Licence de mathématiques, professeur certifié, professeure agrégée, docteur et doctorant en didactique des mathématiques, docteure en sciences de l'éducation) et en fonction de leurs connaissances en didactique des mathématiques, de grandes différences ont été relevées tant au niveau des contenus qu'au niveau de la manière dont ils étaient proposés.

Ces principes ont animé mes recherches depuis la première que j'ai réalisée dans le cadre de mon mémoire de DEA jusqu'à celles que je mène actuellement, même si ce n'est qu'aujourd'hui que je peux identifier leur marquage aussi explicitement. Ils permettent également d'éclairer les différents travaux que je propose d'exposer dans la présente note de synthèse.

Les travaux que j'ai choisi d'exposer ne concernent qu'une partie de ceux que j'ai pu mener depuis ma thèse. J'ai en effet préféré consacrer cette note de synthèse à la présentation du cadre didactique de l'évaluation des apprentissages mathématiques pour à la fois préserver une unité thématique à cet écrit et à la fois rendre compte de mes derniers travaux et présenter ceux actuels et à venir.

Dans un premier temps, je présenterai les travaux majeurs qui m'ont permis "d'entrer en évaluation" et qui m'ont influencée pour définir ce cadre didactique qui structure aujourd'hui mes recherches et les formations que je dispense. J'évoquerai les travaux issus de chercheur·e·s francophones (belges, suisses, français·e·s, canadien·ne·s et même luxembourgeois·e·s) qui ont grandement participé au développement de ce champ scientifique, mais aussi bien évidemment des chercheur·e·s anglophones comme Black Wiliam, Schoenfeld ou Stiggins qui, à partir de la fin des années 90, ont été les précurseurs du courant *Assessment for Learning* central pour penser conjointement évaluation et apprentissages des élèves.
Je présenterai ensuite les travaux que j'ai menés sur l'évaluation des apprentissages mathématiques des élèves de fin d'école primaire à partir de l'étude d'un bilan national de fin d'école (2008) et de plusieurs expérimentations (en 2012 et 2013), ainsi que les résultats de la recherche collaborative que j'ai menée entre 2014 et 2016 sur les pratiques évaluatives en mathématiques de vingt-cinq professeur·e·s des écoles.

L'exposé de ces différents travaux (anglophones, francophones et personnels) me permettra de définir et de justifier le cadre didactique de l'évaluation que je présenterai dans une seconde partie. Ce cadre me permet de traiter à la fois la question de l'évaluation des apprentissages mathématiques des élèves et à la fois la question des pratiques évaluatives en mathématiques des professeur·e·s. Il vise à proposer un nouveau cadre de l'évaluation qui prend davantage en compte les apprentissages mathématiques évalués et s'appuie sur des travaux en Didactique des mathématiques, tout en intégrant des résultats produits en Sciences de l'éducation.

En conclusion, j'évoquerai mes recherches en cours sur la formation des professeur·e·s des écoles à l'évaluation des apprentissages mathématiques des élèves dans le cadre du LéA[1] (EvalNumC2) ainsi que les perspectives de recherche que je souhaite développer notamment sur la problématique des inégalités scolaires en lien avec l'évaluation et sur celle de l'articulation entre évaluations externes et évaluations internes.

[1] Les LéA ont été définis dans le programme scientifique de l'IFÉ comme des lieux à enjeux d'éducation, rassemblant un questionnement des acteurs, l'implication d'une équipe de recherche, le soutien du pilotage de l'établissement et la construction conjointe d'un projet dans la durée. EvalNumC2 est l'acronyme de "l'évaluation au service des apprentissages numériques au cycle 2".

PARTIE I : REVUE DE LITTÉRATURE ANGLOPHONE & FRANCOPHONE

J'ai choisi de réaliser cette revue de littérature sur l'évaluation et ses pratiques en distinguant les travaux anglophones des travaux francophones pour deux raisons. La première : ces travaux ont été développés de manière indépendante, par des chercheur·e·s anglophones (Black, Bloom, McMillan, Scriven, Stiggins, Wiliam, etc.) et francophones (Allal, Cardinet, De Ketele, De Landsheere, Mottier Lopez, Perrenoud, Rey, etc.). La seconde : l'évaluation est fortement liée à la culture évaluative d'un pays, les différentes cultures de l'évaluation s'accommodent assez bien de la distinction entre les pays anglophones et francophones. Force est de constater que si les travaux des chercheur·e·s anglophones se réfèrent rarement aux travaux des chercheur·e·s francophones, la réciproque est également vraie, bien qu'elle soit moins marquée. Ce constat que j'ai pu éprouver tout au long de mes lectures, a pour conséquence que les orientations scientifiques portées par les chercheur·e·s de ces deux sphères linguistiques sont différentes, comme je vais tacher de le montrer.

A- REVUE DE LITTÉRATURE ANGLOPHONE

Pour rendre compte des travaux anglophones réalisés autour de l'évaluation des apprentissages mathématiques, j'ai d'abord cherché à répertorier les articles publiés par des chercheur·e·s majeur·e·s du champ de l'évaluation dans différentes revues internationales telles que *Educational Measurement : Assessment in Education : Principles, Policy & Practice*, *Educational Research & Evaluation, Issues and Practice* (NCME), *Educational Measurement, Teaching and teacher Education*, ou des revues spécifiquement mathématiques telles que ESM (*Educational Studies in Mathematics)*, ou IJSME (*International Journal of Sciences and Mathematics Education*) ou encore JMTE (*Journal of Mathematics Teacher Education*). Les auteur·e·s de ces articles travaillent pour la plupart soit dans des universités (James McMillan, Pamela Moss, Laurie Shepard, Denisse Thompson) ou des entreprises privées (Richard Stiggins, Susan Brookhart) américaines, soit dans des universités du Royaume-Uni (Paul Black, Dylan Wiliam, Jeremy Hodgen, Maddalena Taras, Candia Morgan, Anne Watson) ou du Canada anglophone (Christine Suurtamm). Je ferai également référence à des chercheur·e·s d'autres pays ayant publié des articles dans les revues anglophones mentionnées ci-dessus : Mogens Niss (Danemark), Marja van den Heuvel-Panhuizen et Michiel Veldhuis (Pays-Bas), Leonor Santos et Jorge Pinto (Portugal), Guri Nortvedt (Norvège), Ana Remesal (Espagne).
Pour réaliser cette revue de littérature, je n'ai pas cherché l'exhaustivité du champ de l'évaluation car cette quête aurait été aussi vaine qu'impossible tant les travaux qui s'y réfèrent sont nombreux et variés scientifiquement. J'ai choisi de retenir les articles et les ouvrages, fréquemment cités, qui m'ont paru centraux du point de vue de la problématique de l'évaluation en lien avec les apprentissages des élèves et les pratiques des professeur·e·s d'abord d'un point de vue général, puis en mathématiques.
Je n'ai délibérément pas retenu, dans cette présentation, les articles relatifs aux tests à grande échelle (*Large-Scale Assessment*) car, même si je m'y suis référée dans mes premiers travaux sur l'évaluation, ils tiennent aujourd'hui une moindre place dans mes travaux et sont secondaires par rapport au cadre didactique de l'évaluation en mathématiques que je propose dans la présente note de synthèse.

Avant de présenter ces travaux, il me semble important d'évoquer l'existence de groupes institutionnellement structurés et reconnus aux États-Unis et au Canada anglophone (NCTM,

National Council of teachers of Mathematics ou le NAEP, *National Assessment of Educational Progress*) et au Royaume-Uni (AAIA, *Association for Achievement and Improvement through Assessment* et l'ARG, *The Assessment Reform Group*) qui produisent des rapports et ressources ayant une grande influence dans l'approche de l'évaluation et de ses pratiques pour les chercheur·e·s et praticien·ne·s de ces pays.

De même, il me paraît également important de prendre en compte les contextes institutionnels et politiques (réformes et orientations) dans lesquels s'inscrivent les différents travaux référés, car ils ont forcément un impact sur l'approche de l'évaluation des praticien·ne·s et des chercheur·e·s, comme le rappelle Schoenfeld (2007) dans son livre *Assessing Mathematical Proficiency*. Aux Etats-Unis, la loi NCLB (*No Child Left Behind*) votée en 2001 a eu de fortes incidences sur la façon de penser l'évaluation des apprentissages des élèves. La place et le rôle important des *Scholastic Achievement Test* (SAT) et les dérives de *teach to the test* qui peuvent en découler sont également des éléments qu'il faut avoir en tête pour mieux comprendre les travaux des chercheur·e·s américain·ne·s. Au Royaume-Uni, l'influence du groupe ARG de chercheur·e·s en Éducation de plusieurs universités du Royaume-Uni, dont Paul Black et Dylan Wiliam a été déterminante, dès sa création en 1989, et a fortement orienté la politique éducative en matière d'évaluation dans le système scolaire anglais.

Je présenterai cette revue de littérature anglophone en deux parties : une première partie où je rendrai compte de travaux concernant l'évaluation d'un point de vue général (évaluation & apprentissages), puis une seconde partie où les travaux évoqués seront plus directement en lien avec les mathématiques (évaluation & apprentissages mathématiques).

1. Évaluation & Apprentissages

Pour rendre compte des travaux sur l'évaluation produits par des chercheur·e·s anglophones en lien avec les apprentissages des élèves, j'ai choisi de les présenter dans des paragraphes correspondant à la désignation en anglais des principales entrées permettant de traiter la question de l'évaluation d'un point de vue général. Ces paragraphes s'enchaînent dans une logique qui m'a semblée correspondre à l'articulation naturelle des entrées qui les déterminent.

Les premiers travaux visant à considérer l'évaluation dans une perspective de développement des apprentissages datent de Scriven qui a été le premier en 1967 à utiliser le terme *Formative Evaluation* pour qualifier les processus d'évaluation ayant un rôle dans "*the on-going improvement of curriculum*" (p. 41) et à distinguer les évaluations sommatives et des évaluations formatives. Deux ans plus tard, Bloom a repris cette distinction dans son livre intitulé "*Learning for Mastery*", mais en utilisant cette fois le terme *Formative Assessmen*t pour décrire "*any assessment before the big one*" (p. 48). En 1971, Bloom, Hasting et Madaus ont produit un Handbook "*Formative and Summative Evaluation*", dans lequel ils montraient comment l'évaluation formative, intégrée aux processus d'apprentissage pouvaient améliorer les apprentissages des élèves quelle que soit la discipline. Les livres de ces chercheurs ont eu un impact majeur sur tous les travaux sur l'évaluation produits par la suite, aussi bien par des chercheur·e·s anglophones que francophones.

a- Summative Assessment

Les évaluations sommatives sont celles qui, comme leur dénomination l'indique, permettent d'évaluer les élèves à l'issue d'un processus d'enseignement ou de formation. Elles ont fait l'objet de nombreuses recherches spécifiques répertoriées de manière très efficace par Moss

(2013) dans le chapitre du Handbook de McMillan qui lui sont dédiées ("*Research on classroom summative assessment*"). Dans ce chapitre, Moss évoque plusieurs recherches portant sur l'impact des évaluations sommatives sur la motivation des élèves, notamment celles de Brookhart avec Durkin (2003), avec Bronowicz (2006) et avec Walsh & Zientarski (2006), mais celles que j'ai retenues sont les recherches de Wiliam et Taras qui étudient les usages de l'évaluation sommative dans les classes et la façon dont les professeur·e·s les articulent avec les autres types d'évaluation car elles sont plus directement en lien avec l'approche que je souhaite défendre.

Wiliam (2000), dans une contribution au TSG 10 intitulée "*Integrating formative and summative functions of assessment*" présentée du 10ème colloque ICME (Japon), a très justement montré que cette distinction entre évaluation formative et évaluation sommative ne devait pas se faire au détriment des apprentissages des élèves. Il refuse donc l'incompatibilité supposée de ces deux types d'évaluation et propose de la dépasser en adoptant une vision de l'évaluation qui s'articule en trois phases clé (*assessment cycle*) : *Elicitation, Inference & Action*. La première phase concerne la prise d'informations à partir de différents types d'évaluation, la deuxième l'interprétation de ces informations alors que la troisième phase fait référence aux actions que le professeur doit mener suite aux deux phases précédentes. Il défend donc l'idée de "*build up a comprehensive picture of the overall achievements of a pupil by aggregating, in a structured way, the separate results of a set of assessments designed to serve a formative purpose*" (Wiliam, 2000, p. 9).

Taras (2005) a décrit très précisément les relations entre évaluation formative et évaluation sommative et considère que ces deux types d'évaluation sont interdépendantes et au cœur des évaluations de classe, l'une pour sa fonction pédagogique, l'autre pour sa fonction certificative. Néanmoins, après avoir étudié la façon dont cinquante chercheur·e·s d'une même institution (un département d'éducation dans une université anglaise) les définissent et les lient entre elles, elle conclut que la vision de ces deux types d'évaluation et leurs relations ne sont pas toujours partagées et que, par conséquent, il faut d'abord clarifier ce qui est entendu à travers chacune d'elles et entre elles pour décrire plus explicitement les évaluations de classe et les étudier.

Ces deux approches de l'évaluation sommative, parce qu'elles montrent que leur opposition et leur articulation ne sont pas si évidentes lorsqu'on étudie les évaluations de classe sont celles que je retiens pour élaborer mon cadre didactique de l'évaluation en mathématiques et que je développerai dans la seconde partie de cette note de synthèse.

Pour finir, il me semble également important de retenir dans cette recension d'articles relatifs à l'évaluation sommative, l'étude de Hiebert (2003) et Van de Walle (2006), qui indiquent qu'historiquement, les pratiques d'évaluation dans les classes de mathématiques dans les "*urban school*" sont, de manière dominante, traditionnelles et routinières et produisent ce que Haberman (1991) appelle "*the pedagogy of poverty*", c'est-à-dire peu stimulantes du point de vue des apprentissages et "adaptées" à la représentation que les professeur·e·s ont de leurs élèves.

b- Formative Assessment

Chez Scriven comme chez Bloom, l'évaluation est pensée à travers des tests permettant de mesurer, en cours d'enseignement (*formative evaluation*) ou en fin d'enseignement (*summative evaluation*) les apprentissages des élèves. Il a ensuite fallu environ vingt ans pour que la vision de *Formative Assessment* recouvre les processus d'évaluation dans une acception plus large prenant en compte le rôle des enseignant·e·s, mais aussi celui des élèves dans ces processus (Sadler, 1989 ; Torrance, 1993).

En 1998, Black et Wiliam ont répertorié près de deux cent cinquante travaux sur l'évaluation formative en classe dans un article majeur publié dans la revue *Assessment Education : Principles, Policy and Practice*. En 2005, ils ont de nouveau publié une revue de littérature sur l'évaluation formative pour l'OCDE, avec l'ambition de promouvoir de manière plus effective les pratiques d'évaluation formative dans les classes. Ils ont ainsi répertorié les recherches autour de *Formative Assessment* dont un certain nombre concerne les apprentissages mathématiques, mettant en avant le caractère bénéfique de l'évaluation formative, notamment pour les élèves les plus faibles, même s'ils soulignent la difficulté de distinguer si cela résulte du type d'évaluation adopté ou de l'enseignement plus globalement dispensé. Ils s'intéressent également à la nature des feedback produits dans le cadre de l'évaluation formative suivant qu'ils portent sur les tâches ou les élèves, mais aussi suivant qu'ils sont adressés par les professeur·e·s, les élèves eux-mêmes ou leurs pairs.

Dans leur synthèse, Black et Wiliam ont donné une définition de *Formative Assessment* qui servira de référence pour de nombreux travaux par la suite :

> We use the general term assessment to refer to all those activities undertaken by teachers—and by their students in assessing themselves—that provide information to be used as feedback to modify teaching and learning activities. Such assessment becomes formative assessment when the evidence is actually used to adapt the teaching to meet student needs. (p. 140)

L'article qui a eu le plus d'influence dans les mondes éducatif et scientifique est incontestablement celui que Black et Wiliam ont écrit la même année et qu'ils ont intitulé "*Inside the black box : raising standards through classroom assessment*" (1998). Dans cet article, ces chercheurs anglais défendent l'idée que le développement d'une approche formative de l'évaluation est au service des apprentissages des élèves, particulièrement pour ceux qui sont le plus en difficulté.
Pour ces auteurs et pour beaucoup d'autres (Shepard, 2009 ; Kahl, 2005 ; Looney, 2005), ce qui importe est de considérer que l'évaluation formative a pour principal objectif d'orienter les décisions de l'enseignant·e durant le processus d'enseignement, à partir des retours des élèves, pour améliorer leurs apprentissages (Wiliam, 2014, p. 2). Shepard (2005) a d'ailleurs fait un parallèle entre l'évaluation formative et le concept de ZPD développé par Vygotski (1987), notamment en matière d'étayage (*scaffolding*) des apprentissages.

c- Assessment for Learning

Le terme de *Assessment for Learning* (AfL) est apparu pour la première fois en 1986 dans un chapitre d'un livre de Black, mais c'est en 1999, qu'il a été repris et popularisé par l'ARG (The Assessment Reform Group) qui l'a utilisé pour qualifier, plus largement, tout type d'évaluation visant à améliorer les apprentissages des élèves. Ils ont considéré que le terme de *Formative Assessment* ne permettait pas d'aider suffisamment les professeur·e·s à concevoir des évaluations au service des apprentissages des élèves et ont ainsi défini dix principes "opérationnels". Pour eux (Broadfoot, Daugherty, Gardner, Harlen, James, & Stobart, 2002, p. 2–3), ces évaluations doivent :

- be part of effective planning of teaching and learning
- focus on how students learn
- be recognised as central to classroom practice
- be regarded as a key professional skill for teachers
- be sensitive and constructive because any assessment has an emotional impact
- take account of the importance of learner motivation
- promote commitment to learning goals and a shared understanding of the criteria by which they are assessed
- learners should receive constructive guidance about how to improve

- develops learners' capacity for self-assessment so that they can become reflective and self-managing
- recognise the full range of achievements of all learners

En 2009, lors d'un colloque international portant sur *Assessment for Learning*, la définition suivante a été adoptée :

> *Assessment for Learning* is part of everyday practice by students, teachers and peers that seeks, reflects upon and responds to information from dialogue, demonstration and observation in ways that enhance ongoing learning. (Klenowski, p. 264)

On le voit, la distinction entre *Formative Assessment* et *Assessment for Learning* relève d'une vision plus ou moins large de la façon dont l'évaluation peut être conçue pour promouvoir les apprentissages des élèves. Dans *Formative Assessment*, il s'agit principalement d'utiliser l'évaluation pour concevoir et adapter son enseignement alors que dans *Assessment for Learning* tout ce qui peut concourir à promouvoir les apprentissages des élèves est pris en compte, jusqu'à la motivation des élèves et l'impact émotionnel de l'évaluation.

En France, le terme d'*Assessment for Learning* n'a pas d'équivalent littéral, même si on peut l'assimiler à l'évaluation positive préconisée dans de nombreuses prescriptions institutionnelles (programmes, circulaires, loi). Néanmoins, il n'a jamais été indiqué, avec autant de précisions, ce que recouvre ce qualificatif subjectif. Dans un questionnaire récemment diffusé sur les pratiques évaluatives des professeur·e·s des écoles en mathématiques (voir plus loin), la principale difficulté retenue par les enseignant·e·s l'ayant renseigné est concevoir une évaluation "positive".

d- Self-regulated learning

Le concept de *Self-regulated learning* a été défini par Butler and Winne (1995) comme étant :

> A style of engaging with tasks in which students exercise a suite of powerful skills: setting goals for upgrading knowledge; deliberating about strategies to select those that balance progress toward goals against unwanted costs; and, as steps are taken and the task evolves, monitoring the accumulating effects of their engagement (p. 245).

Ce concept s'est ensuite considérablement développé, en parallèle de celui de *Formative Assessment*. En 2005, des chercheur·e·s américains (Stiggins, Arter, Chappuis & Chappuis, 2005) ont défini sept stratégies au service de *Assessment for Learning* accordant une place importante aux élèves et visant à permettre au professeur·e se repérer dans son enseignement à partir de trois questions clé :

Q1 : Where am I going ?
- Provide students with a clear and understandable vision of the learning target.
- Use examples and models of strong and weak work.

Q2 : Where I am now ?
- Offer regular descriptive feedback.
- Teach students to self-assess and set goals.

Q3 : How can I close the gap ?
- Design lessons to focus on one learning target or aspect of quality at a time.
- Teach students focused revision.
- Engage students in self-reflection and let them keep track of and share their learning.

La même année, les anglais Leahy, Lyon, Thompson and Wiliam ont proposé de conceptualiser l'évaluation formative à partir de cinq "*key strategies*" qui accordent également une place majeure aux élèves. En effet, ces stratégies prennent en compte trois processus centrés sur l'élève (where the learner is going, where the learner is right now, and how to get there) et intègrent le fait que l'enseignant·e autant que l'élève ou ses pairs ont un rôle à jouer dans l'évaluation (figure 1).

<table>
<tr><th></th><th>Where the learner is going</th><th>Where the learner is now</th><th>How to get there</th></tr>
<tr><td>Teacher</td><td rowspan="3">Clarifying, sharing and understanding learning intentions and success criteria</td><td>Engineering effective discussions, tasks and activities that elicit evidence of learning</td><td>Providing feedback that moves learning forward</td></tr>
<tr><td>Peer</td><td colspan="2">Activating students as learning resources for one another</td></tr>
<tr><td>Learner</td><td colspan="2">Activating students as owners of their own learning</td></tr>
</table>

Figure 1: Five "key strategies" of formative assessment (Leahy et al., 2005)

L'idée principale que recouvre ce concept est que les élèves ont une part active à jouer dans leurs propres apprentissages, ce qui doit donc être pris en compte pour penser et concevoir l'évaluation formative (Wiliam, 2014).

En France, les pratiques évaluatives sont loin d'accorder une telle place aux élèves (Rapport de l'IGEN, 2013), même si des expérimentations tentent de les encourager. Ainsi, dans le projet ASSIST-ME, l'évaluation formative proposée, qui a fortement intéressé les professeur·e·s de mathématiques engagé·e·s dans cette expérimentation, a amené les élèves à se positionner en "d'accord/pas d'accord/ne sait pas" puis argumenter sur les réponses collectées des élèves de la classe. Les élèves français·e·s se sont révélé·e·s très actif·ve·s dans les différentes phases et se sont bien investi·e·s dans les évaluations entre pairs, même s'ils·elles ont rencontré des difficultés pour se positionner sur les réponses de leurs camarades et surtout pour revenir ensuite sur leurs propres réponses (Coppé, 2015).
Une autre expérimentation menée dans l'académie de Créteil peut également être considérée comme relevant de ce type d'évaluation, même si elle ne s'affiche pas en tant que telle. Il s'agit d'une variante de l'EPCC (Evaluation Par Contrat de Confiance, Antibi, 2007), appelée EPCC participative (Quiquempois, 2016) qui propose de faire prendre en charge par les élèves, en petits groupes, la fiche réussite qui leur permet de préparer le "contrôle". Ainsi, les élèves doivent non seulement s'interroger sur ce qu'ils ou elles doivent savoir, mais aussi en confronter leurs points de vue entre eux·elles. L'enseignant·e n'est pas exclu·e de ce dispositif puisque*, in fin*e, c'est lui ou elle qui arbitre la fiche finale.

Quel que soit le type d'évaluations pratiqué en classe et les usages qui en sont faits, il semble important de considérer ce qui se passe dans la réalité des pratiques de classes, *inside the black box* comme l'ont préconisé Black et Wiliam (2001), en cherchant à comprendre les difficultés que rencontrent les enseignant·e·s pour mettre en œuvre l'*AfL* (*Assessment for Learning*) et les résistances qu'ils ou elles y opposent. Stiggins (2001) a, de son côté, préconisé cinq conditions visant l'amélioration des pratiques évaluatives des professeur·e·s aux Etats-Unis : repenser les croyances, viser une perspective internationale, évoluer vers une évaluation équilibrée, porter davantage d'attention aux classes et aux partenaires avec des programmes de formation des enseignant·e·s.

e- Classroom Assessment

Les recherches sur l'évaluation n'ont pas tout de suite porté sur les évaluations dans les classes (*day-to-day classroom*) car, au début, elles se sont principalement focalisées sur les évaluations externes nationales ou internationales (McMillan, 2003 ; Stiggins, 2001).

Cizek et Fitzgerald avec Shawn et Rachor (1995) et avec Racher (1996), Brookhart (1993), Frary, Cross, et Weber (1993) ont été parmi les premier·e·s à mener des recherches sur les évaluations de classe. Cizek, Fitzgerald, Shawn et Rachor (1995) ont ainsi montré que les pratiques d'évaluation des professeur·e·s étaient très variables et imprédictibles car dépendantes de trop nombreux paramètres tels que le genre, les années d'expérience, le niveau d'enseignement, la familiarité avec les prescriptions locales ou nationales (p. 159). Ils suggèrent même que le niveau d'enseignement est moins important que les caractéristiques individuelles et les croyances des professeur·e·s. Ces chercheurs ont également montré que les professeur·e·s utilisent généralement une variété de facteurs objectifs et subjectifs pour maximiser la probabilité d'attribuer de bonnes notes à leurs élèves.

McMillan a dégagé des profils de pratiques évaluatives de professeur·e·s du secondaire en 2001 (grades 6-12) et de professeur·e·s du primaire (2003). Les résultats de ces études montrent que, même si les professeur·e·s du secondaire ont intégré certains éléments de la réforme sur l'évaluation (promouvant l'*AfL*), ils ou elles gardent cependant des pratiques d'évaluation très traditionnelles. Concernant les professeur·e·s du primaire dont il a étudié les décisions prises en matière d'évaluation durant les séances de classe, il a identifié plusieurs facteurs déterminants : des facteurs relatifs à la notation, des facteurs liés aux stratégies d'évaluation et des facteurs liés aux niveaux cognitifs des évaluations des élèves. Il a ainsi montré que les professeur·e·s du primaire sont influencé·e·s par des facteurs externes et internes à la classe (davantage que les professeur·e·s du secondaire) et qu'ils utilisent un "*hodgepodge*" de facteurs pour évaluer leurs élèves (McMillan, Myran & Workman, 2002).

Veldhuis et Van den Heuvel-Panhuizen ont également cherché à dégager des profils de professeur·e·s du primaire à travers une étude quantitative (N= 960), mais aux Pays-Bas et seulement en mathématiques. Partant du principe que les pratiques d'évaluation des professeur·e·s sont intimement liées à l'enseignement qu'ils ou elles dispensent en classe et qu'elles y sont intégrées, ces chercheur·e·s considèrent que la façon dont les professeur·e·s rendent compte des apprentissages de leurs élèves dépend de plusieurs facteurs : leurs croyances sur l'évaluation de classe, les méthodes d'évaluation qu'ils ou elles choisissent et la conception des apprentissages sur laquelle s'appuie leurs évaluations (Veldhuis, Van den Heuvel-Panhuizen, Vermeulen, Eggen, 2013). A partir d'un questionnaire prenant en compte ces différents facteurs, ils ont identifié quatre profils d'évaluateur·rice·s : l'évaluateur·rice enthousiaste (ayant une vision positive de l'évaluation et de multiples usages, 28,5%), l'évaluateur·rice moyen·ne (ayant des pratiques traditionnelles, avec des caractéristiques majoritairement partagées, 35,3%), l'évaluateur·rice non-enthousiaste (ayant une vision négative de l'évaluation et en usant le moins possible, 25,8%) et l'évaluateur·rice alternatif (autres, 10,3%). Ces profils, selon les auteur·e·s, participent à une meilleure connaissance des évaluations de classe et peuvent servir de base pour concevoir le développement professionnel des évaluateur·rice·s, selon leur profil.

Brookhart (2004), à partir d'une recension de travaux portant sur "*classroom assessment*" effectuée sur vingt ans (1984-2004), situe les pratiques d'évaluation des professeur·e·s à l'intersection de trois fonctions professionnelles : l'enseignement, la gestion de classe et l'évaluation. Elle propose de développer une théorie spécifique des évaluations de classe qui s'appuie donc sur trois champs scientifiques : l'étude des différences individuelles (psychologie cognitive, théorie d'apprentissage et de motivation), l'étude des groupes (sociologie, *social learning*) et l'étude de la mesure (validité & fiabilité, évaluations formative & sommative).

Des recherches ont, par ailleurs, montré (Kilpatrick, 1993 ; Shepard, 2000) que les pratiques d'évaluation et de notation des professeur·e·s n'étaient pas toujours très cohérentes, notamment avec leurs croyances, et qu'il était important d'en rendre compte pour les faire évoluer. Suurtamm et Koch (2014) ont, pour leur part, identifié quatre sortes de dilemme auxquels sont confrontés les professeur·e·s quand ils évaluent les apprentissages de leurs élèves, également en lien avec leurs croyances et leurs conceptions de l'apprentissage. Elles proposent de les prendre comme point de départ pour faire évoluer les pratiques évaluatives :

- des dilemmes conceptuels : pourquoi doit-on évaluer ? quel sens a l'évaluation ? quelle place de l'évaluation dans l'enseignement et l'apprentissage ?
- des dilemmes pédagogiques : comment évaluer ? quelle stratégie adopter ? quels outils utiliser ?
- des dilemmes culturels : comment s'adapter à de nouveaux courants d'évaluation éloignés de sa pratique habituelle, mais prescrits par l'institution ou le contexte ?
- des dilemmes politiques : quelle influence des évaluations standardisées nationales ou internationales ? comment prendre en compte les prescriptions institutionnelles ou locales autour de l'évaluation ?

Remesal (2007) a, de son côté, montré l'influence des politiques "*d'accountability*" dans les pratiques d'évaluation en classe et que ces pratiques étaient d'autant plus variées que ces politiques étaient faibles (comme en France).

Il me semble donc important de retenir de ces travaux (et de bien d'autres que je n'ai pas évoqués, mais qui convergent vers les mêmes résultats), que les pratiques évaluatives des professeur·e·s sont très variées, pas toujours cohérentes et contraintes aussi bien institutionnellement que professionnellement. La question de leur validité et celle de leur fiabilité doivent donc être étudiées pour appréhender leur impact sur les apprentissages des élèves.

f- Validité & fiabilité

Les concepts de validité et de fiabilité ont été définis dans une approche psychométrique, dans le cadre d'évaluations externes. Comme le stipule Bonner (2013) dans l'introduction du chapitre du Handbook de McMillan intitulé "*validity in classroom assessment : purposes, properties and principles*" qu'elle a pris en charge, ce n'est pas parce que l'approche psychométrique de la validité n'a pas permis de dégager des réponses aux besoins des enseignant·e·s utilisant des méthodes informelles d'évaluation de classe, (p. 88) que ce concept ne garde pas sa pertinence pour étudier les évaluations de classe.

De nombreux articles traitant de la validité font référence aux travaux de Messick (1989, 1995). Ce chercheur définit la validité comme suit :

> Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *interpretations* and *actions* based on test scores or other modes of assessment (Messick, 1989, p. 6).

Il précise (1995) que la validité n'est pas la propriété des tests, mais qu'elle se réfère plutôt à leur signification qui peut être différente en fonction des personnes. Il estime que les questions de pertinence, d'utilité et de conséquences sociales liées à la validité de l'interprétation des scores sont entremêlées et à multiples facettes.

Pour Van den Heuvel-Panhuizen et Hodgen (2013) la validité se réfère au fait qu'une évaluation mesure bien ce qu'elle prétend mesurer alors que la fiabilité se réfère au fait que l'évaluation répétée à un autre moment doit produire les mêmes résultats. Van den Heuvel-Panhuizen (1996) avait déjà évoqué les notions de fiabilité et de validité dans son livre présentant l'approche RME (*Realistic Mathematics Education*). Elle y avait précisé que, dans cette approche, la fiabilité était difficile à déterminer selon les critères psychométriques usuels

du fait de la nature même des évaluations qu'elle préconise qui amènent souvent des réponses complexes d'élèves. Par ailleurs, l'évaluation en RME n'est pas compatible avec l'idée que ces derniers doivent trouver un même résultat si on les soumet de nouveau à l'évaluation puisque par principe, l'évaluation est partie prenante du processus d'apprentissage. Néanmoins, elle considère qu'une évaluation est plus valide lorsqu'on évalue, par une mesure directe, les performances des élèves que quand on le fait à travers des QCM. Dans cette perspective, elle évoque le travail de Linn, Baker et Dunbar (1991) qui ont vision élargie de la validité qui doit, de leur point de vue, être pensée en lien avec la fiabilité. Ces chercheur·e·s ont formulé huit critères devant être pris en compte lorsqu'on parle de validité d'une évaluation : (i) les conséquences prévues ou imprévues pour l'enseignement et les apprentissages des élèves, (ii) la justice, (iii) le transfert et la généralisabilité, (iv) la complexité cognitive, (v) la qualité du contenu, (vi) la couverture du contenu, (vii) la signifiance, (viii) le coût et l'efficacité. Pour eux, comme pour Van den Heuvel-Panhuizen, la qualité d'un test ne doit pas être éprouvée à travers l'objectivité de sa notation, mais à travers la qualité de son contenu. L'étude approfondie des réponses des élèves doit permettre de dépasser cette notion de qualité pour porter plus d'attention à ce que ces réponses révèlent des connaissances des élèves. Pour témoigner du fait que la validité d'une évaluation de classe est l'affaire de l'enseignant·e, qui ne peut donc être totalement pensée de manière objective, elle cite Wiliam (1993) qui considère que *"a test is valid to the extent that you are happy for a teacher to teach towards the test* " (p. 7).

Pour Brookhart (2003), les concepts de fiabilité et de validité doivent être totalement repensés dans la mesure où les évaluations de classe ne s'inscrivent pas dans la même perspective que les évaluations externes. Les évaluations de classe procurent des informations sur les connaissances des élèves qui doivent tout de suite être utilisées pour faire avancer le cours, elles n'ont donc pas la même fonction. L'opposition entre un échantillon à grande échelle qui caractérise une évaluation externe et la taille d'une classe qui caractérise une évaluation interne rend caduque toute adaptation, de son point de vue. La fiabilité qui est censée assurer que les paramètres de temps et de forme notamment n'ont pas d'influence sur les résultats des élèves n'a donc plus de sens dans le cadre des évaluations de classe. Les erreurs des élèves sont, dans les évaluations de classe, une fenêtre sur les apprentissages des élèves et non des éléments à corriger. Elles peuvent advenir un jour, mais être corrigées le jour suivant suite aux feedbacks de l'enseignant·e, c'est pourquoi la question de leur fiabilité ne peut se poser dans les mêmes termes que pour les évaluations externes (Shepard, 2001). La fiabilité pourrait être pensée en termes d'écarts entre le travail des élèves et le travail "idéal" tel que défini aussi bien par l'enseignant·e que par les élèves.
Brookhart estime que, dans les évaluations de classe, les croyances des professeur·e·s, leurs pratiques, leur compréhension à la fois du savoir et des élèves (incluant les différences culturelles et linguistiques) sont des critères pertinents de validité. Les critères de validité doivent, pour elle, porter sur les conditions qui permettent de rendre l'évaluation utile du point de vue des apprentissages des élèves. Pour cette chercheure, la validité d'une évaluation de classe est donc dépendante du contexte de la classe et n'a pas de valeur dans l'absolu. L'évaluation faisant partie intégrante de l'enseignement, une évaluation valide est donc "*an episode of genuine learning*". Dans son article, elle cite Black (1998, p. 123) qui estime que la validité de l'évaluation formative est déterminée en partie par le modèle d'apprentissage sur lequel se fonde l'enseignant·e pour produire ses feedbacks. Ces feedbacks peuvent être de nature différente (positifs ou négatifs, centrés sur les points forts ou les lacunes), avec des objectifs différents (évaluatif ou descriptif) et une utilité formative également variable (Tunstall & Gipps, 1996, cité par Brookhart).

Pour Cizek (2009) et Kane (2006), la validité d'une évaluation est également liée aux interprétations de l'enseignant·e qu'ils abordent du point de vue des inférences qu'il ou elle fait à partir des réponses des élèves. Ces inférences peuvent être correctes ou non, leur validité dépend des justifications fournies par l'enseignant·e à partir d'arguments logiques ou de preuves tangibles. Bonner (2013) propose cinq principes pour que l'enseignant·e fasse des inférences "valides" et souhaite ainsi participer au projet d'élaboration d'une "*measurement theory of classroom assessment*" souhaité par Brookhart (2003) :

- l'évaluation doit être alignée avec l'instruction
- les biais doivent être minimaux à toutes les phases du processus d'évaluation
- le processus d'évaluation doit permettre d'engager l'élève dans un processus pertinent du point de vue de ses apprentissages
- les effets des interprétations produites doivent être évalués
- la validation doit être produite par des preuves issues de plusieurs intervenants (les élèves, les parents, les autres professeur·e·s, le ou la directeur·trice)

Pour Moss (2003) également, la validité d'une évaluation réside dans les usages et interprétations que l'enseignant·e en fait. Pour elle, l'unité d'analyse d'une évaluation de classe est l'élève, dans sa singularité cognitive et socio-culturelle, mais en prenant en compte le contexte (notamment de la classe et de l'école) et les facteurs qui ont pu l'amener à produire les résultats. Elle conçoit la validité des évaluations de classe principalement à partir des usages et des interprétations que l'enseignant·e en fait. Le travail de cette chercheure qui s'appuie sur son expérience d'enseignante (et le revendique) pour penser les questions de validité et de fiabilité me semble important prendre en compte car, au-delà de l'utilisation de ces concepts par les chercheur·e·s, c'est bien aux professeur·e·s qu'ils doivent être utiles. Schoenfeld (2007) s'interroge d'ailleurs sur la signification de ces concepts pour les professeur·e·s et doute qu'ils s'y intéressent ou qu'ils en aient même entendu parler.

Pour finir, je citerai Morgan et Watson (2002) qui alertent sur le fait que l'évaluation n'est pas, pour elles, sans danger car des différences de jugement entre des évaluations pratiquées en classe et des évaluations standardisées, dues à des questions de validité et de fiabilité, peuvent générer de l'iniquité.

g- Grading

En conclusion de son chapitre intitulé "*grading*" dans le Handbook de McMillan, Brookhart a explicitement lié la question de notation à la question de la validité, en faisant référence aux travaux de Kane (2006) et de Messick (1989). A travers une recension des recherches menées entre 1985 et 1994 sur les pratiques de notation des professeur·e·s qu'elle a effectuée pour ce chapitre, Brookhart a mis en évidence la convergence de résultats. Notamment sur le fait que les professeur·e·s, quel que soit leur niveau d'enseignement mixent des composantes liées à l'effort et au comportement avec des composantes cognitives pour noter leurs élèves, plus spécialement encore pour les élèves en difficulté. Elle cite plusieurs études à grande échelle, notamment celles de McMillan (2001, 2002) et de Myran et Workman (2002) qui témoignent du fait que les professeur·e·s s'appuient sur des observations informelles de leurs élèves (participation, effort, comportement, etc.) en plus des résultats à des évaluations plus formelles (tests, évaluations sommatives, etc.) pour noter leurs élèves. McMillan (2003) a pourtant indiqué que les enseignant·e·s ont du mal à fournir des justifications concernant leurs pratiques de notation. Ces chercheurs ont identifié des différences entre le secondaire et le primaire. Dans le primaire, McMillan, Myran et Workman (2002) ont montré que les professeur·e·s mixent davantage encore que dans le secondaire de multiples critères pour noter leurs élèves. Ces chercheurs ont également montré que les variances inter-personnelles étaient plus fortes que les variances inter écoles pour expliquer les différences de pratiques de

notation. Cross et Frary (1999) ont montré que les professeur·e·s sont très résistant·e·s à l'idée de noter leurs élèves exclusivement à partir de leurs résultats formels, sans avoir réellement conscience des conséquences négatives de leurs pratiques (en termes d'équité, notamment), même si Brookhart (1994), à partir de la recension qu'elle a effectuée, a conclu que les professeur·e·s accordent beaucoup d'importance au fait de noter équitablement leurs élèves.

Pour conclure cette partie, je retiens que les chercheur·e·s anglophones ayant travaillé sur l'évaluation ont développé des concepts scientifiques fondamentaux qu'ils ont cherché à affiner (*Formative Assessment*, *Assessment for Learning*, etc.) ou à spécifier selon des contextes (pour la classe/externe, milieu social, etc.). Ils ou elles ont tou·te·s accordé une place importante aux facteurs externes pouvant avoir une influence sur les différents processus en jeu dans l'évaluation et sur les acteurs et actrices concerné·e·s (professeur·e·s, élèves, etc.). Ces différents travaux sont centraux pour étudier l'évaluation dans une perspective d'apprentissage. Je ne manquerai pas de m'y référer pour présenter et justifier le cadre didactique de l'évaluation que je propose.
Je retiens également qu'au delà de tous les résultats exposés, plusieurs auteur·e·s (Shepard, 2006, 2008 ; Shavelson, 2008), dont Bennett (2011), considèrent qu'il est important de prendre en compte la discipline pour évaluer précisément les connaissances des élèves et rendre plus effectif le concept de Assessment for learning. En référence à Shulman (1986) qui a fait valoir l'importance de la spécificité des contenus pour appréhender les apprentissages et l'enseignement, mais aussi à Ball, Thames et Phelps (2008) et Hodgen et Marshall (2005) pour ce qui concerne les mathématiques, Bennett considère que :

> To be maximally effective, formative assessment requires the interaction of general principles, strategies, and techniques *with* reasonably deep cognitive-domain understanding. (p. 15)

Je vais donc, dans la partie suivante, répertorier les travaux anglophones sur l'évaluation qui se sont plus particulièrement penchés sur l'évaluation en mathématiques.

2. Évaluation & Apprentissages mathématiques

Au début des années 1990, plusieurs ouvrages ont été publiés autour de l'évaluation en mathématiques. En 1992, un livre "*Assessment and learning of Mathematics*" édité par Gilah Leder et publié par l'*Australian Council for Educational Research* a rassemblé des contributions s'intéressant aux liens entre apprentissages mathématiques des élèves et méthodes d'enseignement et d'évaluation dans différents pays (Australie, Etats-Unis, Pays-Bas et Royaume Uni). Dans cet ouvrage, plusieurs chercheur·e·s se sont intéressé·e·s aux effets de la perception qu'ont les enseignant·e·s de leur rôle dans les apprentissages mathématiques des élèves et à l'influence des contextes sociaux sur les pratiques d'enseignement et d'évaluation des professeur·e·s. En 1993, suite à la sixième étude ICMI (International Commission on Mathematical Instruction) portant sur "*Assessment in Mathematics Education and Its Effects*", Mogens Niss a coordonné la publication de deux ouvrages autour de l'évaluation en mathématiques. Le premier "*Cases of Assessment in Mathematics Education*" présente des études de cas d'évaluation dans différents pays (Espagne, Chine, Caraïbe, pays arabes, Etats-Unis, Royaume uni, Norvège, Danemark, Pays-Bas, Australie) alors que le second "*Investigations into Assessment in Mathematics Education*" propose davantage une analyse de l'évaluation en mathématiques et de ses effets à partir de différentes approches (historique, psychologique, sociologique, épistémologique, idéologique et politique). Dans son chapitre introductif, Niss fait une claire distinction entre les termes *Assessment* (centré sur les capacités mathématiques des élèves) et *Evaluation*

(centré sur les systèmes éducatifs). Il précise également trois entrées qu'il juge fondamentales pour étudier les questions d'évaluations en mathématiques : *provision of information*, *taking decisions and actions* et *shaping of social reality*. Dans ces deux livres, de nombreux chercheur·e·s dressent un portrait désastreux des pratiques évaluatives dans leur pays sans proposer d'alternatives ou méthodes permettant réellement de les améliorer. La question des méthodes d'évaluation qui ne soient pas trop lourdes à utiliser par les professeur·e·s dans le quotidien de leur(s) classe(s) en mathématiques se pose clairement à travers ces présentations. En 1994, Kenneth Ruthven a également répertorié des travaux sur l'évaluation dans un article publié dans *Educational Studies in Mathematics*. Il évoque le principe "WYTIWYG" (*What you teach is what you get*) pour justifier trois clés qu'il propose pour repenser l'évaluation en mathématiques : *increasing the authenticity and realism of assessment tasks* ; *increasing the interpretive quality of assessment information* et *increasing the integration of processes of teaching, learning and assessment*. Il faut noter que, dans ces ouvrages de synthèse sur l'évaluation en mathématiques, le contexte français n'est quasiment jamais présenté et que Antoine Bodin est le seul français à y avoir contribué.

Dans son livre "*Assessing Mathematical Proficiency*", Schoenfeld (2007) a consacré plusieurs chapitres ou sous-chapitres à tenter de répondre à des questions du type : "*What is Mathematical Proficiency and How can it be assessed*" ? ou encore "*Mathematical Proficiency : what is important? how can it be measured*?". Ces questions sont en effet capitales pour appréhender l'évaluation en mathématiques. Morgan (1999) a proposé une réponse qui me semble très juste et qui permet de bien comprendre les enjeux de la question de l'évaluation en mathématiques, tant du point de vue des apprentissages des élèves que des pratiques des professeur·e·s :

> Assessing students in the subject of mathematics is a complex endeavor that relies on different understandings of the purposes of assessment, as well as what it means to know and/or do mathematics and whether and how this knowledge and these activities can be observed and evaluated (p. 8).

Pour répertorier les articles et travaux relatifs à l'évaluation en mathématiques, je suis également partie de questions qui m'ont semblées pertinentes pour éclairer l'approche que je défends.

a- Première question

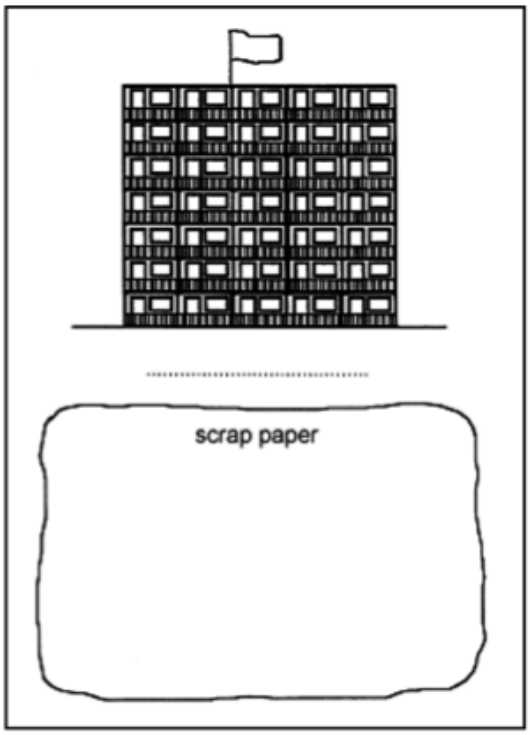
La première question qui a orienté une partie de ma recherche documentaire est : comment évaluer finement (c'est-à-dire en prenant en compte le caractère multidimensionnel des savoirs mathématiques) et justement (c'est-à-dire en en rendant réellement compte dans ses aspects positifs et négatifs) les apprentissages des élèves en mathématiques ?

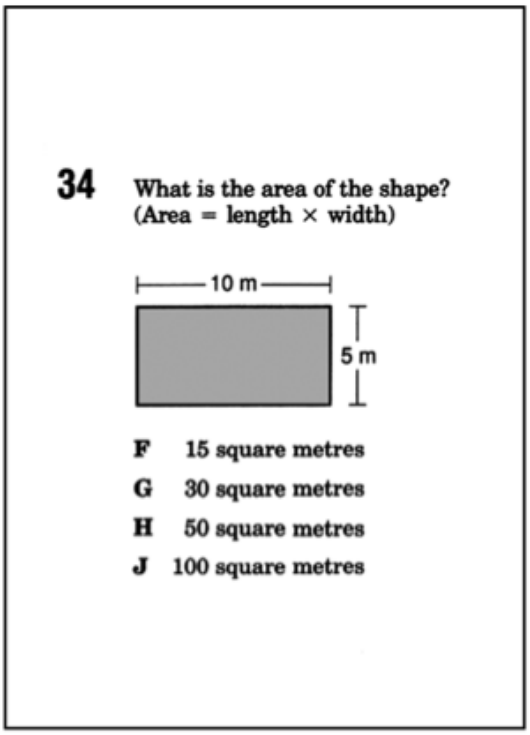
Partant du constat que les instruments et les tests standardisés utilisés dans les classes en mathématiques sont souvent limités dans leur portée et n'évaluent seulement que quelques aspects et composantes des savoirs et compétences mathématiques des élèves (McLean, 1982; Fiske, 1997 ; De Lange, 2007), des chercheur·e·s en *Mathematics Education* ou en *Didactics* ont cherché à répondre à la question de la spécificité de l'évaluation en mathématiques.
Le modèle didactique pour l'évaluation proposé par Van den Heuvel-Panhuizen (1996) est une réponse possible à cette question. Elle l'a développé dans le cadre de son approche *Realistic Mathematics Education* (RME) et dans la continuité des travaux de Freudenthal dans lesquels elle s'inscrit. Le titre de son livre "*Assessment and Realistic Mathematics Education*" témoigne du lien inextricable que cette chercheure conçoit entre évaluation et apprentissages mathématiques. Elle emploie d'ailleurs le terme de "*didactical assessment*" pour qualifier l'évaluation, en précisant que "*this is assessment that is intended as a support to teaching and*

learning process" (p. 2). Elle indique également que cette expression témoigne du fait que le but de l'évaluation, aussi bien que le contenu, les méthodes ou encore les instruments utilisés doivent tous s'inscrire dans une perspective didactique et reprend à son compte la prescription de Shepard (2000) concernant l'évaluation en mathématiques qui doit être "épistémologiquement cohérente" avec la Didactique des mathématiques. Pour Van den heuvel-Panhuizen et Becker (2003), il est évident que ce "*didactical assessment*" doit donc se baser sur la Didactique des mathématiques et non sur la Psychométrie (p. 699) car on doit s'intéresser à la fois aux réponses des élèves et à la fois à leurs procédures. Dans leur article, ces chercheurs se réfèrent à Osterlind (1998) qui a étudié l'approche psychométrique des tests pour dénoncer le fait que cette approche ne permet pas de prendre en compte l'aspect multidimensionnel du savoir, ni de traiter les réponses des élèves efficacement, du point de vue de la réponse à apporter. Dans l'approche psychométrique :

- un item évalue un seul trait ou une seule compétence
- une seule réponse est possible ou attendue
- on évalue seulement ce qui a été enseigné ou appris
- les énoncés doivent être clairs et sans ambiguïté
- la réponse de l'élève est le seul indicateur de sa connaissance
- la bonne réponse doit toujours pouvoir être déterminée

Pour illustrer leur propos, ces chercheur·e·s présentent deux items permettant d'évaluer les connaissances des élèves dans le domaine "grandeurs & mesures" au grade 5 (CM2) autour d'une même situation "*the Flag*". Cette situation permet de bien montrer en quoi les deux approches (psychométrique et didactique) diffèrent et l'intérêt de développer la seconde pour promouvoir les apprentissages des élèves en mathématiques. Dans les deux cas, il s'agit pour l'élève de déterminer les dimensions d'un drapeau :





Dans la présentation de gauche, l'absence d'informations relatives à la mesure effective du drapeau permet à l'élève de montrer plus largement ses connaissances et compétences dans le domaine, notamment grâce au "brouillon" joint à l'énoncé du problème. Dans la classique

présentation de droite, l'élève doit choisir LA bonne réponse, sans possibilité de montrer ce qu'il ou elle sait au-delà de la formule de l'aire d'un rectangle.
Cet exemple reflète la méconnaissance, que Van den Heuvel-Panhuizen et Beckers qualifient de sévère, de la nature même des mathématiques et de ce que peut être un problème mathématique (p. 705). Il donne également à voir ce qu'ils appellent un "riche environnement" pour penser et concevoir l'évaluation des apprentissages des élèves qui place la résolution de problèmes au coeur de l'évaluation et des apprentissages mathématiques (Freudenthal, 1973). Van den Heuvel-Panhuizen a d'ailleurs répertorié, dans son livre (1996), les caractéristiques d'un "bon" problème d'évaluation :

- Problems should be balanced
- Problems should be meaningful and worthwhile
- Problems should involve more than one answer and higher-order thinking
- Concerns about open·ended problems
- Problems should elicit the knowledge to be assessed
- Problems should reveal something of the process
- More concerns about open·ended problems
- Good problems can have different appearances
- does not exist

Ruthven (1994) avait de son côté dégagé des caractéristiques de tâches évaluatives pertinentes pour repenser l'évaluation en mathématiques (p. 441) :
- Assessment tasks should be good exemplars of the practice of mathematics beyond the school;
- Assessment tasks should make human sense to students, engaging their interest and involvement;
- Assessment tasks should call for students to use their mathematical knowledge with insight and imagination;
- Assessment tasks should be framed to encourage the formulation of questions, and the interpretation of conclusions;
- Assessment tasks should be capable of being tackled in a variety of ways, exploiting a range of mathematical knowledge;
- Assessment tasks should be conducted under normal working conditions, and presume student access to resources such as calculators, computers, texts and consultants.

Thompson, avec Kaur (2011) et Bleiler (2012, 2013) défendent également le principe d'une approche multidimensionnelle de l'évaluation en mathématiques. Elle suggère que, quel que soit le domaine traité, les professeur·e·s proposent des tâches permettant d'évaluer les connaissances mathématiques des élèves selon quatre dimensions : *Skills* (S), compétences relatives aux algorithmes et aux procédures, *Properties* (P), propriétés du savoir en jeu, *Uses* (U), usages avec un focus sur les applications et *Représentations* (R), relatives aux diagrammes, images ou représentations visuelles de concepts. Voilà une illustration de cette approche concernant la multiplication décimale proposée au grade 5 (CM2) :

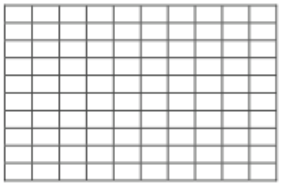
Skills	Evaluate: 0.6×1.5.
Properties	If $0.6 \times A = 15$, what would $0.06 \times A$ equal? How do you know? What about $60 \times A$?
Uses	Find the cost of 0.6 kilos of meat costing 1.50 euros per kilo.
Representations	 If the figure represents 1 unit, illustrate 0.6×1.5.

Figure : multiplication décimale évaluée selon l'approche SPUR

Peu de recherches ont étudié l'évaluation en mathématiques à partir de la spécificité du contenu évalué. Dans son livre "*Assessing Mathematical Proficiency*", Schoenfeld (2007) a consacré deux chapitres à deux contenus spécifiques : l'algèbre et les fractions. Le cas de l'algèbre a été étudié par McCallum et Foster, celui des fractions par Ball et Fisher. Fisher a montré comment différentes tâches évaluatives autour des fractions pouvaient révéler différentes (in)compréhensions et Ball a donné à voir l'intérêt de creuser finement la façon dont un élève de grade 6 (Brandon) se représentait les fractions, à partir d'une interview de 2004 retranscrite dans le chapitre qu'elle a écrit. Ces chercheur·e·s se sont appuyé·e·s sur les spécificités de ces contenus (lien avec les fonctions et rôle des symboles pour l'algèbre, lien avec l'écriture décimale et les ordres de grandeur pour les fractions) pour promouvoir des évaluations adaptées et constructives du point de vue des apprentissages des élèves.

b- Seconde question

L'autre question qui a orienté ma recherche documentaire concerne plus particulièrement la mise en œuvre de l'évaluation dans les classes : comment proposer une évaluation qui soit réellement intégrée au processus d'enseignement et d'apprentissage des élèves?

Cette question est à mettre en lien avec les réformes qui ont, dans de nombreux pays, promu l'évaluation formative ou l'évaluation "*for learning*" et leur impact sur les pratiques effectives des enseignant·e·s.

Des chercheurs (Nortvedt, Santos & Pinto 2015 ; Suurtamm, Koch & Arden, 2010) ont étudié plus particulièrement la façon dont les professeur·e·s de mathématiques mettaient en œuvre l'AfL (*Assessment For Learning*) dans leur classe, suite à de nouvelles directives nationales et de nouveaux programmes promouvant l'évaluation formative en Ontario (Canada). Dans une autre étude sur le même thème (Nortvedt, Santos & Pinto, 2015), mais cette fois comparative entre la Norvège et le Portugal, des résultats similaires (peu de changements effectifs dans les pratiques évaluatives et peu d'usages d'évaluations formatives) ont été trouvés concernant l'adoption de l'*AfL* dans ces deux pays à l'école primaire, du fait des difficultés rencontrées par les professeur·e·s de ces deux pays pour utiliser des critères de jugements sur le niveau de connaissances mathématiques de leurs élèves. Les auteur·e·s insistent donc sur l'importance de prendre en compte la discipline et ne pas se contenter d'une approche générale pour étudier le développement professionnel des enseignant·e·s relatif à l'évaluation. Quand Black et Wiliam (2001) promeuvent l'*AfL* et l'usage de feedbacks adaptés à chaque élève suivant son travail, avec des conseils sur ce qu'il ou elle doit améliorer et éviter les comparaisons avec les autres, ils ne spécifient pas explicitement ce que ces feedbacks doivent être en fonction des

difficultés et obstacles rencontrés en mathématiques. La difficulté pour les professeur·e·s réside bien dans le fait d'identifier les problèmes des élèves à travers leurs productions et de fournir des retours pertinents et adaptés. Cela implique donc que les professeur·e·s aient les connaissances disciplinaires et didactiques permettant de réaliser ce type de feedbacks.
Pour l'étude canadienne (Ontario), les chercheur·e·s ont montré que les professeur·e·s, dans leur grande majorité, s'appuient encore sur des tests ou des quizz pour évaluer les apprentissages de leurs élèves et les noter (étude quantitative par questionnaire), mais que d'autres pratiques plus innovantes avaient également été observées (forum des mathématiques, portfolio, autoévaluation, etc.) lors des études de cas. Ces chercheur·e·s ont donc conclu que, même si les pratiques évaluatives en mathématiques des professeur·e·s évoluent peu statistiquement, ces derniers adhèrent globalement aux fondements de la réforme (NCTM, 2000) et cherchent davantage à comprendre leurs élèves pour mieux les faire réussir. Ils·elles soulignent l'importance de délivrer des messages clairs et cohérents en matière d'évaluation dans les programmes et les ressources qui les accompagnent, mais aussi de promouvoir le développement de réseaux professionnels pour aider les enseignant·e·s à mettre en œuvre la réforme de manière plus effective dans leurs classes car l'intégration de nouveaux standards d'évaluation n'est pas simple[2]. Dans la mesure où peu de formations accompagnent ces nouvelles orientations en matière d'évaluation, cette question mérite d'être étudiée à partir d'un cadre permettant de le faire.

Dans leur chapitre intitulé "*classroom assessment in mathematics*" du Handbook de McMillan portant sur les "*Research on classroom Assessment*" (2013), McGatha et Bush ont répertorié plusieurs recherches centrées sur l'étude d'évaluations formatives mises en place dans des classes de mathématiques, à partir de quatre entrées : *integrated practices*, *isolated practices*, *student work analysis* and *technology*. Elles ont montré que les professeur·e·s mettent en oeuvre des approches variées de l'évaluation formative (allant de l'usage de quizz quotidiens à des feedbacks variés ou à l'autoévaluation) et que globalement, ils ou elles utilisent davantage les productions des élèves avec une visée formative. Elles ont retenu que les pratiques d'autoévaluation ont généralement un impact positif sur la réussite des élèves (Fontana & Fernandes, 1994 ; Ross, 1995 ; Ross, Hogaboam-Gray & Rolheiser, 2002) pourvu que les élèves soient engagé·e·s dans l'élaboration des critères d'évaluation et aient appris à les utiliser. Elles ont également montré qu'un travail spécifique autour des tâches proposées en évaluation est bénéfique pour aider les professeur·e·s à mieux comprendre le fonctionnement cognitif des élèves en mathématiques et à développer de meilleures stratégies de réponse à leurs erreurs et incompréhensions (Lin, 2006). C'est ce que je propose également de faire dans la recherche en cours EvalNumC2 sur les pratiques d'évaluation en mathématiques des professeur·e·s des écoles enseignant au cycle 2 (CP/grade1 au CE2/grade3) qui sera présentée ultérieurement. Concernant l'usage des technologies pour promouvoir l'évaluation formative, elles n'ont évoqué qu'une seule recherche (Shirley, Irving, Sanalan, Pape & Owens, 2011), indiquant que c'était une entrée relativement récente, ayant eu des résultats positifs en termes de suivi individuel et de réussite globale des élèves de six classes du secondaire équipées (*Connected Classroom Technology*) et utilisant le système TI-navigator. McGatha et Bush ont conclu cette partie en évoquant des recherches mettant en avant la nécessité de développer les PCK (*Pedagogical Content Knowledge*, Shulman, 1986) des professeur·e·s en mathématiques pour leur permettre d'utiliser les informations issues de l'évaluation formative pour adapter et programmer leur enseignement en fonction des besoins

[2] Si l'on s'intéresse à la France, on peut se demander dans quelle mesure les professeur·e·s français·e·s appliqueront les nouvelles préconisations sur l'évaluation qui ont été faites dans les différents textes institutionnels proposés à la suite de la loi de la Refondation de l'école de 2013 (Programmes 2015 & 2016, circulaires de rentrée depuis 2014, Socle commun de connaissances, de compétences et de culture, 2015).

repérés (Hodgen & Marshall, 2005 ; Heritage, Kim, Vendlinski & Herman, 2009). Elles ont relevé une phrase qui synthétise très clairement ce fait :

> Even though the pedagogy involved in formative assessment seem generic, the teachers' content knowledge was the crucial element for creating effective assessment tasks and providing useful feedback (p. 453).

Elles ont également formulé deux préconisations sous forme de défis à relever pour favoriser le développement de l'évaluation formative dans les classes de mathématiques :

- the importance of extended professional development as opposed to one-day workshop.
- providing support to teachers through professional learning communities.

J'ai retenu ces préconisations pour concevoir la méthodologie de recherche et de formation de la recherche EvalNumC2 déjà évoquée. Cette recherche collaborative s'appuie sur différents collectifs d'enseignant·e·s engagés sur trois ans pour faire évoluer leurs pratiques évaluatives en mathématiques.

Morgan et Watson (2002) insistent également, sur la nécessité de former les professeur·e·s à interpréter correctement les réponses des élèves et à bien juger de leurs performances pour promouvoir l'équité dans l'évaluation des apprentissages des élèves, en mathématiques. Elles précisent qu'il est plus difficile d'interpréter correctement les productions des élèves lorsque ces dernier·e·s sont confronté·e·s à des tâches complexes et que les jugements émis dépendent de ressources professionnelles et personnelles spécifiquement liées aux mathématiques (p. 84) :

- Teachers' personal knowledge of mathematics and the curriculum, including affective aspects of their personal mathematics history.
- Teachers' beliefs about the nature of mathematics, and how these relate to assessment.
- Teachers' expectations about how mathematical knowledge can be communicated.

Mais également de ressources plus générales :

- Teachers' experience and expectations of students and classrooms in general.
- Teachers' experience, impressions, and expectations of individual students.
- Teachers' linguistic skills and cultural background.

Dans le même ordre d'idée Baxter, Shavelson, Herman, Brown, et Valadez (1993, p. 213) proposent de définir une "*reliable measure of mathematical achievement*".

Ainsi, de nombreux chercheur·e·s ont étudié la spécificité de l'évaluation en mathématiques et ont produit des résultats qui permettent d'éclairer certaines problématiques s'y rattachant. Ils ou elles ont montré l'importance de prendre en compte la spécificité du savoir évalué et la nécessité, pour les enseignant·e·s, de disposer de connaissances disciplinaires et didactiques indispensables pour, à la fois concevoir des évaluations permettant d'évaluer les savoirs mathématiques dans toutes leurs dimensions et, à la fois être en capacité d'interpréter et de prendre en compte les réponses des élèves.

Conclusion de la partie littérature anglophone

Pour effectuer ce travail de recension, j'ai d'abord commencé par rechercher dans les revues "généralistes" les articles incontournables puis ceux qui s'intéressaient plus particulièrement aux mathématiques. Malgré la grande qualité de ces articles, il s'est avéré que les mathématiques y sont souvent anecdotiquement présentées ou émanant d'une opportunité disciplinaire saisie par des chercheur·e·s pour leur appuyer leurs travaux sur l'évaluation. J'ai ensuite recherché, dans les revues spécialisées en mathématiques, les travaux qui traitaient de l'évaluation et cette fois, il m'est apparu que des dimensions qui apparaissaient comme centrales dans les travaux "généralistes" étaient souvent absentes. Il s'agit plus

particulièrement des dimensions liées aux élèves et à la prise en compte de la façon dont ils ou elles percevaient les évaluations auxquelles ils ou elles étaient confronté·e·s (sauf dans le vol 2 de la sixième étude ICMI qui porte sur l'évaluation et ses effets). De même que même si les professeur·e·s sont bien présent·e·s dans les travaux en *Mathematics Education* ou en *Didactics*, la dimension professionnelle de leur pratique est souvent négligée, peu voire pas traitée.
C'est pourquoi je propose un cadre didactique de l'évaluation qui ambitionne de prendre en compte les différentes dimensions liées à l'évaluation (professionnelle, disciplinaire, sociale et institutionnelle) que je présenterai dans la deuxième partie de cette note de synthèse.

B- REVUE DE LITTÉRATURE FRANCOPHONE

Pour faire suite à cette revue de littérature anglophone sur l'évaluation, j'ai choisi de répertorier les travaux de chercheur·e·s francophones qui, tout en s'inscrivant dans les différents courants de l'évaluation développés par des chercheurs majeurs comme Scriven, Bloom, Black, et Wiliam, ont développé des orientations spécifiques, liées au contexte de leur pays ou de leur culture d'évaluation.

Avant d'effectuer ce travail, je souhaite donner quelques éléments liés au contexte scientifique francophone qui a permis l'émergence des travaux que je citerai par la suite. En effet, même si l'on ne peut établir de lien causal direct, il me semble important de préciser que les premiers travaux scientifiques francophones que l'on peut rattacher à l'évaluation datent de la fin des années 1880 et qu'ils ont initialement portés sur le "surmenage " en milieu scolaire. C'est donc une préoccupation liée à la qualité d'enseignement qui a permis l'émergence des premiers travaux réalisés dans le champ de l'évaluation. Dans un livre intitulé "la fatigue intellectuelle" datant de 1898 Binet, alors directeur du laboratoire de psychologie physiologique de la Sorbonne, s'est intéressé au "surmenage" dû aux longues journées d'étude et au caractère encyclopédique des savoirs à acquérir à l'école. Il y dénonçait déjà les formes d'évaluation pratiquées par les professeur·e·s à l'époque. En étudiant, plus particulièrement le Certificat d'études, il mettait déjà en évidence le caractère arbitraire des épreuves de ce diplôme ne permettant pas de savoir ce qui était véritablement évalué.
Les travaux de Binet s'inscrivent dans le cadre de la psychologie expérimentale qui s'était alors focalisée sur l'étude des "différences individuelles" et "la recherche de méthodes précises permettant de caractériser les individus et de confronter les différences en les chiffrant" (Martin, 2002). C'est, dans ce cadre, que le concept d'aptitude décrite comme "ce qui différencie, sous le rapport du rendement, le psychisme des individus" (Claparède, 1922) est apparu. Pour estimer ces aptitudes, la méthode des tests s'est alors imposée et a prévalu pour le classement des individus à partir d'une échelle de rendements construite pour différentes aptitudes. Trois grands types de test ont alors été développés : les tests de développement qui "mesurent l'évolution mentale" ; les tests d'aptitudes proprement dit, qui "révèlent certaines dispositions naturelles" et les tests de connaissances, qui "permettent d'explorer l'acquis scolaire d'un élève, soit en général, soit en une matière particulière" (Duthil, cité par Martin, 2002). Le fait que ces tests aient été élaborés au même moment et dans le même champ scientifique va, de mon point de vue, durablement et irrémédiablement lier l'évaluation à une visée de mesure, ce qui ne sera pas sans impact sur la vision de l'évaluation qui perdure encore aujourd'hui en milieu scolaire. Ne parle-t-on pas, plus volontiers, en France, de l'évaluation des élèves plutôt que de l'évaluation des apprentissages des élèves ? Cet amalgame de mesures de développement, d'aptitudes ou de connaissances reste prégnante et caractérise, encore aujourd'hui chez certains professeur·e·s, leur vision de

l'évaluation. Cette vision normative et parfois anxiogène de l'évaluation pourrait expliquer pourquoi certains chercheur·e·s francophones ont parlé de pression évaluative (Merle, 2005), de peur (Hadji, 2012) ou encore de menace de l'évaluation (Butera, Buchs & Darnon, 2011).

Je vais à présent exposer, comme je l'ai fait pour la partie exposant les travaux anglophones, les principaux travaux des chercheur·e·s francophones, en distinguant ceux relatifs à l'évaluation "en général" et ceux portant plus particulièrement sur l'évaluation en mathématiques.

1- Sur l'évaluation en général

a- Les apports complémentaires aux travaux anglophones

Dans son ouvrage intitulé "évaluations formative et certification des apprentissages", Mottier Lopez (2015) décrit l'émergence du domaine de l'évaluation des apprentissages des élèves et précise que la pédagogie de maîtrise ("*mastery learning*") développée par Bloom et ses collègues (Hasting & Madaus) en 1971, dans le Handbook évoqué dans la partie précédente portant sur les évaluations formatives et sommatives, a eu un impact significatif au plan international dans les communautés scientifiques en éducation. Crahay (1999) a précisé que cette nouvelle conception de l'évaluation s'intégrant dans les processus d'enseignement, n'a pas été simple à accepter pour tout le monde (francophone, notamment). Néanmoins, Perrenoud (1995) a décrit comment cette nouvelle vision de l'évaluation, dépassant la vision normative qui avait cours jusque-là, s'inscrivait davantage dans un courant de justice sociale porté par "les forces de Gauche", qui défendait l'idée d'une école plus "juste", au service de la réussite de tous les élèves, même si Kahn (2010) a nuancé cette analyse. Mottier Lopez (2015) précise également dans son livre que les travaux francophones d'Europe vont d'abord s'employer à décortiquer les différentes fonctions de l'évaluation des apprentissages des élèves. C'est ainsi que Cardinet (1983) a proposé de distinguer une troisième fonction de l'évaluation en plus des fonctions de régulation et de certification, la fonction d'orientation. Cette dernière, que l'on retrouve également sous les appellations de pronostique ou de prédictive, a une visée diagnostique cherchant à prévoir la réussite d'une personne à une formation ou un enseignement. Elle est distincte de la fonction certificative dans la mesure où la fonction de diagnostic, en lien avec des critères, n'a de sens que si elle débouche sur une action appropriée (Perrenoud, 1989).

Dans les revues de littérature que Mottier Lopez a réalisées entre 1978 et 2007, avec Allal (2005) et Laveault (2008) sur les travaux en langue française réalisés autour de l'évaluation, il s'est avéré que la majorité d'entre eux portait sur l'évaluation formative (Mottier Lopez, 2015). Cette chercheuse suisse précise, pour justifier le fait que les problématiques autour de l'évaluation formative se sont multipliées dans les années 1990-2000, que ces travaux se sont développés en lien avec les réformes qui ont eu lieu notamment en Belgique et en Suisse romande qui visaient à promouvoir l'évaluation formative dans les classes. On peut noter à cette occasion qu'en France, encore à ce jour, on ne trouve pas de trace explicite d'évaluation formative dans les prescriptions institutionnelles (programmes, circulaires, socle), même si de nombreuses recommandations en matière d'évaluation des apprentissages des élèves peuvent s'apparenter à cette vision de l'évaluation.

Mottier Lopez précise que l'analyse des travaux qu'elle a effectuée avec Allal en 2005 pour l'OCDE a permis de dégager quatre mouvements principaux de développements conceptuels autour de l'évaluation formative. L'approche didactique de l'évaluation étant plus naturellement rattachée à l'évaluation formative, je choisis de retenir cette catégorisation des travaux sur l'évaluation, qui me permettra de présenter les concepts qui me semblent les plus

intéressants, plutôt que celle de Bonniol et Vial (les modèles de l'évaluation, 1996) ou celle de De Ketele (les paradigmes de l'évaluation, 1993, 2016).
Les mouvements retenus par Mottier Lopez et Allal se sont focalisés sur (p. 26) :
1) **les instruments pour évaluer**, visant le développement de techniques d'évaluation instrumentées avec des tests dont la validité est toujours recherchée de manière optimale, des grilles d'observation ou d'autoévaluation, des échelles d'appréciation ou des portfolios, etc. Bain (1988) a, à propos de ces travaux, prévenu d'un risque d'illusion instrumentale.
2) **l'étude de pratiques existantes dans leurs contextes**, même si comparativement aux travaux anglophones, peu de recherches ont effectivement été menées dans ce domaine entre 1978 et 2002. A partir des années 2000, la mise en place de l'approche par compétences et l'apparition des tâches complexes a néanmoins favorisé l'émergence de travaux visant l'étude de l'application de ces courants menés par des chercheur·e·s principalement suisses, canadiens et belges tels que Kahn, Rey, Fagnant, Roegiers ou Tardif.
3) **l'implication de l'élève dans l'évaluation formative** avec le courant de l'évaluation formatrice développé principalement par Bonniol (1986), Nunziati (1990) et Vial (1995) visant la construction d'un "modèle personnel d'action" permettant à l'élève de s'impliquer dans l'évaluation de ses apprentissages à travers la construction d'une référence qui lui permette de se représenter la performance à réaliser, mais aussi les moyens d'y arriver (Campanale, 2001). Le courant de l'autoévaluation s'est également développé à travers les travaux de Allal notamment qui, avec Michel (1993) a défini trois types de modalités d'organisation de situations d'autoévaluation : l'autoévaluation au sens strict, les évaluations mutuelles réciproques et les co-évaluations). Laveault (1999 et 2007) a plutôt parlé d'autorégulation ce qui, de son point de vue, doit être un moyen pour l'élève d'apprendre, pourvu qu'il ait fait l'objet d'un apprentissage.
4) **le développement de cadres théoriques**, a principalement occupé les chercheur·e·s francophones (Allal, Cardinet, Perrenoud, De Ketele, Rey, Mottier Lopez, etc.), notamment autour de trois grandes "conceptions de l'évaluation" (Cardinet, 1990) associées à des "modèles de sciences" différents (p. 38) : l'évaluation externe et objective, l'évaluation interne et subjective et l'évaluation négociée et interactive.

C'est principalement dans ce dernier mouvement que se trouvent les travaux les plus exploitables du point de vue de la didactique, même si les travaux de Mottier Lopez autour de l'apprentissage situé et la microculture de la classe (Cobb & Yackel, 1998), également importants pour l'approche didactique de l'évaluation et les travaux que je souhaite développer, s'inscrivent dans le deuxième mouvement.

b- Des concepts nouveaux ou revisités

Voilà donc quelques concepts développés par des chercheur·e·s francophones qui viennent enrichir et/ou compléter les travaux des chercheur·e·s anglophones.

- ***La régulation***

Les travaux de la chercheure suisse Linda Allal, très productive dans le champ de l'évaluation depuis plus de trente ans, ont fortement influencé le domaine de l'évaluation dans les pays francophones. Son approche de l'évaluation intégrant davantage les contenus des apprentissages, elle a donc produit des travaux didactiquement plus intéressants. En effet, même si c'est Cardinet qui a introduit le premier la notion de régulation de l'enseignement et des apprentissages en 1977 (en s'inspirant de l'analyse des systèmes cybernétiques), c'est bien Allal qui l'a conceptualisée le plus, permettant ainsi une vision élargie de l'évaluation formative telle que proposée par Bloom, Hasting et Madaus en 1971. C'est ainsi que cette

chercheure a distingué trois formes de régulation associées à l'évaluation formative (1979, 1988) :

- **La régulation interactive** a lieu quand l'évaluation formative est fondée sur des interactions de l'élève avec les autres composantes de la situation, c'est-à-dire avec l'enseignant·e, avec les autres élèves et/ou avec du matériel permettant une auto-régulation de l'apprentissage.
- **La régulation rétroactive** intervient lorsque l'évaluation formative est réalisée à la fin d'une phase d'enseignement.
- **La régulation proactive** intervient lorsque différentes sources d'information permettent l'élaboration de nouvelles activités d'enseignement/apprentissage conçues pour prendre en compte les différences entre les élèves.

Allal et Mottier Lopez précisent, dans leur synthèse de 2005, que les approches novatrices de l'évaluation formative combinent toutes ces trois formes de régulation et que les activités d'enseignement s'organisent autour de plusieurs modalités de régulation interactive fondées sur des démarches d'évaluation informelle (observation, discussion, échanges) auxquelles s'ajoutent des évaluations plus formelles (contrôles, interrogations écrites ou orales, etc.) en vue d'une régulation rétroactive plus efficace. Ce qui me semble important de relever, au delà de l'évolution du concept de feedback vers le concept de régulation plus riche et plus développé, c'est que dans cette approche, les rôles de l'élève et de l'enseignant·e sont redistribués, laissant une place bien plus importante à l'élève dans l'évaluation et la régulation de ses apprentissages.

D'autres chercheur·e·s se sont également intéressés au concept de régulation. Perrenoud (1998) a spécifié que, pour lui, le terme de régulation des apprentissages concernait aussi bien la régulation de l'activité de l'élève, que celle de ses processus cognitifs et celle effective de ses apprentissages. Laveault (2007) a lui travaillé sur les dysfonctionnements possibles des régulations. Il en a identifié trois : des régulations insuffisantes, des régulations inadaptées et des régulations excessives. Ces régulations ne permettent pas, voire peuvent empêcher, des apprentissages de se réaliser. Elles peuvent advenir quand un·e professeur·e ne maîtrise pas suffisamment les savoirs qu'il ou elle doit enseigner ou quand il ou elle ne dispose pas de connaissances didactiques lui permettant de les opérer efficacement. Ce point me semble particulièrement intéressant à retenir, notamment pour étudier les pratiques évaluatives des professeur·e·s des écoles qui ne disposent pas toujours, du fait de leur cursus généralement non-scientifique, de connaissances suffisantes en mathématiques.

On le voit, le concept de régulation, tel que développé par les chercheur·e·s francophones, s'inscrit naturellement dans une approche didactique des apprentissages dans la mesure où, la régulation qu'elle soit interactive, rétroactive ou proactive sera d'autant plus efficace qu'elle prendra en compte la spécificité des contenus. Mottier Lopez (2012) résume très bien ce qu'il est important de retenir de ce concept :

> S'intéresser à la régulation des apprentissages en classe implique ainsi un questionnement à la fois sur l'élève et les situations d'enseignement et d'apprentissage – dont les interventions de l'enseignant font partie – susceptibles d'orienter positivement l'autorégulation de l'élève. (p. 9)

- ***L'évaluation formatrice***

Les travaux menés autour de l'évaluation formatrice ont été précurseurs d'une démarche de recherche de type collaborative puisque, dans le cadre d'une expérimentation menée dans le lycée de Marseilleveyre de Marseille à la fin des années 70, des chercheur·e·s (Bonniol et son équipe de l'université de Provence) et des enseignant·e·s (principalement Nunziati) ont cherché à définir les principes de ce dispositif spécifique d'évaluation. Le concept d'évaluation formatrice se caractérise par une démarche de régulation conduite par celui qui apprend, c'est-à-dire l'élève plutôt que l'enseignant·e. Pour les chercheur·e·s et

enseignant·e·s engagé·e·s dans cette approche, l'évaluation formatrice s'inscrit dans le domaine de la didactique dans la mesure où elle vise à introduire davantage de rationalité dans les apprentissages et une redistribution de la place des élèves et des professeur·e·s dans le processus d'évaluation. Cette redistribution caractéristique de l'évaluation formatrice passe par l'appropriation, par les élèves, des outils d'évaluation des enseignants et par la maîtrise par l'apprenant des opérations d'anticipation et de planification (Nunziati, 1990). Les promoteur·rice·s de l'évaluation formatrice partent du principe que, pour qu'un dispositif pédagogique soit "opérationnel" et garantisse la réussite du plus grand nombre, il doit fournir à l'élève les outils nécessaires à la représentation correcte des buts fixés, à la planification rationnelle de l'action, à l'autocorrection et enfin à l'autoévaluation. Dans son ouvrage de 2012 qui répertorie les différents modèles de l'évaluation, Vial décrit très précisément le contexte dans lequel a émergé l'évaluation formatrice, les principes qui la pilotent ainsi que les outils emblématiques de cette approche spécifique de l'évaluation (carte d'étude, fiche critériée, analyse didactique, etc.).

- ***Validité, fiabilité, pertinence***

Les concepts de validité, fiabilité et pertinence d'une évaluation ont été établis, pour la communauté scientifique francophone, par De Ketele et Roegiers, en 1993 puis repris et développés par De Ketele et Gérard en 2005 dans le cadre de l'approche de l'évaluation par compétences.

La pertinence correspond au caractère plus ou moins approprié de l'épreuve avec les objectifs visés par les évaluateur·rice·s (De Ketele, Chastrette, Cros, Mettelin & Thomas, 1989), c'est "son degré de "compatibilité" avec les autres éléments du système auquel elle appartient" (Raynal & Rieunier, 1997, 2003 cités par De Ketele & Gérard, 2005, p. 2). La validité concerne le degré d'adéquation entre ce que l'on déclare vouloir mesurer et ce que l'on mesure effectivement (Laveault & Grégoire, 1997, 2002).

La fiabilité est "le degré de confiance que l'on peut accorder aux résultats observés : seront-ils les mêmes si on recueille l'information à un autre moment, avec un autre outil, par une autre personne, etc. ?" (De Ketele & Gérard, 2005, p. 2). Ces définitions sont celles adoptées par de nombreux chercheur·e·s francophones qui s'y réfèrent fréquemment voire même quasi exclusivement, dans leurs travaux.

La validité, telle que définie ci-dessus, ne coïncide pas exactement avec celle des chercheur·e·s anglophones qui défendent plutôt l'idée d'une validité liée à l'interprétation que l'enseignant·e fait des productions de l'élève (sauf Van den Heuvel-Panhuizen & Beckers, 2003), mais il est vrai qu'elle n'a pas été définie dans le cadre d'évaluation de classe. Scallon (2004), qui a défendu l'approche de l'évaluation par compétences, que j'évoquerai au prochain paragraphe, est celui qui s'en est le plus rapproché en soulignant l'importance du jugement porté par l'enseignant·e à l'endroit de chaque compétence. J'indiquerai, par la suite, d'autres conceptions de la validité qui ont été développées par des chercheur·e·s francophones, en fonction de leur approche (par compétences ou psycho-didactiques, notamment).

- ***L'approche par compétences***

L'approche par compétences et l'évaluation des compétences ont fait l'objet de vifs débats dans les communautés scientifiques. La communauté des didacticien·ne·s ne l'a pas investie pour de multiples raisons qu'il n'est pas utile de présenter dans cette note, même si la dissolution du savoir dans cette approche en est certainement la principale. Néanmoins, cette approche ne peut être mise de côté tant son importance est grande dans la communauté francophone du fait de l'orientation de nombreux programmes scolaires qui s'y rattachent et donc de l'évaluation qui en découle. Certain·e·s de ses défenseur·e·s arguent même d'une

plus grande facilité d'évaluation des apprentissages des élèves, définis en termes de compétences et en types d'actions qu'elles rendent possibles. Rey (2014) précise d'ailleurs que "en évaluant les apprentissages sur la base de ce qu'ils sont capables de faire, il semble que l'on se donne un indicateur à la fois facile à observer et objectif" (p. 87). Les gages de fiabilité et de validité d'évaluation sont souvent avancés pour défendre l'approche par compétences. Il est vrai que, dans ce type d'approche, l'élève qui doit témoigner de l'acquisition d'une compétence doit le faire dans une situation complexe et inédite et donc, si la situation de l'évaluation est bien choisie (validité), elle sera intrinsèquement fiable. Rey (2014) précise d'ailleurs que, dans le cas de l'approche par compétences, l'évaluation ne peut excéder une à deux situations de ce type car sa réalisation par l'élève requiert forcément un temps important. Il y a donc un vrai paradoxe (et/ou une grande difficulté) pour les professeur·e·s, à devoir évaluer les apprentissages de leurs élèves quand ils sont définis en termes de compétences à travers des évaluations de classe. Ces évaluations sont en effet, traditionnellement et/ou classiquement, conçues à partir d'une série de tâches plus ou moins complexes et plus ou moins articulées entre elles, permettant d'évaluer des connaissances ou des procédures. Au-delà de la question de la complexité des tâches, l'évaluation par compétences se heurte, dans le cas des évaluations de classe, à la question du caractère inédit auquel elles doivent recourir. En effet, quelle que soit la vision que l'on a de l'évaluation, un·e professeur·e a à cœur d'évaluer les apprentissages de ses élèves à partir de l'enseignement qu'il ou elle a dispensé (Sayac, 2016), qui doit être conforme aux programmes scolaires. D'ailleurs, Rey (2014) le signale également :

> Pourtant, bien que la notion de compétence domine aujourd'hui les curriculums de nombreux pays, beaucoup de dispositifs d'évaluation, institutionnellement construits comportent surtout des évaluations de procédures et de connaissances élémentaires. (p. 90)

Pour faire face à cette difficulté, Carette, Rey, Defrance et Kahn ont développé en 2003 un dispositif d'évaluation par compétences en trois phases qui a eu un grand retentissement dans la communauté francophone belge et au-delà. Dans la première phase, l'élève est confronté à une tâche complexe qui exige la mobilisation d'un certain nombre de procédures et d'éléments de savoir qu'il est censé posséder. Dans la deuxième phase, l'élève est confronté à la même tâche complexe, avec étayage alors que dans la dernière phase il est confronté à une batterie de procédures de bases impliquées dans la tâche complexe qu'il est censé avoir automatisées. L'évaluation se fonde ainsi sur la façon dont l'élève s'est comporté face à ces trois séries de tâches et permet une analyse fine de la maîtrise (ou non) des compétences évaluées. De nombreuses recherches ont été menées pour évaluer des tâches complexes en mathématiques, à partir de ce dispositif (Fagnant, Demonty, Dierendonck, Dupont & Marcoux, 2014).

De Ketele et Gérard (2005) ont étudié plus particulièrement la validation des épreuves d'évaluation selon une approche par compétences, vues comme un ensemble de ressources permettant de résoudre une situation-problème appartenant à une famille de situations (De Ketele, 2000, 2001 ; Perrenoud, 1997 ; Scallon, 2004 ; Rey, 1996). Ils ont ainsi défendu, par opposition à une approche "classique" de la validité, l'usage de différents types de validation pouvant être convoqués selon cette approche : une validation *a priori* par recours à des juges pour "vérifier que les paramètres de la famille de situations proposée sont bien présent dans la situation concrète (qui évalue la compétence) et seulement eux" (p. 10), une validation empirique interne à partir de critères à définir spécifiquement pour chaque compétence évaluée et une validation empirique externe en fonction de critères externes (comparaison avec un groupe témoin ayant bien appris les ressources nécessaires à la résolution du problème, mais n'ayant pas appris à les mobiliser sur ce type de situation). Dans cette même approche des compétences, Bodin (2007) a précisé que, dans la mesure où les compétences

sont associées à des classes de situations, il est exclu de penser évaluer une compétence particulière par une seule situation particulière. Il préconise donc de concevoir des épreuves qui, centrées sur des compétences particulières, permettraient plutôt de repérer, chez les élèves, l'état de développement de ces compétences.

En mathématiques, c'est Gérard Vergnaud qui s'est le plus intéressé au concept de compétence. Pour lui, "il est impossible, sans la théorie mathématique et sans l'étude des phénomènes d'apprentissage des mathématiques, d'analyser les différentes compétences susceptibles d'être développées par les élèves. [...]. Il advient toujours des moments où les connaissances sous-jacentes aux compétences doivent être explicitées pour être situées les unes par rapport aux autres, dans un système d'ensemble cohérent" (1997, p. 9). Dans le prolongement de cette vision, il décrit l'évaluation des compétences comme résultant d'une analyse combinée du résultat de l'activité de l'élève et de l'organisation de cette activité (2001). Il propose ainsi une méthode d'analyse qu'il qualifie de "descente vers le cognitif", en précisant qu'il ne peut exister d'analyse des compétences sans analyse de l'activité et réciproquement que l'on ne peut faire une analyse de l'activité sans analyser les conceptualisations sous-jacentes (2001).
Winsløw (2005) s'est également emparé de cette notion et a proposé de distinguer les compétences mathématiques générales des compétences mathématiques spécifiques, en relation avec ce qu'il appelle "la matière". Il s'est inscrit, tout en étant critique, dans la tradition scandinave portée par Niss (1999) qui avait défini une compétence mathématique, comme une composante de l'expertise mathématique : la puissance d'agir avec intelligence et d'une façon convenable dans des situations comportant une certaine forme de défi mathématique (Niss & al., 2002, p. 43). Le modèle de Niss[3] a servi de base à l'élaboration du cadre théorique de la partie *mathematical literacy* du programme PISA (OCDE, 1999) et l'on trouve, sur le site du ministère de l'Éducation nationale français (Eduscol), des ressources s'y référant[4].
Schneider (2006) a également abordé la question des compétences en posant celle du transfert de connaissances d'un problème à un autre à l'intérieur (ou non) d'une même classe de situations. Pour elle, les compétences doivent être pensées à partir des formes spécifiques que prennent des compétences transversales, selon la discipline. En mathématiques, elle a identifié trois compétences transversales : faire preuve d'esprit critique, formuler et vérifier des hypothèses et communiquer (2004).
La définition de compétence que j'ai proposée avec Grapin (2015) a été conçue pour permettre l'étude d'une évaluation standardisée nationale en mathématiques proposée en fin d'école primaire. Ce bilan visait à évaluer le niveau de connaissances et de compétences des élèves de fin d'école en mathématiques. Dans ce cadre, nous avons défini la compétence comme "une capacité d'agir de manière opérationnelle face à une tâche mathématique qui peut s'avérer inédite, en s'appuyant sur des connaissances que l'élève mobilise de façon autonome" (p. 113). Cette définition nous a permis de déterminer plusieurs niveaux de compétences pour analyser les items de l'évaluation, en lien avec les caractéristiques de la tâche mathématique (niveaux de mise en fonctionnement de connaissances associés, adaptations, complexité, etc.) et la façon dont l'élève l'appréhende. Je reviendrai plus loin sur l'usage qui en a été fait, du point de vue de l'évaluation étudiée.

[3] Parfois appelé « fleur des compétences », en référence aux 8 compétences mathématiques qui le caractérisent.
[4] https://cache.media.eduscol.education.fr/file/Formation_continue_enseignants/15/7/Competences_en_maths_aux_cycles_2_et_3_527157.pdf

- ***Le jugement professionnel en évaluation***

La notion de jugement professionnel en évaluation découle des travaux docimologiques qui ont mis en évidence le fait qu'il existait des "défaillances" au niveau du jugement scolaire en évaluation : disparités suivant les évaluateur·rice·s (Piéron, 1963), biais (De Landsheere, 1971 ; Bressoux, 2006), effets (Amigues & Zerbato-Poudou, 1996), arrangements évaluatifs (Merle, 1996, 2015). Elle est actuellement développée par de nombreux chercheur·e·s francophones (Allal & Mottier Lopez, 2008 ; Laveault, 2008 ; Mottier Lopez & Tessaro, 2016 ; Tourmen, 2009, 2014) qui la pensent souvent en lien avec l'évaluation formative et les différentes régulations qui lui sont associées, c'est pourquoi cette notion me semble particulièrement intéressante pour étudier l'évaluation en milieu scolaire avec une approche didactique.

Indépendamment de la question de la mesure qui rentre souvent en jeu dans l'évaluation, la question du jugement évaluatif est centrale dans toute situation d'évaluation. Pelletier (1997) l'exprime d'ailleurs très clairement :

> Ce qui différencie fondamentalement l'évaluation de la mesure, c'est le jugement de valeur, la référence à un critère permettant d'interpréter une mesure pour lui donner une valeur (une signification) dans un contexte donné (p. 90).

La façon dont un·e professeur·e va donc émettre un jugement sur la production d'un élève et la façon dont il ou elle va l'utiliser pour adapter son enseignement sont essentielles lorsque l'on s'intéresse à l'évaluation comme outil au service des apprentissages des élèves. Klenowski et Wyatt-Smith (citées par Mottier Lopez, 2014) indiquent d'ailleurs que la finalité de l'évaluation en classe étant, d'abord et avant tout, d'être au service de l'apprentissage de l'élève et de sa reconnaissance et validation institutionnelles, il est primordial de s'intéresser au jugement professionnel des professeur·e·s. D'après Lafortune et Allal (2008), ce jugement intervient dans l'ensemble des activités du professeur·e, aussi bien dans le choix et l'agencement des situations didactiques qu'il propose, que dans toutes les fonctions et étapes de l'évaluation qu'il met en place dans sa classe (choix d'outils, de critères, des appréciations, d'exploitations de résultats, etc.). Mottier Lopez (2014) précise également, en se référant à Laveault (2008), que :

> Le modèle conceptuel du jugement professionnel accorde une importance toute particulière aux dimensions *éthiques* et *critiques* qui interviennent dans le choix et la nature des informations retenues pour l'évaluation, dans la construction du sens et de la valeur de l'objet évalué, ainsi que dans les processus de communication et de décision qui en découlent. (p. 100)

Tourmen (2009, 2014) insiste sur le fait que le jugement professionnel en évaluation est un processus dynamique qui se situe entre des jugements provisoires et finaux. Allal et Mottier Lopez (2009) inscrivent également le jugement professionnel en évaluation dans le processus d'enseignement. Elles ont ainsi dégagé trois actions principales constitutives du jugement professionnel en évaluation :

- confronter et mettre en relation plusieurs sources d'information.
- interpréter la signification des indices recueillis par rapport à un référentiel et des repères pluriels qui font sens.
- anticiper et apprécier les conséquences probables de plusieurs actions envisagées au regard des résultats de l'évaluation.

Scallon (2004) a montré la nécessité d'un "bon" jugement pour évaluer des compétences, en ayant recours à des "outils de jugement qui mettent en évidence la qualité de ce traitement tant du point de vue du produit fini (les productions réalisées) que de celui de la capacité de faire appel à ses connaissances et à ses habiletés" (p. 24). Il spécifie aussi combien cette approche de l'évaluation est éloignée de l'approche des évaluations qui se pratiquent en classe.

Pour ma part, j'indiquerai ultérieurement ce que j'entends par *jugement professionnel et didactique en évaluation* une acception qui, tout en s'inscrivant dans la lignée des travaux cités ci-dessus, prend en compte les connaissances disciplinaires et didactiques des professeur·e·s. Je retiens également de ces travaux l'importance des dimensions éthique et responsable (du point de vue des enjeux d'apprentissage) du jugement porté par un·e professeur·e dans une situation d'évaluation de classe.

c- Des travaux avec une orientation didactique

Pour finir cette revue de littérature francophone, je citerai quelques travaux qui se sont explicitement inscrits dans une approche didactique de l'évaluation.

Dès 1979, Brun proposait de penser l'évaluation formative au service d'un enseignement différencié en mathématiques, à partir de l'analyse des productions des élèves et des tâches didactiques. Pourtant, c'est principalement en didactique du français que des travaux s'inscrivant dans cette approche se sont développés, notamment avec Bain et Schneuwly. Un ouvrage intitulé "*Évaluation formative et didactique du français*" publié en 1993 sous la direction de Bain, Allal et Perrenoud regroupe la majorité des contributions mêlant évaluation et didactique de cette discipline. Perrenoud y préconise de réfléchir à la place de la régulation des processus d'apprentissage dans les dispositifs didactiques. Il précise que cette régulation doit être prise en charge par l'enseignant·e qui, en estimant à la fois le chemin déjà parcouru par chacun et celui qui reste à parcourir, peut optimiser les processus d'apprentissage en cours. C'est donc dans le cadre d'une régulation proactive (Allal, 1988), que Perrenoud conçoit l'enjeu d'une approche didactique de l'évaluation. Bain et Schneuwly, à partir d'un résumé de texte produit par une élève et comportant de nombreuses fautes de grammaire, syntaxe et orthographe ont, de leur côté, montré la nécessité de se conformer à un modèle didactique (pour le texte présenté celui des activités langagières) pour savoir "quoi faire" d'une telle production, en termes d'évaluation formative et de régulation des apprentissages. Ils ont défini une validité didactique qu'ils voient "comme la pertinence et l'utilité du contrôle évaluatif pour la régulation de l'enseignement et de l'apprentissage dans le cadre d'une séquence didactique" (p. 70-71). Dans la conclusion de leur article, ces deux didacticiens du français ont émis un avis qui me semble encore pertinent aujourd'hui :

> S'il peut être utile ou opportun lors d'une formation didactique d'attaquer les problèmes par le biais de l'évaluation, c'est pour interroger immédiatement par ce moyen les objectifs fixés à la séquence, les concrétiser ou opérationnaliser par le contrôle envisagé et surtout les intégrer à la séquence. (p. 72).

Thouin (1993), en inscrivant ses travaux sur l'évaluation en mathématiques dans une perspective constructiviste, a également revendiqué une approche didactique de l'évaluation. Pour lui, étudier les schèmes cognitifs des élèves et évaluer leurs apprentissages ne peut se faire sans avoir recours à la Didactique. Il fait ainsi référence aux trois grands types de difficultés auxquelles sont confrontés les élèves (les difficultés conceptuelles, les obstacles épistémologiques et les erreurs didactiques), qui ont été identifiées par Vergnaud (1989) pour construire des instruments diagnostiques lui permettant de mesurer et évaluer les apprentissages mathématiques des élèves à l'école primaire.

S'inscrivant dans le courant des approches psycho-didactiques des apprentissages (Weil-Barais, 1996), Vantourout et Goasdoué (2014) ont nommé approche psycho-didactique de l'évaluation (APDE), l'approche qu'ils ont développée dans le cadre de différents travaux réalisés dans leur laboratoire EDA *Education, Discours & Apprentissages* (Vantourout, 2004 ; Vantourout & Maury, 2006). Cette approche combine des éléments (concepts, notions, résultats, méthodes, etc.) issus des didactiques disciplinaires et de la Psychologie cognitive dans le but de développer le cadre d'une évaluation au service des apprentissages des élèves,

c'est-à-dire, comme ils le précisent, un cadre qui repose en priorité sur la qualité du diagnostic élaboré par l'évaluateur ou l'évaluatrice. Vantourout et Goasdoué préconisent donc de s'intéresser aussi bien aux contenus impliqués dans les évaluations (analyse des tâches) qu'au fonctionnement cognitif des élèves et aux processus de réponse convoqués par les tâches. A cette fin, ils ont aussi développé la notion de validité psycho-didactique des tâches proposées en évaluation qui tient compte de la façon dont l'élève a produit sa réponse en fonction de la tâche pour asseoir la qualité du diagnostic à produire. Ce qui est intéressant dans cette approche, c'est que même si ces chercheurs indiquent que l'APDE peut être perçue comme une forme d'évaluation formative, leur approche "ne se définit pas en relation avec une fonction de l'évaluation, mais par la volonté de produire un diagnostic de qualité" (p. 11). Je reviendrai ultérieurement sur ce point.
La thèse de Christophe Blanc (2017) s'est inscrite dans cette perspective de validité psycho-didactique. Ce chercheur s'est intéressé aux évaluations proposées par les enseignant·e·s de Cours Préparatoire (grade1) en Français et en Mathématiques et a étudié leur validité suivant trois dimensions : une dimension curriculaire (conformité aux référents institutionnels), une dimension psycho-didactique (tâches données et réponses des élèves) et une dimension épistémo-didactique (couverture didactique des domaines concernés). Il a étudié les livrets scolaires remplis par 16 professeur·e·s de CP ainsi que les épreuves et tests qu'ils ou elles avaient données à leurs élèves durant deux trimestres d'une année scolaire. La conclusion majeure de son travail est que la validité des épreuves et des tests proposés par les professeur·e·s n'est pas assurée, dans bien des cas. Cela le conduit également à remettre en cause les jugements évaluatifs prononcés suite aux passations de l'évaluation.

2. Sur l'évaluation en Mathématiques

Pour rendre compte les travaux francophones autour de l'évaluation en mathématiques, je propose de distinguer ceux qui ont utilisé l'évaluation comme outil au service d'autres problématiques de ceux qui l'ont traitée comme objet spécifique d'étude. Cette entrée outil/objet est en effet pertinente pour distinguer ces travaux. Elle permet également de témoigner du fait que, jusqu'à présent, les didacticien·ne·s des mathématiques n'ont pas porté une grande attention à la problématique de l'évaluation, en tant qu'objet d'étude. Je reviendrai ultérieurement sur ce fait car il résulte, de mon point de vue, d'une épistémologie propre à la Didactique française et à ses théories. J'évoquerai également les travaux portant sur évaluations standardisées dans la mesure où ils ont permis l'émergence des travaux actuels développés en didactique des mathématiques. Je finirai par évoquer quelques grands projets de recherche qui sont actuellement en cours, ainsi que les travaux que j'ai développés jusqu'à ce jour.

a- Les travaux qui utilisent l'évaluation comme outil au service d'une autre problématique

Les travaux qui s'inscrivent dans cette partie sont ceux qui utilisent l'évaluation soit en amont d'un dispositif d'enseignement ou de formation (Grugeon-Allys, Pilet, Chesné) pour mieux le penser et le concevoir, soit en aval pour en rendre compte (Horoks, Chesnais). Ils visent donc soit à améliorer des pratiques d'enseignement autour de domaines spécifiques (algèbre, numération, calcul), soit à analyser les pratiques d'enseignement de professeur·e·s, sur un sujet précis (triangles semblables, symétrie axiale, nombres & calcul) et/ou dans un contexte particulier (éducation prioritaire).

- ***L'évaluation en amont d'un enseignement ou d'une formation***

Brigitte Grugeon-Allys (1995, 1997) a été la première à concevoir un modèle d'analyse multidimensionnelle d'un domaine mathématique (l'algèbre) ayant une visée diagnostique à partir de l'identification de cohérences de fonctionnement chez les élèves en algèbre qu'elle propose d'appeler "profil de l'élève en algèbre". Elle s'appuie sur les travaux de Chevallard (1985, 1989) pour modéliser l'activité mathématique de l'élève en termes de praxéologies mathématiques, c'est-à-dire, à partir des types de tâches et de techniques les résolvant (savoir-faire), chaque technique étant justifiée par un discours technologique, lui-même justifié par une théorie (savoir). Ainsi, la praxéologie épistémologique de référence du domaine algébrique qu'elle a définie est structurée à partir des praxéologies de *calcul* relatives aux expressions algébriques et aux formules (*calculer*, *substituer*, *reconnaître, développer*, *factoriser*) et aux équations (*reconnaître, résoudre une équation*) et des praxéologies d'*usage de l'outil algébrique,* praxéologies pour *généraliser*, *modéliser, mettre en équation*, *traduire*, *prouver*.

Saisissant l'opportunité d'un projet s'inscrivant dans une problématique EIAH (Environnements Informatiques pour l'Apprentissage Humain) réunissant des informaticien·ne·s et des didacticien·ne·s, Grugeon-Allys a pu concevoir un prototype automatisé de son diagnostic (*Pépite*), initialement conçu dans une version papier-crayon, pour aider les enseignant·e·s à établir plus efficacement un profil des compétences en algèbre de leurs élèves à l'entrée en 3ème et à ainsi réguler leurs apprentissages. Pour décrire plus précisément la démarche de cette évaluation, je reprendrai la description faite par Pilet (2015) :

> Cette évaluation est composée de dix tâches (questions à choix multiples ou énoncés plus ou moins ouverts) recouvrant les différents problèmes de l'algèbre. Les réponses des élèves sont codées en plusieurs étapes selon une analyse *a priori* fondée sur le modèle multidimensionnel de la compétence algébrique (Grugeon-Allys 1997). Au moyen de ce codage, le diagnostic établit les principaux traits caractéristiques des compétences de l'élève en algèbre en lui attribuant un niveau sur une échelle définie *a priori* pour trois composantes : "Calcul algébrique", "Usage de l'algèbre" et "Traduction algébrique" (p. 35).

Dans le cadre de différents projets de recherche ou de formation (LINGOT, PépiMeP, groupe IREM) des enseignant·e·s ont eu l'occasion d'expérimenter *Pépite* dans leurs classes avec leurs élèves. Ces dernier·e·s ont alors exprimé la difficulté qu'ils ou elles avaient rencontrée pour exploiter les profils construits par le logiciel, notamment pour proposer des situations d'apprentissages adaptés à leurs élèves (Delozanne et al., 2002). Pour répondre à cette difficulté, Grugeon-Allys a travaillé sans cesse à l'amélioration et au développement du logiciel *Pépite*, en collaboration avec de nombreux chercheur·e·s (Coulange, Delozanne, Prévit, Chenevotot, Pilet) appartenant à des champs scientifiques variés tels que la Didactique des mathématiques, l'Ergonomie cognitive, l'Informatique et les Sciences de l'éducation. Aujourd'hui encore, *Pépite* est utilisé dans de nombreuses classes (notamment via la plateforme LaboMeP de Sésamath*)* pour améliorer les apprentissages des élèves, en algèbre à partir du diagnostic qu'il permet de réaliser et des profils d'élèves qu'il dégage. Depuis 2014, le projet ANR NéoPraEval (Nouveaux Outils pour de nouvelles PRAtiques d'EVALuation et d'enseignement des mathématiques) dont Brigitte Grugeon-Allys est responsable (voir plus loin), lui permet de revisiter les problématiques développées dans les projets précédents en prenant davantage en compte la dimension "évaluation" de son dispositif (Grugeon-Allys, 2015). Elle s'intéresse plus particulièrement à la validité didactique des évaluations développées fondées sur la même approche théorique que *Pépite*, une validité qu'elle cherche à définir en prenant en compte à la fois des apports en Didactique des mathématiques (avec Grapin) et à la fois des apports en Sciences de l'éducation (avec Vantourout et Goasdoué).

Les travaux de Grugeon-Allys ont montré la nécessité d'une étude didactique et épistémologique fine du savoir mathématique pour permettre un diagnostic fiable et valide du

point de vue des connaissances et compétences des élèves et utile pour les enseignant·e·s. Après s'être initialement intéressée à l'évaluation à visée pronostique dans le but de soutenir les apprentissages des élèves en algèbre et outiller les enseignant·e·s pour les améliorer, cette didacticienne des mathématiques cherche aujourd'hui à utiliser les travaux qu'elle a développés antérieurement pour promouvoir l'évaluation formative en classe et la différenciation *via* des parcours différenciés.

Dans la continuité des travaux de Grugeon-Allys, Pilet (2012) a, dans sa thèse, cherché à concevoir des parcours différenciés en lien avec une évaluation diagnostique en algèbre élémentaire pour des élèves de Troisième et de Seconde, plus précisément sur les expressions algébriques polynomiales de degré inférieur ou égal à 3 à coefficients réels. Pour cela, elle a fait passer le test *Pépite* à 289 élèves de Troisième et de Seconde ce qui lui a permis de définir des profils d'élèves en les caractérisant par des praxéologies apprises. L'analyse multidimensionnelle des réponses des élèves à l'évaluation lui a permis de distinguer trente-six profils d'élèves en algèbre, qu'elle a ensuite regroupés selon quatre praxéologies apprises liées aux niveaux de maîtrise pour chaque composante de la compétence algébrique. La mise en regard de la praxéologie à enseigner avec les praxéologies apprises lui a permis de mettre à jour des besoins d'apprentissage des élèves et ainsi de définir un modèle de Parcours d'Enseignement Différencié (PED) sur les expressions algébriques qu'elle a fait tester par une enseignante en charge d'une classe de Troisième. La façon dont Pilet a conçu son modèle de parcours différencié est intéressante à mettre en perspective avec la vision que l'on peut avoir de l'évaluation au service des apprentissages des élèves. En effet, les conditions qu'elle a retenues pour qu'un enseignement différencié soit favorable aux apprentissages témoignent du fait que l'approche de l'évaluation par les fonctions (notamment formative et certificative) n'est pas toujours opérationnelle dans la réalité des classes. Les trois conditions qu'elle énonce (2014, p. 20-21) sont :

- La différenciation, pour être favorable aux apprentissages, doit être intégrée au processus d'enseignement et ne pas être conçue uniquement comme une remédiation qui arriverait en fin d'enseignement.
- Le repérage des besoins d'apprentissage des élèves et les choix de différenciation (tâches, moments) doivent reposer sur un référent épistémologique solide du savoir enseigné (Bolon, 2002 ; Charnay, 1995 ; Bosch & Gascon, 2005).
- La différenciation de l'enseignement ne peut consister à proposer des tâches portant sur des objectifs très différents selon les élèves.

J'adhère totalement à ces conditions qui s'inscrivent pleinement dans la vision de l'évaluation au service des apprentissages que je défends. Il convient néanmoins de relever qu'au-delà de ces conditions, la mise en œuvre dans les classes de tels parcours dépend également des connaissances disciplinaires et didactiques des professeur·e·s qui ne pourront jamais être à la hauteur de celles d'un·e chercheur·e en didactique des mathématiques capable de définir une praxéologie de référence pour un domaine mathématique donné. L'expérimentation réalisée à partir du modèle de parcours différencié en trois étapes proposé par Pilet (2015), montre bien les limites rencontrées par un tel dispositif, notamment au regard de l'écart entre les praxéologies habituelles de l'enseignante et celles proposées par la chercheure (p. 31). Le fait que seulement quelques élèves aient tiré un bénéfice de ce parcours différencié témoigne des difficultés à mettre en œuvre un tel modèle dans les classes, bien qu'il soit impossible de tirer une généralité d'un seul exemple. Pilet préconise de prévoir un temps nécessaire d'appropriation des parcours par les enseignant·e·s. La thèse de Bedja (2016) s'inscrit dans la continuité des travaux de Grugeon-Allys et Pilet puisque cette chercheure a étudié, dans le cadre d'un groupe IREM, la façon dont le test de diagnostic en algèbre calcul algébrique disponible sur la plateforme LaboMeP ainsi que les parcours d'enseignement différencié développés dans la thèse de Pilet peuvent modifier les pratiques des enseignant·e·s ainsi que

leurs représentations personnelles sur ce nouveau type d'enseignement. De mon côté, c'est en termes de développement du jugement professionnel et didactique des enseignant·e·s que je conçois l'appropriation de tels modèles. J'y reviendrai dans la seconde partie de ma note de synthèse.

Pour finir la présentation des travaux qui se rattachent à cette partie et qui me semblent importants, je présenterai le travail que Chesné a réalisé dans le cadre de sa thèse en 2014 et qui se différencie des autres travaux présentés ci-dessus par le fait que ce chercheur a utilisé l'évaluation en amont de séances de formation et non d'enseignement. Il avait également utilisé des tests à l'issue de son dispositif de formation pour montrer quels effets il avait pu avoir sur les apprentissages des élèves selon que leur professeur·e ait été "associé·e", "témoin" ou "correspondant·e" au sein de ce dispositif. Le dispositif PACEM (Projet pour l'Acquisition de Compétences par les Élèves en Mathématiques) proposé par Chesné s'appuie sur l'exploitation d'évaluations standardisées à des fins diagnostiques (pour les élèves) et formatives (pour les professeur·e·s). Ce dispositif innovant, testé à grande échelle durant plusieurs années consécutives, s'appuie sur un test constitué d'items relevant du domaine "Nombres & calcul" du programme de 2008 (écriture des nombres décimaux et passage au quotient, calcul et résolution de problèmes) destiné à des élèves de CM2 (grade 5) et sixième (grade 6). Ce test proposé aux élèves en début d'année permet d'une part de mesurer l'impact de la formation dispensée selon les différentes modalités du dispositif (directe ou indirecte) et d'autre part servir de support pour des séances de formation s'appuyant sur les résultats des élèves à ce test. Je ne développerai pas plus précisément le dispositif de formation proposé car, dans cette partie de ma note de synthèse, je m'intéresse plus particulièrement à l'évaluation, mais je souhaite souligner l'originalité d'une telle expérimentation en France. En effet, jusque là les évaluations standardisées à visée diagnostique réalisées en France n'étaient pas "accompagnées" de formations. L'opportunité d'exploiter les résultats des élèves à ces tests est laissée à la discrétion des professeur·e·s de mathématiques qui n'en font pas toujours un usage constructif en classe. L'élaboration des tests (pré-test et post-test) est également un aspect du travail de Chesné qui m'intéresse car ils ont été conçus à des fins de passation à grande échelle pour les élèves, mais dans un but de formation pour les enseignant·e·s, ce qui est très innovant. Chesné le souligne d'ailleurs lui-même en précisant que :

> Tout d'abord, la formation est ancrée sur des évaluations standardisées : elle les intègre comme des objets de culture professionnelle pour les enseignants. En les décryptant avec eux et en faisant émerger des utilisations potentielles, le formateur s'en sert comme des outils pour faire évoluer, voire déconstruire en partie, certaines représentations des enseignants, puis pour accéder à leurs pratiques effectives et les faire évoluer dans une zone proximale de leur développement des pratiques. (p. 268)

C'est cette perspective qui lie évaluation et formation dans une visée de développement professionnel des enseignant·e·s que je défends également dans mes travaux et dans les formations que je dispense à l'ÉSPÉ de Créteil. D'ailleurs, les résultats positifs de PACEM en termes d'effet sur les apprentissages des élèves dans le domaine considéré (et même au-delà pour les élèves des professeur·e·s ayant participé deux ans au dispositif de formation) témoignent de l'intérêt d'une telle démarche, même si PACEM a bénéficié de conditions de développement et d'expérimentation à grande échelle assez inédites et difficilement reproductibles.

- ***L'évaluation en aval d'un enseignement ou d'une formation***

Dans sa thèse, Horoks (2006) a étudié et comparé les pratiques de cinq professeur·e·s autour de l'enseignement des "triangles semblables", en classe de Seconde (grade 10). Elle a cherché à mettre en relation l'organisation des enseignements sur les triangles semblables choisie par trois professeur·e·s dans leur classe avec les apprentissages qui pourraient en découler chez leurs élèves. Pour cela elle s'est intéressée aux contrôles donnés par les professeur·e·s en fin

de séquence comme un moyen de renseignement sur les connaissances acquises par les élèves sur cette notion géométrique, suite à l'enseignement dispensé. Elle a analysé les résultats des élèves au contrôle donné par les différent·e·s professeur·e·s en fin de chapitre et essayé de comprendre pourquoi et comment ils échouent ou réussissent. Pour cela, elle a comparé les exercices du contrôle avec tous ceux qui ont pu être proposés en classe précédemment, notamment du point de vue de la complexité des tâches proposées durant ces différents moments. A partir de l'étude approfondie de l'enseignement dispensé par les trois professeur·e·s qu'elle a étudié·e·s, elle a établi que :

- les erreurs des élèves sont en rapport avec le niveau de mise en fonctionnement des propriétés nouvelles évaluées au contrôle. Les élèves réussissent généralement moins bien lorsque les tâches sont plus difficiles, même si elles ont été travaillées précédemment en classe.
- les exercices de contrôle les mieux réussis par les élèves sont ceux qui ont été traités en classe, et plus particulièrement les applications qui ont, le plus souvent, été proposées.
- un travail répétitif d'adaptation d'une propriété ne garantit pas la réussite au contrôle d'une tâche similaire, bien qu'un peu plus difficile.

A l'issue de l'étude des contrôles donnés par les professeur·e·s de son échantillon Horoks a, tout en précisant que ce n'était pas l'objet de sa recherche et qu'elle ne pouvait les traiter, posé des questions qui sont toujours d'actualité :

> Comment le professeur conçoit l'énoncé de son contrôle et ce qu'il souhaite réellement tester chez les élèves ? Comment mesure-t-il la complexité d'une telle évaluation, et quels sont ses moyens pour en déduire ce qu'elle révèle de l'état réel des apprentissages des élèves ? (p. 161).

Chesnais (2014), de son côté, a cherché à caractériser la diversité et les régularités de pratiques d'enseignement autour de la symétrie axiale en classe de Sixième (grade 6) selon que les professeur·e·s enseignent dans des établissements d'éducation prioritaire ou non. Elle s'est intéressée aux scénarios proposés par les professeur·e·s et au déroulement de leurs séances en recherchant le poids des contraintes liées aux différents contextes d'enseignement (éducation prioritaire ou "ordinaire"). Elle a également recherché l'effet des pratiques observées sur les apprentissages des élèves en étudiant leurs résultats aux évaluations de fin de séquence sur la symétrie axiale. En étudiant ces évaluations et les tâches qui les constituaient, Chesnais a montré que les résultats des classes "ordinaires" étaient en moyenne meilleurs que ceux des classes en éducation prioritaire, mais que cela n'était pas systématique. En effet le caractère plus ou moins ambitieux des scénarios choisis par les professeur·e·s peut modifier ces résultats dans un sens ou dans un autre. Ainsi, un scénario ambitieux retenu par un·e professeur·e en éducation prioritaire peut avoir des répercussions positives sur les apprentissages des élèves, alors qu'un scénario peu ambitieux porté par un·e professeur·e dans un établissement "ordinaire" peut amener à des résultats en deçà des autres établissements quels qu'ils soient (prioritaire ou "ordinaire").

b- Les travaux sur l'évaluation comme objet d'étude

On peut considérer Jean Brun comme le précurseur de l'approche didactique de l'évaluation en mathématiques. En effet, dès 1979, il s'est intéressé à l'évaluation formative en lien avec les apprentissages des élèves c'est-à-dire "au sens où elle s'attache à repérer les états d'organisation, ou niveaux de représentation qu'ont construits les élèves et à partir desquels ils donnent un sens à tout nouveau problème" (p. 179). Il s'est appuyé sur les travaux de Vergnaud et Durand (1976) pour étudier à quelles conditions l'évaluation formative pouvait réellement favoriser les apprentissages des élèves en mathématiques et plus spécifiquement en résolution de problèmes additifs.

Chevallard et Feldman (1986) ont également été des précurseurs de cette approche. En écrivant que "(le didacticien) doit se rendre à l'évidence : les faits d'évaluation qu'il peut alors y (la classe) observer ne sont pas simplement un existant contingent, un mal nécessaire que l'on pourrait ignorer, mais bien l'un des aspects déterminants du processus didactique qui règle et régule tout à la fois les comportements de l'enseignant·e comme l'apprentissage des élèves" (préface, p. 2), ils ont posé les bases d'un cadre didactique pour l'évaluation des apprentissages mathématiques. Dans leur document, ils ont étudié finement à la fois les épreuves de contrôles d'un échantillon de professeur·e·s de mathématiques enseignant en collège et à la fois les fluctuations des notes des élèves d'une classe et ainsi défini les trajectoires des élèves au cours d'un trimestre. Ils ont montré comment la comparaison des variantes (moyennes, écart-types) permettait d'observer le travail de négociation des professeur·e·s et que "l'enseignant/évaluateur n'est nullement assimilable à un appareil de mesure susceptible d'opérer indéfiniment ; son univers de référence est clos, ses mesures ne sont nullement indépendantes entre elles, elles dépendent les unes des autres et des objectifs qu'il s'assigne (et qu'il modifie d'ailleurs) au cours de la conduite du processus didactique" (p.121). Ils parlent du mythe fondateur qui considère que la notation serait une opération de mesurage et la note une mesure et concluent en indiquant que "la note assignée n'est pas mesure, mais message. Ce message intervient dans une négociation, ou une transaction, qui signe un rapport de forces entre l'enseignant, les enseignés, à propos du savoir enseigné" (préface, p. 3).
La même année (1986), Noirfalise s'est intéressé aux attitudes des enseignant·e·s et à leur impact sur les résultats des élèves en mathématiques. Il a étudié leur influence sur les attitudes des élèves à l'égard des mathématiques, leurs connaissances en mathématiques et leurs capacités à résoudre des problèmes. Il a conclu qu'en géométrie les enseignant·e·s plus centré·e·s sur les élèves obtenaient de meilleurs résultats alors qu'en algèbre, c'était les professeur·e·s plus centrés sur les contenus qui avaient un meilleur impact.
En 1988, dans une publication de l'IREM d'Aix-Marseille (n°13) intitulée "Note sur la question de l'échec scolaire", Chevallard a longuement traité de l'échec scolaire et de la place que la Didactique pouvait avoir dans cette problématique. Il a indiqué que cette place se trouve dans l'étude des situations d'attribution d'échec (hétéro ou auto-attribution) qu'il a plus précisément nommés "verdicts d'échec". Il a souligné combien les enseignant·e·s, les élèves, les parents sont dépourvu·e·s face à de telles situations et que ce désarroi mène souvent à des conduites personnelles dont les effets, incontrôlés et empreints d'assujettissements personnels, sont négatifs. En conclusion, Chevallard précise que "l'analyse didactique des "échecs" et des dysfonctionnements, à tous les niveaux de l'institution scolaire, constitue un champ immense d'investigation, fondé sur les notions d'objets didactiques" (p.75). Il oppose cette analyse aux interprétations personnelles et individuelles auxquelles peuvent recourir les professeur·e·s (idéologie des dons, souvent prégnante en mathématiques) livré·e·s à eux ou elles-mêmes et démuni·e·s.
Dans un texte datant de 1990, Chevallard étudie plus spécifiquement le rôle de l'enseignant·e "évaluateur". Il y rappelle combien l'évaluateur ou l'évaluatrice est assujetti·e à l'institution à laquelle il ou elle appartient et que l'évaluation engage des sujets (élèves et professeur·e·s) dans leur singularité d'individus concrets. Il souligne que "ce qui importe avant toute chose, pour qu'un "fait évaluatif" soit reconnu comme tel, c'est la connexion qu'on voit s'y établir entre deux institutions ou, plus largement, deux objets institutionnels – l'objet "élève" et l'objet "professeur" par exemple –, incarnés en certains individus concrets" (p. 10).

Dans un texte publié en 1990, Chevallard, en s'appuyant sur la théorie anthropologique du didactique qu'il a développée, revient sur l'évaluation. Il s'intéresse à la situation d'examen où une personne (un·e examinateur·rice) doit évaluer les connaissances d'une autre personne

(candidat·e) par rapport à un objet o, à partir de la conformité du rapport R(x, o) au rapport institutionnel R_i (p, o) où p est la position que x est censé·e occuper au sein de I. Le R_i(p, o) est en fait "le rapport à l'objet o jugé nécessaire dans I pour assumer le rôle dévolu aux sujets en position p, c.à.d. pour prendre sa part, de manière idoine, dans l'ensemble des tâches dont l'accomplissement, par les techniques orthodoxes dans I, entraîne à la fois l'activation de o et l'intervention des sujets de I en position p" (p. 11-12). La non-conformité résulte du défaut d'usage de "la bonne technique selon I" alors même que x sait accomplir la tâche proposée, et qu'il ou elle connaît donc o à sa façon. Le problème réside justement dans ce Ri(o, p) que Chevallard qualifie lui-même de flottant et susceptible d'une négociation entre x et y, ou entre y et ses collègues, mais aussi entre les autres candidat·e·s x_i. Si on relit cette proposition théorique de l'évaluation des élèves à l'aune des effets dénoncés par les chercheur·e·s en docimologie (effets de source, d'ancrage, d'ordre, etc.) et des contraintes pesant sur l'enseignant·e (l'institution, les collègues, les parents, etc.), on perçoit la richesse et la pertinence de son approche car ces effets et contraintes pèsent simultanément sur l'enseignant·e en position d'évaluateur·rice et sur les élèves en position d'apprenant·e et ne sauraient être négligés. Dans l'étude de la conformité du rapport R(x, o) au rapport institutionnel R_i (p, o), on peut également voir une référence à l'évaluation critériée (Hadji, 1989), mais là encore l'approche de Chevallard permet d'aller plus loin et d'étudier plus finement en quoi l'évaluation des apprentissages des élèves est assujettie à la fois aux personnes, aux savoirs et aux institutions.
Il est dommage que Chevallard n'ait pas continué à étudier l'évaluation car son approche didactique, ancrée dans la réalité des pratiques de classes et des institutions, aurait pu influencer durablement et de façon plus marquante la communauté scientifique des didacticien·ne·s des mathématiques sur les questions d'évaluation.

Sylvie Coppé s'est également très tôt intéressée à l'évaluation, mais à partir d'une entrée singulière, centrée sur les élèves. Dans sa thèse (1993), elle a étudié les processus de vérification que développent les élèves lorsqu'ils ou elles sont en situation de devoir surveillé en mathématiques. Elle a élaboré une typologie des vérifications dans laquelle elle distingue les processus de vérification interne (mettant en jeu des savoirs ou des savoir-faire typiquement mathématiques, ne dépendant pas nécessairement de la situation dans laquelle on les utilise) et les processus de vérification externe (convoquant des connaissances portant sur d'autres savoirs ou savoir-faire du type "trouver un résultat entier", "tomber sur un point spécifique" en géométrie, etc.). Elle a ainsi montré que les pratiques de vérification lors des devoirs de mathématiques restent de l'ordre du travail privé (les vérifications faites sur le brouillon étaient rarement recopiées) et ne sont donc pas forcément données à voir à l'enseignant·e (1998). Elle considère que le devoir surveillé est un moment d'apprentissage important qui participe à l'institutionnalisation car il permet de montrer plus clairement, à certains élèves, ce que l'enseignant·e voulait qu'ils apprennent à travers son cours ou les exercices proposés.
Aujourd'hui, Sylvie Coppé (2015) s'intéresse à la place, la fonction et la nature de l'évaluation formative dans les pratiques de classe en mathématiques et s'interroge sur la nécessité à laquelle ce type d'évaluation peut correspondre dans le cours d'une étude. Dans le cadre d'un projet européen (ASSIST-ME, voir plus loin) elle a expérimenté des outils d'évaluation formative implantés dans une classe de Sixième et en a étudié les usages par les élèves et par l'enseignant·e. Elle en a conclu que leur viabilité suppose d'abord la mise en place en classe d'un contrat didactique laissant aux élèves la responsabilité de la validation de leurs réponses. Elle affirme donc que les méthodes d'évaluation formative doivent porter à la fois sur les connaissances mathématiques et sur les compétences en résolution de problèmes et que la diffusion de nouveaux outils ne peut se faire sans conditions sur les pratiques déjà en

place. C'est une des rares chercheur·e·s, en didactique des mathématiques, qui s'intéresse au concept de *Self-regulated learning* (Butler & Winne, 1995) évoqué dans la partie des travaux anglophones.

Dans sa thèse, Vantourout (2004) s'est, lui, intéressé aux compétences évaluatives des professeur·e·s des premier et second degrés en mathématiques, à partir des échanges et jugements évaluatifs qu'ils ou elles ont pu produire lors de situations "simulées" d'évaluation formative autour de la notion de proportionnalité et des représentations graphiques. Il a montré que les jugements de ces professeur·e·s, apparemment identiques, peuvent en fait reposer sur une grande diversité de connaissances et de processus et que les connaissances disciplinaires jouent un rôle essentiel dans la qualité des jugements émis. Avec Maury (2006), il a plus particulièrement étudié les connaissances disciplinaires en jeu lors de ces situations fictives d'évaluation et montré que des connaissances disciplinaires et didactiques des professeur·e·s acquises lors de leur formation initiale ou de leur cursus étaient indispensables pour conduire une analyse didactique pertinente des productions des élèves. Les professeur·e·s ayant un niveau de connaissances mathématiques et didactiques insuffisant (notamment les professeur·e·s des écoles n'ayant pas de cursus scientifique) sont les plus à même de produire des jugements erronés menant à des régulations inadaptées.

c- Les travaux qui s'intéressent aux évaluations standardisées

Antoine Bodin est indéniablement le didacticien qui a le plus étudié les évaluations standardisées nationales et internationales en mathématiques, en France et au-delà. En 1987, dans le cadre de l'APMEP (Association des Professeur·e·s de mathématiques de l'Enseignement Public), il a créé un observatoire de l'enseignement des mathématiques (EVAPM), chargé de recueillir et d'analyser des informations sur les conditions d'enseignement et sur l'état des acquis des élèves en mathématiques, de façon continue, en France. L'observatoire a ainsi commandité de nombreuses études qui ont permis d'accumuler un grand nombre d'observations et de résultats, notamment sur le niveau de maîtrise des élèves, à un moment donné, pour chacune des notions des programmes, sur l'évolution de cette maîtrise dans le temps de l'élève et dans le temps tout court ; sur les relations entre différents domaines (géométrique et numérique par exemple) et également sur les différences de réussite entre redoublant·e·s et non-redoublant·e·s, entre élèves destiné·e·s à des orientations différentes, entre garçons et filles, etc.

L'organigramme de cet observatoire montre comment le travail réalisé en son sein irrigue à la fois les pratiques des enseignant·e·s de mathématiques et les questionnements des chercheur·e·s.

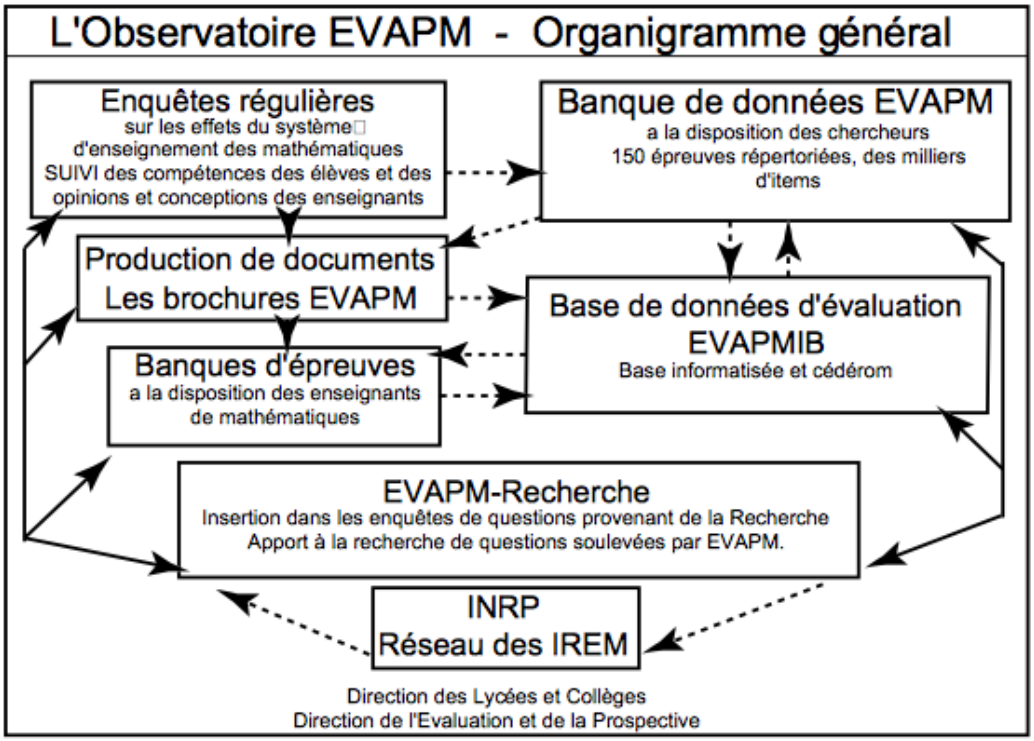


Figure 2 : organigramme général de l'EVAPM (daté de 2012)

Bodin (2007) a également mis en évidence le manque de cohérence de l'ensemble des actions d'évaluation menées dans et hors les classes et la nécessité de réduire la dissonance constatée entre ces différentes actions. Il a souligné l'importance de distinguer "enseigner" et "apprendre" ainsi que la nécessité de préciser les places et les rôles respectifs de chacun (enseignant·e et enseigné·e) au préalable de toute réflexion sur l'évaluation.
Du côté des évaluations standardisées internationales, Bodin a également réalisé beaucoup d'analyses des résultats du PISA, de TIMSS, etc. Dans la synthèse qu'il a produite en 2006 (et également en 2010), il a présenté les cadres théoriques sur lesquels s'appuient ces enquêtes, ainsi que leur évolution. Il s'est intéressé à leur validité et à la signification qu'il convenait de donner aux résultats produits dans ces évaluations standardisées. Ses analyses ont l'intérêt de donner une visibilité nationale à ces enquêtes internationales, de montrer leur complémentarité. Elles participent à la "démystification" de ces enquêtes (2006), à mieux les connaître pour apprécier, à leur juste valeur, les résultats qu'elles produisent. Elles ont aussi le mérite de montrer la place qu'occupe la France dans ces enquêtes et d'être "une bonne occasion de s'interroger sur la solidité et sur la qualité de notre système" (2006, p. 27). En novembre 2016, il a été l'auteur principal d'un rapport d'envergure publié par le CNESCO sur TIMSS et PISA permettant de mieux appréhender la spécificité de ces enquêtes internationales. Les analyses des items de ces deux enquêtes ont été réalisées à partir du cadre habituel adopté pour les évaluations de l'EVAPM (taxonomie de Gras, revue par Bodin), mais enrichi par des outils issus de la Didactique des mathématiques : la dialectique outil/objet de Douady (1986) et les niveaux de mise en fonctionnement des connaissances de Robert (1998). Roditi et Salles (2015) ont également exploité les niveaux de mise en fonctionnement des connaissances mathématiques de Robert pour analyser les items PISA 2012. Ils ont ainsi montré que les classifications utilisées par l'OCDE ne permettent ni de recenser précisément les connaissances acquises des élèves, ni d'estimer le niveau d'acquisition de ces connaissances.

Au-delà de la qualité de ses analyses, Bodin interpelle les chercheur·e·s en didactiques pour qu'ils s'emparent des évaluations standardisées pour "prendre de la distance par rapport aux contextes locaux et gagner une compréhension globale des phénomènes liés à l'enseignement et l'apprentissage des mathématiques" (2006, p. 17). C'est justement ce qu'a fait Nadine Grapin dans le cadre de sa thèse que j'ai co-dirigée avec Brigitte Grugeon-Allys.

Dans sa thèse, Grapin (2015) a abordé la question de l'évaluation sous deux angles : l'étude de la validité des évaluations externes et le développement d'un modèle d'analyse multidimensionnelle des connaissances numériques des élèves. Elle a développé une méthodologie d'analyse de la validité de dispositifs d'évaluation en articulant des approches didactique, épistémologique et cognitive, en complément d'approches psychométriques, spécifiques aux évaluations à grande échelle. Grapin s'est plus particulièrement focalisée sur les connaissances des élèves en fin d'école primaire dans le domaine de la numération évaluées dans le bilan CEDRE 2008 et 2014 pour éprouver la validité de cette enquête nationale. Le modèle multidimensionnel qu'elle a développé pour étudier les connaissances numériques des élèves lui a permis de concevoir une évaluation diagnostique de fin de cycle 3 visant, comme Pilet (2012), à définir des parcours différenciés pour la classe. Ce modèle s'appuie sur une double approche anthropologique et cognitive et nécessite une caractérisation des connaissances numériques des élèves en plusieurs dimensions qu'elle a réalisée en s'inspirant des travaux de Grugeon-Allys (1995, 1997), notamment de *Pépite* en algèbre (voir plus haut). Les dimensions ont été décrites à partir des technologies impliquées qui les caractérisent et selon quatre modes technologiques (technologies correctes et attendues, correctes mais non attendues, incorrectes mais faisant référence à des technologies attendues, incorrectes sans référence à des technologie attendues). À partir de l'analyse du domaine réalisée à l'aide de ce modèle multidimensionnel, elle a défini des profils d'élèves présentant des cohérences de fonctionnement se traduisant par des technologies dominantes.
Le travail de Grapin a non seulement permis de rendre compte, d'un point de vue didactique, de la validité d'une telle évaluation, mais il a aussi permis de montrer l'intérêt qu'il y avait à le faire puisque sa visée finale est de définir des profils d'élèves pour penser et concevoir un enseignement différencié concernant les connaissances numériques. Cette approche qui part de l'étude d'évaluations externes en mathématiques pour développer des outils au service de l'enseignement des mathématiques pour la classe est une approche à laquelle j'adhère totalement. Elle permet de penser conjointement évaluation et enseignement dans une perspective d'amélioration des apprentissages des élèves. Je pense en effet que les évaluations externes pourraient être davantage exploitées en tant que ressources pour l'enseignement et l'évaluation des apprentissages mathématiques des élèves en classe. Je reviendrai ultérieurement sur ce point pour expliciter ce que j'entends par là.

Ruminot Vergara (2014) a également utilisé des outils et des cadres didactiques pour étudier une évaluation externe nationale. Dans sa thèse, elle a étudié l'impact d'une évaluation standardisée au Chili (SIMCE, *Système de Mesure de la Qualité de l'Éducation*) sur les pratiques des enseignant·e·s de huitième année d'enseignement dans le domaine de la géométrie (et des grandeurs) et sur leur formation. Elle s'est appuyée sur la Théorie Anthropologique du Didactique (Chevallard, 1992, 1999, 2002) et notamment sur les différents niveaux de codétermination didactique pour mieux comprendre les effets de cette évaluation standardisée nationale sur les pratiques des enseignant·e·s et prendre en compte la diversité des conditions et contraintes en jeu dans les classes au Chili. Pour étudier les items du domaine retenu, elle s'est appuyée sur la notion de paradigme géométrique définie par Houdement et Kuzniak (2006). Pour appréhender plus spécifiquement les caractéristiques de l'évaluation SIMCE, elle a réalisé une comparaison avec d'autres évaluations à grande échelle auxquelles sont également soumis les élèves chiliens (PISA, TIMSS et SERCE). Pour réaliser

cette comparaison elle a utilisé un outil proposé par Artigue et Winsløw (2010) qui associe à la hiérarchie des niveaux de codétermination dix niveaux de comparaison possibles et qui permet d'effectuer des comparaisons d'évaluations internationales selon les contextes (comparaison de type horizontal entre deux ou plusieurs contextes pour un même niveau ou de type vertical entre les niveaux dans chaque contexte). Cette comparaison s'est faite à partir de trois catégories structurelles (les domaines mathématiques, les processus cognitifs et les niveaux de réussite) qui ont permis de caractériser plus précisément l'évaluation SIMCE. Ce qui est également intéressant dans ce travail, c'est que Ruminot Vergada a cherché à comprendre et à analyser la vision des enseignant·e·s sur cette évaluation nationale, mais aussi si elle influençait d'une façon ou d'une autre leurs pratiques. Elle a distingué des profils d'enseignant·e·s suivant quatre dimensions qui les caractérisent en situation et par rapport à SIMCE : la formation et l'expérience, le positionnement dans l'institution, la relation à SIMCE et la sensibilité didactique. Cette étude participe donc explicitement à l'exploitation didactique des interactions entre évaluations externes et enseignement que j'appelle de mes vœux.

Pour finir, j'évoquerai brièvement les travaux de Nathalie Loye qui s'est plus particulièrement intéressée aux difficultés en mathématiques des élèves en formation professionnelle au Québec. A partir d'un modèle statistique de classification diagnostique (MCD) qu'elle a développé (Loye, 2010), elle a cherché à identifier les différents attributs (habiletés, connaissances, savoirs et savoir-faire) d'une tâche mathématique et ainsi mieux comprendre les difficultés qu'elle pouvait engendrer. Elle a conçu des tests mathématiques permettant un diagnostic des besoins et des difficultés de ces élèves spécifiques.

Pour compléter cette présentation de travaux francophones sur l'évaluation, je vais à présent brièvement exposer les principaux grands projets terminés ou en cours autour des problématiques d'évaluation en mathématiques.

d- Quelques grands projets de recherche sur l'évaluation en Mathématiques

- ***Le projet ASSIST-ME*** *(Assess Inquiry in Science, Technology and Mathematics Education)*

Ce projet européen qui s'est terminé à la fin de l'année 2016 a eu pour objectifs d'analyser l'influence de nouveaux dispositifs d'évaluations formatives en lien avec les évaluations sommatives dans le cadre de démarches d'investigations, sur les apprentissages et les pratiques enseignantes en Sciences, Mathématiques et Technologie, mais également de concevoir et de diffuser des méthodes d'évaluation formative dans sept pays et dans différents niveaux de classe. L'enjeu de ce projet était de faire évoluer les pratiques enseignantes vers un enseignement scientifique fondé sur l'investigation – EMI – de manière à favoriser la pratique scientifique des élèves, et ainsi leur faire apprendre des mathématiques. L'évaluation formative est, dans ce projet, vue comme un moyen de développer la mise en œuvre par les enseignants d'un EMI.
Dans le cadre de ce projet, Sylvie Coppé et son équipe ont réalisé plusieurs expérimentations sur l'évaluation entre pairs en utilisant l'argumentation à partir de séances de résolution de problèmes complexes en mathématiques (Coppé & Moulin, soumis ; Coppé, Moulin & Roubin, à paraître ; Gandit & Levasseur, à paraître). Dans un premier temps du projet, ces chercheur·e·s ont élaboré une séquence de classe dans laquelle les élèves de deux classes de CM1-CM2 (élèves de 9-10 ans) et quatre classes de 6^{e} (élèves de 11-12 ans) devaient résoudre un problème, puis se prononcer sur la validité des réponses des autres élèves de la classe par écrit avant et après un débat argumentatif. Ils·elles ont étudié la question des

conditions en termes de milieu et contrat didactique et de l'organisation à mettre en place par l'enseignant·e pour une mise en commun productive. Ils et elles ont ainsi pu montrer que les effets des débats ne sont pas toujours positifs et que l'enrichissement du milieu par les réponses des élèves ne donne pas suffisamment de rétroactions pour avoir des effets sur les apprentissages (les fractions dans le cas du problème proposé). Ces chercheur·e·s ont donc conclu qu'il était nécessaire de réfléchir à la forme et à l'organisation de ces débats lors de séances de ce type. Lors de cette première expérimentation, ils et elles ont également étudié, à partir des vidéos des séances, les feedbacks des professeur·e·s en cherchant à déterminer leurs effets sur les élèves d'une part (en termes de connaissances mais aussi d'attitudes et de compétences) et sur les professeur·e·s d'autre part (quelles décisions de régulation sont prises ? Comment le cours de la séance est-il modifié ? Quel traitement différencié peut être réalisé ?).

Dans un deuxième temps du projet, ces chercheur·e·s ont souhaité étudier ces méthodes d'évaluation formative (évaluation entre pairs et débats argumentatifs) dans le cadre de séances plus ordinaires portant sur une séquence d'algèbre élémentaire incluant des phases de dévolution et d'institutionnalisation. Ils·elles ont plus spécifiquement étudié deux activités utilisant des programmes de calcul dans des classes de 4^{e} et 3^{e} de collège (la traduction d'un programme de calcul en écriture algébrique et une tâche de preuve en algèbre et sur la technique permettant de la réaliser). Ces activités visent à initier un changement de contrat didactique à travers une redistribution des responsabilités entre l'enseignant·e et ses élèves et à réintroduire dans le milieu des éléments technologiques (au sens de Chevallard, 1999) permettant aux élèves de vérifier ou prouver leurs productions.

Le projet, qui s'est achevé à l'automne 2016, a abouti à une liste de recommandations pour la mise en œuvre d'une évaluation formative dans l'enseignement scientifique. Concernant les pratiques de classe, il est ainsi préconisé de :

- Donner l'habitude aux élèves de s'évaluer et d'évaluer les autres en s'appropriant les critères d'évaluation et donc les enjeux des savoirs et savoir-faire (par exemple, mettre en œuvre l'évaluation par les pairs, dans la continuité de l'enseignement).
- S'assurer de la cohérence des évaluations formatives avec les évaluations sommatives (mêmes critères d'évaluation, mêmes compétences en jeu, etc.).
- Au lycée, expliciter les sous-compétences aux élèves pour qu'ils s'approprient les critères sur lesquels ils sont évalués, les savoirs et savoir-faire à comprendre et acquérir.
- Au collège, décliner les compétences du socle en indicateurs selon plusieurs niveaux (quatre ou plus) pour quelques compétences clés en équilibrant les compétences spécifiques à la séquence et les compétences transversales.
- Utiliser des tableaux de progression (des grilles d'évaluation, des fiches de critères) intégrés dans l'activité même des élèves.
- Élaborer des fiches d'aide, des coups de pouces, permettant à tous les élèves de comprendre le minimum requis.
- Au lycée, décliner les compétences attendues en sous-compétences (capacités) en les reliant aux savoirs en jeu dans les différentes activités, cours, exercices, TP.
- Concevoir des activités qui permettent en commençant le chapitre d'établir des connaissances et un vocabulaire communs pour que, dès le début, tous les élèves puissent contribuer au travail sur les savoirs et savoir-faire. Cela facilitera l'aide aux élèves par une évaluation sur le champ à toutes les étapes du chapitre (partie, thème).
- Anticiper au maximum les réponses que les élèves pourraient donner de manière à se préparer à rétroagir, sans se faire surprendre par des réponses qui pourraient paraître déroutantes.

- ***Le projet européen FaSMEd*** (*Formative Assessment in Science and Mathematics Education*)

Ce projet traitant des apports des technologies dans le processus d'évaluation formative regroupe huit pays européens, dont la France. Gilles Aldon et Monica Panero, qui sont les chercheur·e·s français·e·s engagé·e·s dans ce projet, émettent l'hypothèse que la technologie numérique favorise le processus d'évaluation formative notamment en ce qui concerne le recueil des données, leur traitement et le renvoi d'informations à l'élève. Ils·elles ont étudié, dans le cadre du LéA *Parc Chabrières*, les remédiations "à chaud" et "à froid" que permettent les propriétés des technologies concernant en particulier les possibilités de partage et d'analyse des données recueillies dans la classe. Ils·elles s'appuient sur un modèle de transposition *Méta*-Didactique (Arzarello, Robutti, Sabena, Cusi, Garuti, Malara & Martignone, 2014) fondé sur la Théorie Anthropologique du Didactique pour décrire et interpréter les processus qui se mettent en place quand des communautés de chercheur·e·s et des enseignant·e·s interagissent ensemble dans un but de développement professionnel. Leur conclusion souligne l'importance de prendre en compte l'aspect didactique dans l'évaluation formative.

- ***Le projet ANR NéoPraEval*** (*Nouveaux Outils pour de nouvelles PRAtiques d'EVALuation et d'enseignement des mathématiques*)

Ce projet financé par l'Agence National de Recherche et porté par Brigitte Grugeon-Allys rassemble des chercheurs de plusieurs champs scientifiques (Didactique des mathématiques, Informatique, Sciences de l'Éducation). Il vise à outiller les enseignant·e·s pour gérer l'hétérogénéité des apprentissages mathématiques de leurs élèves en mettant à leur disposition des outils d'évaluation diagnostique automatique utilisables dans leurs classes ainsi que des ressources appropriées aux besoins repérés des élèves. Il s'appuie sur divers travaux scientifiques développés sur le dispositif *Pépite* (Delozanne, Prévit, Grugeon-Allys & Chenevotot, 2010 ; Grugeon-Allys, Pilet, Chenevotot & Delozanne, 2012 ; Pilet, 2012), sur l'évaluation du dispositif CEDRE en mathématiques (Sayac & Grapin, 2013, 2014) et sur les pratiques enseignantes (Robert & Rogalski, 2002 ; Roditi, 2011 ; Horoks, 2006). Ce projet d'envergure, programmé sur trois ans (2014-2017) s'articule autour de trois tâches :

Tâche 1 : Développer une expertise pour étudier la validité des outils d'évaluation et concevoir des dispositifs d'évaluation.
Tâche 2 : Utiliser cette expertise pour étendre des dispositifs d'évaluation existants.
Tâche 3 : Analyser les pratiques d'évaluation des enseignant·e·s en classe.

Je ne développerai pas, dans cette note, les travaux attachés aux tâches 1 & 2 qui concernent des dispositifs d'évaluation dans des environnements informatiques car je m'intéresse plus spécifiquement à l'évaluation "ordinaire" en classe. Je rendrai donc compte de l'avancée des travaux de la tâche 3 auxquels je participe et qui s'inscrit plus directement dans l'approche didactique de l'évaluation en classe de mathématiques que je souhaite promouvoir.
Cette tâche, placée sous la responsabilité de Éric Roditi, a trois objectifs (tels que définis dans le projet en 2014) :
- Le premier objectif consiste à enrichir et affiner, en fonction des contenus mathématiques enseignés, les catégories permettant de décrire les pratiques enseignantes quant à la programmation de leur enseignement, quant à la régulation des situations d'apprentissage en classe et quant aux évaluations sommatives et formatrices qu'ils mettent en œuvre.
- Le deuxième objectif est la mise en relation des observables relevant de ces catégories afin de mettre au jour diverses cohérences des pratiques enseignantes.
- Le troisième objectif consiste à établir des critères portant sur la formation mathématique et didactique des enseignant·e·s comme sur leur connaissance des difficultés des élèves.

Les chercheur·e·s contribuant à cette tâche (Eric Roditi, Brigitte Grugeon-Allys, Mariam Haspekian, Julie Horoks, Michella Kiwan, Julia Pilet et moi-même) sont engagé·e·s dans différents projets de recherche et notamment deux LéA (Lieux d'Éducation Associé) qui leur permettent de produire des résultats en lien avec les objectifs de cette tâche. Je présenterai donc brièvement les travaux de Roditi, Kiwan et Haspekian sur la notion de feedback ainsi que ceux de Horoks, Pilet et Haspekian sur les pratiques d'évaluation en algèbre, au collège. Je développerai ceux qui me concernent autour des pratiques d'évaluation en numération à l'école élémentaire dans la partie afférente à mes travaux.

Après avoir réalisé une revue de littérature sur les concepts d'évaluation formative, de régulation et de feedback, Haspekian et al. (2016) ont montré la nécessité d'une référence théorique sur l'apprentissage et sur l'activité de l'enseignant·e pour mettre en lumière la dimension évaluative des échanges entre élèves et professeur·e·s sans en perdre la dimension de transmission de savoirs mathématiques. Ils et elles ont élaboré une telle référence permettant de fonder leurs analyses sur une base théorique cohérente pour l'enseignant·e et pour l'élève. Leur approche est à la fois didactique et écologique car ils et elles souhaitent analyser les pratiques d'évaluation formative des enseignant·e·s en lien avec les apprentissages visés et telles qu'elles sont à l'œuvre dans le quotidien de l'enseignement. Ils·elles étudient les interactions enseignant·e/élèves en classe et les considèrent, du point de vue de la théorie de l'activité à laquelle ces chercheur·e·s se réfèrent, comme un processus dynamique adaptatif qu'ils·elles ont désigné par le terme de "régulation didactique". Ces interactions sont porteuses de feedback ainsi que d'évaluations formatives "on-line" ou d'hétéro-régulations improvisées. Ils précisent :

> Cette façon d'analyser chaque interaction élève/enseignant correspondant à une hétéro-régulation conduit à identifier, pour l'activité d'évaluation formative de l'enseignant, un couple information-feedback dont l'information comme le feedback peuvent être associés à un résultat (R), une procédure (P) ou un état de connaissance (C). Chaque couple information-feedback peut ainsi être classé dans le tableau à double entrée suivant où figurent les neuf possibilités de couples RR, RP, RC, PR, PP, PC, CP, CP, CC. (p. 21)

Pour déterminer la tendance de pratique d'évaluation formative d'un·e professeur·e, ces chercheur·e·s proposent donc de réaliser une analyse didactique qualitative de chaque interaction qui se déroule dans sa classe et de leur associer un type de régulation parmi les neuf possibles. En comparant la somme des différents types d'interactions associées à chaque professeur·e, ils ou elles peuvent ainsi déterminer des différences inter enseignant·e·s, mais également une éventuelle variabilité de pratique pour un·e même enseignant·e c'est-à-dire des différences intra enseignant·e·s.

Les premières analyses réalisées avec cet outil leur ont permis de mettre à jour une grande variabilité de pratiques d'évaluation formative des enseignant·e·s qu'ils et elles ont étudié·e·s ainsi qu'une certaine concentration des couples information-feedback autour de cinq ou six des neuf types possibles, le couple RR (Résultat/Résultat) s'étant avéré dominant chez tous les professeur·e·s. Parmi les différentes "régulations didactiques" possibles, ils et elles ont distingué les "régulations didactiques horizontales" où les enseignant·e·s proposent des feedbacks qui sont au même niveau que l'information reçue c'est-à-dire au même niveau que la réponse de l'élève (RR, PP ou CC) et les "régulations didactiques verticales" où les enseignant·e·s, dans leur retour aux élèves, changent de niveau (RP, RC, PR, PC, CR, CP). Cette classification les a amené·e·s à s'interroger sur le type de régulations (horizontales ou verticales) qui favoriserait le plus les apprentissages des élèves. Cet outil didactique nouveau et prometteur gagnera à être expérimenté à plus large échelle pour valider les hypothèses de ces chercheur·e·s et pourrait également s'avérer utile en formation pour promouvoir une évaluation au service des apprentissages en mathématiques des élèves.

Dans le cadre du LéA Pécanuméli, Horoks, Pilet et Haspekian (2015) ont, de leur côté, étudié les pratiques de plusieurs professeur·e·s autour de l'enseignement et de l'évaluation de l'algèbre élémentaire. Elles se sont plus particulièrement focalisées sur l'évaluation formative et ont cherché à comprendre et à étudier comment ce type d'évaluation peut se réaliser dans les classes. Pour cela, elles ont pris en compte à la fois les spécificités des contenus enseignés, l'algèbre élémentaire du collège (Grugeon-Allys & al., 2012), et à la fois les activités de l'enseignant·e en classe, plus spécialement la manière dont ils ou elles impliquent les élèves dans leurs apprentissages. Elles ont étudié les retours proposés par les enseignant·e·s de leur LéA sur les productions des élèves, et, plus largement, le contrat didactique mis en place pour responsabiliser les élèves dans les processus d'apprentissage. Elles se sont heurtées à la difficulté de repérer des traces de cette évaluation formative diffuses dans l'ensemble des pratiques de l'enseignant·e et dans les interactions entre élèves et professeur·e. Pour appréhender plus efficacement ces indices d'évaluation formative, elles ont choisi de se pencher sur des moments de résolution de tâches et plus particulièrement sur la mise en commun des productions des élèves, après la recherche d'une solution à un problème algébrique posé. Elles ont montré que l'évaluation formative effectuée par les enseignant·e·s du second degré se déroule, pour une part très importante, durant les échanges en classe avec leurs élèves et plutôt de manière informelle. Plus précisément, elles ont montré que, pour certain·e·s enseignant·e·s, ces moments d'évaluation avait une fonction purement régulative, alors que pour d'autres elles avaient aussi une autre fonction pour les élèves (fonction d'apprentissage et fonction de correction/compréhension des erreurs, fonction de comparaison des méthodes ou des procédures, fonction de réflexion sur celles qui sont les plus économiques, etc.). Elles ont conclu que, de leur point de vue, la fonction de l'évaluation n'est pas déterminée à l'avance par les enseignant·e·s, mais que cette fonction dépend de ce qui a été récolté comme informations au cours de la séance.

Les travaux que je mène actuellement s'inscrivent également dans la tâche 3 du projet ANR NéoPraEval, mais en complément des travaux de mes collègues, ils concernent les pratiques d'évaluation plutôt sommative des enseignant·e·s du premier degré. Je les préciserai dans le paragraphe suivant, après avoir indiqué comment ils se situent par rapport à mes premiers travaux sur l'évaluation en mathématiques.

C. MES TRAVAUX SUR L'ÉVALUATION

1. Première étude : analyse des items du bilan CEDRE 2008 en Mathématiques

En 2007, j'ai été sollicitée pour participer à un groupe d'expert·e·s de la DEPP chargé·e·s d'élaborer les items du bilan CEDRE 2008 mathématiques de fin d'école. J'ai intégré ce groupe en tant que formatrice IUFM et non en tant que chercheure, mais au fur et à mesure que les items étaient élaborés et testés sur des élèves, je m'interrogeais sur leur validité didactique (que je n'appelais pas comme cela à cette époque). Lorsqu'en 2009 les résultats des bilans CEDRE école & collège ont été présentés officiellement, je me suis aperçue qu'une autre collègue didacticienne (Nadine Grapin) avait participé au groupe d'expert·e·s du collège et que les deux groupes avaient travaillé indépendamment l'un de l'autre. Ce constat surprenant (parce que les apprentissages mathématiques de l'école et du collège doivent être pensés dans une même logique et dans une certaine continuité) et l'opportunité de travailler

avec cette collègue m'ont amenée à m'emparer de ces bilans et de leurs résultats pour les étudier d'un point de vue didactique.
Dans un premier temps, nous avons travaillé à l'élaboration d'un outil d'analyse des items du bilan CEDRE 2008 de fin d'école afin de déterminer leur complexité *a priori* du point de vue de facteurs intégrant des résultats issus de la Didactique des mathématiques, mais aussi d'autres facteurs. En effet, nous avons été interpellées par certains items CEDRE 2008 qui, du point de vue de la tâche mathématique à réaliser, pouvaient être perçus comme équivalents alors que leurs résultats étaient très différents. Pour illustrer ce fait, voici deux problèmes issus items du bilan CEDRE 2008 de fin d'école primaire :
Problème 1 : Monsieur Paul achète 9 rosiers à 4€ et 3 sapins à 17€ pièce. Quel est le montant de sa dépense ?
Problème 2 : Monsieur Jacques achète 8 cahiers et 5 stylos. Le prix d'un cahier est de 3€. Le prix d'un stylo est de 2€. Quel est le montant de sa dépense ?
Alors que ces problèmes se résolvent tous deux par le calcul de sommes algébriques (9x 4 + 3 x 17 et 8 x 3 + 5 x 2) et même si nous avions anticipé un score de réussite moindre pour le premier problème parce que les nombres en présence conduisent à des opérations un peu plus "difficiles" pour les élèves, nous avons été surprises de l'écart de réussite important entre les deux problèmes (le premier problème a été réussi par 62,95%, alors que le second l'a été par 80,73% des élèves).

Les facteurs de complexité
L'outil que nous avons conçu a donc cherché à prendre en compte différents paramètres intervenant dans la complexité *a priori* d'un item (Sayac & Grapin, 2015). Il se compose de trois facteurs (deux facteurs de complexité et un facteur de compétences) se déclinant eux-mêmes en trois niveaux (du moins au plus complexe). Ces trois niveaux prennent en compte les niveaux scolaires, c'est-à-dire que des niveaux de complexité différents peuvent être attribués à une même tâche selon le niveau de classe dans lequel la tâche a été donnée :
- Facteur de complexité 1 (FC1) : ce facteur est lié à la façon dont les élèves sont amenés à comprendre la tâche mathématique qu'ils doivent réaliser. Dans ce facteur, le niveau de langue de l'énoncé ainsi que la nature des informations à traiter (texte, graphique, schéma, etc.) sont pris en compte pour déterminer comment l'élève est amené à comprendre, de façon plus ou moins aisée, la tâche qu'il doit réaliser. Est également pris en compte pour ce facteur le contexte de l'énoncé suivant qu'il soit plus ou moins proche de la "vie réelle" de l'élève.
Pour illustrer, ces trois niveaux, voilà trois façons plus ou moins complexe de demander aux élèves de décomposer un nombre par unités de numération :
Niveau 1 : décompose chaque nombre comme dans l'exemple : 4 567 = 4000 + 500 + 60 + 7
Niveau 2 : écris chaque nombre en le décomposant unité par unité (millier, centaine, dizaine, unité).
Niveau 3 : décompose chaque nombre en écrivant sa décomposition additive donnant la valeur de chaque chiffre.

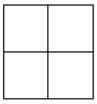
- Facteur de complexité 2 (FC2) : Ce facteur est directement lié au savoir mathématique en jeu. De ce point de vue, la tâche à réaliser peut être plus ou moins complexe, nous nous référons aux divers travaux effectués en Didactique des mathématiques dans les différents domaines concernés par l'évaluation pour déterminer son niveau de complexité. Dans ce facteur, sont également pris en compte les variables didactiques, ainsi que les distracteurs proposés dans les items car ils peuvent avoir une influence non négligeable sur la réussite des élèves, dans un sens positif ou négatif.

Par exemple, pour la tâche "écrire en chiffres les nombres" qui peut être plus ou moins complexe selon les nombres d'unités proposés, impliquant ou non des conversions ou la présence de zéro(s), voilà les niveaux proposés[5] :
Niveau 1 : ni conversion, ni zéro (ex : 5 centaines, 3 dizaines et 4 unités)
Niveau 2 : (zéro à au moins une des places) (ex : 7 centaines, 3 unités)
Niveau 3 : (conversion) (ex : 8 centaines, 12 dizaines et 45 unités)

Le troisième facteur est relatif aux compétences en jeu pour résoudre la tâche. La définition que nous avons adoptée a été spécialement conçue pour analyser la complexité des items du bilan CEDRE, mais elle me semble valable au-delà car elle fait sens pour n'importe quelle tâche mathématique, qu'elle soit proposée dans le cadre d'une évaluation ou d'une activité de classe.
- Facteur de compétences (FC) : ce facteur est lié aux différents niveaux de mise en fonctionnement des connaissances (technique, mobilisable, disponible) de Robert et Rogalski (2002) et aux adaptations de Robert (2008). Il intègre également une dimension liée au caractère inédit ou non de la tâche.
Niveau 1 : pour les tâches qui amènent à des applications immédiates des connaissances, c'est-à-dire simples (sans adaptation) et isolées (sans mélange), où seule une connaissance précise est mise en œuvre sans aucune adaptation, mis à part la contextualisation nécessaire. Ces tâches sont, par ailleurs, généralement usuelles.
Niveau 2 : pour les tâches qui nécessitent des adaptations de connaissances qui sont en partie au moins indiquées. Ces tâches peuvent être plus ou moins usuelles.
Niveau 3 : pour les tâches qui nécessitent des adaptations de connaissances qui sont totalement à la charge de l'élève. Ces tâches sont souvent inédites.

Pour illustrer ce facteur, voici trois niveaux de complexité autour d'une tâche de mise en correspondance entre une surface à hachurer et une fraction donnée.

Niveau 1 : Hachurez la surface correspondant à la fraction ¼ dans la figure ci-contre :

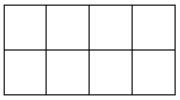


Cette tâche relève d'un bas niveau de compétences puisque l'élève doit simplement appliquer ses connaissances pour réaliser une tâche extrêmement familière pour lui.

Niveau 2 : Hachurez la surface correspondant à la fraction ¼ dans la figure ci-contre :

L'élève doit cette fois prendre à sa charge le découpage en quatre parties égales du carré, mais la figure proposée le permet aisément et il a pu déjà être confronté à ce type de tâche auparavant.

Niveau 3 : Hachurez la surface correspondant à la fraction $\frac{5}{10}$ dans la figure ci-contre :



Cette fois, l'élève est confronté à une tâche relevant d'un haut niveau de compétences puisqu'il doit d'abord réaliser que la fraction 5/10 est égale à la fraction ½ avant de pouvoir

[5] Tous les exemples proposés pour illustrer les différents niveaux de complexité proviennent d'évaluations récoltées au cours de mes recherches.

hachurer les quatre carreaux correspondants à la moitié du rectangle. Pour réaliser cette tâche l'élève n'est en rien guidé dans sa démarche et n'a certainement pas été confronté à ce type de tâche.

Cet outil a été appliqué à tous les items relatifs aux "Fractions & décimaux" et "Grandeurs & mesures" du bilan CEDRE 2008 afin d'étudier dans quelle mesure le bilan était équilibré du point de vue des connaissances et des compétences de fin d'école dans ces deux domaines. Les tableaux ci-dessous rendent compte de la répartition des items dans les deux domaines en fonction des différents facteurs de l'outil :

FACTEUR DE COMPLEXITE 1 : FC1			**FACTEUR DE COMPLEXITE 2 : FC2**			**FACTEUR DE COMPETENCES : FC**		
Niveau 1	Niveau 2	Niveau 3	Niveau 1	Niveau 2	Niveau 3	Niveau 1	Niveau 2	Niveau 3
12 %	76%	12%	44%	42%	14 %	82 %	18%	0%

***Tableau 1** : items relatifs au domaine "fractions & décimaux"*

FACTEUR DE COMPLEXITE 1 : FC1			**FACTEUR DE COMPLEXITE 2 : FC2**			**FACTEUR DE COMPETENCES : FC**		
Niveau 1	Niveau 2	Niveau 3	Niveau 1	Niveau 2	Niveau 3	Niveau 1	Niveau 2	Niveau 3
19 %	71 %	10%	74%	19%	7%	69 %	19%	12%

***Tableau 2** : items relatifs au domaine "Grandeurs & Mesures"*

L'utilisation de cet outil permet de montrer que le bilan CEDRE 2008 est peu équilibré du point de vue des niveaux de complexité des trois facteurs considérés, dans les deux domaines étudiés. En effet, on constate qu'il y a une grande majorité d'items de faible niveau de complexité pour les trois facteurs, quelques items de haut niveau et peu d'items intermédiaires. Par ailleurs, les items relevant du niveau le plus élevé du facteur de compétences sont également, pour plus de la moitié d'entre eux, d'un niveau élevé du point de vue du deuxième facteur (FC2). En termes d'évaluation, il est donc difficile de déterminer si l'échec de l'élève relève d'un manque de compréhension de l'énoncé, d'un manque de connaissances mathématiques ou d'un manque de compétences adaptées (ou les trois). Par ailleurs, un·e élève qui réussit une tâche complexe est *a priori* plus à même de réussir une tâche qui l'est moins, mais il est pertinent de se demander si les connaissances et compétences qui permettent à un·e élève de réussir une tâche peu complexe sont suffisantes ou suffisamment résistantes pour réussir une tâche un peu plus complexe, voire très complexe. C'est tout l'enjeu de cet outil qui permet ainsi d'évaluer plus finement les apprentissages des élèves en mathématiques.

En nous intéressant plus spécifiquement aux items mettant en jeu des fractions (autres que les fractions décimales), nous avons montré qu'ils sont évalués seulement à partir de deux types de tâche, reconnaître ou représenter une fraction donnée à partir d'un partage d'unité, et qu'ils relèvent tous du niveau de compétences 1. Ces items se distinguent par les fractions choisies (fractions supérieures ou non à l'unité), la nature de l'unité à partager (segments – disques, etc.) et le lien entre la fraction et le partage existant de l'unité (par exemple, une unité partagée en huit et une fraction à représenter en quart) ; ainsi pour une même tâche, la complexité émane, non pas de la situation proposée, mais des conversions entre différents registres que l'élève doit réaliser pour effectuer la tâche, sans pour autant que ces conversions aient été réellement pensées lors de l'élaboration des items. L'évaluation des connaissances des élèves sur les nombres décimaux n'a porté que sur un nombre restreint de tâches : comparer des nombres décimaux, transformer ou reconnaître une écriture (passage écriture décimale à fraction décimale et réciproquement), placer un nombre décimal sur une droite graduée, intercaler un décimal entre deux nombres, poser des opérations faisant intervenir des nombres décimaux. Par ailleurs, presque aucun de ces items n'est contextualisé. Seul un item portant sur un calcul de périmètre a fait intervenir des décimaux dans le cadre de la résolution de

problèmes ainsi que deux autres items mettant en jeu des nombres décimaux dans des problèmes, mais ils ont exclusivement évalué la reconnaissance de l'opération mise en jeu (et aucunement des connaissances sur les décimaux).
Concernant les items relatifs au domaine "Grandeurs & mesures", nous avons noté leur faible nombre (42 items au total) qui ne permet pas de véritablement évaluer l'ensemble des connaissances relatives à ce domaine et parmi ceux-là, très peu d'items ont concerné des conversions d'unités alors que les items relatifs à la lecture d'heures sont surreprésentés (12 sur 42). Par ailleurs, il faut également noter que la nature de l'évaluation CEDRE ne permet pas d'évaluer la capacité des élèves à utiliser correctement des instruments de mesure.
De même que nous avons constaté que le découpage des groupes d'élèves déterminés par la DEPP était très arbitraire et qu'ils ne témoignaient que de performances locales, pas forcément révélatrices de ce que savaient les élèves. Ces groupes, statistiquement définis à partir des réponses des élèves, regroupent en effet des items très hétérogènes du point de vue du contenu mathématique et de la forme qui permettent de caractériser que très globalement et de manière seulement descriptive les acquis et/ou difficultés des élèves.

Cette première étude, qui a fait l'objet de plusieurs publications (Sayac & Grapin, 2013, 2014, 2015), m'a confortée dans l'idée qu'il était important de porter un regard didactique sur ces évaluations standardisées afin de savoir véritablement ce qu'elles peuvent nous apprendre des connaissances mathématiques des élèves, à des niveaux de scolarité clés.

2. Deuxième et troisième études : QCM et stratégies de réponses

Comme convenu avec la DEPP qui nous avait autorisées à travailler sur les données du bilan CEDRE 2008, les résultats de notre étude ont été présentés au responsable chargé des évaluations nationales des élèves. Malgré les critiques émanant des résultats de cette première étude, il nous a été demandé de travailler à la préparation du bilan CEDRE suivant (2014) pour en améliorer la conception. Dans cette perspective, il m'a semblé indispensable d'explorer, en amont du travail demandé, la façon dont les élèves de fin d'école appréhendent le format d'une évaluation externe, majoritairement composée de QCM, car peu d'études ont porté sur ce thème, à ce niveau scolaire, en France.
Deux autres études ont donc été menées, avec Nadine Grapin, avec des objectifs distincts.

a- Deuxième étude : les stratégies des élèves face aux QCM

Une étude réalisée en 2012 portant sur les stratégies que développent les élèves de 10-11 ans lorsqu'ils sont confrontés à des QCM en mathématiques (Sayac & Grapin, 2014) : nous avons élaboré un test comportant des items de niveaux de complexité variés qui a été proposé à 155 élèves que nous avons observés individuellement lors de sa passation. Leurs stratégies de réponse ainsi que les degrés de certitude qu'ils ou elles accordent à leurs réponses ont été étudiés. Plusieurs résultats ont pu être dégagés par cette expérimentation, au-delà du constat déjà établi que les élèves français·e·s rencontrent, en fin d'école primaire, de grandes difficultés avec les nombres décimaux et les fractions. Un des résultats que je retiens concerne les stratégies que développent les élèves pour répondre à ces QCM (Sayac & Grapin, 2014). Il s'est avéré que les élèves "faibles"[6] et les élèves "moyens" en mathématiques utilisent, dans des proportions équivalentes (plus d'un élève sur trois), des stratégies de repli/substitution ou des stratégies mixtes pour répondre à des QCM en mathématiques alors que les élèves les plus fort·e·s les utilisent beaucoup moins (moins d'un élève sur quatre). Ce résultat peut paraître

[6] La catégorisation « faible », « moyen » et « fort » a été déterminée à partir des résultats des élèves à l'évaluation nationale de fin de CM2 de l'année d'expérimentation.

évident *a priori*, mais ce qui est surprenant, c'est que les élèves "moyens" adoptent le même type de stratégies que les élèves les plus faibles et que ce n'est qu'à partir d'un certain niveau que les élèves mobilisent pleinement leurs connaissances lorsqu'ils ou elles sont confronté·e·s à des tâches à résoudre. A noter que les élèves français·e·s de fin d'école primaire ne s'autorisent que très peu à répondre au hasard. Il semblerait qu'un contrat didactique s'opère dans le cadre de ce type d'évaluation, ne permettant pas aux élèves d'envisager cette stratégie comme scolairement acceptable. Le résultat le plus inquiétant est celui issu de l'utilisation des échelles de degrés de certitude. Il s'est en effet avéré que, pour les items relatifs aux décimaux, les élèves les plus faibles ayant eu les taux de réussite les plus bas, y avaient répondu avec le niveau de certitude le plus élevé. Ces élèves ont donc un double handicap car au-delà de leurs faibles connaissances dans ce domaine, ils ou elles ont aussi le handicap de ne pas en avoir conscience (ignorance ignorée). Ce constat m'inspire une réflexion que je n'avais pas eue à l'époque, mais qui m'apparaît évident aujourd'hui : il est indispensable d'impliquer davantage les élèves, et plus particulièrement les élèves les plus en difficulté, dans l'évaluation de leurs propres connaissances à travers le développement d'une évaluation formatrice ou à travers des dispositifs d'autoévaluation ou d'évaluation entre pairs. Nous avions d'ailleurs conclu notre article par une inquiétude (Sayac, Grapin, 2014) :

> Si un enseignant estime que certains élèves échouent par manque de confiance en eux, ils peuvent être amenés à leur proposer des situations mathématiquement moins riches qui produisent plus facilement des réussites (logique de réussite immédiate, Butlen, Peltier, Pézard, 2002), mais qui au final réduisent les apprentissages effectifs des élèves. Les travaux du groupe RE.S.E.I.D.A (Rochex & Crinon, 2011) ont montré que les difficultés des élèves pouvaient être, en partie, due à une différenciation pédagogique inadaptée conçue par des enseignants ayant pourtant une réelle volonté d'aider leurs élèves. (p. 197).

Cette inquiétude, au cœur de mes préoccupations scientifiques (principe 1), m'a amenée à choisir d'engager la recherche que je mène actuellement dans des établissements en REP+[7] (voir plus loin).

L'étude spécifique des items du bilan CEDRE 2008 relevant de la résolution de problèmes nous a également conduites à considérer que les QCM pourraient être utilisés comme des vecteurs possibles d'autocontrôle par les élèves à travers les rétroactions qu'ils permettent parfois de réaliser, même si leur usage en évaluation bilan modifie l'activité de l'élève et par conséquent influe sur les résultats (Sayac & Grapin, 2014).

b- Troisième étude : les formats de question

Une étude comparative des résultats en mathématiques des élèves suivant le format des items (QCM ou questions ouvertes) a été réalisée en 2013 : deux tests ont été élaborés dans les deux formats, avec des tâches mathématiques équivalentes, c'est-à-dire se résolvant avec des procédures identiques et avec les mêmes niveaux de complexité. Les résultats des 195 élèves de notre échantillon aux deux tests ont été comparés, ce qui nous a permis de constater que le format des items avait des incidences variables selon la nature de la tâche mathématique et les distracteurs proposés (peu de différences de scores pour des problèmes complexes, des écarts importants pour les tâches plus basiques). Cette étude nous a également permis d'étudier la validité d'une évaluation par QCM à partir des stratégies utilisées par les élèves pour répondre, mais aussi à partir de la comparaison entre les différents formats des tests. Par exemple, nous avons montré que concernant la traduction d'une écriture fractionnaire en écriture décimale (Q→ D) ou l'inverse (D → Q), la cohérence des réponses proposées par les

[7] REP+: Réseau d'Éducation Prioritaire de rang 1 ; ces réseaux regroupent des établissements scolaires primaires et secondaires (collèges) qui bénéficient d'aménagements visant à mieux accompagner la formation continue des enseignant·e·s et le travail en équipe, notamment.

élèves est variable et le format de la question a un impact sur leurs réponses. Dans le sens D → Q, nous avons pu observer que le QCM ne favorise pas systématiquement la réussite, puisque 11 % des élèves réussissent ces items en question ouverte et se trompent aux QCM correspondants. Dans le sens Q → D, il n'en est pas de même puisque dans ce cas, beaucoup d'élèves qui se trompent en question ouverte modifient pour beaucoup leur réponse pour choisir la bonne réponse en QCM.
Nous avons également étudié les différences de stratégies, de performances et de niveaux de certitude attribués selon le sexe des élèves (Sayac & Grapin, 2016). Il ressort de cette étude que, à niveau de performance égal (bas ou élevé), les garçons sont plus assurés dans leurs réponses que les filles. Ce résultat converge avec ceux issus des évaluations PISA ou CEDRE, mais il signifie également que les garçons les plus faibles sont dans une ignorance ignorée plus importante que les filles, ce qui n'est donc pas un atout pour eux.

Ces deux expérimentations réalisées dans des conditions où j'ai pu être au plus près des élèves lorsqu'ils ou elles répondaient à un test m'ont fortement interpellée, notamment sur le fait qu'il est en réalité bien difficile, même outillée scientifiquement, d'appréhender le cheminement cognitif et individuel des élèves, leur activité mathématique et leurs stratégies étant complexes et pas toujours cohérentes (du point de vue d'un adulte). Elles m'ont également confortée dans l'idée qu'évaluer ce que les élèves savent ou ne savent pas véritablement devait être davantage étudié d'un point de vue didactique car ce type d'analyse permet d'aller au-delà de simples statistiques, en dégageant des résultats qui aboutissent à une meilleure compréhension de l'activité des élèves lorsqu'ils ou elles sont évalué·e·s en mathématiques. Par ailleurs, j'ai été interpellée par le fait que, même en observant les élèves individuellement lors de ces expérimentations, ils ou elles semblaient avoir des comportements identiques suivant la classe à laquelle ils appartenaient. Ce constat empirique m'a alors donné envie de revenir à mes premières préoccupations scientifiques, c'est-à-dire m'intéresser aux pratiques des enseignant·e·s en mathématiques, mais cette fois spécifiquement lorsqu'ils ou elles évaluent les apprentissages de leurs élèves.

3. Première étude sur les pratiques évaluatives des enseignant·e·s

Le réseau RCPE (Recherches collaboratives sur les Pratiques Évaluatives) de l'ADMEE que j'ai intégré en 2013 m'a convaincue de l'intérêt d'expérimenter cette forme de recherche qui m'a parue très adaptée à mon objectif d'étudier les pratiques d'évaluation des professeur·e·s et à ma volonté de faire travailler ensemble praticien·ne·s et chercheur·e·s pour une meilleure compréhension des phénomènes étudiés. J'ai donc conçu une recherche collaborative visant à étudier ce qui se passe, en matière d'évaluation en mathématiques, dans les classes des professeur·e·s des écoles. Quatre formatrices "terrain" et deux directrices d'école ont accepté de s'engager dans la recherche qui visait à :

- étudier les pratiques évaluatives des professeur·e·s des écoles en mathématiques
- co-construire des dispositifs de formation à l'évaluation à partir des résultats obtenus

Dans un premier temps, les évaluations ainsi que toutes les traces d'activités mathématiques autour d'une séquence de leur choix en Numération (devoirs, exercices, fiches de préparation) de vingt-cinq professeur·e·s des écoles ont été récoltées (du CP/grade 1 au CM2/grade 5). Il est important de préciser que, alors qu'aucune indication n'a été donnée sur le type d'évaluation à récolter, les professeur·e·s n'ont fourni que des évaluations de fin de séquence.
Ces évaluations ont été analysées à partir de l'outil "facteurs de complexité et de compétences" qui a permis de montrer que les professeur·e·s de notre échantillon proposent à leurs élèves des tâches de bas niveaux de complexité et de compétences qui correspondent (presque) exactement à celles travaillées durant les séances ayant précédé l'évaluation. Par

contre, la forme, la nature et le contenu de ces évaluations extrêmement variés d'un·e professeur·e à un·e autre ont nécessité quelques explications/justifications que nous avons obtenues à l'aide d'entretiens menés avec chacun·e d'eux et d'elles. Il en est ressorti que les pratiques évaluatives des professeur.e.s des écoles en mathématiques, telles que nous avons pu les appréhender à partir de notre échantillon, sont :

- homogènes du point de vue de la complexité des tâches proposées en évaluation :

- ✓ Les professeur·e·s sont soucieux de ne pas proposer à leurs élèves des tâches qu'ils ou elles ne seraient pas à même de comprendre, même si parfois certaines consignes ne sont pas toujours très claires. Ce qui se traduit par des évaluations majoritairement d'un niveau faible de complexité du point de vue du premier facteur et ce, quel que soit leur cycle d'enseignement.
- ✓ Le niveau de complexité des tâches données en évaluation varie suivant le cycle d'apprentissage considéré. Au cycle 3[8], on trouve davantage de tâches complexes du point de vue des mathématiques en jeu, principalement dues à la présence de zéros, à la taille des nombres et à l'identification des classes d'unités.

- hétérogènes du point de vue de leur conception et de la notation utilisée :

- ✓ Le contenu des évaluations proposées est souvent corrélé au type de notation choisi par l'enseignant·e. Pour ceux ou celles qui choisissent de mettre une note chiffrée ou une lettre conditionnée par un nombre de bonnes réponses, le nombre et la nature des tâches proposées dans l'évaluation dépendent de ce choix.
- ✓ Les professeur·e·s des écoles élaborent leurs évaluations de manière très individuelle et à partir d'une grande diversité de ressources (sites Internet, manuels divers, manuel de la classe, etc.).

Cette première étude des pratiques évaluatives des professeur·e·s des écoles en mathématiques a témoigné de la grande diversité d'évaluations proposées aux élèves et de leurs usages. Même si la méthodologie adoptée n'a pas forcément permis de récupérer des évaluations autres que sommatives, très peu de professeur·e·s ont évoqué d'autres formes d'évaluation durant leur entretien (quatre professeur·e·s pour l'évaluation formative, deux professeur·e·s pour l'évaluation diagnostique).

Cette recherche a également montré à quel point les pratiques évaluatives des professeur·e·s des écoles sont peu construites professionnellement du fait que les professeur·e·s ne bénéficient pas (ou très peu) de formation à l'évaluation et que ces pratiques sont très personnellement conçues et pensées, comme divers travaux ou enquêtes l'avaient également souligné (TALIS, 2013).

La nécessité de continuer d'explorer ces pratiques s'est très naturellement imposée à moi et je me suis donc engagée dans un autre projet de recherche, intégrant toujours une dimension collaborative, avec une équipe élargie de chercheur·e·s et de praticien·ne·s : le LéA EvalNumC2.

4. Recherche en cours sur les pratiques évaluatives des enseignant·e·s en REP+

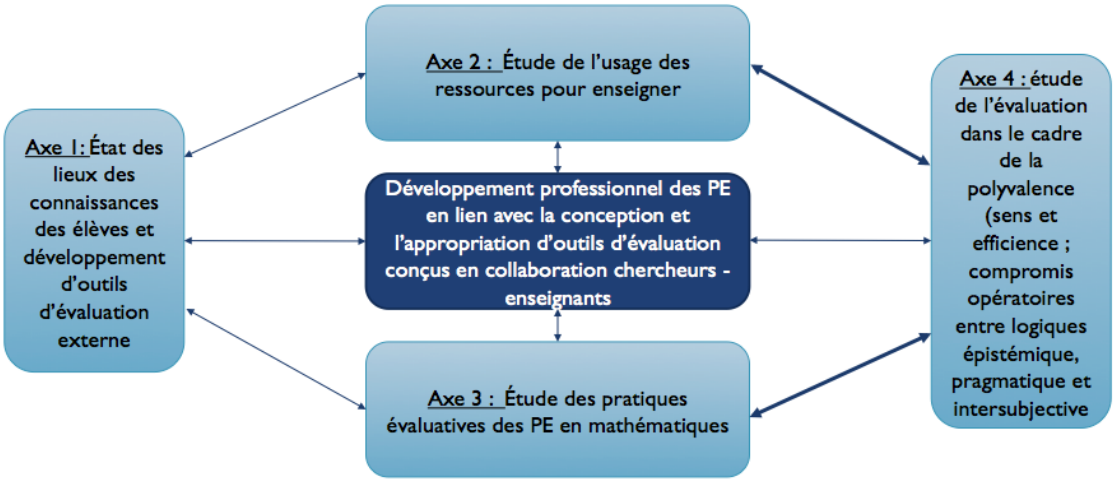
Ce projet de recherche (2016-2019), dont je suis responsable pour l'IFÉ (Institut Français d'Éducation), a été conçu avec deux autres chercheur·e·s en Didactique des mathématiques (Nadine Grapin et Eric Mounier) et une chercheure en Sciences de l'Éducation (Aline Blanchouin). Il est structuré autour de quatre axes scientifiques :

[8] Du CE2 (grade 3) au CM2 (grade 5), au moment de l'étude.

- Axe 1 : la conception de tests externes pour évaluer les apprentissages numériques des élèves au cycle 2,
- Axe 2 : les usages d'une ressource pour enseigner la numération,
- Axe 3 : les pratiques évaluatives des enseignant·e·s,
- Axe 4 : la polyvalence dans les différents moments d'évaluation.

Les tests réalisés dans le cadre de l'Axe 1 permettent d'étudier la robustesse d'une ressource (Axe 2) et servent également d'outil pour enrichir les pratiques d'évaluation des professeurs des écoles engagés dans les axes 3 et 4.

Le travail engagé dans les quatre axes vise à promouvoir le développement professionnel des professeur·e·s des écoles, mais aussi des chercheur·e·s engagé·e·s dans ce projet de recherche-formation. Le schéma ci-dessous permet de synthétiser le projet et de montrer les liens entre les différents axes de recherche.



Concernant l'Axe 3 qui m'incombe plus particulièrement, plusieurs objectifs sont visés, tant d'un point de vue de recherche que d'un point de vue de formation.

Du point de vue de la formation, il s'agit principalement de :
- Engager les professeur·e·s des écoles (enseignant dans les écoles de la circonscription concernée) à développer une réflexion sur leurs pratiques évaluatives en mathématiques.
- Permettre à ces professeur·e·s des écoles de concevoir des évaluations en mathématiques plus équilibrées du point de vue des contenus et plus variées du point de vue des tâches proposées.
- Outiller ces professeur·e·s des écoles pour leur permettre de choisir ou de concevoir des tâches d'évaluation en mathématiques valides et pertinentes au regard des objectifs qu'ils ou elles se fixent.

Du point de vue de la recherche, il s'agit de :
- Continuer d'étudier les pratiques évaluatives des professeur·e·s des écoles en mathématiques, dans un temps plus long et en étant en capacité, via le dispositif choisi, de récolter des données plus fines et plus importantes que dans la recherche précédente.
- Expérimenter l'apport de tâches issues d'évaluations externes pour faire évoluer les évaluations internes et ainsi étudier les conditions d'un enrichissement constructif pour les deux types d'évaluation (projet de recherche en cours avec Nadine Grapin).

Cette recherche venant de débuter, je ne peux présenter de résultats dans cette note de synthèse, mais je peux d'ores et déjà préciser que :

- Deux collectifs d'enseignant·e·s (un avec 7 professeur·e·s de CE1 et de CE2 et l'autre avec 12 professeur·e·s du CP au CM2) sont engagés dans le projet et bénéficient de temps de formation différents (plus long pour le premier collectif, plus court pour le second), mais avec des contenus et des objectifs identiques. Ce choix de deux collectifs vise à inscrire ce projet dans la logique de la recherche menée par Veldhuis et Van Den Heuvel Panhuizen sur des formations à l'évaluation en mathématiques aux Pays-Bas et en Chine (Zhao, Van Den Heuvel Panhuizen & Velhuis, 2016, 2017) et permettra des comparaisons (prévues avec ces chercheur·e·s). Pour ces deux collectifs, l'outil "facteurs de complexité et de compétences" est utilisé pour analyser les évaluations produites par les professeur·e·s (individuellement et collectivement), mais il est également présenté aux professeur·e·s pour les aider à construire leurs évaluations.
- Le développement professionnel des différents acteurs et actrices (professeur·e·s des écoles, formatrices terrain et enseignant·e·s-chercheur·e·s) est étudié à partir du triangle d'Engeström (2001) qui est adapté à ce type de recherche. Ce choix s'inscrit dans un projet de collaboration avec une chercheure en didactique des mathématiques grecque (Despina Potari) visant à éprouver ce cadre théorique dans un contexte de recherche-formation sur les pratiques évaluatives d'enseignant·e·s.

Pour finir de décrire cette recherche, je préciserai qu'elle se situe dans un REP+ de la région parisienne parce que d'une part, les enjeux en termes d'apprentissages des élèves y sont plus forts et que d'autre part, les professeur·e·s des écoles enseignant dans ce type d'écoles sont souvent débutant·e·s et peu outillé·e·s pour faire face aux difficultés qu'ils ou elles rencontrent pour évaluer les apprentissages de leurs élèves.

Cette recherche, parce qu'elle se déroule sur trois ans et qu'elle se réalise dans un contexte d'écoles en éducation prioritaire, mais aussi parce qu'elle me permet d'éprouver le cadre didactique et la méthodologie de formation à l'évaluation que je propose pour analyser les pratiques évaluatives des enseignant·e·s en mathématiques, occupe une place importante dans mon cursus scientifique. Le chapitre suivant, où ce cadre sera présenté et explicité, permettra de mieux comprendre les enjeux et les orientations de cette recherche.

Conclusion de la partie I

Ces différents travaux sur l'évaluation évoqués dans cette Note de synthèse et plus spécifiquement ceux concernant les apprentissages mathématiques ont tous permis des avancées scientifiques plus ou moins importantes néanmoins, ils n'ont pas suffisamment permis, de mon point de vue, à faire avancer la "cause didactique" de l'évaluation, du moins en France. Par "cause didactique" de l'évaluation, j'entends le fait qu'en France, le concept de "*Assessment for learning*" ne s'est pas suffisamment développé, ce qui peut expliquer pourquoi on n'y trouve pas d'expression équivalente. L'"évaluation positive" prônée dans les nouveaux programmes français ne recouvre pas vraiment tout ce qui est entendu derrière l'expression anglaise et les enseignant·e·s français·e·s peinent à concevoir l'évaluation comme un outil au service des apprentissages de leurs élèves. Il y a quelques années, dans le cadre du concours de recrutement des professeur·e·s des écoles (CRPE) et aujourd'hui encore dans le cadre des concours du second degré, les candidat·e·s doivent proposer une séquence d'enseignement autour d'un sujet donné. Les candidat·e·s, mais aussi les sujets "corrigés", proposent des séquences constituées d'un certain nombre de séances autour du sujet, la dernière étant immanquablement étiquetée "évaluation". Cette organisation, cohérente en soi, impose une vision strictement sommative de l'évaluation qui n'aide pas les futur·e·s enseignant·e·s à la penser autrement.

Des différents travaux présentés ci-dessus, je retiens que ceux proposés en Sciences de l'éducation, même ceux qui se rapportent à des apprentissages mathématiques, ne prennent pas suffisamment en compte les contenus mathématiques pour ce qu'ils sont, avec leurs

spécificités. Quand Allal (1979, 1988) évoque les différents types de régulation, même si elle est une des rares chercheur·e·s à souligner l'importance du contenu, elle ne précise pas explicitement ce qu'ils pourraient être en fonction des différents savoirs en jeu. Or comment, par exemple, concevoir de nouvelles activités d'enseignement/apprentissages qui prennent en compte les différences entre les élèves dans le cadre d'une régulation pro-active sans prendre en compte les variables didactiques liées au savoir en jeu ? Comment identifier les différences entre élèves et les prendre vraiment en compte sans une analyse didactique rigoureuse de leurs productions ? Les chercheur·e·s en Sciences de l'éducation qui étudient des évaluations en mathématiques n'exploitent généralement qu'en surface les savoirs en jeu et ignorent souvent les concepts et méthodes développés en Didactique qui leur permettraient d'analyser plus finement leurs données.

D'un autre côté, les travaux de didacticien·ne·s ne prennent pas suffisamment en compte des résultats pourtant cruciaux développés en Sciences de l'éducation. En effet, quand ils ou elles s'intéressent à l'évaluation, c'est souvent en utilisant *a minima* des concepts de l'évaluation développés en Science de l'Éducation en tant qu'outils au service de leurs objets d'étude habituels (savoir spécifique comme l'algèbre ou la numération, pratiques enseignantes des premier et second degré, etc.).

L'évaluation formative, naturellement la plus intéressante pour ces chercheur·e·s car, par définition, plus naturellement en lien avec les apprentissages des élèves, fait l'objet de recherches de plus en plus nombreuses en Didactique des mathématiques, mais elle est souvent étudiée avec difficulté et sans prendre en compte tous les paramètres qui s'y rattachent. Il y est souvent question soit de rechercher dans des pratiques enseignantes ordinaires des traces d'évaluation formative (Coppé, Roditi et al., Horoks et al.), soit d'expérimenter des dispositifs spécifiquement conçus pour développer ce type d'évaluation (Coppé, Aldon), mais dans tous les cas, les chercheur·e·s engagé·e·s dans de telles recherches ont souligné la difficulté qu'ils rencontraient pour identifier ces moments d'évaluation formative et la nécessité d'avoir des outils didactiques permettant de le faire efficacement. L'approche didactique me semble, de fait, épistémologiquement différente de celles qui se sont développées autour de l'évaluation, j'y reviendrai ultérieurement. En effet, même si aujourd'hui les paradigmes de l'évaluation formative et de l'évaluation-accompagnement (De Ketele, 2016) font consensus pour penser évaluation et apprentissages dans une même vision, ils ont longtemps été disjoints et opposés. Pendant longtemps, seuls les apprentissages ont été centraux pour les didacticien·ne·s, sans que leur évaluation soit considérée comme pouvant participer à leur développement. Par ailleurs, les recherches portant sur les pratiques enseignantes en classes "ordinaires" ont commencé bien plus tard en Didactique qu'en Sciences de l'éducation et ont été centrées sur les situations potentiellement apprenantes, sur l'encadrement de l'activité de élèves par l'enseignant·e et peu sur la composante "évaluation" de l'activité de l'enseignant·e. La difficulté principale des didacticien·ne·s est, en fait, qu'il manque ce que Brookhart (2003) appelle une "*measurement theory of classroom assessment*", véritablement conçue pour penser et concevoir une évaluation au service des apprentissages des élèves prenant en compte les contenus mathématiques des évaluations proposées en classe, dans la réalité des pratiques "ordinaires".

C'est ce que je propose de faire dans la partie suivante de cette note de synthèse et, même si je n'ai pas l'ambition de proposer une véritable théorie, je propose de développer un cadre didactique pour penser et analyser les "faits évaluatifs" (Chevallard, 1986) qui conjuguent efficacement savoirs scientifiques en évaluation (dans la diversité des champs scientifiques concernés) et savoirs en Didactique des mathématiques. Ce cadre propose à la fois des concepts qui mixent ou complètent les approches (validité, régulation, contrat didactique en évaluation) et à la fois des concepts qui sont définis ou redéfinis pour être adaptés à l'étude de

"faits évaluatifs" : *épisodes évaluatifs*, *jugement professionnel et didactique en évaluation*, *contrat didactique en évaluation*.

PARTIE II : UN CADRE DIDACTIQUE POUR L'ÉVALUATION EN MATHÉMATIQUES

"L'analyse de l'évaluation ne peut aboutir si elle n'est pas d'abord une analyse didactique de l'évaluation, c'est-à-dire une analyse des fonctions didactiques de l'évaluation" (Chevallard & Feldmann, 1986, p. 66). Cette affirmation, qui date de plus de trente ans, me paraît encore pertinente aujourd'hui, même si elle nécessite d'être actualisée par l'explicitation de certains points.

Dans l'introduction, j'ai précisé les principes qui m'animent en tant qu'enseignante-chercheure et qui fondent à la fois mes travaux, mais aussi leur ambition. Je rappelle ces principes en débutant cette seconde partie de ma note de synthèse :

- Je suis une chercheure engagée, animée par des problématiques de réussite des élèves et de formation des enseignant·e·s en mathématiques (principe 1).
- Je suis une chercheure qui considère qu'éprouver les frontières d'autres champs scientifiques permet d'enrichir l'approche en Didactique des mathématiques (principe 2).
- Je suis une chercheure convaincue de la nécessité de davantage prendre en compte la composante personnelle des individus, acteurs et actrices de l'enseignement et de la formation des enseignant·e·s en mathématiques (principe 3).

Les travaux en Sciences de l'éducation présentés dans la partie précédente ont nourri le cadre didactique de l'évaluation que je vais proposer ici et ont permis de conforter mon deuxième principe. Ils ont été conçus par des chercheur·e·s ayant tous pour but d'appréhender l'évaluation de manière plus constructive du point de vue des apprentissages des élèves, une évaluation "pour apprendre". Ils m'ont permis d'éprouver leur pertinence (*Assessment for Learning*, jugement professionnel en évaluation, etc.) et leurs limites (l'entrée par les fonctions de l'évaluation, la prise en compte insuffisante des contenus) du point de vue du cadre que je propose. J'en retiens certains concepts et résultats alors que d'autres me semblent moins intéressants, voire inappropriés. Je les préciserai au fur et à mesure de la présentation de mon cadre.

Au-delà des analyses et des résultats nouveaux que ce cadre cherche à produire, il vise également le développement d'outils pour la formation à l'évaluation des enseignant·e·s pour amener *in fine* des retombées en termes de réussite des élèves (principe 1). Il a donc été conçu avec cette double visée (scientifique et professionnelle), mais également, sans que ce soit ce qui importe le plus, pour faire évoluer la vision des mathématiques que peuvent avoir les différent·e·s acteurs et actrices du monde scolaire (élèves, parents, enseignant·e·s, etc.). En effet, des enseignant·e·s formé·e·s et outillé·e·s pour évaluer en mathématiques pourraient générer une évaluation plus constructive du point de vue des apprentissages et ainsi avoir un impact sur la façon dont les élèves appréhendent cette discipline souvent perçue comme facteur de sélection et/ou de discrimination (notamment celles liées aux stéréotypes de sexe), en France. La prise en compte de la vision des mathématiques des différents acteurs, notamment des élèves, s'inscrit donc dans une démarche d'*Assessment for Learning*, telle que préconisée par Broadfoot et al. (2002). Il me semble important de préciser que cette prise en compte des conséquences de l'évaluation du point de vue de la vision des mathématiques des élèves ne suffit évidemment pas à produire une évaluation "pour apprendre". En effet, si l'enseignant·e n'est pas suffisamment outillé·e pour évaluer les apprentissages de ses élèves, cette prise en compte s'opérera à partir d'éléments subjectifs liés à sa propre vision des mathématiques, parfois négative notamment chez les professeur·e·s du premier degré, avec des conséquences néfastes au niveau des apprentissages des élèves.

Les travaux sur l'évaluation explorant les liens entre évaluation et apprentissages des élèves sont principalement axés sur les différentes fonctions de l'évaluation, mais certain·e·s chercheur·e·s ont soit exprimé le fait qu'il fallait dépasser l'opposition entre telle ou telle fonction de l'évaluation (Wiliam, 2000 ; Taras, 2005), soit choisi de ne pas entrer par ces différentes fonctions (Vantourout & Goasdoué, 2014), soit montré la difficulté de soumettre leurs analyses à cette catégorisation (Horoks et al., 2015). De mon côté, même si Hadji (1989) a indiqué que "il ne faut pas concevoir la fonction de l'évaluation comme quelque chose d'unidimensionnel où serait enfermé tout le sens d'une pratique" (p. 61), je considère que ces fonctions ne doivent pas être au cœur d'un cadre didactique de l'évaluation car d'une part elles enferment l'évaluation dans des catégories distinctes qui ne le sont pas forcément dans la réalité des pratiques de classe, d'autre part, elles peuvent amener les professeur·e·s à concevoir/penser l'évaluation comme disjointe des processus d'enseignement et d'apprentissage. Wiliam (2000) a clairement dénoncé ce risque en expliquant comment l'étiquetage des différentes fonctions de l'évaluation avait participé à opposer évaluation et apprentissages. Par ailleurs, en France, même si les professeur·e·s sont certainement tous capables de distinguer théoriquement les différentes fonctions de l'évaluation (certificative, formative, diagnostique) et que Issaieva et al. (2011) ont évoqué une "désirabilité sociale de l'évaluation formative" des professeur·e·s français·e·s, ils et elles sont en réalité souvent peu outillé·e·s et peu formé·e·s pour mettre en place d'autres formes d'évaluation que les évaluations sommatives qui leur permettent de remplir leur "mission" institutionnelle d'évaluation (remplir des livrets, rendre compte de niveaux de connaissances des élèves, etc.). De son côté, Perrenoud dès 1991 a indiqué que les enseignant·e·s soucieux et soucieuses de bien évaluer, étaient pris·e·s entre deux modèles "un modèle didactique séduisant, mais qui ne dit pas grand-chose de l'évaluation et un modèle d'évaluation formative transdisciplinaire [...], qui s'est développé indépendamment de la didactique et du curriculum spécifique d'une discipline" (p. 56).

Il me semble donc nécessaire de proposer un cadre didactique permettant en premier lieu d'étudier les "faits évaluatifs", mais qui puisse également être exploité en formation et être "séduisant" (Issaieva et al., 2011) pour les enseignant·e·s (principe 1). Pour que ce cadre soit opérationnel aussi bien d'un point de vue scientifique que professionnel, il doit s'inscrire dans la réalité des pratiques d'évaluation telles qu'elles existent dans les classes or, comme de nombreux chercheur·e·s l'ont souvent précisé (Cizek et al., 1995 ; McMillan, 2001 ; Brookhart, 2004 ; Sayac, 2016), ces pratiques sont très individuelles, variables et imprédictibles car dépendantes de nombreux paramètres personnels et contextuels (ce qui conforte mon principe 3). Si l'on souhaite les comprendre et les faire évoluer, il convient donc de concevoir des outils qui intègrent cette diversité et prennent en compte les éléments qui la constituent.

Le cadre didactique de l'évaluation que je propose ne retient donc pas une entrée par les différentes fonctions de l'évaluation. Dans ce cadre, c'est le couplage entre le moment où l'évaluation est proposée et ce qui la caractérise en termes de contenu, de gestion et d'enjeux qui permet de la caractériser. Je propose, pour intégrer ces différents traits caractéristiques, la notion d'*épisode évaluatif* qui permet d'inscrire davantage l'évaluation dans un temps didactique et de dépasser l'entrée par les fonctions. Ces épisodes sont régis par un contrat didactique spécifique que je propose d'appeler *contrat didactique en évaluation* et dont je préciserai la nature. La validité de ces épisodes évaluatifs fait également l'objet d'une interprétation adaptée à ce cadre.

La définition d'un tel cadre ne peut faire l'économie, pour répondre au principe 3, d'une prise en compte de facteurs humains d'ordres divers qui s'inscrivent dans une "logique évaluative"

propre à chaque professeur·e. Je propose d'appréhender cette "logique évaluative" notamment à partir de la notion de *jugement professionnel et didactique en évaluation* (JPDE) que je développe dans la continuité des travaux sur le jugement professionnel en évaluation revus avec une approche didactique (principe 2). Je définirai précisément ce jugement spécifique et montrerai son rôle dans l'évaluation "pour apprendre" (principe 1). Cette "logique évaluative" est également appréhendée à travers la façon dont les enseignant·e·s conçoivent le contenu des épisodes évaluatifs qu'ils proposent à leurs élèves (ressources utilisées, modes, etc.) et leur gestion, mais aussi la notation (au sens large) qu'ils ou elles utilisent pour rendre compte des apprentissages de leurs élèves.

Dans une première partie je définirai les premiers éléments constitutifs de ce cadre puis je montrerai, dans une deuxième partie, comment il permet de penser, de concevoir et de faire évoluer les pratiques évaluatives des professeur·e·s en mathématiques.

A. L'ÉVALUATION DES APPRENTISSAGES EN MATHÉMATIQUES

Pour aborder la question de l'évaluation dans une perspective didactique, la prise en compte de la temporalité des apprentissages est incontournable. Dans son approche des praxéologies didactiques du professeur (leçon 1), Chevallard (1998) indique que l'évaluation est le sixième moment de l'étude[9] et précise qu'"un moment de l'étude se réalise généralement en plusieurs fois, sous la forme d'une multiplicité d'épisodes éclatés dans le temps de l'étude" (p. 19-20). Schubauer-Leoni (1991) perçoit également une dimension temporelle répétitive dans les moments d'évaluation puisqu'elle indique que "les actes d'évaluation participent à scander le rythme des situations didactiques en marquant institutionnellement la progression du savoir". Je retiens de ces deux propositions que l'évaluation des apprentissages des élèves se réalise tout au long de l'étude, de différentes manières et à différents moments, à travers ce que je vais nommer des *épisodes évaluatifs*.
Si l'on s'intéresse au terme d'épisode on trouve, dans le Larousse (2017), deux entrées qui me paraissent très opportunes pour le définir. Une entrée comme "partie d'une œuvre narrative ou dramatique s'intégrant à un ensemble, mais ayant ses caractéristiques propres" et une autre comme "événement qui fait partie d'une action plus générale et qui se distingue par tel ou tel caractère". Ces deux entrées caractérisent ce qui me semble important de retenir dans le concept d'épisode évaluatif et qui réfère à la fois à une dimension globalisante (partie d'un tout) et à la fois à une dimension de singularité (caractère propre). À travers ce terme, l'évaluation peut être véritablement pensée comme faisant partie intégrante du processus didactique, avec des spécificités propres. Mercier (1992, 1995) avait également utilisé ce terme pour identifier un moment spécifique de la biographie didactique d'un élève, qu'il qualifiait *d'épisode didactique* ou *épisode biographique* en se référant au "fonctionnement adidactique d'une relation didactique, pour cet élève au moins, relativement à un objet de savoir pour lequel cet élève a rencontré de l'ignorance comme un besoin de savoir, et des moyens de satisfaire ce besoin", mais ces moments ne concernent qu'un élève sur un moment spécifique alors que les épisodes évaluatifs qui m'intéressent se rapportent aussi bien aux élèves qu'à l'enseignant·e et qu'ils s'intègrent au processus didactique global.
On a pu voir que Brookhart (2003) avait également utilisé le terme d'épisode lorsqu'elle a évoqué la validité des évaluations de classe. Ce terme me semble donc approprié pour définir ces moments particuliers qui s'inscrivent dans un processus didactique, à condition de le spécifier à l'évaluation.

[9] Il s'agit évidemment ici de l'étude pour l'élève (sujet didactique).

Les épisodes évaluatifs peuvent être de natures différentes et correspondent à des catégories d'évaluations formelles que l'on retrouve dans le système scolaire français sous les dénominations d'interrogation, de contrôle, de devoir à la maison, etc. À ces catégories s'ajoutent des moments plus informels associés généralement à des moments d'évaluation formative pouvant se réaliser dans le temps de l'étude à partir d'échanges entre l'enseignant·e et ses élèves (oraux ou par observation de production d'élève) qui génèrent une prise d'information, un traitement et un feedback (Black & Wiliam 1998, Allal 1979, 1988, Van Den Heuvel Panhuizen 1996).
Je vais, à présent, préciser ce que j'entends par épisode évaluatif et comment ce concept participe d'une approche didactique de l'évaluation en mathématiques.

1. Les épisodes évaluatifs

Un *épisode évaluatif* correspond à un moment spécifique choisi par un·e professeur·e ou une institution et auquel des élèves sont confronté·e·s afin d'évaluer, avec toutes les dimensions que ce terme peut recouvrir (négociatrice, régulatrice, certificative, diagnostique, véridictionnelle, etc.), un état de connaissances relatif à des savoirs ou des savoir-faire prescrits ou enseignés.
Pour étudier un épisode évaluatif, il faut prendre en compte le moment où il est proposé, mais aussi les tâches évaluatives qui le constituent, les enjeux qui lui sont associés ainsi que la gestion qui l'organise. Il convient également de prendre en compte le contrat qui lie professeur·e et élèves à propos du savoir en jeu dans l'épisode évaluatif concerné, ce que je préciserai ultérieurement.

a- Ce qui caractérise les épisodes évaluatifs

Pour caractériser les épisodes évaluatifs, il convient de prendre en compte le moment où ils sont proposés, la nature des tâches qui les constituent ainsi que le recouvrement du domaine d'étude par ces tâches. Ces informations sont indispensables pour étudier le rôle qu'ils jouent dans le processus d'apprentissage en cours et comment ils s'y inscrivent.

- ***Le moment où ils sont proposés***

Le moment où un épisode évaluatif est proposé renseigne sur le rôle qu'il joue dans l'avancée du temps didactique de la classe. Pour Chopin (2006), le temps didactique est structuré par des événements didactiques, s'inscrivant dans un temps institutionnel, assurant la production d'un nouveau savoir. Les épisodes évaluatifs, dans la diversité de leur nature et du moment où ils apparaissent, participent donc de ce temps didactique (Chevallard & Mercier, 1987 ; Chevallard, 1991 ; Mercier, 1992) qui permet de faire progresser la classe dans l'étude d'un savoir. Un épisode évaluatif proposé en début d'étude se distingue logiquement d'un autre situé en fin ou même en milieu d'étude puisque les rapports aux savoirs anciens et nouveaux évoluent au cours de l'étude. Ainsi, le moment où un épisode évaluatif est proposé peut nous renseigner sur le rôle qu'il est supposé assurer du point de vue de la chronogénèse : passerelle entre savoirs anciens et nouveaux en début d'étude avec des tâches connexes au savoir ancien, régulation dans le cours d'étude avec des tâches s'y inscrivant localement, véridiction en fin d'étude avec des tâches représentatives, autant que faire se peut, de l'ensemble du savoir enseigné ou certification, dans les moments de transition scolaire avec des tâches correspondant aux attentes institutionnelles de ces moments. Le moment où un épisode évaluatif est proposé peut être programmé à l'avance par l'enseignant·e (on pourra alors le qualifier de formel) ou bien advenir au cours d'une séance si l'enseignant·e juge opportun d'en proposer un à ce moment là (on pourra alors le qualifier d'informel).

- ***La nature des tâches qui leur sont associées***

Au-delà du moment où se situent les épisodes évaluatifs, il est important de prendre en compte la nature et la gestion des tâches évaluatives qui leur sont associées afin d'en évaluer la pertinence du point de vue du processus didactique dans lequel ils s'inscrivent. L'étude de ces tâches peut se réaliser à partir d'outils émanant de cadres didactiques retenus par les chercheur·e·s comme étant les plus adaptés aux épisodes considérés et à leur(s) objectif(s) scientifique(s). Dans les trois exemples ci-dessous, je fais apparaître différents cadres didactiques qui ont été utilisés par des chercheures pour analyser les tâches proposées lors de moments d'évaluation qu'elles ont étudiés et que je considère *a posteriori* être des épisodes évaluatifs :

- Dans leur étude des pratiques enseignantes en collège autour de l'Algèbre, Horoks et al. (2016) se sont intéressées aux moments de "mise en commun" des productions d'élèves qui relèvent, de leur point de vue, de l'évaluation formative. Ces chercheures ont réalisé une analyse épistémologique des notions en jeu et de la façon dont elles avaient été mises en fonctionnement dans les tâches proposées par les professeur·e·s lors de la "mise en commun" et ont étudié les retours faits aux élèves relativement à ces tâches (gestion des erreurs, notamment). La complexité des tâches a été déterminée à partir notamment des différents niveaux de mise en fonctionnement des connaissances (Robert, 1998).
- Dans sa thèse, Grapin (2015) a étudié un bilan évaluatif externe proposé à des élèves de CM2, en fin d'année scolaire. La gestion de ce type d'évaluation standardisée ne présentant guère d'intérêt d'un point de vue didactique, elle a focalisé son étude sur les tâches associées au domaine numérique, comprenant les nombres entiers à travers la numération décimale, les relations arithmétiques entre les nombres, le calcul et les problèmes numériques. Elle s'est située dans une approche anthropologique et cognitive afin de définir, sur le domaine étudié, un référent épistémologique à partir duquel il lui a été possible d'analyser le contenu des évaluations à partir des tâches évaluatives proposées dans les items concernés et d'interpréter les résultats des élèves. La complexité des tâches a été déterminée à partir de six niveaux hiérarchisés d'intervention des types de tâche qui ont été définis en lien avec les travaux de Castela (2008) et Robert (1998).
- Dans l'étude que j'ai réalisée sur les pratiques évaluatives en mathématiques à l'école primaire, je me suis plus particulièrement intéressée aux épisodes évaluatifs proposés par des professeur·e·s des écoles à la fin d'une séquence de Numération. Pour comparer les pratiques des professeur·e·s de mon échantillon, j'ai étudié toutes les tâches évaluatives qu'ils ou elles avaient proposées à leurs élèves à partir de l'outil développé avec Grapin (Sayac & Grapin, 2013, 2014, 2015) émanant de différents cadres didactiques et autres.

La méthodologie utilisée pour cette étude a permis de caractériser les niveaux de complexité et de compétences des tâches évaluatives proposées par les vingt-cinq professeur·e·s de l'échantillon, mais elle ne m'a pas permis d'avoir accès ni à la gestion de ces épisodes, ni au contrat didactique en évaluation associé. Or, même si l'entretien post-analyse réalisé après l'analyse des tâches évaluatives a permis de conclure que, dans la plupart des cas, aucun retour aux élèves n'avait été proposé, on ne sait pourquoi les professeur·e·s ont fait ce choix, ni comment ils ou elles ont exploité l'épisode évaluatif considéré, ni comment les élèves ont réagi. Cette étude aurait pu être plus instructive du point de vue des pratiques évaluatives si j'avais pu prendre en compte la gestion des épisodes proposés ainsi que le contrat didactique en évaluation qui y était attaché. Le cadre didactique que je propose aujourd'hui devrait permettre d'étudier de manière plus efficace car plus globale les pratiques évaluatives des enseignant·e·s en mathématiques.

- ***Le recouvrement des tâches qui leur sont associées***

La question du recouvrement d'un domaine mathématique par les tâches évaluatives proposées dans les différents épisodes évaluatifs qui ponctuent son étude est également une question déterminante pour comprendre ce qui se joue du point de vue des apprentissages des élèves, aussi bien dans la classe qu'en dehors. Selon celui ou celle qui la considère (chercheur·e, institution, enseignant·e), le recouvrement est mis au regard d'un référent adapté : généralement épistémologique pour les didacticien·ne·s, curriculaire pour les enseignant·e·s ou défini plus spécifiquement pour les autres.

Ce qu'il me semble important de souligner par rapport au recouvrement d'un domaine mathématique par des tâches évaluatives c'est que :

- dans le cadre d'évaluations internes, l'étude du recouvrement d'un domaine par des tâches évaluatives ne peut se concevoir que dans la globalité des épisodes évaluatifs proposés dans le cours de l'étude de ce domaine.
- dans le cadre d'évaluations externes, le recouvrement d'un domaine par des tâches évaluatives peut être plus ou moins scientifiquement établi, selon les commanditaires.

Dans les deux cas néanmoins, cette question est à relier à celle de la validité des épisodes évaluatifs considérés.

b. La validité des épisodes évaluatifs

La question de la validité des épisodes évaluatifs est complexe. En 1989, Chevallard avait déjà appréhendé cette complexité en précisant que :

> Quelles que soient les précautions prises, on pourra disputer à l'infini. On observera, surtout, que la validité de l'évaluation obtenue pourra être contestée sans toutefois que sa nature d'évaluation le soit, sauf à lier dogmatiquement l'une à l'autre – exigence déraisonnable et, dans les faits, largement ignorée. (p. 9)

Pour autant, la question de validité des épisodes évaluatifs ne peut être absente d'un cadre didactique de l'évaluation, même si elle doit y être spécifiquement définie (voir Brookhart, 2003).

Dans la partie précédente, j'ai fait état de deux conceptions scientifiques de la validité d'une évaluation portées par différent·e·s chercheur·e·s. Pour les un·e·s (Moss, 2003 ; Kane, 2006 ; Cizeck, 2009) s'inscrivant dans la continuité des travaux de Messick (1989, 1995), la validité se conçoit en lien avec la qualité du jugement ou de l'interprétation émis à partir des réponses des élèves. Pour les autres, la validité se conçoit telle que définie par De Ketele et Roegiers (1993) qui considèrent qu'une évaluation est valide si elle évalue bien ce qu'elle prétend évaluer. Il convient également de prendre en compte le point de vue de chercheures comme Brookhart (2003) ou Rémond (2006) qui ont une approche que l'on peut qualifier de plus didactique et qui estiment que les critères de validité doivent porter sur les conditions qui permettent de rendre l'évaluation utile du point de vue des apprentissages des élèves.

Je propose une approche de la validité qui combine ces différentes orientations, en y ajoutant les validités curriculaire et pédagogique que distingue De Landsheere (1988) pour dégager une validité didactique adaptée aux épisodes évaluatifs et donc aux évaluations de classe. Ainsi, dans le cadre didactique de l'évaluation que je propose, la validité d'un épisode évaluatif est éprouvée à partir :

- d'éléments liés aux dimensions épistémo-didactique et/ou curriculaire du savoir en jeu dans l'épisode.
- d'éléments liés à la gestion de cet épisode par l'enseignant·e.

Pour qu'un épisode évaluatif soit valide, il faut donc à la fois qu'il s'inscrive dans le processus didactique en cours c'est-à-dire que l'enseignant·e lui attribue un rôle spécifique dans le processus d'enseignement, que les tâches qui lui sont associées soient suffisamment variées et représentatives du thème et à la fois qu'il soit géré de manière à permettre aux

élèves “d’apprendre” au cours de cet épisode évaluatif. J’entends par là qu’il faut que l’enseignant·e le considère comme un outil au service des apprentissages des élèves.

Si l’on s’intéresse plus particulièrement aux tâches évaluatives, on ne peut se restreindre à l’étude d’un seul épisode qui n’aurait qu’une validité locale et réduite, mais l’on doit prendre en compte l’ensemble des tâches évaluatives proposées dans le cadre des épisodes évaluatifs relatifs à l’étude d’un savoir spécifique. On pourra ainsi disposer d’un nombre non négligeable de tâches évaluatives et se poser la question du “caractère représentatif de l'échantillon d’items par rapport à l'univers de référence” (De Ketele & Dufays, 2003) et ainsi avoir une vision globale de la validité des épisodes évaluatifs relatifs à l’étude d’un savoir dans toute son étendue. Les points suivants concourent à l’édification de preuves de la validité d’un ou de plusieurs épisodes évaluatifs :

- *Le recouvrement des tâches évaluatives* aussi bien du point de vue curriculaire que du point de l’univers de référence didactique car comme Grapin (2015) le précise dans sa thèse “apporter des preuves de validité didactiques sur le contenu de l’évaluation demande d’abord à ce que, sur un domaine mathématique et un niveau donnés, les savoirs à évaluer soient définis.” (p. 171)
- *La cohérence des tâches entre elles***,** au-delà de ce recouvrement. J’entends par là que les tâches doivent constituer un tout cohérent du point de vue du savoir à évaluer (cf. la validité épistémo-didactique de Grapin 2015 ou Grugeon-Allys & Grapin 2016) et de celui de l’activité des élèves. Par exemple, certaines tâches évaluatives peuvent parfois être sur-représentées ou dupliquées quasi à l’identique sans que cela permette réellement d’apporter des informations complémentaires relatives à l’état de connaissances des élèves.
- *Les différents niveaux de complexité de ces tâches*, déterminés à partir d’outils didactiques (tels les niveaux de mise en fonctionnement des connaissances, les changements de cadres, les relations entre les registres de représentations sémiotiques, etc.) doivent permettre de rendre compte, finement, de ce que les élèves ont acquis en termes de connaissances, de savoirs et de compétences.

c- La gestion d’un épisode évaluatif

Concernant la gestion d’un épisode évaluatif, je propose de l’appréhender à travers les notions de contrat didactique en évaluation que je présente ci-après et de jugement professionnel et didactique en évaluation que je développerai plus loin.

Ce que je souligne dès à présent c’est que, tout autant que le choix approprié des tâches évaluatives proposées dans le cadre d’un épisode évaluatif, sa gestion est déterminante pour le considérer comme relevant d’une évaluation “pour apprendre” ou non. Cette gestion relève pour moi des pratiques évaluatives d’un·e enseignant·e et sera donc étudiée dans la partie correspondante.

2. Contrat didactique en évaluation

La notion de contrat est au fondement des approches didactiques de la TSD (Théorie des Situations) et de la TAD (Théorie Anthropologique du Didactique). Chez Brousseau (1980), le contrat principal est le contrat didactique liant enseignant·e et enseigné·e, explicitement pour une petite part, mais surtout implicitement autour de la connaissance mathématique visée. Pour Chevallard (1988b), le contrat est l’ensemble des rapports institutionnels qui se mettent en place, dans une institution donnée, à un moment donné. L’évaluation ayant une forte valeur institutionnelle à la fois pour les élèves et pour les professeur·e·s, je propose donc d’étudier plus spécifiquement le contrat didactique en jeu lors des épisodes évaluatifs proposés par un·e enseignant·e à ses élèves au cours de l’étude d’un savoir.

Dans sa note de synthèse sur le contrat didactique (p. 106), Sarrazy (1995) s'étonne que les didacticien·ne·s se soient peu intéressés aux écarts de réussite des élèves pointés par les psychosociologues (Roux, Andreucci, 1989) ayant travaillé sur le "contrat didactique", même s'il relève quelques travaux intégrant une dimension différentielle (Schubauer-Leoni, 1986 ; Balacheff & Laborde, 1985 ; Krummheuer, 1988). Je considère, à la suite de Sarrazy, qu'il faut effectivement s'emparer de la notion de contrat didactique et la travailler plus spécifiquement dans le cadre de l'évaluation. Ce point de vue semble partagé par Schubauer-Leoni (1991) qui considère également que "les moments d'évaluation vont être des lieux privilégiés pour donner la(les) preuve(s) qu'ils (les élèves) ont bien compris à la fois le jeu et les règles du jeu (instaurés par le maître)" (p. 80).
Inscrivant l'émergence du concept de contrat didactique chez Brousseau entre deux courants (systémique et interactionniste) Sarrazy (1995) met également l'accent sur l'aspect communicationnel du contrat didactique qu'il voit comme "le produit d'un mode spécifique de communication didactique (lié à l'épistémologie de l'enseignant·e) instaurant un rapport singulier de l'élève au savoir mathématique et à la situation didactique" (p. 93), certains contrats pouvant générer des "bruits" faisant obstacle à la communication de connaissances (voir travaux de la partie précédentes). Cette définition à partir d'une communication didactique entre l'élève et le savoir médiée par l'enseignant·e impose, de mon point de vue, d'étudier spécifiquement le contrat didactique lors des moments d'évaluation.

a- Contrat didactique et évaluation dans différentes théories

Pour situer mon propos et l'inscrire dans un cadre didactique, je propose de revisiter à l'aune de l'évaluation différents travaux produits en Didactique des mathématiques autour de la notion de contrat didactique.

- ***Du côté de la TSD***

Il est intéressant de rappeler, comme le précise Sarrazy dans sa note de synthèse sur le contrat didactique (1995), que la notion de "contrat didactique" est apparue lors d'une recherche portant sur les échecs électifs. Intéressant parce que l'échec peut se manifester, soit par une résistance de l'élève à s'engager dans une situation non conforme à sa représentation de l'apprentissage de l'élève comme dans le cas de Gaël (Brousseau, 1980, 2009), soit par une non-conformité aux réponses attendues par l'enseignant·e dans le cas d'une évaluation. Dans les deux cas, on trouve la nécessité de préciser le contrat en jeu pour expliquer les phénomènes d'échecs spécifiques aux mathématiques, Brousseau (1980) ayant d'emblée précisé que les causes de l'échec sont à chercher dans le rapport de l'élève au savoir et aux situations didactiques et non dans ses aptitudes ou dans ses caractéristiques permanentes générales.
Dans la continuité de cette vision, Grenier (1998) a aussi indiqué qu'"une intervention de l'enseignant consistant à invalider une action d'élève peut être analysée comme un moyen d'enrichir le milieu de la situation du point de vue de la rétroaction, mais aussi comme la poursuite de la négociation d'un contrat avec les élèves" (p. 129). Elle considère que c'est au milieu de permettre l'invalidation de productions erronées d'élèves et que si le milieu n'est pas en mesure de le faire, l'enseignant est en difficulté pour intervenir. De son point de vue, "les interventions du professeur peuvent être analysées comme palliatifs aux insuffisances du milieu, dans le but de créer artificiellement un rapport idoine entre élèves et contenus de connaissances visées, autrement dit un contrat didactique conforme à ses objectifs" (p. 130).
On voit bien à travers ces différentes considérations que l'évaluation, même pensée dans sa dimension formative, n'a pas de place réelle dans le cadre de la TSD car c'est le milieu, avec

les rétroactions potentielles qu'il permet, qui théoriquement assure ce rôle de régulation des apprentissages.

- ***Du côté de la TAD***

Pour Chevallard & Feldman (1986), "le pilotage du processus didactique par l'enseignant se trouve médié par le contrat didactique qui fixe l'exigence d'une progression dans le savoir, dont il pose globalement le principe et dont il gère à chaque instant les contenus et les modalités" (p. 68). Pour ces auteurs, la négociation didactique qui met aux prises l'enseignant·e et les élèves à propos du savoir est au cœur du processus didactique et la notation participe de cette négociation. Ils indiquent également que "le contrat didactique (...) devra préciser, à côté de la place occupée par le professeur (...) la place allouée aux élèves, c'est-à-dire le lieu que ceux-ci devront venir occuper vis-à-vis de l'élément de savoir enjeu momentané de la négociation didactique" (p. 68). L'autoévaluation, l'évaluation formatrice ou entre pairs peuvent, dans cette approche, être pensées comme une redéfinition des places de l'enseignant·e et des élèves en termes de contrat.
Chevallard et Feldmann précisent, très finement, comment cette négociation peut s'opérer à partir, notamment, de la notation du professeur étudié, mais ils ne développent pas véritablement le lien entre le contrat didactique et l'évaluation des élèves.

- ***Du côté de théorie de l'action conjointe***

Dans le cadre de la théorie, Sensevy (2011) reprend la notion de contrat didactique de la TSD dont il étend quelque peu les caractéristiques. Pour lui, ce contrat évolue et se modifie au cours des différents moments de la leçon, mais il n'attache pas une attention particulière au moment de l'évaluation. Il introduit également la notion de contrats didactiques différentiels variant selon les élèves avec leur histoire personnelle et précise que ces contrats s'installent au cours de l'apprentissage et rendent manifestes des incertitudes quant à l'efficacité des connaissances initiales. Il décrit comment les signes reconnus et interprétés par les uns et les autres permettent la régulation de l'action conjointe du professeur et des élèves, signes qui peuvent aller d'un simple hochement de tête de l'enseignant·e pour valider une réponse d'élève à une note apposée sur une copie.
Cette approche de contrats différenciés et locaux permet d'envisager la prise en compte d'un contrat spécifique pour chaque épisode évaluatif qui viendrait s'adjoindre au contrat didactique global instauré entre le professeur et ses élèves, au sein d'une classe. Ce n'est ni un avenant car il ne modifie pas les termes du contrat, ni une clause puisqu'il ne concerne qu'un moment spécifique du processus d'apprentissage. Il est néanmoins intimement lié au contrat didactique puisque les acteurs et actrices sont les mêmes et que les épisodes évaluatifs participent des apprentissages. Il convient donc, au-delà de la nature des tâches évaluatives dont est constitué un épisode évaluatif, du moment où il est proposé et de sa gestion, de prendre en compte la nature du contrat didactique en jeu lors du moment de l'étude où il s'inscrit.

Pour préciser la nature du contrat didactique en jeu dans un épisode évaluatif, l'approche de Perrin-Glorian et Hersant (2003) me semble très intéressante. Les différentes composantes du contrat didactique qu'elles définissent s'appliquent assez bien à la notion d'épisode évaluatif telle que je l'ai définie puisqu'elles concernent le domaine mathématique (tâches évaluatives), le statut didactique du savoir (moment), la nature et les caractéristiques de la situation didactique ainsi que le partage de responsabilité vis-à-vis du savoir entre le ou la professeur·e et les élèves (gestion). Les trois niveaux de structuration du milieu qu'elles proposent s'adaptent parfaitement au contrat didactique que je propose pour les épisodes évaluatifs. En

effet, le niveau de *mésocontrat* défini à l'échelle d'une activité ou de la résolution d'un exercice, semble correspondre plus particulièrement à ce qui se joue lors d'un épisode évaluatif, même si le niveau micro peut également convenir. La notion de partage de responsabilité, utilisée dans un sens très local par Perrin-Glorian et Hersant, permet également d'envisager la part plus ou moins grande que peuvent prendre les élèves dans un épisode évaluatif et ainsi de rattacher un épisode aux courants de l'évaluation formatrice porté par Bonniol (1986) et Nunziati (1990) ou du *self-regulated learning* porté par Butler et Winne (1995). D'ailleurs Perrin-Glorian et Hersant précisent que :

> A l'intérieur de ces *mésocontrats* didactiques, le professeur prend une part plus ou moins grande de la responsabilité par rapport au savoir qu'il désire enseigner, d'une part dans la production de la connaissance et, d'autre part, dans l'évaluation et la validation des réponses des élèves" (p. 244).

Ces différentes approches du contrat didactique contribuent, de mon point de vue, à envisager comme pertinente la prise en compte spécifique d'un contrat didactique lié à l'évaluation.

b- Contrat didactique en évaluation

Je considère donc que le contrat didactique rattaché à un épisode évaluatif s'apparente à un mésocontrat, voire microcontrat, tels que définis par Perrin-Glorian et Hersant (2003) et qu'il est spécifique d'un moment particulier de l'étude parce que les dimensions de pouvoir (Brousseau) et de négociation (Chevallard) de chacun des protagonistes (professeur·e et élèves), du fait ses finalités et de ses enjeux, y sont plus aiguës. Je propose de nommer ce contrat *Contrat Didactique en Évaluation* pour permettre d'appréhender les spécificités de ce contrat particulier.

Dans ce contrat didactique en évaluation les notions fondamentales du contrat didactique décrites par Sarrazy (1995) ne sont pas du même ordre car les enjeux et les finalités ne sont pas forcément les mêmes entre une situation didactique ou a-didactique et un épisode évaluatif qui serait, par exemple, proposé en fin de séquence. En effet, entre ces deux situations, le rapport aux connaissances est différent et l'asymétrie des contractant·e·s face au savoir n'est pas la même. Dans les situations didactiques ou a-didactiques, l'enseignant·e vise l'acquisition de connaissances que l'élève ne possède pas encore alors que dans les épisodes évaluatifs de fin de séquence, l'enseignant·e vise à permettre à l'élève de rendre compte des connaissances qu'il ou elle a ou non acquises. Le *jeu*, la *notion d'incertitude* ainsi que la *négociation* ne jouent pas le même rôle, même s'ils restent utiles pour penser ce contrat de nature spécifique. Les notions de *communication didactique* et de *bruit* venant perturber cette communication sont, par contre, très pertinentes pour étudier le contrat didactique en évaluation associé à un épisode évaluatif.

Selon la nature de l'épisode évaluatif considéré, le contrat peut être différent, aussi bien pour l'enseignant·e que pour les élèves. Pour chaque épisode évaluatif, il convient donc de considérer le contrat didactique en évaluation en jeu car ce contrat n'est pas sans incidence sur la façon dont l'élève va être amené·e à produire une réponse et la façon dont le ou la professeur·e va en tenir compte pour son enseignement. Dans un des chapitres de son livre intitulé "*the 'didactical contract' versus the 'assessment contract'*", Van Den Heuvel Panhuizel (1996) estime également que l'évaluation a des conséquences sur les interactions entre l'enseignant·e et ses élèves et cite Elbers (1991b, 1991c, 1992) qui inclut un "contrat d'évaluation" dans le "contrat didactique" constituant le fondement de la communication au sein d'une classe, avec des accords qu'elle qualifie d'implicites.

Le contrat didactique en évaluation s'inscrivant dans la logique du contrat didactique, les effets de contrat pointés par Brousseau (1983) gardent leur pertinence et pourraient même l'accroître du fait de la spécificité du moment et de la négociation qu'ils sous-tendent. Brousseau évoque lui-même, à propos de l'effet Topaze, "une sorte de négociation dérivant

du fait que Topaze propose successivement des problèmes de moins en moins difficiles et finit par vider la question de son sens" (p. 2). Il est vrai que l'exemple de la dictée de M. Topaze illustrant l'effet de contrat qui porte son nom est, de fait, un épisode évaluatif et qu'il correspond à une pratique évaluative assez répandue. L'effet Jourdain est un autre effet que l'on peut aisément retrouver dans de nombreux épisodes évaluatifs. Par exemple, quand un·e professeur·e considère comme correcte une réponse d'élève qui ne témoigne que partiellement de l'acquisition d'une connaissance, mais que, pour une raison qui lui est propre il ou elle choisit de valider.
Je reviendrai ultérieurement sur ces effets, mais ils m'incitent à spécifier plus explicitement ce qui peut se jouer aussi bien pour un·e professeur·e que pour les élèves lors d'un épisode évaluatif. Pour ce faire, je propose d'évoquer deux moments dont les professeur·e·s de mathématiques font usage de manière contingente et que je considère comme étant des épisodes évaluatifs.

- ***Les interrogations "surprise"***

Les interrogations "surprise" font partie de ces épisodes évaluatifs qui ne sont pas anodins et dont le contrat didactique en évaluation mérite d'être étudié. Quel est l'objectif de l'enseignant·e qui fait le choix d'en faire usage et comment les élèves appréhendent-ils ce type d'épisodes évaluatifs ? Quelle négociation engendrent-t-ils ?
Si l'on étudie ce type d'épisodes évaluatifs du point de vue de leur contrat didactique en évaluation, sans tenir compte de leur contenu, on peut les interpréter selon deux scénarios opposés, même si évidemment, dans la réalité d'une classe il y en a bien d'autres possibles :

1er scénario : le ou la professeur·e souhaite s'assurer que ses élèves ont "appris leur cours" et ajuster son enseignement en fonction des réponses obtenues. Pour cela, il ou elle propose quelques questions assez "simples", avec des niveaux de mise en fonctionnement de connaissances plutôt techniques. Cet épisode évaluatif peut alors s'apparenter à une évaluation formative qui permet éventuellement au professeur·e de récolter une note qu'il intégrera à la notation finale qu'il doit fournir en fin de session pour chaque élève. De leur côté, les élèves savent que leur professeur·e pratique ce type d'épisode évaluatif et veillent donc à "apprendre" régulièrement leur cours. Ils ou elles n'apprécient pas forcément le fait de ne pas être prévenus à l'avance, mais ils ont appris à anticiper le moment où leur professeur·e les propose, ni trop tôt, ni trop tard et sont familiers du format homogène adopté. Cela leur permet également d'obtenir assez facilement des notes plutôt élevées qui permettront d'augmenter leur notation de fin de trimestre, ce qui les satisfait. Le contrat didactique en évaluation de cet épisode évaluatif est alors explicite quand bien même il intègre des éléments qui ne le sont pas forcément, mais il peut s'accorder avec le principe d'une évaluation "pour apprendre".
2ème scénario : le ou la professeur·e propose ce type d'épisode évaluatif pour faire valoir sa position dominante vis-à-vis du savoir et en use au gré de l'étude, de manière contingente, pour contraindre ses élèves à systématiquement revoir ce qui a été fait en classe. Il n'a pas d'autre usage de cet épisode évaluatif que celui d'exercer une pression constante sur les élèves. De leur côté, les élèves craignent ces moments d'évaluation car ils n'ont pas de repère pour les anticiper et redoutent leur caractère hétérogène. Le contrat didactique en évaluation de ce type d'épisode évaluatif est contre-productif du point de vue des apprentissages des élèves, pour la majorité d'entre eux à n'en pas douter. Il concourt également, très certainement, à une vision négative des mathématiques et de la façon dont on les apprend.

On voit comment, à travers ces deux scénarios différents et le contrat qui leur est associé, le rôle que peut jouer un épisode évaluatif dans le processus d'apprentissage et la nécessité de

prendre en compte, au-delà des contenus proposés, le contrat didactique en évaluation associé à chaque type d'épisode évaluatif proposé dans le cours de l'étude par l'enseignant·e.

- ***Les interrogations de début de séance***

En début de séance, certains professeurs ont l'habitude d'interroger quelques élèves de leur classe pour rappeler ce qui a été vu précédemment tout en proposant une transition avec la séance à venir. Le moment et la gestion de ce type d'épisode évaluatif pourraient en faire des moments d'évaluation "pour apprendre", mais des jeux de contrat peuvent biaiser cette orientation. En effet, un·e élève sait très bien que s'il ou elle vient d'être interrogé·e par son enseignant·e, il ou elle a peu de chance de l'être à nouveau au cours des prochaines séances et inversement, un·e élève n'ayant pas été interrogé·e depuis un certain temps prévoit qu'il ou elle va certainement l'être prochainement. Le but du "jeu" pourrait donc être, pour un·e élève, d'anticiper le moment où il ou elle sera interrogé·e plutôt que de rentrer dans celui du professeur·e qui veut l'amener à s'inscrire dans le processus d'apprentissage qu'il lui propose. Cet épisode évaluatif permet d'appréhender les décalages pouvant advenir suivant les interprétations et finalités personnelles des un·e·s (professeur·e) et des autres (élèves) et les conséquences en termes d'apprentissage qui peuvent en résulter.

Entre jeu de pouvoir dominant-dominés ou partenaires d'un jeu "gagnant-gagnant", on peut voir à travers ces deux *épisodes* évaluatifs particuliers, comment les points de vue des professeur·e·s et des élèves peuvent varier, mais plus généralement, voyons ce qui peut les sous-tendre.

Du côté de l'enseignant·e

La façon dont l'enseignant·e conçoit les apprentissages des élèves dans un cadre scolaire et les représentations qu'il a de l'évaluation sont au cœur des contrats didactiques en évaluation attachés aux différents épisodes évaluatifs proposés dans le courant de l'étude. Dans son rapport sur l'évaluation des étudiants à l'université, Romainville (2002) fait référence à Samuelowicz et Bain (2002) qui ont établi trois profils d'évaluateurs ou d'évaluatrices qui me semblent intéressants à considérer du point de vue de la notion de contrat didactique en évaluation.

- Dans le **premier profil**, les professeur·e·s cherchent surtout à mesurer la capacité de l'élève à reproduire de l'information, telle qu'elle a été présentée durant l'étude.
- Dans le **deuxième profil** se trouvent les professeur·e·s qui cherchent à mesurer la capacité de l'élève à reproduire de l'information et à l'appliquer à des situations nouvelles.
- Au sein du **troisième profil**, les professeur·e·s conçoivent des épreuves qui mesurent la capacité de l'élève à intégrer, transformer et utiliser de manière personnelle des connaissances.

Même si ces profils ont été établis pour des professeur·e·s enseignant à l'université je considère que, du fait du manque de formation et de cadrage institutionnel autour de l'évaluation en France, ils sont emblématiques de pratiques d'évaluation plus générales et qu'ils gardent une certaine pertinence pour décrire une réalité évaluative correspondant à des épisodes évaluatifs de fin de séquence dans le premier ou second degré.

Le **premier profil** correspond aux professeur·e·s qui veillent à proposer en évaluation des tâches évaluatives correspondant exactement (ou presque) aux tâches proposées dans le temps de l'étude. Pour ces professeur·e·s, les élèves ont appris s'ils sont capables de répondre correctement à des tâches auxquelles ils ou elles ont été confronté·e·s durant l'étude. Le contrat plus ou moins explicite passé avec les élèves sera donc de cet ordre et si l'enseignant·e donne une tâche "nouvelle" ou "inédite", il y a rupture de contrat. On peut aisément comprendre comment l'évaluation par compétences peut être parfois difficile à concevoir et réaliser par les professeur·e·s ayant un tel profil, la dimension inédite de la tâche revendiquée dans certaines approches de la compétence (Beckers 2002, Rey, Carette, Defrance et Kahn,

2003) étant ainsi rendue inopérante. Dans cette catégorie se trouvent certainement davantage des professeur·e·s des écoles qui, comme j'ai pu le constater lors de ma première étude des pratiques évaluatives en mathématiques à l'école ont été nombreux (plus de 80%) à revendiquer de donner à leurs élèves "exactement" ce qu'ils proposent en cours, ce que j'ai d'ailleurs pu constater en confrontant les tâches qu'ils ou elles avaient données en évaluation à celles données durant la séquence de numération étudiée.

Le **deuxième profil** correspond peut-être davantage à des professeur·e·s exerçant à partir du secondaire. Pour eux, l'apprentissage implique une transformation ou une réinterprétation des connaissances acquises de manière à répondre aux exigences d'une tâche nouvelle. Le contrat didactique en évaluation en jeu intègre donc, pour ces professeur·e·s et pour leurs élèves, une part d'incertitude plus élevée et une négociation moins évidente que dans le profil précédent.

Le **troisième profil** est certainement le moins répandu dans l'enseignement primaire et secondaire. Il correspond à des épisodes évaluatifs qui génèrent indubitablement une négociation plus difficile, voire inexistante, entre l'enseignant·e et les élèves. Les recommandations institutionnelles de "bienveillance" ne semblent pas compatibles avec le type de contrat didactique en évaluation correspondant aux professeur·e·s de ce profil, à moins qu'une gestion particulièrement soutenue puisse *in fine* en faire un moment d'apprentissage.

Au-delà des enjeux d'apprentissage et de négociation de ces apprentissages, nul ne peut ignorer qu'une relation de pouvoir est implicitement embarquée dans chaque épisode évaluatif et que l'enseignant·e occupe, par la nature même de sa fonction, la position haute. Libre à lui ou elle de l'utiliser à des fins d'apprentissage ou à d'autres fins. Dans ces trois profils, le niveau d'exigence vis-à-vis des connaissances à acquérir est posé, plus ou moins explicitement, par l'enseignant·e et les élèves doivent s'y soumettre s'ils veulent réussir les différents épisodes évaluatifs.

Par ailleurs, d'autres composantes interviennent dans la nature des contrats didactiques en évaluation qui ont cours lors des différents épisodes évaluatifs proposés par un·e professeur·e à ses élèves et notamment la composante sociale. Dans les classes ou établissements socialement défavorisés, les professeur·e·s ont souvent à cœur de ne pas mettre les élèves en échec et de renvoyer une image valorisante de leur classe pour contrer une fatalité sociale et culturelle qu'ils veulent combattre. Ils peuvent alors être amenés à revoir à la baisse leurs exigences en termes de savoirs et ainsi mener une négociation plus "molle" avec des conséquences souvent néfastes pour les élèves. Ainsi, lors de l'expérimentation que j'ai menée sur les stratégies de réponses des élèves à des QCM, qui intégrait des degrés de certitude permettant de rendre compte de la confiance que les élèves accordaient à leurs réponses, il s'est avéré que les élèves des classes de REP qui ont le moins réussi les items proposés ont souvent été ceux ou celles qui étaient le plus assuré·e·s de leurs (mauvaises) réponses. Ce constat relève, à tous les niveaux (didactique, social, institutionnel), d'une dérive potentielle du contrat didactique et plus spécifiquement des contrats didactiques en évaluation instaurés dans ces classes. Les travaux des didacticien·ne·s sur les pratiques enseignantes en milieu défavorisé (Butlen, Peltier-Barbier, Pézard, 2004 ; Coulange, 2013 ; Chesnais, 2014) ont bien montré comment les professeur·e·s sont amené·e·s à réduire leurs exigences au niveau des contenus proposés aux élèves quand les contraintes sociales sont trop fortes. Partant du principe que les pratiques évaluatives des professeur·e·s sont intégrées à ces pratiques globales, je considère que le constat est valable pour les pratiques d'évaluation et que l'étude des contrats didactiques en évaluation est une entrée pertinente pour essayer de comprendre pourquoi, en France, les inégalités scolaires sont plus fortement corrélées aux

inégalités sociales, culturelles et migratoires (PISA 2012, 2015, rapport CNESCO, 2016).

Du côté des élèves

Concernant l'évaluation, il faut avoir en tête que le graal des élèves, au-delà des apprentissages qu'ils doivent réaliser dans le cadre de leurs études, est d'avoir une "bonne note", quels que soient les épisodes évaluatifs auxquels ils ou elles sont soumis. Coppé (1998) a d'ailleurs éprouvé cette réalité lors de son étude sur les devoirs surveillés au lycée puisqu'elle a dû reconsidérer la définition du problème de l'élève qu'elle s'était donnée pour prendre en compte le fait que, pour les élèves, l'enjeu principal était "d'avoir la meilleure note possible dans le temps donné" (p. 140). C'est donc aussi à partir de ce principe qu'il faut analyser ce qui se joue, pour les élèves, dans les contrats didactiques en évaluation associés aux différents épisodes évaluatifs qui leur sont proposés et ce, quel que soit le niveau scolaire considéré.

Pour réussir un épisode évaluatif, l'élève développe des stratégies qui relèvent aussi bien de ses apprentissages que d'éléments extérieurs. Au-delà des réponses qu'il ou elle doit fournir et que j'étudierai ultérieurement, l'élève peut être amené à faire des vérifications. Coppé (1993) a distingué les processus de vérification internes (aux mathématiques) des processus de vérification externes (qui se rapportent au contrat didactique) pour rendre compte des différentes vérifications que les élèves étaient amené·e·s à faire lors d'une activité de résolution de problèmes mathématiques. Elle a souligné l'importance de prendre en compte les rapports institutionnels et personnels au savoir des élèves lors des moments de contrôles qui s'apparentent, dans le cadre présenté ici à des épisodes évaluatifs particuliers. Elle a décrit le rapport institutionnel à un objet de savoir à partir :

- des "savoirs officiels" décrits dans le cours.
- des savoir-faire relatifs aux objets de savoir en jeu qui ont été durant le cours soit, désignés par l'enseignant·e, soit identifiés comme tels par les élèves.
- d'autres rapports institutionnels qui ne sont pas en jeu au moment de l'épisode évaluatif.
- des savoirs relatifs aux conditions d'utilisation de ces savoirs que Chevallard (1985) a qualifié de protomathématiques.

Coppé (1998) a également décrit d'autres rapports entrant en jeu dans l'épisode évaluatif qu'elle a étudié. Il s'agit :

- du rapport personnel à l'objet "exigences du maître"
- du rapport personnel à des objets dépendant du contrat didactique plus général tels que, par exemple, l'idée que l'élève se fait de ce qu'est une démonstration.
- le rapport personnel à l'objet "limitation du temps".
- le rapport qualifié de croyance de l'élève sur lui-même vis-à-vis d'objets qui peuvent être à la fois des savoirs spécifiques et à la fois des situations scolaires de type résolution de problème, interrogation, travail en groupe, etc.

Ces éléments ainsi que les notions de composantes privée et publique dans le rapport personnel aux objets de savoir (Chevallard, 1988b) me semblent très pertinents pour rendre compte de ce qui se joue, pour un élève, dans les contrats didactiques en évaluation auxquels il ou elle est confronté·e lors des épisodes évaluatifs qui ponctuent ses apprentissages.

c- Exemples

Je donnerai pour terminer cette partie, deux exemples d'épisodes évaluatifs de nature très différente (l'un relevant d'une évaluation externe, l'autre d'une évaluation interne

particulière) pour illustrer mon propos, mais avant je souhaite évoquer une expérience personnelle que je décode *a posteriori* en termes de contrat didactique en évaluation :
Lorsque j'ai été amenée à faire passer des tests dans des classes pour étudier les réponses et les stratégies des élèves de CM2 pour des items mathématiques présentés sous la forme de QCM, j'ai veillé à établir (inconsciemment) un *contrat expérimental* (Schubauer-Leoni, 1988) qui correspond tout à fait à un contrat didactique en évaluation. En effet, j'ai précisé aux élèves que le test proposé ne participait en aucune façon à leur évaluation de classe et qu'il ne "compterait pas" dans leur livret, afin de ne pas "parasiter" mon expérimentation par des éléments du contrat didactique établi entre les professeur·e·s et leurs élèves que je ne connaissais pas, mais qui auraient pu immanquablement interférer dans les réponses des élèves.
Dans le premier exemple, je souhaite mettre en avant l'importance que peut avoir le rapport qualifié de croyance de l'élève par Coppé (1998) autour d'une problématique qui me tient à cœur et que j'explore régulièrement dans mon travail (égalité scolaire entre filles et garçons) alors que dans le second, je souhaite montrer comment l'explicitation d'un contrat didactique en évaluation peut avoir un impact sur la façon dont les élèves peuvent être impliqués dans un épisode évaluatif.

Exemple 1 : *l'expérimentation de Spencer, Steele & Quinn*[10] (1999)
Dans cette expérimentation, des chercheurs ont proposé à deux groupes mixtes d'élèves de reproduire un dessin constitué de figures géométriques imbriquées. Dans l'un des groupes, ils ont précisé que cet exercice relevait d'une évaluation en mathématiques alors que dans le second, ils ont évoqué une séance d'Arts Plastiques. Les résultats de cette expérimentation, qui peut être assimilée à un épisode évaluatif externe, ont été pour le moins surprenant puisque les filles et les garçons ont réalisé des scores différents suivant l'étiquetage de la séance, les premières réussissant moins bien que les seconds à l'évocation des mathématiques. Du point de vue du contrat didactique en évaluation on voit comment les représentations des élèves vis-à-vis de cet épisode évaluatif varient suivant le sexe des élèves, même si évidemment chaque élève peut individuellement en adopter un qui lui est propre.
La façon dont les élèves se représentent les tâches auxquelles ils ou elles sont confronté·e·s durant un épisode évaluatif ne peut donc être négligée et doit être prise en charge par l'étude du contrat didactique en évaluation associé.

Exemple 2 : *l'EPCC "l'évaluation par contrat de confiance"*
L'EPCC est un dispositif d'évaluation spécifique proposé par Antibi (2003). C'est un épisode évaluatif proposé en fin de séquence avec des tâches prescrites à l'avance par l'enseignant·e (pour environ ¾ d'entre elles, en général) et travaillées par les élèves en amont de l'épreuve. Le contrat didactique en évaluation est clairement spécifié dans le cadre de ce dispositif, même si tous les termes ne sont pas précisés explicitement (notamment autour de la partie non programmée de l'évaluation). Élèves et professeur·e savent explicitement ce qu'ils ou elles doivent faire et attendre de l'autre par rapport au savoir en jeu. Dans ce dispositif, le jeu, les parts d'incertitude et de négociation ainsi que la rupture sont totalement revus et peuvent engendrer des effets pervers (restriction du type de tâches évaluées/travaillées) ou constructifs (accroissement du travail des élèves) selon le regard que l'on porte sur ce dispositif d'évaluation particulier.

A travers ces deux exemples et les différents contrats didactiques en évaluation qui peuvent advenir durant les épisodes évaluatifs décrits, je souhaite montrer la nécessité de prendre en compte, d'un point de vue didactique, les éléments externes aux mathématiques et notamment

[10] Reprise par Huguet, Brunot et Monteil en 2001, puis Huguet et Régner en 2007.

la composante personnelle pouvant interférer de manière contingente sur les apprentissages lors de moments d'évaluation en classe (principe 3). La prise en compte de ces éléments participe pleinement de la logique de l'*Assessment for Learning*, notamment au niveau de la motivation des élèves.

d- Explicitation du contrat

Dans son livre Van Den Heuvel Panhuizel (1996), faisant référence à Elbers et Kelderman (1991), indique que l'ignorance des buts et des règles de l'évaluation peut avoir des répercussions importantes sur les résultats des élèves. Je suis également de cet avis et, bien que Sarrazy ait décrit les tentatives d'explicitation de contrat comme des dérives inappropriées du concept initial, je considère que le *contrat didactique en évaluation* gagnerait à être davantage explicité pour que les *épisodes évaluatifs* soient plus clairement intégrés au processus d'apprentissage.

- ***Évaluation & apprentissages***

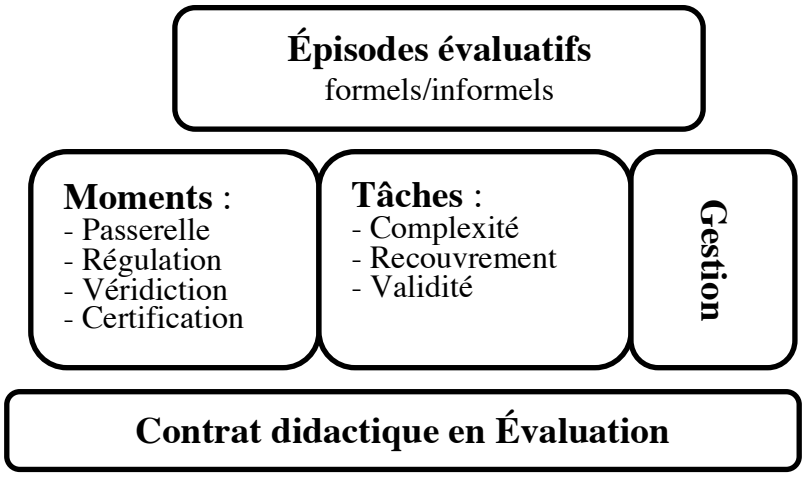
Les moments d'évaluation sont souvent perçus comme des moments détachés des moments d'apprentissage, aussi bien par les professeur·e·s que par les élèves, quand bien même ils s'y rattachent naturellement puisqu'ils se rapportent à des savoirs enseignés. Expliciter leur lien avec les autres moments du processus d'apprentissage et préciser les attentes en termes de savoirs et savoir-faire à acquérir favoriserait une forme de "dévolution évaluative" qui permettrait aux élèves de s'engager de manière plus constructive dans les *épisodes évaluatifs* auxquels ils ou elles sont confronté·e·s périodiquement en acceptant la responsabilité de la situation ainsi que les conséquences de ce transfert (Brousseau, 1988). Cela s'inscrit également dans l'implication plus grande des élèves préconisée dans les approches de l'*AfL* ou de l'évaluation formative, voire formatrice. Scallon (1997) a d'ailleurs évoqué à propos de l'autoévaluation "une démarche visant à la fois à responsabiliser les individus et à les placer aux premières loges du feed-back dans un contexte d'évaluation formative." (p. 29)
Parmi les variantes de l'EPCC, il y en a une que je trouve particulièrement intéressante et qui s'inscrit pleinement dans cette préconisation, il s'agit de l'EPCC participative qui se pratique notamment dans l'académie de Créteil (Quiquempois, 2016). Cette variante consiste à faire prendre en charge par les élèves, par petits groupes, la fiche réussite qui leur permet de préparer le "contrôle". Ce travail est doublement constructif de mon point de vue puisqu'il permet non seulement aux élèves de s'interroger sur ce qu'ils ou elles doivent savoir (ce qui rejoint une des préconisations de Broadfoot et al. 2002 concernant l'*AfL*), mais en plus, ils ou elles le font en confrontant leurs points de vue (ce qui rejoint le concept de *Self-regulated learning* et s'inscrit dans les "*key strategies*" de Leahy et al. 2005).

- ***Relation professeur·e/élèves***

Du point de vue des relations entre le ou la professeur·e et ses élèves, l'explicitation du contrat didactique en évaluation permettrait également de faire évoluer le "partage des responsabilités et des tâches, en fonction de certaines dispositions, capital ou potentiel 'à savoir' des sujets mathématiques" (Ligozat & Leutenegger, 2008, p.329), et donc la topogénèse d'un épisode évaluatif. Il serait alors plus difficile pour le ou la professeur·e d'abuser de sa position dominante puisque, s'il précise ses attentes et les savoirs que les élèves doivent acquérir, ces derniers seront en mesure de dénoncer les ruptures de contrat pouvant advenir lors d'un épisode évaluatif.

Conclusion de la partie A

Pour résumer l'étude d'un épisode évaluatif, il faut donc à la fois s'intéresser au moment où cet épisode est proposé, ce qui permet de situer cet épisode dans l'avancée du savoir, mais aussi prendre en compte la gestion qui l'accompagne ainsi que le contrat didactique en évaluation qui lui est associé pour en mesurer les enjeux didactiques aussi bien pour l'enseignant·e que pour l'élève. Comme on l'a vu, la question de la validité d'un épisode évaluatif doit également faire partie de cette étude, en lien avec les tâches évaluatives proposées et la gestion de cet épisode. Le schéma ci-dessous caractérise donc un épisode évaluatif et soutient l'étude de sa validité didactique telle que définie précédemment :



L'étude de ces épisodes évaluatifs n'a de sens qu'à travers une étude plus globale des pratiques évaluatives des enseignant·e·s en mathématiques. Comment les conçoivent-ils·elles? Comment les intègrent-ils·elles dans leur processus d'enseignement ? Quelle négociation didactique engendrent-ils ? etc. Ces questions sont fondamentales pour comprendre ce qui se joue du point de vue des apprentissages des élèves, c'est pourquoi je propose de les prendre en compte dans l'étude des pratiques évaluatives des enseignant·e·s, en mathématiques.

B. LES PRATIQUES D'ÉVALUATION EN MATHÉMATIQUES

Roditi et Chevallard ont tous deux évoqué l'évaluation comme un moment ou une activité spécifique que l'enseignant·e doit assurer dans le cadre de l'enseignement qu'il ou elle dispense à ses élèves. Pour Roditi (2011), l'évaluation est une des activités organisatrices des pratiques enseignantes intégrant des dimensions personnelle, institutionnelle et professionnelle. Pour Chevallard (1999), l'évaluation est le sixième moment de l'étude qui se situe dans le groupe IV "contrôles", même s'il précise également que "l'ordre indiqué est largement arbitraire" (p. 254). Étudier les pratiques évaluatives en mathématiques des professeur·e·s revient donc à étudier une des activités caractérisant les pratiques enseignantes en mathématiques, activité spécifique se réalisant de manière plurielle et contingente dans le courant de l'étude d'un thème mathématique.

Pour étudier cette activité spécifique, la double approche de Robert et Rogalski (2002) est adaptée puisqu'elle permet de prendre en compte, à la fois l'activité des professeur·e·s et à la fois les apprentissages des élèves qui sont au cœur de toute évaluation. L'approche psycho-ergonomique de la double approche telle précisée par Rogalski (2007, p. 3) peut permettre

d'analyser cette activité d'évaluation déterminée par les propriétés de la situation de travail (compétence 7 du référentiel de compétences professionnelles des professeur·e·s) et par les caractéristiques et états de l'enseignant·e. L'idée de double régulation de l'activité qui a un effet à la fois sur la situation (et donc sur les élèves) et à la fois sur l'enseignant·e coïncide ainsi en partie avec l'idée de régulation des apprentissages telle que préconisée dans l'évaluation formative. L'idée de prendre en compte ses effets (productions, réponses des élèves) par rapport à des attendus (critères d'évaluation) pouvant éventuellement conduire à modifier l'activité (d'enseignement), soit dans le moment même de l'action, soit à plus long terme (Leplat, 1997) peut s'interpréter en termes de régulation des apprentissages. On le voit, cette approche est pertinente pour étudier l'activité d'évaluation des professeur·e·s. Néanmoins, elle ne suffit pas toujours à rendre compte de ce qui se joue, pour tous les acteurs et actrices concerné·e·s, autour de l'évaluation dans la réalité des pratiques de classe :
- Elle ne permet pas d'étudier avec une égale pertinence tous les épisodes évaluatifs dans la diversité de leur nature et de leur réalisation dans les classes. En effet, les épisodes évaluatifs proposés en fin d'étude tels que les contrôles ou les bilans ne s'inscrivent pas forcément dans une perspective de double régulation telle que décrite ci-dessus et l'activité de l'enseignant·e lors de ce type d'épisode ne relève pas de la même dynamique d'action que dans un épisode proposé dans le cours de l'étude. Ces épisodes évaluatifs de fin d'étude sont pourtant ceux qui sont majoritairement proposés dans les classes en France, à tous les niveaux d'enseignement.
- Même si la composante personnelle est bien présente dans la double approche, elle n'est pas suffisamment prise en compte pour analyser les pratiques d'évaluation des enseignant·e·s alors que de nombreux travaux (voir partie précédente) ont montré l'importance de la prise en compte des facteurs personnels dans l'activité d'évaluation. La façon dont les enseignant·e·s conçoivent les épisodes évaluatifs qu'ils ou elles proposent à leurs élèves et les documents évaluatifs associés sont des éléments qui méritent d'être particulièrement étudiés pour analyser, comprendre et faire évoluer les pratiques évaluatives des professeur·e·s. La double approche intègre bien des dimensions liées aux choix ou non choix des professeur·e·s dans l'analyse des pratiques enseignantes, mais elle ne les étudie pas spécifiquement or, la conception et la programmation des épisodes évaluatifs ainsi que la conception des documents associés me semblent des éléments importants à prendre en compte pour étudier les pratiques évaluatives des professeur·e·s en mathématiques. De même que les jugements que les enseignant·e·s sont amené·e·s à émettre lors des épisodes évaluatifs doivent être spécifiquement étudiés car ils sont au cœur des enjeux didactiques de l'évaluation en mathématiques.
- Elle ne permet pas non plus de prendre en compte les pratiques de notation qui jouent pourtant un rôle essentiel dans la négociation didactique de nombreux épisodes évaluatifs. Même s'il n'est pas question d'étudier la notation dans une approche édumétrique, mais bien de la considérer dans une perspective didactique, il est indispensable de prendre en compte cette dimension dans un cadre didactique de l'évaluation.

Le cadre didactique de l'évaluation que je propose s'inscrit donc dans la continuité des travaux de Robert et Rogalski (2002) et de Roditi (2011) sur les pratiques enseignantes, mais il intègre des éléments développés dans d'autres cadres (principe 1) et accorde une place importante à la dimension personnelle des pratiques d'évaluation (principe 3).
Pour analyser ces pratiques, je propose d'étudier les différents épisodes évaluatifs proposés par un·e enseignant·e à partir de la nature des tâches les constituant, des moments où ils sont proposés, leur gestion ainsi que les contrats didactiques en évaluation qui leur sont associés (cf. partie précédente), mais comme je retiens également les principes de complexité et stabilité de ces pratiques en lien avec les travaux de Robert & Rogalski et Roditi, je cherche

aussi à repérer la "logique évaluative" qui anime l'enseignant·e lors de son activité d'évaluation et qui peut s'apparenter à la "raison" évoquée par Perrenoud (1997).
La notion de "logique évaluative" permet ainsi d'appréhender la dimension personnelle des pratiques d'évaluation des enseignant·e·s en mathématiques (principe 3), de mieux comprendre ce qui se joue en matière d'évaluation dans la classe et d'envisager des actions de formation pour la faire évoluer (principe 1). Elle se révèle à travers des indicateurs, plus ou moins observables selon le type d'épisode évaluatif étudié. Dans l'optique indiquée ci-dessus, j'ai choisi de prendre comme indicateurs la façon dont les professeur·e·s conçoivent les épisodes évaluatifs qu'ils ou elles proposent à leurs élèves et les documents évaluatifs associés, les jugements que les professeur·e·s sont amené·e·s à émettre dans le cadre de ces épisodes évaluatifs et la notation qu'ils ou elles adoptent, en lien avec la négociation qu'elle engendre (Chevallard, 1986).
C'est donc l'ensemble des *épisodes évaluatifs* proposés par un·e professeur·e à ses élèves qui constituent, en s'articulant les uns aux autres à partir d'une "logique évaluative" intégrant des dimensions personnelles, institutionnelles et professionnelles, les pratiques d'évaluation en mathématiques d'un·e professeur·e.

Je vais à présent donner quelques précisions sur les indicateurs que je retiens pour étudier la "logique évaluative" des enseignant·e·s.

1. La conception des épisodes évaluatifs

On sait en réalité très peu de choses sur la façon dont les professeur·e·s conçoivent et élaborent les différents épisodes évaluatifs qui ponctuent l'étude d'un savoir mathématique.
L'enquête TALIS *(Teaching And Learning International Survey)* de 2013 a pointé quelques caractéristiques françaises autour de l'évaluation des apprentissages scolaires. Elle a notamment témoigné du fait que plus de 90% des professeur·e·s français·e·s avaient des pratiques d'évaluation très variées et que 86% d'entre eux concevaient leurs propres évaluations (alors que la moyenne des autres pays de l'enquête se situe à 68%).
Dans la recherche que j'ai menée sur les pratiques évaluatives des professeur·e·s des écoles en mathématiques (Sayac, 2016a, 2016b, accepté), j'ai également pu constater une grande variété dans les pratiques de conception des épisodes évaluatifs proposés par les professeur·e·s de l'échantillon, variété rendue possible par l'institution scolaire française. En effet, même si des injonctions nouvelles et peut-être plus affirmées ont été faites récemment à l'adresse des professeur·e·s notamment dans le cadre de la loi de la refondation de l'école de juillet 2103, la liberté pédagogique, au fondement de l'institution scolaire française, alliée à une formation réduite, voire inexistante, à l'évaluation des apprentissages des élèves favorisent cette grande diversité d'approches et de conceptions.

a- Les ressources utilisées

Quel que soit leur niveau d'enseignement, les professeurs ne sont pas des utilisateurs passifs de ressources, mais ils sont des concepteurs actifs, créant et partageant leurs propres ressources, des ressources vivantes (Gueudet, Pépin & Trouche, 2012). Pour élaborer les supports de l'évaluation des apprentissages de leurs élèves, les professeur·e·s utilisent une multitude de ressources de différente nature (numériques, professionnelles, institutionnelles, personnelles, etc.) qu'ils assemblent, transforment et recomposent pour les adapter à leur programmation didactique. Le travail documentaire des professeur·e·s a fait l'objet de nombreux travaux (Gueudet & Trouche, 2008, 2010 ; Wozniak & Margolinas, 2009, 2010), mais dans les faits, peu autour de l'évaluation des apprentissages des élèves qui est pourtant identifiée comme "un des moments principaux d'une leçon" (Gueudet & Trouche, 2010).

Au cours de genèses documentaires (Gueudet & Trouche, 2010), l'enseignant·e est amené·e à développer des documents, en fonction de ses convictions professionnelles et de ses pratiques habituelles (Gueudet, 2013), qui serviront de supports aux épisodes évaluatifs qu'il ou elle envisage de proposer à ses élèves lors d'une étude. Concernant les épisodes évaluatifs, je propose de nommer *document évaluatif* le document développé par un·e professeur·e lors d'une genèse documentaire et destiné à être utilisé lors d'un épisode évaluatif. La liberté pédagogique accordée aux professeur·e·s en France pour évaluer les apprentissages de leurs élèves rend encore plus pertinente l'utilisation de cette dénomination spécifique qui intègre à la fois une grande diversité de ressources et à la fois des schèmes d'utilisation propres à la nature spécifique de chaque épisode évaluatif. Les dimensions institutionnelle, sociale, personnelle et professionnelle, particulièrement marquées lors des moments d'évaluation, confèrent aux documents évaluatifs élaborés par les professeur·e·s un caractère particulier. Ils sont intimement liés aux représentations des professeur·e·s sur l'évaluation et les apprentissages, mais dépendent également de leurs connaissances disciplinaires, didactiques et professionnelles ; ce sont "des entités hybrides, composées de ressources réorganisées et de schèmes d'utilisation structurés par des invariants opératoires" (Gueudet & Trouche, 2010, p. 2). La question de la cohérence mathématique (Margolinas & Wozniack, 2010) de ces documents évaluatifs est d'autant plus pertinente que, pour évaluer les apprentissages de leurs élèves, les professeur·e·s de mathématiques disposent aujourd'hui d'une offre foisonnante de ressources en ligne (Artigue & Gueudet 2008) qui touche également le premier degré (Bueno Ravel & Gueudet, 2014).
Ces documents évaluatifs sont de nature diverse et sont adaptés à l'épisode évaluatif auquel ils sont destinés et aux finalités attendues par l'enseignant·e. Il ou elle les aura plus ou moins conçus et finalisés en amont de l'épisode évaluatif, mais même dans le cas où ils ne l'auraient pas été, ils seront le fruit de genèses documentaires antérieures.

b- Les documents évaluatifs

Dans l'institution scolaire française, ces documents évaluatifs peuvent prendre différentes formes plus ou moins utilisées ou familières des professeur·e·s selon leur niveau d'enseignement, leur formation, leurs préférences, etc. Comme indiqué précédemment, ils peuvent également être plus ou moins élaborés. Un épisode évaluatif peut très bien avoir été programmé avec un objectif précis sans pour autant que l'enseignant·e ait explicitement posé sur sa préparation les différentes tâches évaluatives qui le constituent ou bien, un·e professeur·e peut tout à fait improviser un épisode évaluatif informel (oral ou écrit) au cours d'une séance, s'il ou elle en éprouve le besoin à un moment donné.
On le voit, la diversité des documents évaluatifs correspondant aux différents épisodes évaluatifs peut être très grande. Il serait vain de vouloir tous les identifier car, dans la réalité des pratiques évaluatives des professeur·e·s, ils sont désignés de multiples façons, plus ou moins consensuelles et partagées. Ce qui me semble toutefois important de préciser, c'est ce qu'ils recouvrent par rapport au savoir en jeu et du point de vue de la négociation didactique qu'ils génèrent. Je propose de distinguer trois catégories de documents évaluatifs correspondant aux documents majoritairement utilisés lors des différents épisodes évaluatifs pouvant advenir en classe, même s'ils ne recouvrent certainement pas tous les documents existants : les documents évaluatifs diagnostiques qui permettent aux enseignant·e·s de confronter les savoirs anciens et nouveaux de leurs élèves, les documents évaluatifs partiels conçus pour évaluer des connaissances en cours d'acquisition sur un domaine mathématique et les documents évaluatifs bilan qui recouvrent l'ensemble des savoirs et connaissances visées par le domaine étudié.

Dans la première catégorie se trouvent les documents souvent proposés en amont de l'étude d'un savoir et qui ont une double visée : indiquer à l'enseignant·e les connaissances mathématiques qui pourraient éventuellement manquer aux élèves, celles dont ils ou elles disposent et avec quel niveau de disponibilité, mais aussi préciser aux élèves les connaissances qui vont leur être nécessaires pour l'étude du savoir à venir. Ils sont le support d'épisodes évaluatifs souvent qualifiés de diagnostiques dans la littérature scientifique ou scolaire. D'un point de vue de négociation didactique, ils peuvent s'avérer problématiques dans la mesure où le message implicite de l'enseignant·e qui dit à ses élèves "vous devez savoir ça", peut être entendu par certain·e·s comme "si je ne sais pas ça, je ne pourrai pas apprendre la suite" et donc avoir un impact négatif sur les apprentissages programmés. Dans une approche d'évaluation pour apprendre qui s'inscrit dans celle d'*Assessment for learning (*Broadfoot et al.), cette dimension ne doit pas être négligée c'est pourquoi il me semble important d'évoquer ce type d'impact possible. C'est donc autant par sa gestion que par l'explicitation du contrat didactique en évaluation en jeu, qu'un tel épisode évaluatif pourra avoir une incidence positive ou négative sur les apprentissages des élèves.

Dans la deuxième catégorie se trouvent les documents composés de tâches évaluatives ne recouvrant qu'une partie des savoirs ou des savoir-faire en jeu dans l'étude et/ou avec un niveau de complexité intermédiaire et/ou une variété réduite. On peut y trouver des "interrogations" écrites ou orales proposées aux élèves dans le courant de l'étude, des "contrôles continus" ou bien encore des tests sous forme de QCM ou avec des clickers comme cela se pratique dans certaines classes du secondaire ou du Supérieur.

Dans la dernière catégorie se trouvent les documents qui ambitionnent de rendre compte des connaissances des élèves sur le domaine entier en fin d'une étude. On peut y trouver ce qui est communément appelé "contrôle", "bilan", "examen" ou encore "devoir surveillé" (Coppé, 1998). La question de leur validité se pose en termes de cohérence et en lien avec les paramètres de validité définis précédemment pour les épisodes évaluatifs.

c- Les méthodes

La circulaire de rentrée scolaire 2016, ainsi que de nombreux textes institutionnels parus en France ces derniers temps, enjoignent aux professeur·e·s d'élaborer leurs évaluations de manière collective, mais dans les établissements, cette pratique est contingente et en réalité peu répandue. Dans la recherche que j'ai conduite en 2014-2016 sur les pratiques évaluatives des professeur·e·s des écoles en mathématiques, 72% des professeur·e·s de mon échantillon ont déclaré élaborer leurs documents évaluatifs seul·e·s, même si lors de l'entretien, ils ou elles ont souvent indiqué collaborer avec leurs collègues pour évaluer les apprentissages de leurs élèves (Sayac, 2016a, soumis).

Il va sans dire que la façon dont les professeur·e·s élaborent leurs documents évaluatifs n'est pas neutre, aussi bien du point de vue des savoirs évalués que de la négociation didactique en jeu. Margolinas et Wozniak (2009) ont montré à quel point les professeur·e·s avaient des usages variés de la documentation scolaire et qu'ils ou elles avaient à cœur de s'emparer de diverses ressources pour construire leur *œuvre*. Il est donc indispensable d'en tenir compte dans l'étude des pratiques évaluatives en mathématiques des professeur·e·s. Gueudet et Trouche (2010) ont également montré combien les genèses documentaires au sein de communautés de pratiques pouvaient être transformées et enrichies par les échanges et les interactions entre les différents membres de ces communautés. Dans l'étude que je mène actuellement auprès des professeur·e·s des écoles enseignant en REP+, je recherche des traces de l'impact de la formation dispensée dans le cadre du LéA EvalNumC2 dans les documents évaluatifs élaborés par les professeur·e·s impliqué·e·s dans ce projet de recherche-formation. Je ne suis pas encore en mesure de présenter les résultats de cette recherche en cours, mais

j'espère bien y constater une évolution des tâches évaluatives proposées dans le sens d'un enrichissement en termes de nature, de forme, de complexité et donc de validité.
Dans la perspective d'élargir l'horizon didactique des enseignant·e·s pour penser et concevoir les épisodes évaluatifs qu'ils ou elles proposent à leurs élèves, notamment dans leurs documents évaluatifs, je fais l'hypothèse que les ressources évaluatives correspondant à des épisodes évaluatifs de nature externe (tests, épreuves nationales ou internationales, *high-stakes tests*) peuvent y contribuer, qu'ils soient élaborés par des institutions, des collectifs ou des chercheur·e·s. En effet, l'apport de documents évaluatifs externes peut permettre d'enrichir et de diversifier les tâches évaluatives conçues par les professeur·e·s en en proposant d'autres, potentiellement différentes du point de vue de leur nature, de leur forme ou de leur complexité. C'est ce que je propose d'étudier, dans le cadre du LéA EvalNumC2. J'ai ainsi été amenée à présenter aux professeur·e·s d'un des collectifs des tests en numération ayant une validité didactique certifiée et explicitée par Nadine Grapin afin d'étudier si ces professeur·e·s allaient s'en emparer et le cas échéant, comment les tâches proposées allaient être intégrées, identiques ou transformées, aux documents évaluatifs qu'ils ou elles conçoivent dans le cadre ordinaire de leurs pratiques évaluatives.
Comme je l'ai indiqué précédemment, il ne s'agit bien évidemment pas de tomber dans les dérives du "*teach to the test*", ni même de se conformer strictement à des injonctions institutionnelles ou scientifiques, mais plutôt de s'inscrire dans l'idée de se re-sourcer, au sens donné par Gueudet et Trouche (2010), avec des ressources évaluatives produites dans d'autres communautés. L'étude de l'intégration de telles ressources dans les pratiques évaluatives des professeur·e·s, et notamment les facteurs pouvant la faciliter ou au contraire l'empêcher, fait partie du projet de recherche en cours.
De leur côté, Horoks et Pilet (2016) ont également constaté des évolutions au niveau des pratiques d'évaluation formative des professeur·e·s qu'elles ont étudié·e·s, suite au travail collaboratif réalisé dans le cadre du LéA Pécanuméli. Leur démarche se différencie quelque peu de la mienne dans la mesure où le travail contractualisé avec les enseignant·e·s du LéA Pécanumeli se focalise principalement sur l'enseignement de l'algèbre au collège et vise, de manière contingente, une évolution de leurs pratiques d'évaluation formative, alors que le mien se focalise plus directement sur les pratiques évaluatives des professeur·e·s que je souhaite faire évoluer plus directement, même si bien évidemment en tant que didacticienne, c'est par le contenu (les apprentissages numériques à l'école) que j'aborde ce travail avec les professeur·e·s du LéA EvalNumC2.

2- Le jugement professionnel et didactique en évaluation

De nombreux chercheur·e·s ont mis en évidence l'importance des jugements émis par les professeur·e·s dans leur activité d'évaluation (Morgan & Watson, 2002 ; Scallon, 2004 ; Vantourout, 2004 ; Lafortune & Allal, 2008), certain·e·s conditionnant même la validité d'une évaluation à la qualité de ces jugements (Messick, 1989, 1995 ; Linn & Al. 1991 ; Brookhart, 2003 ; Kane, 2006 ; Cizek, 2009). Je considère également que les jugements portés par les professeur·e·s sur les productions de leurs élèves, la façon dont ils ou elles les émettent et les utilisent sont au cœur d'une approche didactique de l'évaluation. En effet, tout jugement émis par un·e professeur·e devrait porter exclusivement sur des contenus, des procédures et des techniques et non sur des personnes ou sur tout autre considération même si, on l'a vu, de nombreux paramètres peuvent interférer dans ce jugement. Ainsi, dans le cadre didactique de l'évaluation que je propose, la "logique évaluative" d'un·e professeur·e intègre les jugements émis lors des différents épisodes évaluatifs qu'il ou elle programme dans sa classe parce que ces jugements ont un rôle à jouer dans les apprentissages des élèves. Ces jugements évaluatifs se rattachent à des dimensions professionnelle (pour leur rôle dans les

apprentissages) et didactique (par rapport à la nature même de ces apprentissages), c'est pourquoi je propose de les qualifier de jugements professionnels et didactiques en évaluation.

Ces jugements professionnels et didactiques en évaluation s'inscrivent bien évidemment dans la lignée des travaux des chercheur·e·s francophones sur le jugement professionnel en évaluation précédemment cités (principe 2). Je retiens particulièrement la vision de Mottier Lopez et Allal (2009) qui voient ce jugement comme "un acte de discernement et une capacité à construire une intelligibilité des phénomènes d'évaluation en situation" (p. 26), mais accompagnée d'une dimension éthique liée à la spécificité de l'activité d'évaluation (Mottier Lopez & Allal, 2010 ; Schoenfeld, 2007). Cette dimension éthique impose au professeur·e de ne pas porter un tel jugement à partir de données restreintes, mais de l'établir à partir d'un certain nombre d'informations récoltées lors d'épisodes évaluatifs multiples et variés. Mottier Lopez et Allal (2009, 2010) ont utilisé la notion de triangulation des informations recueillies lors d'évaluations comme outil au service du jugement professionnel en évaluation permettant d'augmenter la validité et la fiabilité des données de l'évaluation, mais aussi pour augmenter la pertinence du jugement produit (dimension éthique).
Le jugement professionnel et didactique en évaluation se rattache à la notion de *vigilance didactique* (Charles-Pézard, 2010), mais spécifiquement appliquée à l'évaluation. Cette vigilance se situe à la fois du côté du savoir mathématique, des connaissances didactiques et à la fois de leur mise en fonctionnement dans l'acte d'enseigner (Butlen, Masselot, Pézard, 2011), donc dans l'activité d'évaluation des professeur·e·s. Les professeur·e·s activent leur jugement professionnel et didactique en évaluation lorsqu'ils ou elles doivent :

- Émettre un avis sur l'état des connaissances mathématiques de leurs élèves individuellement et collectivement, à partir des informations récoltées lors de différents épisodes évaluatifs.
- Articuler les différents moments de l'étude entre eux (notamment intégrer les épisodes évaluatifs aux autres moments de l'étude), en fonction des informations récoltées lors des différents épisodes évaluatifs.
- Gérer les épisodes évaluatifs de manière à favoriser les apprentissages de leurs élèves, c'est-à-dire en prenant en compte des paramètres liés à la négociation didactique en jeu et à la motivation de leurs élèves.

Ce jugement professionnel et didactique en évaluation est lié à différents paramètres du professeur·e et joue un rôle fondamental pour penser, concevoir et gérer les différents épisodes évaluatifs qui ponctuent une étude. Parmi les paramètres importants à prendre en compte pour étudier ce jugement se trouvent les connaissances disciplinaires, didactiques et professionnelles des professeur·e·s et des facteurs individuels.

a- Les connaissances disciplinaires, didactiques et professionnelles des professeur·e·s

De nombreux travaux ont montré les liens forts qui pouvaient exister entre les connaissances mathématiques et didactiques des professeur·e·s et leur pratique d'enseignement (Clivaz, 2011, 2014 ; Coulange, 2001 ; Hill et al., 2005 ; Ball, 2005 ; Bloch, 2009 ; etc.). Les travaux de Emprin (2007) ont montré que les enseignant·e·s peuvent mettre en place une situation sans conscience de l'enjeu précis des tâches qu'ils proposent et que la mauvaise connaissance des enjeux mathématiques peut conduire l'enseignant·e à dévoyer la situation qu'il met en œuvre. Concernant spécifiquement l'activité d'évaluation, Vantourout (2004) a mis en lumière le rôle essentiel que jouent les connaissances disciplinaires des professeur·e·s lors de la réalisation d'une évaluation à visée formative en mathématiques ainsi que dans l'activité d'évaluation de productions d'élèves en mathématiques (Vantourout & Maury, 2006). Ces chercheur·e·s ont identifié trois grands pôles de connaissances ("disciplinaire", "évaluation" et "socio-psycho- pédagogique") intervenant dans l'activité d'évaluation des professeur·e·s.

De leur côté, Butlen, Pézard et Masselot (2012) ont souligné que pour exercer une vigilance didactique, la maîtrise des contenus à enseigner est indispensable. Ces dernier·e·s ont précisé que :

> Pour exercer cette vigilance didactique, certes la maîtrise des contenus mathématiques enseignés est nécessaire, mais aussi et surtout que la maîtrise de concepts relatifs à *l'enseignement de ces contenus* est indispensable. Cette dernière implique notamment des connaissances didactiques sur les cheminements cognitifs des élèves, sur les situations qui les accompagnent, sur des résultats de recherches [...], sur les grands types d'erreurs susceptibles d'apparaître lors de l'apprentissage d'une notion donnée, sur les critères permettant d'établir des hiérarchies de procédures susceptibles d'être mobilisées par les élèves, mais aussi sur les outils nécessaires à la mise en œuvre (ou à la lecture) d'une analyse a priori des situations proposées par les ressources accessibles aux enseignants, leur permettant de se les approprier (reconnaissance des enjeux des situations, des intentions des auteurs...) et de les adapter. (p. 80-81)

Appliquées à l'activité d'évaluation des professeur·e·s, ces précisions suggèrent donc qu'au-delà de leurs connaissances mathématiques les professeur·e·s doivent également disposer des compétences évaluatives spécifiques leur permettant de :

- ***Concevoir des épisodes évaluatifs s'intégrant pleinement dans le cours de l'étude et coordonnés entre eux*** pour favoriser un jugement professionnel et didactique en évaluation qui soit valide et efficace (du point de vue des apprentissages des élèves). Cette compétence intègre donc l'élaboration de tâches évaluatives à la fois riches et variées du point de vue des connaissances en jeu (différents cadres, recouvrement du domaine, etc.) et de leur complexité (différents niveaux de mise en fonctionnement, jeu de variables didactiques, etc.) et de documents évaluatifs cohérents et adaptés à ces différents épisodes.
- ***Analyser les réponses des élèves et les intégrer au processus didactique***, qu'elles soient conformes ou non aux attentes de l'enseignant·e et quel que soit l'épisode évaluatif considéré. Je n'ai pas distingué ces deux actions car elles sont pour moi intimement liées et fondamentales pour penser une évaluation au service des apprentissages des élèves. Lorsque DeBlois et Squalli (2002) ont étudié la façon dont des professeur·e·s stagiaires analysent et interprètent les erreurs des élèves, ils ont bien mis en évidence la difficulté que ces professeur·e·s pouvaient rencontrer pour identifier les composantes des concepts en jeu, les raisonnements plausibles d'élèves et élaborer des pistes d'intervention possible, même si ces constats ont été fait dans le cadre de la formation initiale des professeur·e·s et que l'ambiguïté de la posture épistémologique adoptée par des professeur·e·s stagiaires ne permet pas de généraliser ce constat.

Dans le cas d'un épisode évaluatif proposé en cours d'étude, ces deux actions participent de la régulation des apprentissages. Mon propos étant plus large, je considère que cette double action (analyser et intégrer) doit se réaliser pour tout type d'épisode évaluatif.

- ***Effectuer un retour constructif aux élèves,*** du point de vue de leurs apprentissages, en fonction des réponses qu'ils ou elles ont fournies lors de leur confrontation aux différentes tâches évaluatives. On pourrait penser que cette action est intégrée aux précédentes or elle est distincte dans la mesure où un·e professeur·e peut bien analyser les réponses de ses élèves et les intégrer au processus didactique sans pour autant que ce ou ces derniers se sente(nt) directement concerné(s). Bloch (2006, 2009) a spécifiquement travaillé sur la capacité du professeur·e à renvoyer à ses élèves des réactions mathématiquement pertinentes, cette action relève pour moi de cette capacité, spécifiquement mise en œuvre dans les épisodes évaluatifs qui ponctuent une étude. Cette compétence évaluative se rattache à la notion de feedback développée par de nombreux·ses chercheur·e·s (voir partie précédente).

Ainsi, le jugement professionnel et didactique en évaluation relève bien de compétences à la fois professionnelles et didactiques, même si d'autres paramètres, plus personnels interfèrent également dans ce type de jugement.

b- Des facteurs individuels dans la "rationalité" des pratiques évaluatives

Perrenoud considère que "l'évaluation passe par les pratiques d'acteurs, individuels ou institutionnels, qui sont rarement dépourvus de raison et de raisons, mais dont les rationalités sont limitées et diverses, parfois contradictoires." (Perrenoud, 1997 p.16). J'ai bien souvent éprouvé la justesse de ce propos lors de mes échanges avec des professeur·e·s engagé·e·s dans mes recherches. Une anecdote vécue lors d'une rencontre récente avec des enseignant·e·s illustre bien ce que Perrenoud met en évidence et témoigne de pratiques d'évaluation bien réelles et certainement très discriminantes. Un professeur, enseignant en REP+, se plaignait des écarts qu'il était souvent amené à constater entre les réponses de ses élèves aux évaluations sommatives et celles qu'ils ou elles produisaient au cours de la séquence qui les avait précédées. Lorsque je lui ai demandé ce qu'il faisait en réaction, il a indiqué que souvent, il faisait refaire les évaluations sommatives "ratées" aux élèves qu'il estimait compétent·e·s. Ce professeur, qui a à cœur de faire réussir tous ses élèves et plus particulièrement ceux ou celles qui sont le plus en difficulté, donne pourtant, de fait, une deuxième chance aux élèves qui ont potentiellement le moins de difficulté au détriment de ceux ou celles qui en ont le plus.

Les facteurs individuels pouvant influencer les jugements et les pratiques des enseignant·e·s sont nombreux. Je propose de retenir ceux qui sont généralement reconnus en tant que tels et étudiés par des chercheur·e·s qui se sont plus particulièrement intéressé·e·s aux mathématiques (Thompson, 1992 ; Niss, 1993 ; Raymond, 1997 ; Schoenfeld, 1998 ; Beswick, 2006 ; Philipp, 2007) : les croyances et représentations sur l'évaluation (*beliefs* dans la littérature anglophone), les conceptions et les expériences vécues par les professeur·e·s en matière d'évaluation.

- ***Croyances et représentations sur l'évaluation***

Les croyances et les représentations des professeur·e·s sur l'évaluation sont socialement, professionnellement et personnellement construites, en lien avec leurs conceptions sur les apprentissages (Skott, 2015). Schoenfeld (2007) considère que les professeur·e·s agissent en fonction de leurs ressources (connaissances, matériel, etc.), leurs buts conscients ou inconscients et leurs orientations (croyances, valeurs, dispositions, etc.). Dans la recherche menée avec des chercheur·e·s engagé·e·s comme moi dans l'axe 4 de RE.S.E.I.D.A, les croyances et les doxas qui circulent chez les enseignant·e·s ont été investiguées à travers une enquête par questionnaire (Sayac, Crinon & Fersing, 2013). Nous avons montré le lien étroit qui existait entre les stratégies d'enseignement des professeur·e·s telles que perçues à travers leurs réponses au questionnaire et leurs croyances, associées à des éléments d'épistémologie pratique (Marlot, Théry, Sayac & Pironom, accepté).

Par ailleurs, comme le soulignent Rey et Feyfant (2014) en se référant aux travaux de Maulini (2012) :

> Pour de nombreux acteurs sociaux, les classements sont un mal nécessaire dans une société inégalitaire. Dès lors, l'enjeu des évaluations scolaires est de se rapprocher autant que possible de principes d'équité et de justice pour que les classements scolaires, qui donnent accès à des positions sociales différenciées, soient aussi clairs et explicites que possible aux yeux des parents, les élèves et les enseignants.

Les mathématiques étant une discipline jouant un rôle très discriminant dans le système scolaire français, on peut aisément imaginer comment ce type de croyance sur l'évaluation peut influencer les pratiques évaluatives des enseignant·e·s qui la partagent.

- ***Conceptions sur l'apprentissage***

Les conceptions sur l'apprentissage des élèves sont également déterminantes pour asseoir le jugement professionnel et didactique en évaluation des professeurs. De Ketele (1993) a bien montré, lorsqu'il a listé les différents paradigmes de l'évaluation, comment les premiers promoteurs de l'évaluation formative (Bloom, 1971 ; de Landsheere, 1973, 1980), s'inscrivaient dans un modèle d'apprentissage spécifique, celui de la pédagogie de maîtrise

dont le postulat de base est que l'élève peut réussir un apprentissage déterminé dans la mesure où il lui consacre le temps nécessaire. Les fondements de cette approche ont amené Bloom (1971) à distinguer les trois types d'évaluation à la base de nombreux travaux sur l'évaluation. De Ketele détaille ce que Bloom a préconisé dans son livre "Caractéristiques individuelles et apprentissages scolaires" (version originale en 1976 ; version française en 1979) et qui est en lien direct avec l'évaluation formative telle que définie au début :

> Préciser clairement les résultats attendus à la fin d'un cours ou d'une séquence d'apprentissage ; préparer les étudiants pour qu'ils puissent entrer avec fruit dans la séquence d'apprentissage ; enrichir l'apprentissage de rétroactions fréquentes et de démarches correctives ; ne pas passer à l'apprentissage ultérieur si l'apprentissage actuel n'est pas suffisamment maîtrisé. (p. 64)

Shepard (2005) a, de son côté, fait un parallèle entre l'évaluation formative et le concept de ZPD (zone proximale de développement) de Vygotsky (1978, 1987), ce qui montre bien combien les conceptions de l'évaluation et celles relatives aux apprentissages sont liées. Par ailleurs, les profils d'évaluateur décrits plus haut par Romainville (2002) sont immanquablement fondés sur des conceptions d'apprentissage différentes et témoignent du lien existant entre ces conceptions et la manière dont un·e enseignant·e conçoit l'évaluation des apprentissages de ses élèves.

La place accordée aux élèves dans l'évaluation de leurs apprentissages est un élément clé pour appréhender les conceptions des professeur·e·s sur l'apprentissage. En France, le faible pourcentage de professeur·e·s pratiquant l'autoévaluation ou l'évaluation entre pairs témoigne d'une vision de l'enseignement où l'élève a peu de place et est peu impliqué·e dans l'évaluation de ses apprentissages. Comme Coulange (2013), il me semble que l'étude de la topogénèse (Chevallard, 1999 ; Sensevy & Mercier, 2007) dans les épisodes évaluatifs est particulièrement intéressante pour comprendre ce qui se joue entre l'enseignant·e et ses élèves du point de vue des apprentissages de ces derniers *via* leur évaluation.

- ***Expériences évaluatives***

Les expériences évaluatives des professeur·e·s, qu'elles soient professionnelles ou personnelles en tant qu'ancien.ne élève ou étudiant·e, vont également être consciemment ou inconsciemment marquantes pour fonder leur jugement professionnel en évaluation (Raymond, 1997). C'est d'abord en tant qu'élève, puis en tant qu'étudiant·e qu'un·e professeur·e est confronté·e à des jugements évaluatifs qui peuvent influencer la façon dont il ou elle va penser son rôle et sa place dans l'évaluation des apprentissages de ses élèves. Son vécu en tant qu'élève ou étudiant·e selon qu'il ait été constructif ou au contraire douloureux pourra être, dans un premier temps, le référent de ses pratiques évaluatives. Certain·e·s professeur·e·s des écoles ont été confronté·e·s à des situations d'échec en mathématiques et en ont gardé des traces qui influencent ou ont influencé de manière contingente, à un moment donné, le jugement professionnel et didactique en évaluation qu'ils·elles ont développé. De la même façon, les situations de réussite en mathématiques auxquelles ont pu être confronté des professeur·e·s dans leur scolarité ou dans leurs études peuvent également avoir une incidence non négligeable sur le jugement professionnel et didactique en évaluation qui façonne leurs pratiques évaluatives en mathématiques. Ces traces peuvent être de nature différente et pas forcément uniformément corrélées au vécu évaluatif des professeur·e·s.

En tant qu'enseignant·e, les professeur·e·s acquièrent également, au cours de leur carrière, des expériences évaluatives qui font évoluer leurs pratiques et leur jugement professionnel. Ces expériences peuvent se réaliser à travers la confrontation à des dispositifs évaluatifs qui les interpellent ou qui les séduisent. C'est le cas par exemple de l'EPCC qui est utilisée, de manière plus ou moins fidèle au dispositif initial (Antibi, 2005), par un nombre croissant de professeur·e·s, las de dispositifs qui ne leur donnent pas entière satisfaction en matière

d'évaluation des apprentissages de leurs élèves et qui trouvent, dans ce dispositif singulier, des réponses à leur quête évaluative.
Il convient de noter que le niveau d'enseignement joue un rôle important dans ces facteurs individuels et qu'il influence grandement le jugement professionnel et didactique en évaluation des professeur·e·s (Cauley & McMillian, 2000). Issaieva, Pini et Crahay, (2011) ont néanmoins souligné que même si l'âge, le sexe ou le nombre d'années d'expérience professionnelle des enseignant·e·s ont une incidence sur les jugements portés, ces caractéristiques ne sont pas liées à un positionnement particulier.

3- La notation

La notation est le troisième indicateur que je retiens pour étudier la "logique évaluative" des enseignant·e·s car elle est centrale pour comprendre ce qui se joue, entre l'enseignant·e et ses élèves, du point de vue des apprentissages. Les usage(s) et les critères de réussite (Bonniol, 1981) adoptés, la désignation de la "bonne" ou "mauvaise" réponse (Chevallard, 1985) et son rôle dans la négociation didactique ainsi que le message adressé aux élèves (Chevallard & Feldmann, 1986) varient d'un·e professeur·e à un·e autre et participent de sa "logique évaluative".
La façon dont la notation des professeur·e·s va être prise en compte dans le processus didactique participe de la détermination du cadre didactique de l'évaluation que je souhaite défendre. Il convient de préciser en quoi et pourquoi étudier la notation, d'un point de vue didactique, est différent et complémentaire des autres approches (édumétrique, psychologique, sociologique, etc.). Ainsi, même si Delcambre (1994) considère que "l'usage de la note comme instrument d'évaluation contribue à brouiller le paysage didactique" (p. 18) et que "la note opère une globalisation de jugement évaluatif qui est néfaste à la compréhension par l'élève de ce qui la motive" (p. 19), il me semble essentiel d'en appréhender les contours, à partir de ses liens avec le contrat didactique en évaluation et avec le jugement professionnel et didactique en évaluation, mais également à travers la façon dont les élèves vont être amené·e·s à répondre en vue de la meilleure attribution de notation possible.
La notation recouvre la façon dont l'enseignant·e va être amené·e, publiquement, à travers des notes ou des commentaires, à rendre compte du jugement évaluatif qu'il ou elle émet à propos des réponses des élèves produites lors des différents épisodes évaluatifs proposés durant une étude. Publiquement signifie que le jugement évaluatif de l'enseignant·e va se traduire par un codage reconnu par les différents acteurs et actrices du système scolaire (élèves, parents, collègues, établissement, etc.) et interprété selon leurs normes respectives. Ce caractère public de la notation s'inscrit dans la logique des composantes publique et privée utilisées par Chevallard (1988b) pour décrire les rapports des individus aux objets de savoir, mais je la reprends ici car elle me paraît pertinente pour inscrire la notation dans sa dimension institutionnelle, même s'il n'y a pas exactement de normes institutionnelles en termes de notation, mais seulement des préconisations.
Je ne prends pas en compte, dans la notation, les appréciations orales qu'un·e professeur·e peut être amené·e à produire lors d'épisodes évaluatifs verbaux, même si elles peuvent influencer le processus d'apprentissage des élèves et doivent également être étudiées à ce titre. En effet, nul ne peut récuser le fait qu'une appréciation du type "ce n'est pas exactement ça, mais tu y es presque" n'a pas le même impact qu'une autre de type "c'est inexact", mais le caractère volatile de ces appréciations ne permet pas d'en faire un enjeu public de notation.

a- En lien avec le contrat didactique en évaluation

La place, l'interprétation et le rôle de cette notation dans le processus d'apprentissage est une des premières caractéristiques du cadre didactique adopté. Le contrat didactique en évaluation que j'ai défini intègre des dimensions de pouvoir et de négociation qui sont au fondement de la notation scolaire. Notation et contrat didactique en évaluation sont donc intrinsèquement liés et s'ajustent mutuellement à chaque épisode évaluatif. Comment la notation participe-t-elle à l'élaboration/évolution du contrat didactique en évaluation établi au niveau du processus didactique global et au niveau de chaque épisode évaluatif ? Réciproquement comment une évolution du contrat didactique en évaluation peut-elle potentiellement amener des changements en matière de notation ? Comment l'attribution d'une notation va-t-elle être adressée à l'élève ? A la classe ?

Chevallard et Feldmann (1986) précisent bien que "la note assignée n'est pas mesure, mais message. Ce message intervient dans une négociation, ou une transaction, qui signe un rapport de forces entre l'enseignant, les enseignés, à propos du savoir enseigné." (p. 71). Ils illustrent leur propos en spécifiant que quand un·e professeur·e attribue une note de 16,5 sur 20, il ou elle ne met pas la note de 16,5 sur 20, mais il ou elle négocie au quart de point. Chevallard et Feldmann (1986) ont également montré comment, à travers la fluctuation des moyennes et la dispersion des notes, un·e professeur·e pouvait orienter la négociation avec ses élèves.

La notation chiffrée est celle qui est majoritairement adoptée par les professeur·e·s en France (à l'université, dans le secondaire et dès le cycle 3), mais d'autres notations sont également utilisées (lettres, échelle de niveaux, symboles, etc.). Dans tous les cas, la notation adoptée n'est neutre ni professionnellement, ni institutionnellement, ni socialement, et elle joue un rôle important dans le contrat didactique. Puisque la notation est liée à la validité ou non l'épisode évaluatif considéré, on pourrait qualifier d'effet Jourdain du contrat didactique en évaluation la propension d'un·e professeur·e à considérer comme correcte la réponse d'un élève qui n'aurait en fait donné qu'un faible indice de preuve d'acquisition de la connaissance évaluée, mais que l'enseignant·e juge valide pour influencer la négociation en cours. Par exemple, un·e professeur·e qui aurait donné des tâches évaluatives trop "difficiles" lors d'un épisode évaluatif pourrait être tenté·e de valider une réponse partiellement correcte pour ne pas avoir à assumer, institutionnellement et professionnellement, un nombre trop important "d'échecs". À l'inverse, un·e professeur·e pourrait invalider une réponse d'élève potentiellement correcte d'un point de vue mathématique, mais qui n'aurait pas été suffisamment explicitée de son point de vue ou qui aurait témoigné de l'usage d'une technique qu'il ou elle récuse à ce moment-là. Ces situations pourraient alors générer une rupture du contrat didactique en évaluation qui ne serait pas sans incidence sur la négociation en cours. Merle (1996) parle très justement d'arrangements évaluatifs que les professeur·e·s sont souvent amené·e·s à faire pour qualifier ce type de situations.

Par ailleurs, certain·e·s chercheur·e·s ont montré que les attentes de l'enseignant·e envers les élèves influencent l'évaluation de leurs apprentissages. Trouilloud et Sarrazin (2002) soulignent que "les attentes joueraient le rôle de filtres interprétatifs, susceptibles de conduire à des distorsions de la réalité lorsque l'enseignant perçoit, interprète et évalue les actions d'un élève" (p. 72). En considérant que ces attentes relèvent d'un contrat didactique en évaluation, on peut donc également considérer qu'elles ont une incidence sur la notation qui va découler de ces distorsions de la réalité.

La façon dont un·e professeur·e va rendre compte à ses élèves de la notation attribuée à chacun·e est également à relier avec le contrat didactique en évaluation. En effet, selon la forme et l'importance qu'il va donner à cette restitution, les dimensions de pouvoir et de négociation vont être plus ou moins activées et produire des effets sur le processus didactique en cours, différenciés suivant chaque élève. La question de l'attribution "d'échec" doit,

comme le défend Chevallard (1988) être prise en charge et étudiée pour ne pas en rester à l'état de constat funeste :

> Ce que la didactique peut et doit étudier, en revanche, ce sont les situations d'attribution d'échec. En amont de l'attribution, il convient ainsi d'élucider les mécanismes de sa genèse, son "étiologie" ; en aval, la conduite des sujets de l'institution didactique (enseignants, enseignés et autres) face à de telles situations, ainsi que, plus largement, la manière dont se met en place, dans un contexte socio-culturel donné, le "traitement" de l'attribution d'échec. L'essentiel est alors que, si l'institution scolaire d'aujourd'hui prononce, à différents niveaux (ceux de la classe, du conseil de classe, etc.), des verdicts d'échec, elle laisse chacun (enseignant, élèves, parent, responsable) fort dépourvu quant à la conduite à tenir face à de telles situations, donnant ainsi libre cours à des conduites toutes personnelles dont les effets, largement incontrôlés et informés seulement par les divers assujettissements des personnes, sont le plus souvent négatifs. (p. 4).

La question récurrente autour de la suppression des notes qui fait régulièrement débat en France est souvent appréhendée d'un point de vue strictement psychologique, mais plus rarement d'un point de vue cognitif ou en termes de contrat didactique, pourtant elle serait intéressante à étudier de ces points de vue-là. Si l'enseignant·e ne disposait (ou choisissait de ne plus disposer) de cet outil de négociation didactique, quel serait l'impact de cette suppression en termes de contrat didactique en évaluation ? Comment l'enseignant·e pourrait-il ou elle signifier à ses élèves à quelle hauteur ils doivent se situer vis-à-vis de telle ou telle connaissance, à quel niveau d'exigences il ou elle souhaite les amener ? Et comment l'enseignant·e pourrait, personnellement, estimer l'impact de son enseignement ? Quelles que soient les réponses apportées à ces questions, ce qui me semble important de considérer, c'est que la négociation aurait bien évidemment toujours lieu, mais qu'elle se situerait à un autre niveau qui rendrait peut-être les verdicts moins préjudiciables pour certains élèves et surtout, plus constructif du point de vue de leurs apprentissages. Mon propos n'est pas ici de promouvoir un enseignement sans note car, à côté des effets négatifs, se trouvent également des effets positifs tels que la motivation ou la stimulation, mais c'est simplement de défendre l'idée que la notation n'est qu'un outil de négociation didactique qui peut se réaliser autrement, notamment à travers un contrat didactique en évaluation plus explicite et co-construit par l'enseignant·e avec ses élèves. Les travaux sur l'évaluation formatrice (Bonniol, 1986 ; Nunziati, 1990 & Vial, 1995 ; Nunziati, 1990) sont à interroger à l'aune de ces questionnements.

b- En lien avec le jugement professionnel et didactique en évaluation

Le jugement professionnel et didactique en évaluation porté par un·e professeur·e sur la production d'un·e élève se situe en amont de toute notation et permet d'en rendre compte publiquement. Chevallard (1989) perçoit d'ailleurs la notation comme l'énonciation d'un jugement. Elle est en quelque sorte une traduction institutionnelle et publique de ce jugement, même si elle n'en est qu'une facette (Bressoux & Pansu, 2003).

Tout professeur·e se doit d'évaluer les progrès et les acquisitions des élèves et d'en rendre compte à ses élèves, à leurs parents et à l'institution. La première injonction relève du jugement professionnel et didactique en évaluation alors que la seconde se réfère à la notation, mais elles sont toutes deux intimement liées. Tout·e professeur·e est donc amené·e à adopter une notation lui permettant de rendre compte institutionnellement et publiquement de l'état des connaissances de ses élèves, mais également d'opérationnaliser cette notation dans le cadre d'une négociation didactique.

- ***Choix de la notation***

Pour Chevallard et Feldmann (1986) "l'enseignant-évaluateur n'est nullement assimilable à un appareil de mesure susceptible d'opérer indéfiniment ; son univers de référence est clos, ses 'mesures' ne sont nullement indépendantes entre elles, elles dépendent les unes des autres et des objectifs qu'il s'assigne (et qu'il modifie d'ailleurs) au cours de la conduite du processus didactique" (p. 121). La notation qu'un·e professeur·e adopte s'inscrit dans sa

“logique évaluative” et résulte soit d’un choix personnel issu de son “univers de référence” propre, soit d’un choix lié à des contraintes institutionnelles ou professionnelles (adoption d’une même notation au sein d’un établissement, ou d’un même niveau scolaire, etc.), mais dans tous les cas, elle intègre une dimension institutionnelle forte. Chevallard (1989) précise d’ailleurs que “tout jugement suppose, chez celui qui le porte, qu’il engage dans son énonciation ce que j’ai appelé une passion institutionnelle. Tout “juge” parle, et soutient son dire, en tant que sujet passionné d’une institution – quand bien même il ne parlerait pas “au nom” de l’institution. L’énonciation du jugement serait-elle privée qu’une telle référence ne saurait être absente, même si nous ne la reconnaissons pas au premier coup d’œil.” (p. 10).

Dans ma thèse (2003), parmi les diverses questions que j’avais posées aux professeur·e·s de mathématiques en lycée pour cerner leurs pratiques, il y en avait une qui avait été donnée à titre heuristique et qui s’est avérée très discriminante du point de vue des pratiques enseignantes, il s’agissait de la question suivante :

Comment élaborez-vous vos moyennes trimestrielles ?
❒ En arrondissant la moyenne des notes au demi-point supérieur
❒ En arrondissant la moyenne des notes au point supérieur
❒ En établissant la moyenne exacte des notes, au dixième près
❒ Autrement

Les réponses des professeur·e·s de l’échantillon à cette question correspondent assez précisément aux cinq catégories de professeur·e·s que j’ai identifiées et qui ont des caractéristiques communes au regard des déterminants retenus (sexe, âge, cursus). Il s’est, par exemple, avéré que les professeur·e·s appartenant au genre “conservateur” avaient davantage tendance à choisir la notation au dixième près alors que ceux ou celles du genre “didacticien” ont majoritairement préféré la notation “au point supérieur”. Ce résultat très marqué m’avait déjà amenée à m'interroger sur le lien entre pratiques enseignantes et pratiques de notation, mais ce n’était pas dans mes préoccupations scientifiques à ce moment-là. Au-delà de la notation adoptée, son usage et sa conception doivent donc être pris en compte pour rendre compte de la “logique évaluative” d’un·e· professeur·e.

- ***Grille de la notation***

Dans son étude sur la fabrication des notes dans le secondaire, Merle (2007) a montré que les notes étaient le fruit d’un “bricolage” ou d’arrangement évaluatif. Ce qu’il me semble important de retenir de ce résultat, c’est que ce “bricolage” ou cet arrangement se fait à partir des tâches évaluatives choisies par l’enseignant·e en lien avec les savoirs enseignés et qu’il n’est donc pas sans incidence sur le processus didactique. Schubauer-Leoni (1991) a, très justement, précisé que “les hésitations du maître dans la construction de la note mettent en lumière des éléments du fonctionnement didactique de l’évaluation dans une classe.” (p. 93). L’élaboration d’un barème a parfois une incidence forte sur le document évaluatif élaboré par les professeur·e·s et peut avoir des conséquences sur sa validité. C’est effectivement ce que j’ai pu constater lors de mon étude sur les pratiques évaluatives en mathématiques de professeur·e·s des écoles. Pour illustrer mon propos, je donnerai l’exemple d’une professeure de CM2 ayant participé à l’étude qui, pour faciliter le comptage des points du document évaluatif bilan qu’elle avait donné à ses élèves après une séquence sur les fractions, a proposé un exercice sur 5 points, comportant 10 fois la même tâche évaluative, créditée chacune d’un demi-point. Le fait que les dix occurrences de cette tâche n’aient pas été conçues pour affiner le niveau de connaissances des élèves et qu’elles aient été redondantes du point de vue des connaissances à évaluer a participé à la remise en cause de la validité du document évaluatif élaboré par cette professeure.

- ***Attribution d'une notation***

Pour attribuer une notation à la production d'un élève, l'enseignant·e doit interpréter ses réponses et leur accorder les crédits correspondant à la grille qu'il ou elle a élaborée. Or, on le sait, quelle que soit la finesse de cette grille, elle ne pourra croiser toutes les procédures et réponses que les élèves peuvent produire (Amigues & Zerbato-Poudou, 1996 ; Merle, 2015). Une harmonisation des notations est systématiquement prévue lors de la correction des épreuves certificatives nationales (DNB, Baccalauréat) et, même si un barème précis, discuté avec les correcteurs et correctrices est toujours établi, elle s'avère souvent nécessaire et indispensable tant les interprétations des professeur·e·s peuvent être variées et subordonnées à la diversité de leur jugement. L'interprétation des réponses des élèves peut, au-delà des connaissances disciplinaires et didactique du professeur·e, être perturbée par la prise en compte de composantes socio-culturelles telles que le sexe de l'élève (Bressoux, 2000 ; Marro, 1995 ; Mosconi, 2001 ; Terrier, 2014) ou l'appartenance socio-professionnelle des parents (Messick, 1989 ; Schubauer Leoni, 1991 ; Delcambre, 1994, etc.). Estimer qu'un élève a fait une "faute d'étourderie" ou qu'il n'a "pas acquis la connaissance" ne relève pas du même jugement évaluatif et n'aboutit certainement pas la même notation.

Les travaux de Vantourout, Goasdoué, Maury et Nabout (2012) à partir de situations mathématiques "aménagées" ont également témoigné du fait que des divergences de notations ne recouvraient pas forcément des divergences de jugements évaluatifs, ce qui tend à montrer que la question de l'attribution d'une notation est un phénomène complexe qui ne dépend pas seulement des connaissances disciplinaires et didactiques des professeur·e·s, même si, bien évidemment, elles sont indispensables. Par ailleurs, Pluvinage (1979) a pointé les risques inhérents à toute situation d'évaluation, notamment ceux pris par l'évaluateur ou l'évaluatrice et qu'il identifie par :

- attribuer un résultat négatif à un individu qui satisfait aux exigences,
- attribuer un résultat positif à un individu qui ne satisfait pas aux exigences.

Ce que je souhaite mettre en évidence ici c'est, qu'au-delà des questions de justice scolaire ou de constante macabre (Antibi, 2003), le processus didactique est perturbé par ces attributions de notation différenciées et qu'il est essentiel que les professeur·e·s puissent le réaliser pour limiter ces perturbations ou bruits.

c- En lien avec les réponses des élèves

Comme le dit Sensevy (2008), "l'Élève doit *témoigner d'un Savoir*. Mais la reconnaissance de ce témoignage, et donc le gain de l'Élève (et donc celui du Professeur) ne viennent pas, la plupart du temps, d'une cause indépendante du jugement du Professeur : ils sont soumis à son verdict (il faudrait plutôt dire à un très grand nombre de verdicts), que ce soit en cours ou en fin d'apprentissage, verdict constitué des jugements que le Professeur porte de manière plus ou moins implicite ou inconsciente sur l'action de l'élève" (p. 46). Le jugement des professeur·e·s s'appuie donc sur les réponses que ses élèves sont amenés à fournir lors d'épisodes évaluatifs sanctionnés par une notation. Il semble donc important de s'intéresser à aux réponses des élèves et à ce qui peut les motiver.

- ***La "course au 20" en évaluation***

Un élève cherchera toujours et avant tout à avoir "une bonne note", la meilleure possible. Cette "course au 20" particulière va l'amener à chercher à répondre, le plus exactement possible, aux attentes de son enseignant·e telles qu'il ou elle les perçoit. Ces attentes peuvent être plus ou moins explicitées dans le cadre d'un contrat établi (ou non) entre l'enseignant·e et ses élèves, mais dans tous les cas, elles sont interprétées par chaque élève. Les notions de composantes "privée" et "publique" du rapport personnel de l'élève à l'objet de savoir enseigné de Chevallard (1988) me semblent pertinentes pour témoigner du fait que l'élève fait des choix de réponses, indépendamment de ce qu'il sait ou ne sait pas, et qu'il adopte des

stratégies qui lui permettent, de son point de vue, de concourir à cette "course au 20". Coppé (1998) qui a cherché comment un élève, en situation de devoir surveillé, détermine ce qu'il ou elle rend public ou garde pour lui ou elle en fonction des attentes supposées du professeur·e, a d'ailleurs conclu que "certains élèves en difficulté sont assujettis au discours et aux actions du maître, c'est-à-dire qu'ils sont sans arrêt dans le domaine du rapport public et qu'ils ont très peu d'autonomie d'action" (p. 140). Il faut donc convenir qu'un·e l'élève oriente toujours ses réponses et qu'il ou elle ne donne à voir, au professeur·e, qu'une partie de ce qu'il ou elle sait ou ne sait pas.

- ***Les stratégies de réponses***

Dans le cadre de l'expérimentation que j'ai menée avec Grapin en 2013, une réponse d'élève m'a tout particulièrement fait réaliser la nécessité d'étudier finement les stratégies de réponse des élèves. Le contrat expérimental établi avec les élèves de la classe était qu'ils ou elles devaient forcément cocher une case du QCM pour chaque item et renseigner le niveau de certitude qu'il ou elle accordait à leur réponse à l'aide d'une échelle à quatre niveaux (pas sûr du tout/ presque sûr /sûr/ sûr et certain). Un élève dit d'emblée que la question est trop difficile, qu'il ne peut y arriver et me signifie qu'il n'a pas vraiment envie d'y répondre. Devant le rappel au contrat que je lui oppose, il finit car cocher 10 (la réponse correcte) et à indiquer être "sûr et certain" de sa réponse. Surprise par ces réponses, je l'interroge et lui demande de les justifier. Il me répond alors "c'est toujours 10" !
A travers cette anecdote et l'étude qui a suivi, j'ai pu réaliser à quel point il était nécessaire d'étudier finement les stratégies de réponses des élèves qui peuvent n'être que partiellement liées à des connaissances effectives (Sayac & Grapin, 2015), et bien souvent conditionnées par cette "course au 20".
Chevallard (1986) et Schubauer-Leoni (1991) ont mis en évidence que les élèves tentent de décrypter et de comprendre les attentes de leurs professeur·e·s pour essayer d'y répondre au plus près. Ils ou elles développent donc des stratégies pour atteindre cet objectif. Par exemple, un·e élève peut préférer chercher à se remémorer une correction d'exercice qu'il ou elle reconnaît avoir fait en classe auparavant plutôt que d'essayer de le résoudre indépendamment. L'élève est alors persuadé·e que sa réponse sera correcte puisqu'elle est conforme à celle qui a été validée antérieurement par son enseignant·e.
Ces stratégies de réponse, que l'on pourrait qualifier d'externes car non strictement liées à des savoirs mathématiques, sont propres aux épisodes évaluatifs sanctionnés par une notation, ce qui signifie que les élèves ne les développent pas forcément lors d'épisodes évaluatifs non assujettis à une note. Elles témoignent du fait qu'un·e professeur·e ne doit pas se contenter de données issues de tels épisodes pour déterminer le degré de connaissances de ses élèves, mais qu'il doit impérativement croiser les données issues de plusieurs types d'épisodes évaluatifs et réaliser une triangulation (Allal & Mottier Lopez, 2008) pour renforcer la qualité de son jugement didactique en évaluation.

- ***La vérification***

Dans la logique (non exclusive) de cette "course au 20", l'élève peut être amené·e à effectuer des vérifications. Coppé (1993) a élaboré une typologie de processus de vérification dans le cadre de devoirs surveillés au lycée qui distingue les processus de vérification internes et externes, suivant qu'ils s'appuient ou non sur des savoirs et savoir-faire mathématiques. La distinction qu'elle établit entre vérifications externes et internes permet de mieux comprendre ce qui se joue au niveau du contrat didactique en évaluation lors de tels processus.
Parmi les stratégies de vérification que les élèves utilisent, il y a celles que j'estime reliées au contrat didactique en évaluation et qui relèvent de normes scolaires intégrées/partagées comme par exemple : le résultat "tombe juste", la valeur est entière, etc. Si à l'issue d'un calcul complexe, l'élève trouve une valeur entière, il ou elle sait que cette réponse a de fortes

chances d'être LA "bonne" réponse, par contre s'il ou elle tombe sur une valeur à partie décimale illimitée, il ou elle vérifiera certainement ses calculs pour trouver d'éventuelles erreurs de calcul ou bien il ou elle reprendra plus globalement sa démarche.
Ces processus de vérification sont activés de manière contingente par les élèves d'une classe et il s'avère que souvent, les élèves en difficulté y ont peu recours (Coppé, 1993). Dans la perspective d'une évaluation au service des apprentissages des élèves, ils pourraient avoir un rôle à jouer et être davantage intégrés au processus didactique. Les contrats didactiques en évaluation pourraient alors évoluer et s'inscrire dans des orientations d'autoévaluation ou d'évaluation entre pairs qui ne seraient pas non plus sans incidence sur le jugement professionnel en évaluation des professeur·e·s.

Conclusion de la partie II

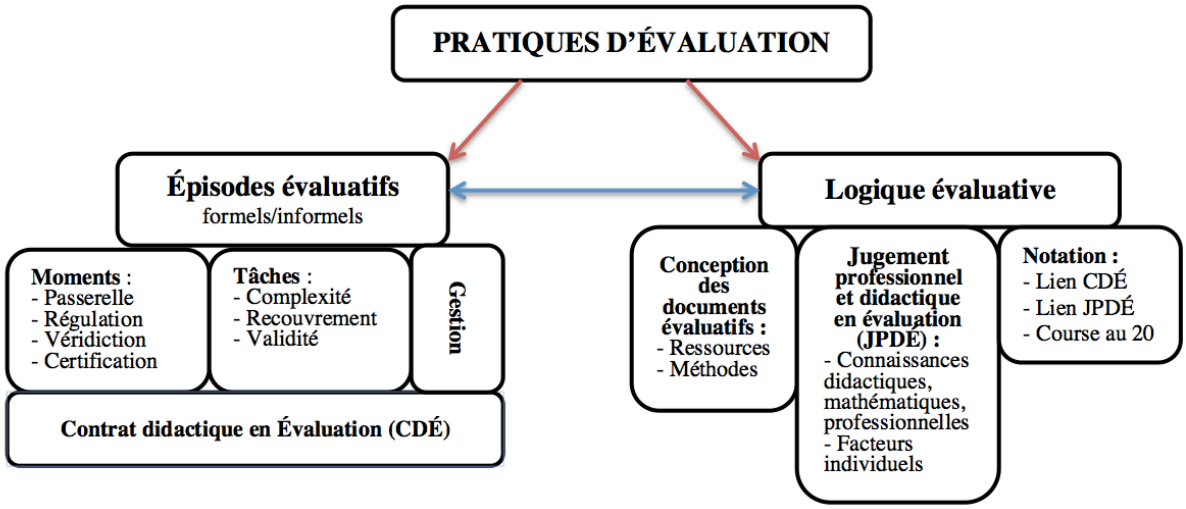
Du point de vue du cadre développé

Dans cette partie, j'ai souhaité présenter et justifier un cadre didactique pour l'évaluation en mathématiques, c'est à dire un cadre qui étudie l'évaluation comme partie prenante des processus de transmission et d'acquisition de savoirs mathématiques et qui propose de décrire et d'expliquer les phénomènes relatifs aux rapports entre l'enseignement, les apprentissages et l'évaluation de ces apprentissages (en référence à la définition de la Didactique de Douady, 1984). Pour définir ce cadre, je me suis appuyée sur différents travaux s'inscrivant dans cette perspective (principe 2), en essayant de combiner les concepts et résultats les plus pertinents pour penser et concevoir une évaluation au service des apprentissages des élèves en mathématiques (principe 1).
Ce cadre articule deux axes d'analyse des pratiques d'évaluation des professeur·e·s en mathématiques, un axe focalisé sur les épisodes évaluatifs proposés aux élèves et un axe structuré par la "logique évaluative" des professeur·e·s. Les notions d'épisode évaluatif, de contrat didactique en évaluation et de jugement professionnel et didactique en évaluation qui s'inscrivent dans la "logique évaluative" d'un·e professeur·e sont au cœur de ce cadre didactique. Elles permettent d'inscrire pleinement l'évaluation dans les processus d'enseignement et d'apprentissages tout en prenant en compte les acteurs et actrices principales que sont les professeur·e·s et les élèves.
Un épisode évaluatif est étudié à partir de ses caractéristiques propres (moments où il est proposé, nature des tâches associées, recouvrement de contenu). La validité d'un épisode évaluatif est éprouvée à partir d'éléments liés aux dimensions épistémo-didactique et curriculaire du savoir en jeu dans l'épisode et d'éléments liés à la gestion de cet épisode par l'enseignant·e.
Le contrat didactique en évaluation associé à chaque épisode évaluatif intègre des dimensions de communication, de pouvoir et de négociation didactiques. C'est un contrat spécifique qui peut s'apparenter à un *mésocontrat* didactique (Hersant & Perrin, 2003) et qui permet de considérer ce qui se joue, pour les élèves et pour l'enseignant·e, dans un épisode évaluatif donné. L'explicitation d'un tel contrat est un enjeu didactique pour penser et concevoir une évaluation "pour apprendre".
La "logique évaluative" d'un·e professeur·e est appréhendée à partir d'indicateurs retenus pour permettre à la fois de décrire cette logique personnelle et à la fois de la faire évoluer (principes 1, 2 et 3). Ces indicateurs sont : la façon dont les professeur·e·s conçoivent et élaborent les différents épisodes évaluatifs qui ponctuent l'étude d'un savoir mathématique (ressources utilisées, documents évaluatifs, méthodes), le jugement professionnel et didactique en évaluation et la notation adoptés par l'enseignant·e. Le jugement professionnel et didactique en évaluation s'appuie sur les travaux développés en Sciences de l'éducation sur

le jugement professionnel en évaluation (Mottier Lopez & Allal, 2009, 2010 ; Tessaro, 2013), mais il est spécifié aux mathématiques et intègre des notions développées en didactique des mathématiques telle que la vigilance didactique (Pézard, 2010). Il dépend des connaissances mathématiques et didactiques des professeur·e·s, mais aussi de facteurs individuels (croyances et représentations sur les apprentissages, sur l'évaluation, expériences évaluatives). Le jugement professionnel et didactique en évaluation des professeur·e·s est activé lorsque qu'ils ou elles doivent émettre un avis sur l'état des connaissances mathématiques de leurs élèves, articuler les différents moments de l'étude entre eux (notamment intégrer les épisodes évaluatifs aux autres moments de l'étude) et gérer les épisodes évaluatifs de manière à favoriser les apprentissages de leurs élèves. Ces trois entrées ont été retenues pour leur pertinence dans l'étude de ce qui se joue en termes d'apprentissages des élèves lors des différents épisodes évaluatifs, mais aussi pour les perspectives qu'elles offrent, en formation, pour faire évoluer les pratiques d'évaluation des professeur·e·s en mathématiques (principe 1).

La notation est également un élément pris en compte dans le cadre didactique de l'évaluation en mathématiques. Elle est liée au contrat didactique en évaluation par la dimension de négociation qu'elle génère et par son incidence sur les réponses des élèves (stratégies, vérifications, "course au 20"). Elle est également en lien avec le jugement professionnel et didactique en évaluation des professeur·e·s à travers leurs choix de notation, de grilles adoptées et d'attribution de notation aux élèves.

Le schéma ci-dessous récapitule comment les éléments constitutifs du cadre didactique proposé s'articulent entre eux et met en évidence les deux axes qui le structurent.



Du point de vue méthodologique

Le cadre didactique de l'évaluation que j'ai développé vise à mieux comprendre et analyser les pratiques évaluatives des professeur·e·s en mathématiques dans la réalité de leur réalisation en classe. Il permet également de repenser, d'un point de vue méthodologique, les recherches sur ce type de pratiques jusque-là étudiées, soit en Sciences de l'éducation en négligeant souvent la spécificité des contenus mathématiques, soit en Didactique à partir d'outils permettant d'analyser les contenus mathématiques prenant peu en compte les acteur·rice·s engagé·e·s dans le processus d'enseignement/apprentissage et les dimensions extra-mathématiques. Dans ce cadre, les pratiques évaluatives des professeur·e·s sont appréhendées non seulement à partir de la façon dont les apprentissages des élèves sont évalués *via* les différents épisodes évaluatifs auxquels ils ou elles sont confronté·e·s, mais aussi à partir de la "logique évaluative" qui pilote les choix et les actions de l'enseignant·e en

matière d'évaluation. La prise en compte simultanée de ces deux axes me semble aujourd'hui indispensable pour accéder à ce qui se passe réellement dans les classes au niveau de l'évaluation des apprentissages des élèves en mathématiques. Étudier les jugements évaluatifs des professeur·e·s, la notation qu'ils ou elles adoptent ou la façon dont ils·elles conçoivent leurs évaluations indépendamment de l'étude des épisodes évaluatifs proposés ne permet pas de comprendre ce qui se joue réellement du point de vue des apprentissages des élèves lors de ces différents moments d'évaluation. Inversement, étudier les épisodes évaluatifs proposés par un·e professeur·e sans prendre en compte la façon dont il ou elle les conçoit, émet des jugements ou note les productions des élèves ne permet que partiellement de comprendre la logique qui l'anime.

Lorsque j'ai analysé les pratiques évaluatives en mathématiques de vingt-cinq professeur·e·s des écoles à travers les tâches évaluatives qu'ils ou elles avaient proposées à leurs élèves dans le domaine de la numération, j'ai réalisé qu'il manquait des éléments liés à ce qui avait piloté au choix des tâches évaluatives proposées pour affiner l'analyse et comprendre comment et pourquoi les professeur·e·s les avaient retenues. La méthodologie utilisée a certes produit des résultats par rapport aux évaluations récoltées, mais elle n'a pas permis de décrire et comprendre finement les pratiques évaluatives des professeur·e·s étudié·e·s. L'étude des productions corrigées d'élèves a permis de renseigner sur la façon dont les professeur·e·s émettent des jugements et notent leurs élèves, mais ces informations n'ont pas été mises en réseau avec celles issues de l'analyse didactique des tâches évaluatives proposées lors des épisodes évaluatifs étudiés. L'entretien mené avec chaque professeur·e, centré sur la compréhension ou la justification des tâches évaluatives, a dégagé des informations insuffisantes pour comprendre ce qui était en jeu au niveau de l'évaluation proposée pour les élèves (apprentissages) et pour l'enseignant·e (enseignement). Cette recherche m'a amenée à comprendre l'intérêt d'étudier la "logique évaluative" des enseignant·e·s en lien avec les tâches qu'ils ou elles proposent en évaluation à leurs élèves. Par ailleurs, j'ai également réalisé combien l'étude d'un seul épisode évaluatif ne peut suffire pour appréhender réellement les pratiques évaluatives des professeur·e·s en mathématiques. Même quand les professeur·e·s avaient donné plusieurs documents évaluatifs, je n'ai pas toujours été en mesure de comprendre la finalité de chacun d'eux, leur articulation les uns aux autres, ni comment ils pouvaient s'inscrire dans les processus d'enseignement et apprentissage. La méthodologie adoptée pour cette recherche n'a permis que partiellement d'avoir accès à la "logique évaluative" des professeur·e·s de l'échantillon et ne s'est pas appuyée sur tous les épisodes évaluatifs proposés lors des séquences de numération.

Le cadre didactique développé dans cette note de synthèse permet, me semble-t-il, de concevoir des méthodologies de recherche sur les pratiques évaluatives des professeur·e·s en mathématiques plus efficaces car structurées par les deux axes qui constituent ce cadre. L'étude de ces pratiques nécessite la récolte de données multiples et variées qui sont indispensables pour les appréhender et comprendre les incidences qu'elles peuvent avoir sur les apprentissages mathématiques des élèves. Cette contrainte de récolte de données permettant à la fois de reconstituer la "logique évaluative" des professeur·e·s et d'étudier l'ensemble des épisodes évaluatifs proposés amène à repenser les méthodologies de recherche ayant pour objectif d'étudier des pratiques d'évaluation en mathématiques. Certaines de ces données peuvent s'avérer difficiles à récolter car, étudier les pratiques évaluatives en recherchant des éléments constitutifs de la "logique évaluative" des professeur·e·s, peut s'avérer intrusif pour ces dernier·e·s. En effet, la liberté pédagogique accordée aux enseignant·e·s français·e·s et les pratiques très individuelles constatées en matière d'évaluation confèrent à ces pratiques une dimension qui relève de "l'intime professionnel" des professeur·e·s. Je me suis d'ailleurs plusieurs fois heurtée à la difficulté de récolter certaines données et j'ai ainsi pu constater combien il n'était pas aisé, pour un·e professeur·e

de fournir les évaluations qu'il ou elle donnait à ses élèves ou de répondre à des questions relatives à ses pratiques d'évaluation en mathématiques. Dans le premier degré, cela peut même s'avérer encore plus difficile dans la mesure où les professeur·e·s des écoles sont, pour la plupart, issu·e·s de cursus non scientifiques qui les rendent souvent peu à l'aise avec les mathématiques. Quand, en tant que chercheur·e, on s'intéresse à l'évaluation des apprentissages des élèves en mathématiques et que pour cela, on est amené·e à étudier les pratiques de leur professeur·e, l'idée que ces pratiques vont elles-mêmes être évaluées peut très naturellement germer dans la tête des professeur·e·s et générer des réserves, voire des inquiétudes.

Ainsi, il me semble-t-il indispensable de bâtir, avec les professeur·e·s dont on souhaite étudier les pratiques d'évaluation, des collaborations de recherche qui permettent de les amener à privilégier une posture de compréhension de phénomènes et de questionnement du monde (Delarue-Breton, 2016) susceptible de réduire leurs craintes de jugements portés sur leurs pratiques. Ces collaborations sont également nécessaires pour garantir l'accès et la qualité des données nécessaires à l'étude de ces pratiques complexes. La double vraisemblance prônée par Dubet (2007) s'inscrit dans cette perspective de travail collaboratif entre chercheur·e·s et praticien·ne·s, puisque :

> Si l'on part de l'hypothèse que les acteurs sont des acteurs et qu'ils possèdent donc des capacités d'action et de réflexion, ce sont elles que le chercheur doit mobiliser plutôt que de se conférer un monopole du sens qui ne risque guère de lui être contesté par ceux qu'il étudie car il est rare qu'ils lisent ses ouvrages ou ses articles. » (p. 24).

Les recherches collaboratives (Desgagné, 1997) sont particulièrement adaptées pour étudier les pratiques d'évaluation des professeur·e·s en mathématiques. Elles ont d'ailleurs été expérimentées par de nombreux·ses chercheur·e·s en didactique des mathématiques (Bednarz, Roditi, Saboya, Barry, Sayac). Dans l'introduction du livre qu'elle a dirigé sur la recherche collaborative, Bednarz (2013) a justement précisé que "cette approche de recherche accorde une place importante aux actions et significations des praticiens (enseignants, conseillers pédagogiques, intervenants divers du monde de l'éducation) et elle produit des données qui fournissent de nouveaux éclairages sur les questions qui se posent aujourd'hui à la profession enseignante" (p.7). Darré (1999) a, de son côté, explicitement indiqué que ce type de recherche permettait d'avoir accès aux "logiques d'actions" des enseignant·e·s qui coïncident avec la logique évaluative des enseignant·e·s que je souhaite appréhender pour analyser et comprendre leur pratique d'évaluation en mathématiques.

Le cadre didactique de l'évaluation que je propose accorde une place importante à la composante personnelle des pratiques (principe 3), c'est pourquoi il impose de créer des conditions spécifiques de récolte de données et d'adopter des méthodologies adaptées. Les recherches collaboratives (Desgagné, 1997 ; Bednarz & Desgagné, 2001, 2005) sont des dispositifs de recherche particulièrement adaptés à ce cadre.

Du point de vue de la formation

Conformément au premier principe qui anime mes travaux, le cadre que je propose permet de concevoir une approche singulière de la formation à l'évaluation en mathématiques des enseignant·e·s. Parce qu'il articule deux axes distincts et complémentaires, ce cadre impose de penser des formations spécifiques à l'évaluation en mathématiques, adaptées à chacun de ces axes. Parce que ce cadre accorde une place importante à la composante personnelle des pratiques évaluatives, les formations qui en découlent doivent s'inscrire dans la ZPDP "Zone proximale de développement des pratiques" (Robert & Vivier, 2013) des professeur·e·s. En effet, comme le précisent Rogalski et Robert (2015) :

> Pour qu'un travail en formation enrichisse les pratiques et pas seulement des connaissances sur les exercices ou les déroulements par exemple, il est important non seulement de faire travailler les

> pratiques, mais encore de s'appuyer sur des éléments de ces pratiques dont les participants ont conscience, soit qu'ils partagent, soit sur lesquels ils ressentent des besoins. (p. 104)

En ce qui concerne la formation à l'évaluation en mathématiques, les expériences que j'ai pu avoir à l'IUFM ou à l'ÉSPÉ de Créteil me font douter que les professeur·e·s ressentent véritablement des besoins dans ce domaine ; c'est pourquoi je préconise plutôt de proposer des formations collectives, autour de problématiques d'évaluation partagées. Je rejoins ainsi Mottier Lopez (2012, 2014) qui a réalisé de nombreuses formations à l'évaluation dans le cadre de la modération sociale (Maxwell, 2002). Pour elle, le dispositif de modération sociale est idéal pour offrir des conditions d'échanges et de construction de repères communs autour de questions sensibles et permet de soutenir le développement de compétences professionnelles (2014). Il permet également d'inciter les enseignant·e·s à confronter leurs choix évaluatifs dans des évaluations concrètes (Mottier Lopez, 2014, p. 94). Pour faire évoluer les pratiques d'évaluation en mathématiques, il semble donc indispensable de permettre aux enseignant·e·s de confronter leur pratique d'évaluation et la logique qui les sous-tend. Dans le cadre que je propose cette logique s'appréhende à travers les modes de conception des documents évaluatifs, les jugements émis sur des productions d'élèves et la notation adoptée par les professeur·e·s qui sont autant de points d'entrée pour favoriser cette confrontation. Dans la recherche en cours menée dans le cadre du LéA EvalNumC2, plusieurs collectifs d'enseignant·e·s ont été constitués pour s'inscrire dans cette perspective de formation. Les enseignant·e·s engagé·e·s dans ces collectifs sont amené·e·s à élaborer des évaluations communes (tests ou évaluations plus formelles), à discuter des productions des élèves confronté·e·s à ces évaluations (les leurs et celles des élèves des autres classes de même niveau) et à échanger sur la notation à adopter pour rendre compte de ces diverses productions. Ces moments de discussion et d'échanges permettent donc à la fois de travailler sur des documents évaluatifs (choix de tâches évaluatives, validité, gestion), mais également de faire émerger les croyances et les représentations des professeur·e·s sur l'évaluation, l'enseignement et les apprentissages en mathématiques. Des échanges, parfois vifs, entre les professeur·e·s les amènent à justifier ou contester des choix, ce qui contribue de mon point de vue à faire évoluer la "logique évaluative" de chacun·e. Ce qui importe avant tout, c'est de créer les conditions permettant de récolter les données nécessaires à l'analyse des pratiques d'évaluation des professeur·e·s sans avoir recours à des méthodologies classiques souvent déficientes pour récolter de telles données. Dans son dispositif PACEM, Chesné (2014) s'est également appuyé sur les résultats d'élèves à des évaluations standardisées en mathématiques pour faire évoluer les pratiques des professeur·e·s ayant participé à cette formation innovante. Lui aussi avait fait le pari de pouvoir décrypter les représentations des enseignant·e·s à partir de leurs propres discours provoqués par les résultats des élèves à un test en mathématiques.

Par ailleurs, les études que j'ai menées autour des pratiques de formation (Sayac, 2013) et de l'impact de la recherche dans les formations d'enseignant·e·s (Sayac, 2012, 2013) m'amènent à penser que la place et le rôle des formé·e·s doivent être repensés vers une implication personnelle plus grande pour espérer avoir un impact sur leurs pratiques. J'ai également défendu le fait que l'initiation à la recherche dans la formation des enseignant·e·s contribue à "contrecarrer la dogmatisation des savoirs" (Crinon, 2012), en développant une approche scientifique et critique des phénomènes de prescriptions et de doxas qui traversent le monde de l'école sans que ni l'institution ni la recherche ne les aient étayés" (Sayac, 2013, p. 4). Le poids des croyances et les doxas qui circulent dans le monde scolaire autour de l'évaluation imposent donc une approche spécifique de la formation dans ce domaine.

Les recherches collaboratives que je préconise pour étudier les pratiques évaluatives des professeur·e·s au plus près de leur réalisation en classe s'inscrivent dans cette perspective. Dans ce cadre, les enseignant·e·s sont engagé·e·s dans des recherches, avec des chercheur·e·s autour de problématiques partagées, étudiées à partir de méthodologies co-construites visant

une co-production de savoirs scientifiques et professionnels (Desgagné, 1997, Bednarz, 2009, 2013). Ils ou elles sont amené·e·s à apporter leur expertise professionnelle et à mobiliser une compréhension des phénomènes étudiés qui ne peut être sans incidence sur leur pratique et donc sans conséquence du point de vue de leur développement professionnel. Bednarz (2013) évoque à propos des recherches collaboratives, l'opportunité d'installer une "zone interprétative partagée" qui participe à la fois à une entreprise de recherche et à la fois à une entreprise de développement professionnel (p. 28). Cette "zone interprétative partagée" est à mettre en parallèle avec la "zone proximale de développement des pratiques" de Robert et Vivier (2013). Cette notion de "zone" me paraît pertinente pour concevoir des formations à l'évaluation en mathématiques susceptibles de faire évoluer les pratiques évaluatives des enseignant·e·s qui sont, comme je l'ai précisé, des pratiques spécifiques plus fortement portées par des dimensions personnelles et institutionnelles. Roditi (2013) a aussi fait valoir que l'objectif et les méthodes des recherches collaboratives visent à la compréhension, voire la transformation des pratiques. Dans sa contribution à l'ouvrage de Bednarz, il cite Bru (2002 p. 65) qui affirme que "s'intéresser aux pratiques enseignantes en choisissant de les appréhender dans des conditions de forte proximité participative de terrain et sur la durée est à n'en pas douter une façon à la fois de les connaître et de contribuer à leur transformation." (p. 353).

Former les professeur·e·s à évaluer les apprentissages de leurs élèves en mathématiques doit donc se concevoir dans le cadre d'un travail collectif (modération sociale, recherche collaborative, etc.) qui permet l'émergence et la confrontation de pratiques d'évaluation. Il est cependant également indispensable de fournir aux enseignant·e·s en formation des outils permettant de développer leurs compétences professionnelles en évaluation et leur vigilance didactique appliquée aux questions d'évaluation. Ces outils doivent être conçus pour les aider à concevoir des épisodes évaluatifs valides, s'intégrant pleinement dans le cours de l'étude et gérés de manière à favoriser les apprentissages des élèves. Ils doivent également les aider à analyser les productions des élèves pour leur adresser un retour constructif. En s'appuyant sur le cadre didactique de l'évaluation que je propose, ces aides pourraient être apportées au niveau des méthodes de conception, des ressources utilisées pour choisir les tâches constituant leurs documents évaluatifs, de la gestion des épisodes évaluatifs ou/et au niveau de la triangulation des informations récoltées lors des différents épisodes évaluatifs.

Dans le cadre du LéA EvalNumC2, les outils proposés aux enseignant·e·s engagé·e·s dans les deux collectifs sont : les facteurs de complexité et de compétences (Sayac & Grapin, 2015), une liste de tâches dans le domaine du nombre au cycle 2 (issue de la thèse de Nadine Grapin, adaptée au cycle 2) et des tests externes élaborés par des chercheures dans le cadre de l'axe 1 du LéA. Je fais le pari, avec Nadine Grapin, que la confrontation à des documents évaluatifs externes permettra d'élargir "l'horizon évaluatif" des professeur·e·s en les confrontant à des tâches différentes de celles qu'ils ou elles ont l'habitude de proposer à leurs élèves et en les exposant à d'autres logiques d'organisation de tests.

Ainsi, dans le cadre que je propose, la formation des professeur·e·s à l'évaluation en mathématiques doit être pensée de manière à s'appuyer aussi bien sur l'axe porté par les épisodes évaluatifs que par celui relatif à la "logique évaluative" des enseignant·e·s. De même que l'analyse des pratiques évaluatives en mathématiques, la formation des enseignant·e·s ne peut se réaliser dans ce cadre sans prendre en compte conjointement les deux axes structurant ces pratiques, ce qui suppose de concevoir des dispositifs spécifiques adaptés et nouveaux.

CONCLUSION DE LA NOTE DE SYNTHÈSE

Il y a quelques années, lorsque j'ai commencé à m'intéresser à l'évaluation, je ne pensais pas que ce thème deviendrait ma préoccupation centrale de recherche. Au moment de conclure ma note de synthèse et de considérer le chemin parcouru depuis mes premiers travaux sur la question, je suis en mesure maintenant de mieux comprendre pourquoi cette thématique a pris tant d'importance dans mon parcours scientifique et comment j'ai pu contribuer à la faire avancer en tant que didacticienne.
Le travail présenté dans cette note de synthèse est double puisqu'il comporte une partie recension des travaux anglophones et francophones sur l'évaluation et une partie où je définis un cadre didactique de l'évaluation. La partie recension m'est apparue indispensable pour rendre compte de la grande diversité et de la richesse des travaux existants dans ce domaine et de faire valoir la spécificité de ma contribution. La seconde partie présente un nouveau cadre de l'évaluation qui s'inscrit dans une histoire scientifique ayant négligé ou peu exploré la dimension didactique de l'évaluation et de ses pratiques en mathématiques.

Concernant la première partie
Deux règles ont présidé au travail de recension réalisé dans la première partie : (1) explorer les différents champs scientifiques sans *a priori*, mais (2) ne retenir que ceux pouvant s'inscrire dans une approche didactique, susceptibles d'ouvrir des perspectives en termes de réussite des élèves et de formation des enseignant·e·s (principe 1).
Cette recension a permis de rendre compte des nombreux travaux élaborés autour de l'évaluation par des chercheur·e·s appartenant à différents champs scientifiques et travaillant dans différents pays. Elle a témoigné du fait que les chercheur·e·s anglophones ont été les premiers à développer une approche de l'évaluation se préoccupant des apprentissages des élèves avec les courants de *Formative Assessment* ou d'*Assessment for Learning*, mais que les chercheur·e·s francophones s'en sont rapidement emparé pour développer des approches spécifiques, s'inscrivant dans d'autres contextes et dans des cultures de l'évaluation différentes.
Cette partie a également permis de situer les travaux des chercheur·e·s français·e·s et de montrer la dynamique qui s'est récemment développée autour de l'évaluation dans le champ de la Didactique des mathématiques.
La présentation de mes travaux a été l'occasion de rendre compte de mon cheminement scientifique autour de l'évaluation en mathématiques. Partant d'un engagement fortuit dans un groupe d'expert·e·s de la DEPP chargé·e·s de concevoir les items d'un bilan national en mathématiques, j'ai peu à peu compris l'intérêt, pour les didacticien·ne·s, de s'emparer des questions d'évaluation des apprentissages des élèves et des pratiques évaluatives des enseignant·e·s en mathématiques. Les études que j'ai réalisées depuis 2011 et l'ouverture à d'autres champs scientifiques que la didactique m'ont amenées aujourd'hui à proposer un nouveau cadre de l'évaluation que j'ai développé dans la seconde partie de cette note de synthèse.

Concernant la seconde partie
Je propose de mettre en exergue, dans cette conclusion, ma contribution au champ de l'évaluation, les perspectives ouvertes par mon travail et ce qu'il reste à investiguer à partir des deux notions clés du cadre didactique de l'évaluation présenté : les épisodes évaluatifs et la "logique évaluative" des professeur·e·s.

- **Les épisodes évaluatifs**

La notion d'épisodes évaluatifs est au cœur du premier axe qui structure le cadre didactique de l'évaluation que je propose. Elle permet de dépasser l'entrée par les fonctions de l'évaluation jusque-là adoptée dans les différents travaux scientifiques sur l'évaluation et de porter un regard plus didactique sur les "faits évaluatifs" (Chevallard, 1986). La définition des épisodes évaluatifs intègre des travaux développés principalement en didactique des mathématiques et permet de prendre davantage en compte les contenus disciplinaires évalués ainsi que le temps didactique. Les épisodes évaluatifs s'intègrent aux processus d'enseignement pouvant advenir de manière formelle ou informelle dans les classes et englobent les différents types d'évaluation étudiés jusque-là par les chercheur·e·s en didactique des mathématiques ou en Sciences de l'éducation, usuellement qualifiés de formatifs, sommatifs ou diagnostiques. La question de leurs caractéristiques (moment où ils sont proposés, nature des tâches proposées), de leur validité (liée au recouvrement, à la pertinence et à la complexité des tâches qui les constituent), de leur gestion et du contrat didactique qui leur est associé contribuent à penser et concevoir l'évaluation dans une approche didactique.

Les perspectives ouvertes par la notion d'épisodes évaluatifs sont importantes dans la mesure où le cadre didactique proposé oblige à ne pas restreindre l'étude de l'évaluation à celles de certains "faits évaluatifs" caractérisés par des fonctions spécifiques qui ne permettent pas toujours de rendre compte de la façon les professeur·e·s évaluent les apprentissages de leurs élèves tout au long du processus d'enseignement. Prendre en compte l'ensemble des épisodes évaluatifs pouvant advenir durant un temps donné permet d'avoir une vision globale de l'activité d'évaluation des professeur·e·s en mathématiques. Pour cela, de nouvelles méthodologies de recherche adaptées à la nécessité de prendre en compte l'ensemble de ces épisodes doivent être conçues et expérimentées. L'étude de ces différents épisodes évaluatifs pourra ainsi plus justement rendre compte de la dynamique existant entre enseignement, apprentissages et évaluation dans la globalité de l'étude d'un thème mathématique. La notion d'épisodes évaluatifs est également riche en perspectives de formation. L'identification et la variété des épisodes évaluatifs qu'un·e professeur·e peut proposer à des élèves tout au long d'une séquence d'enseignement peuvent être étudiées en formation et ainsi permettre d'élargir l'horizon évaluatif des enseignant·e·s. Les épisodes évaluatifs qui sont au cœur des pratiques d'évaluation des professeur·e·s sont généralement le fruit d'un travail individuel, conçu à partir de méthodes et de ressources personnelles. Les étudier en formation, à partir du cadre didactique que je propose (caractéristiques, validité, gestion), ouvre donc des perspectives nouvelles pour faire évoluer les pratiques évaluatives des professeur·e·s en mathématiques et ainsi améliorer les apprentissages de leurs élèves.

La question de savoir quelle durée doit être prise en compte pour étudier les différents épisodes évaluatifs relatifs à un thème mathématique donné est une question qui reste à travailler. Doit-on se restreindre à une séance ou une séquence qui borne généralement le temps scolaire pour étudier un thème mathématique donné ou doit-on plus largement étudier les épisodes évaluatifs proposés autour de ce thème sur une période plus large (trimestre, semestre), voire sur une ou plusieurs années scolaires ? Le jeu d'articulation entre connaissances anciennes et nouvelles qui entretient l'aspect spiralaire des apprentissages plaident pour une étude s'inscrivant dans un temps long, mais la segmentation[11] des évaluations telles que pratiquées par les professeur·e·s des écoles tend plutôt à circonscrire cette étude à un temps plus court. Au-delà des questions de fiabilité et de validité des

[11] Aucune étude, à ma connaissance, n'a traité de cette question, mais les études que j'ai pu mener jusque là m'amènent à penser que les professeur·e·s évaluent les apprentissages de leurs élèves sur une période donnée (séquence généralement), sans y revenir ultérieurement.

évaluations habituellement retenues pour étudier la qualité des évaluations, la question du temps défini pour l'étude des épisodes évaluatifs mérite d'être investiguée et pourrait s'avérer cruciale pour comprendre les cheminements cognitifs des élèves et la façon dont les apprentissages évoluent dans le temps scolaire.

- **La "logique évaluative"**

La "logique évaluative" des professeur·e·s qui structure le deuxième axe du cadre didactique de l'évaluation que je propose est révélée à partir d'indicateurs émanant de différents travaux en didactique des mathématiques et en Sciences de l'éducation présentés dans la première partie de la note de synthèse (principe 2). Les questions de conception des évaluations (ressources et méthodes utilisées, documents évaluatifs produits), de jugement professionnel et didactique en évaluation et de notation permettent d'appréhender cette "logique évaluative" qui pilote les épisodes évaluatifs proposés par les professeur·e·s à leurs élèves. L'étude de ces trois indicateurs permet d'avoir accès à la composante personnelle des pratiques évaluatives des enseignant·e·s (principe 3).

Les perspectives ouvertes par l'étude de la "logique évaluative" des professeur·e·s sont nombreuses et se trouvent parmi les trois indicateurs qui permettent de l'appréhender.

Concernant la conception des évaluations :

Etudier la façon dont les professeur·e·s conçoivent leurs évaluations et élaborent leurs documents évaluatifs est une perspective qui mérite d'être explorée dans un cadre didactique. Le travail documentaire des professeur·e·s a fait l'objet de nombreux travaux (Gueudet & Trouche, 2008, 2010 ; Wozniak & Margolinas, 2009, 2010), mais ils ne se sont pas spécifiquement posés sur l'évaluation. Étudier la genèse documentaire de ces documents spécifiques permettra d'explorer un des déterminants de la "logique évaluative" des enseignant·e·s, de mieux la comprendre et envisager les moyens de la faire évoluer.

Une autre perspective liée à cet indicateur concerne le lien entre évaluations externes et évaluations internes (en classe). D'un côté, les évaluations externes, de plus en plus présentes dans l'institution scolaire, sont conçues par des spécialistes ou expert·e·s qui ne sont pas toujours au fait des évaluations pratiquées dans les classes. D'un autre côté, les évaluations pratiquées en classe sont parfois très éloignées de celles proposées dans le cadre d'évaluations externes, ce qui engendre des écarts qui ne traduisent pas toujours la réalité des apprentissages mathématiques des élèves[12]. Faire travailler les enseignant·e·s français·e·s sur ce type d'items permettrait d'enrichir la palette des tâches évaluatives possibles, aussi bien du point de vue de leur variété que de leur complexité ou de leur forme et participerait ainsi à élargir leur l'horizon évaluatif tout en familiarisant leurs élèves à ce type d'évaluation.

Concernant le jugement professionnel et didactique en évaluation

La plupart des recherches que j'ai menées sur les pratiques évaluatives des professeur·e·s des écoles en mathématiques se sont situées dans des établissements en zones d'éducation prioritaires. J'ai pu réaliser, à cette occasion, combien les choix et les jugements évaluatifs des enseignant·e·s étaient pilotés par des considérations personnelles liées à ce contexte d'enseignement spécifique. Les différents rapports nationaux (CNESCO) ou internationaux (PISA, TIMSS) publiés récemment font état, au-delà des résultats des élèves français·e·s en mathématiques, d'une particularité française préoccupante : le lien entre inégalités sociales et inégalités scolaires. Des chercheur·e·s anglo-saxon·ne·s (par exemple Morgan & Watson, 2002) se sont emparé·e·s de cette question pour la traiter en termes de *Assessment & Equity*,

[12] Les résultats de la dernière enquête TIMSS 2015 ont placé les élèves français·e·s au dernier rang des pays ayant participé à cette évaluation internationale. Même si ce constat traduit un problème qui doit impérativement être traité, il me semble que la nature et la forme de certains items, pas forcément usuelles dans l'enseignement en France, ont pu accentuer les mauvais résultats des élèves français·e·s.

mais en France peu de travaux l'ont réellement fait. Cette perspective me semble donc indispensable à investir aujourd'hui.

Concernant la notation

Chevallard et Feldmann (1986) avaient étudié la variation des notes qu'un professeur de mathématiques avait attribuées au cours d'un trimestre et montré le rôle majeur de la notation dans la négociation didactique. Depuis, peu de chercheur·e·s ont étudié ce rôle et considéré en quoi et comment la notation peut impacter les apprentissages des élèves. À l'heure où les prescriptions institutionnelles préconisent la bienveillance en matière de notation et où des expérimentations de classes sans note sont menées à différents niveaux d'enseignement, il semble indispensable d'étudier ces phénomènes avec une approche didactique. Le cadre didactique de l'évaluation proposé permet de ne pas traiter la question de la notation de manière isolée, mais de la rattacher directement à la "logique évaluative" des professeur·e·s et indirectement aux épisodes évaluatifs et donc à des contenus mathématiques, ce qui caractérise une approche didactique.

Plusieurs questions restent à investiguer en ce qui concerne les pratiques évaluatives des professeur·e·s en mathématiques, notamment le lien qui existe entre la "logique évaluative" des professeur·e·s et les épisodes évaluatifs qu'ils ou elles proposent à leurs élèves. Mieux comprendre comment la "logique évaluative" des professeur·e·s pilote les épisodes évaluatifs et comment la gestion des épisodes évaluatifs et les retours des élèves peuvent faire évoluer la "logique évaluative" des professeur·e·s est un enjeu fort pour faire évoluer les pratiques évaluatives en mathématiques.

Les deux axes qui structurent le cadre didactique de l'évaluation que je propose, parce qu'ils prennent en compte à la fois des contenus disciplinaires et à la fois des réalités professionnelles, permettent de définir une nouvelle approche scientifique de l'évaluation, plus didactique et volontairement ancrée dans la réalité des pratiques en classe. Le cadre présenté dans cette note de synthèse vise ainsi à contribuer à l'émergence de la "*measurement theory of classroom assessment*" chère à Brookhart (2003) et à revendiquer la "cause didactique" de l'évaluation que je défends depuis quelques années. Il impose de repenser les méthodologies de recherche et de formation centrées sur l'évaluation des apprentissages des élèves en mathématiques et les pratiques évaluatives de leurs professeur·e·s. Parce que la dimension personnelle de ces pratiques épisodes évaluatifs tient une plus grande part et que l'évaluation relève de ce que j'ai appelé "l'intime professionnel" des enseignant·e·s il convient, si on veut les étudier ou les faire évoluer, de développer des recherches ou des formations engageant chercheur·e·s et praticien·ne·s dans des projets communs et partagés. De telles recherches ou formations avec une dimension collaborative affirmée sont actuellement en développement dans la communauté scientifique des didacticien·ne·s des mathématiques. Elles contribueront très certainement, en appui au cadre didactique de l'évaluation proposé dans cette note de synthèse, à enrichir le champ scientifique de l'évaluation et de ses pratiques ainsi que les pratiques évaluatives des enseignant·e·s en mathématiques. Je précise pour conclure qu'au-delà des résultats escomptés par ce nouveau cadre, je considère qu'enrichir et développer les pratiques évaluatives des professeur·e·s est un levier efficace pour enrichir et développer plus globalement leur pratique d'enseignement. L'évaluation est un cheval de Troie pour améliorer les pratiques enseignantes et par là même, la réussite de tous les élèves en mathématiques.

RÉFÉRENCES BIBLIOGRAPHIQUES

Allal, L. K., Cardinet, J., & Perrenoud, P. (1979). *L'Évaluation formative dans un enseignement différencié : actes du colloque à l'Université de Genève, mars 1978*. P. Lang.

Allal, L. (1988). Vers un élargissement de la pédagogie de maîtrise : processus de régulation interactive, rétroactive et proactive. *Assurer la réussite des apprentissages scolaires*, 86-126.

Allal, L. & Michel, Y. (1993). Autoévaluation et évaluation mutuelle en situation de production écrite. In L. Allal, D. Bain & P. Perrenoud (Eds.), Évaluation formative et didactique du français, 239-264. Neuchâtel : Delachaux et Niestlé.

Allal, L., & Lopez, L. M. (2005). L'évaluation formative de l'apprentissage : revue de publications en langue française. *L'évaluation formative, Pour un meilleur apprentissage dans les classes secondaires*, 265-299. Paris : OCDE

Allal, L., & Mottier Lopez, L. (2008). Mieux comprendre le jugement professionnel en évaluation. Apports et implications de l'étude genevoise.

Allal, L., & Mottier Lopez, L. (2009). Au cœur du jugement professionnel : en évaluation : des démarches de triangulation. *Les dossiers des sciences de l'éducation*, (22), 25-40.

Amigues, R., Zerbato-Poudou, M. T., & Armogathe, D. (1996). *Les pratiques scolaires d'apprentissage et d'évaluation*. Dunod.

Andreucci, C., & Roux, J. P. (1989). Présentation pratique et numérique de problèmes de volume : une hypothèse socio-cognitive relative aux différents modes de résolution utilisés. *JM Monteil & M. Fayol, La psychologie scientifique et ses applications*, 275-287.

Antibi, A. (2003). *La constante macabre ou comment a-t-on découragé des générations d'élèves ?* Toulouse : Ed. Math'adore.

Antibi, A. (2007). *Pour en finir avec la constante macabre ou L'évaluation par contrat de confiance*. Toulouse : Ed. Math'Adore.

Artigue, M., & Gueudet, G. (2008). Ressources en ligne et enseignement des mathématiques. In *Université d'été de Saint-Flour "Quelle place pour l'enseignement des mathématiques ?"*.

Artigue, M., & Winsløw, C. (2010). International comparative studies on mathematics education: A viewpoint from the anthropological theory of didactics. *Recherches en didactiques des mathématiques*, *30*(1), 47-82.

Arzarello, F., Robutti, O., Sabena, C., Cusi, A., Garuti, R., Malara, N., & Martignone, F. (2014). Meta-didactical transposition: A theoretical model for teacher education programmes. In *The mathematics teacher in the digital Era* (pp. 347-372). Springer Netherlands.

Bain, D. (1988). L'évaluation formative fait fausse route, in INRAP, *Évaluer l'évaluation*, Dijon, INRAP, pp. 167-172.

Bain, D., & Schneuwly, B. (1993). Mécanismes de regulation des activités textuelles : strategies d'intervention dans les sequences didactiques. In L. Allal, D. Bain & P. Perrenoud. *Evaluation formative et didactique du français* (pp. 219-238). Neuchachâtel : Delachaux et Niestlé

Balacheff, N., & Laborde, C. (1985). Langage symbolique et preuves dans l'enseignement mathématique : une approche sociocognitive. *Psychologie sociale du développement cognitif*, 203-224.

Ball, D. L., Hill, H. C., & Bass, H. (2005). Knowing mathematics for teaching: Who knows mathematics well enough to teach third grade, and how can we decide?

Baxter, G. P., Shavelson, R. J., Herman, S. J., Brown, K. A., & Valadez, J. R. (1993). Mathematics performance assessment: Technical quality and diverse student impact. *Journal for Research in Mathematics Education*, 190-216.

Bednarz, N. (2009). Recherches collaboratives en enseignement des mathématiques : Une nouvelle entrée sur la conception d'activités en mathématiques à l'intersection de pratique en classe et recherche. *Actes du 61ème colloque de la CIEAEM (Commission Internationale pour l'Étude et l'Amélioration de l'Enseignement des Mathématiques), publiés dans Quaderni di Ricerca in Didattica Matematica,* (2), 3-18.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5-25.

Beswick, K. (2006). Changes in preservice teachers' attitudes and beliefs: The net impact of two mathematics education units and intervening experiences. *School Science and Mathematics*, *106*(1), 36-47.

Binet, A., & Henri, V. (1898). *La fatigue intellectuelle* (Vol. 1). Librairie C. Reinwald.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, *5*(1), 7-74.

Black, P. (1998). *Testing: Friend or foe? Theory and practice of assessment and testing*. London: Falmer Press

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, *5*(1), 7-74.

Black, M. C., & Williams, P. L. (2001). Preliminary assessment of metal toxicity in the middle Tisza River (Hungary) flood plain. *Journal of Soils and Sediments*, *1*(4), 213-216.

Bleiler, S. K., & Thompson, D. R. (2013). Multidimensional assessment of CCSSM. *Teaching Children's Mathematics*, *19*(5), 292-300.

Bloch, I. (2006). Peut-on analyser la pertinence des réactions mathématiques des professeurs dans leurs classe ? Comment travailler cette pertinence en formation, dans des situations adidactiques ? *Actes du Séminaire National de Didactique des Mathématiques*, Paris : Université Paris 7.

Bloch, I. (2009). Les interactions mathématiques entre professeurs et élèves. *Comment travailler leur pertinence en formation ?,* 25-52.

Bloom, B. S., Hastings, J. T. & Madaus, G. F. (1971). Handbook on Formative and Summative Evaluation of Student Learning. New York: McGraw-Hill Book. Co.

Bloom, B. S., & De Landsheere, V. (1979). *Caractéristiques individuelles et apprentissages scolaires*. F. Nathan ; Labor.

Bodin, A. (2006). Ce qui est vraiment évalué par PISA en mathématiques. Ce qui ne l'est pas. Un point de vue français. *Bulletin de l'APMEP*, *463*, 240-265.

Bodin, A. (2007). Dissonances et convergences évaluatives. De l'évaluation dans la classe aux évaluations internationales : quelle cohérence ? *Bulletin de l'APMEP*, 474, 47-79

Bonniol, J. J. (1986). Recherches et formations : pour une problématique de l'évaluation formative. *L'évaluation : approche descriptive ou prescriptive*, 119-133.

Bolon, J. (2002). Pédagogie différenciée en mathématiques : mission impossible ou défi. *Grand N*, *69*, 63-82.

Bonner, S. M. (2013). Validity in classroom assessment: Purposes, properties, and principles. *Sage handbook of research on classroom assessment*, 87-106.

Bonniol, J. J. (1981). *Déterminants et mécanismes des comportements d'évaluation d'épreuves scolaires* (Doctoral dissertation).

Bonniol, J. J. (1986). Recherches et Formations : Pour une problématique de l'évaluation formative. In J.M. De Ketele (Ed.), L'évaluation : approche descriptive ou prescriptive ? Bruxelles : De Boeck, 119-133.

Bosch, M., & Gascón, J. (2005). La praxéologie comme unité d'analyse des processus didactiques. *Balises pour la didactique des mathématiques. La Pensée Sauvage, Grenoble,* 107-122.

Bressoux, P. (2000). Modélisation et évaluation des environnements et des pratiques d'enseignement. *Habilitation à diriger les recherches. Université Pierre Mendès France. Grenoble.*

Bressoux, P., & Pansu, P. (2003). *Quand les enseignants jugent leurs élèves*. Paris : Presses universitaires de France.

Bressoux, P. (2006). Effet-classe, effet-maître. *Apprendre et faire apprendre*, 213-226.

Broadfoot, P.M., Daugherty, R., Gardner, J., Hareln, W., James, M., & Stobart, G. (2002). Assessment for Learning: 10 principes. Cambridges, UK: University of Cambridge School of Education.

Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied Measurement in Education*, *10*(2), 161-180.

Brookhart, S. M. (2001). The "Standards" and Classroom Assessment Research. *Educational measurement: Issues and Practice*, *18*(1), 23-27.

Brookhart, S. M. (2003). Developing measurement theory for classroom assessment purposes and uses. *Educational measurement: Issues and practice*, *22*(4), 5-12.

Brookhart, S. M., & Durkin, D. T. (2003). Classroom assessment, student motivation, and achievement in high school social studies classes. *Applied Measurement in Education*, *16*(1), 27-54.

Brookhart, S. M. (2004). Classroom assessment: Tensions and intersections in theory and practice. *Teachers college record*, *106*(3), 429-458.

Brookhart, S. M., Walsh, J. M., & Zientarski, W. A. (2006). The dynamics of motivation and effort for classroom assessments in middle school science and social studies. *Applied Measurement in Education*, *19*(2), 151-184.

Brousseau, G. (1980). L'échec et le contrat. *Recherches*, *41*, 177-182.

Brousseau, G. (1983). *Les « effets » du « contrat didactique »*. Actes de la 2ème école d'été de didactique des mathématiques (Olivet).

Brousseau, G. (1988). Les différents rôles du maître. *Bulletin de l'AMQ. Montréal.*, (23), 14-24.

Brousseau G., (1998), *Théorie des Situations didactiques en Mathématiques*. Grenoble : La Pensée Sauvage

Brousseau, G. (2009). Le cas de Gaël revisité (1999-2009).

Bru, M. (2002). Pratiques enseignantes : des recherches à conforter et à développer. *Revue française de pédagogie*, 63-73.

Brun, J. (1979). L'évaluation formative dans un enseignement différencié de mathématiques. *L'évaluation formative dans un enseignement différencié*, 203-215.

Buchs, C., Darnon, C., & Butera, F. (2011). L'évaluation, une menace. *Paris Presses Universitaires de France-PUF*, *140*.

Bush, W. S., & McGatha, M. B. (2010). Teachers' knowledge, beliefs, attitudes, and practices. *Teaching and learning mathematics: Translating research for school administrators*, 19-23.

Bueno-Ravel, L., & Gueudet, G. (2014). Quelles ressources pour les professeurs des écoles et leurs formateurs ? Apports de la recherche en didactique. *Actes du colloque COPIRELEM*, 15-36.

Butlen, D., Peltier-Barbier, M. L., & Pézard, M. (2004). Des résultats relatifs aux pratiques de professeurs débutants ou confirmés enseignant les mathématiques à l'école. *Dur pour les élèves, dur pour les enseignants, dur d'enseigner en ZEP*, 70-81.

Butlen, D., Charles-Pézard, M., & Masselot, P. (2011). Deux dimensions de l'activité du professeur exerçant dans des classes de milieux défavorisés : installer la paix scolaire, exercer une vigilance didactique. In *Colloque international INRP, 16, 17 et 18 mars 2011. Le travail enseignant au XXIe siècle Perspectives croisées : didactiques et didactique professionnelle*.

Butler, D., & Winne, P. (1995). Feedback and self-regulated learning: A theoretical synthesis. Review of Educational Research, 65(3), 245-281.

Campanale, F. (2001). Quelques éléments fondamentaux sur l'évaluation. *IUFM de Grenoble–DPT SHS-Ressources pédagogiques*.

Cardinet, J. (1983). *Des instruments d'évaluation pour chaque fonction*. Institut romand de recherches et de documentation pédagogiques.

Cardinet, J., & Institut romand de recherches et de documentation pédagogiques (Neuchâtel). (1990). *Remettre le quantitatif à sa place en évaluation scolaire*. Institut romand de recherches et de documentation pédagogiques.

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

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.

Cauley, K. M., & McMillan, J. H. (2000). Do teachers grade differently in low SES middle schools. In *annual meeting of the American Educational Research Association, New Orleans, LA*.

Charles-Pézard M. (2010). Installer la paix scolaire, exercer une vigilance didactique. Recherches en didactique des mathématiques, 30 (2), 197-261.

Chesnais, A. (2014). Différenciation dans le processus d'enseignement-apprentissage en mathématiques en éducation prioritaire et ailleurs. *Revue française de Pédagogie,* 176(3), 63-73.

Chesné, J. F. (2014). *D'une évaluation à l'autre : des acquis des élèves sur les nombres en sixième à l'élaboration et à l'analyse d'une formation d'enseignants centrée sur le calcul mental* (Doctoral dissertation, Université Paris 7–Denis Diderot).

Chevallard, Y. (1985). *La transposition didactique* (Vol. 95). Grenoble : La pensée sauvage.

Chevallard, Y. (1986). Vers une analyse didactique des faits d'évaluation. In J-M. De Ketele (Ed.), *L'évaluation : approche descriptive ou prescriptive*, Bruxelles: De Boeck, 31-59.

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

Chevallard, Y., & Mercier, A. (1987). *Sur la formation historique du temps didactique*. Cahier n° 8. Marseille, France : IREM d'Aix-Marseille.

Chevallard, Y. (1988a). *Notes sur la question de l'échec scolaire*. IREM d'Aix-Marseille.

Chevallard, Y. (1988b). *Sur l'analyse didactique : deux études sur les notions de contrat et de situation*, IREM d'Aix-Marseille n°14, 92 pages.

Chevallard, Y. (1989). Évaluation, véridiction, objectivation, Conférence inaugurale donnée lors des Rencontres internationales sur l'évaluation en éducation (Paris, 27-29 septembre 1989). Paru in J. Colomb et J. Marsenach (Eds.), L'évaluateur en révolution (pp. 13-36), INRP, Paris.

Chevallard, Y. (1990). Évaluation, véridiction, objectivation. *L'évaluateur en révolution*, 13-36.

Chevallard, Y. (1991). La transposition didactique. Du savoir savant au savoir enseigné, 2nd eds. La Pensée Sauvage Editions, Grenoble perspectives. *Encyclopedia of mathematics education. Springer, Dordrecht*.

Chevallard, Y. (1992). Concepts fondamentaux de la didactique : perspectives apportées par une approche anthropologique. *Recherches en didactique des mathématiques*, 12(1), 73-111.

Chevallard, Y. (1998). *Analyse des pratiques enseignantes et didactique des mathématiques*, Actes de l'université d'été de La Rochelle, 4-11 juillet 1998, IREM de Clermont-Ferrand, 91-120.

Chevallard, Y. (1999). L'analyse des pratiques enseignantes en théorie anthropologique du didactique. *Recherches en didactique des mathématiques*, *19*(2), 221-265.

Chevallard Y. (2002). Organiser l'étude 1. Structures et Fonctions. In J-L. Dorier & al. (eds) Actes de la 11ième Ecole d'été de didactique des mathématiques -Corps- 21-30 Août 2001 (pp. 3-22). Grenoble : La Pensée Sauvage

Chopin, M. P. (2006). Temps d'enseignement et temps didactique Approche didactique de la question du temps dans l'enseignement des mathématiques au cycle 3 de l'école élémentaire. *Carrefours de l'Education*, (1), 53-71.

Cizek, G. J., Fitzgerald, S. M., & Rachor, R. A. (1995). Teachers' assessment practices: Preparation, isolation, and the kitchen sink. *Educational assessment*, *3*(2), 159-179.

Cizek, G. J. (2009). *Handbook of formative assessment*. Taylor & Francis US.

Claparède, É. (1922). *L'orientation professionnelle : ses problèmes et ses méthodes*. Bureau International du Travail.

Clivaz S. (2011). *Des mathématiques pour enseigner, analyse de l'influence des connaissances mathématiques d'enseignants vaudois sur leur enseignement des mathématiques à l'école primaire*. Thèse de doctorat, Université de Genève.

Clivaz, S. (2014). *Des mathématiques pour enseigner ? quelle influence les connaissances mathématiques des enseignants ont-elles sur leur enseignement à l'école primaire ?* La Pensée sauvage éditions.

Cobb, P., & Yackel, E. (1998). The culture of the mathematics classroom. *The culture of the mathematics classroom*, 158.

Coppé, S. (1993). *Processus de vérification en mathématiques chez les élèves de 1ère scientifique en situation de devoir surveillé,* Thèse de doctorat, Université Claude Bernard. Lyon I.

Coppé, S. (1998). Composantes privées et publiques du Travail de l'élève en situation de devoir surveillé de mathématiques. *Educational studies in mathematics*, *35*(2), 129-151.

Coppé, S. (2015). Développer les pratiques d'évaluation formative pour les professeurs de mathématiques et sciences. In *27e colloque de l'ADMEE-Europe*.

Coulange, L. (2001). Enseigner les systèmes d'équations en troisième une étude économique et écologique. *Recherches en didactique des mathématiques*, *21*(3), 305-353.

Coulange L. (2013) Débuter en collège ZEP : quelles pratiques enseignantes ? Un zoom sur deux professeurs de mathématiques. *Recherches en Didactique des Mathématiques,* 32(3), 361-408.

Crahay, M. (1999). Les enseignants, leurs croyances, leurs pratiques d'évaluation et l'échec scolaire. *Voyage dans un espace multidimensionnel. Hommage à Daniel Bain*, 15-35.

Crinon, J. (2012). Écrire, s'initier à la recherche, se professionnaliser : un triptyque à repenser. R. Lozi et N. Biagioli (dir.), *Actes des Deuxièmes rencontres des chercheurs en interdidactique : l'initiation à la recherche dans la formation des enseignants à l'université*. Nice, France : Université de Nice Sophia-Antipolis

Darré, J. P. (1999). *La production de connaissance pour l'action : arguments contre le racisme de l'intelligence*. Editions Quae.

DeBlois, L., & Squalli, H. (2002). Implication de l'analyse de productions d'élèves dans la formation des maîtres du primaire. *Educational Studies in Mathematics*, *50*(2), 212-237.

De Ketele, J.M., Chastrette, M., Cros, D., Mettelin, P., Thomas, J. (1989). Guide du formateur. Bruxelles, De Boeck Université.

De Ketele, J. M. (1993). L'évaluation conjuguée en paradigmes. *Revue française de pédagogie*, *103*(1), 59-80.

De Ketele, J. M., & Roegiers, X. (1993). L'évaluation : Approche et démarches. *Louvain-la-Neuve, Diffusion Universitaire CIACO*.

De Ketele, J.-M. (2000). En guise de synthèse : Convergences autour des compétences. In C. Bosman, F.-M. Gerard & X. Roegiers (Éds.). *Quel avenir pour les compétences ?* (pp. 187-191). Bruxelles : De Boeck Université.

De Ketele, J. M. (2001). Évolution des problématiques issues de l'évaluation formative. *L'activité évaluative réinterrogée : Regards scolaires et socioprofessionnels*, 102-108.

De Ketele, J. M., & Dufays, J. L. (2003). Vers de nouveaux modes d'évaluation des compétences. *Collès, L., Dufays, J.-L., Maeder, C. Enseigner le français, l'espagnol et l'italien. Les langues romanes à l'heure des compétences. Bruxelles: De Boeck–Duculot*.

De Ketele, J. M. (2016). L'évaluation et ses nouvelles tendances, sources de dilemmes. *Éducation Permanente*, *208*, 19-32.

De Ketele, J. M., & Gerard, F. M. (2005). La validation des épreuves d'évaluation selon l'approche par les compétences. *Mesure et évaluation en éducation*, *28*(3), 1-26.

De Lange, J. (2007). Large-scale assessment. *Second handbook of research on mathematics teaching and learning*, 1111-1142.

De Landsheere, G. (1973). Formes nouvelles de l'évaluation. *Français dans le Monde : Revue Internationale et Francophone des Professeurs de Français*, *100*, 45-53.

De Landsheere, G. (1980). *Examens et évaluation continue. Précis de docimologie*. Bruxelles, Paris : Labor, Nathan.

De Landsheere, V. (1988). *Faire réussir faire échouer : la compétence minimale et son évaluation*. Presses Universitaires de France.

Delarue-Breton, C. (2016). *Inégalité d'accès au savoir, ou inégalité d'accès au questionnement? De l'étude du dialogisme du discours scolaire à l'étude de l'activité dialogique des élèves et des étudiants*. Note de synthèse pour le diplôme d'HDR, soutenue à l'université Paris 8, le 6 décembre.

Delozanne, E., Grugeon, B., & Jacoboni, P. (2002). Analyses de I'activité et IHM pour I'éducation. In *Proceedings of the 14th Conference on l'Interaction Homme-Machine* (pp. 25-32). ACM.

Delcambre, I. (1994). La note : mesure ou message ? *Recherches* (21), 17-24.

Demonty, I., & Fagnant, A. (2005). Evaluation de la culture mathématique dans PISA : un regard neuf sur les compétences des élèves. *Mathématique et Pédagogie*, *54*, 39-56.

Desgagné, S. (1997). Le concept de recherche collaborative : l'idée d'un rapprochement entre chercheurs universitaires et praticiens enseignants. *Revue des sciences de l'éducation*, *23*(2), 371-393.

Desgagné, S., Bednarz, N., Lebuis, P., Poirier, L., & Couture, C. (2001). L'approche collaborative de recherche en éducation : un rapport nouveau à établir entre recherche et formation. *Revue des sciences de l'éducation*, *27*(1), 33-64.

Desgagné, S., & Bednarz, N. (2005). Médiation entre recherche et pratique en éducation : faire de la recherche «avec» plutôt que «sur» les praticiens. *Revue des sciences de l'éducation*, *31*(2), 245-258.

Dierendonck, C., & Fagnant, A. (2014). Approche par compétences et évaluation à large échelle : deux logiques incompatibles ? *Mesure et évaluation en éducation*, *37*(1), 43-82.

Douady, R. (1984). *Jeux de cadres et dialectiques outil-objet dans l'enseignement des Mathématiques. Une réalisation dans tout le cursus primaire,* Thèse de Doctorat, Université Paris VII.

Dubet, F. (2007). Injustices et reconnaissance. In *La quête de reconnaissance* (pp. 15-43). La Découverte.

Emprin, F. (2007). *Formation initiale et continue pour l'enseignement des mathématiques avec les TICE : cadre d'analyse des formations et ingénierie didactique,* Thèse de Doctorat, Université Paris-Diderot-Paris VII.

Engestrom, Y. (2001). Expansive learning at work: Toward an activity theoretical reconceptualization. *Journal of Education and Work*, 14, 133–156.

Elbers, E., & Kelderman, A. (1991). Verwachtingen en misverstanden bij het toetsen van kennis [Expectations and misconceptions in the assessment of knowledge]. *Pedagogische Studiën, 68* (4), 176-184.

Fagnant, A., Demonty, I., Dierendonck, C., Dupont, V., & Marcoux, G. (2014). Résolution de tâches complexes, évaluation « en phases » et compétence en mathématiques. *L'évaluation des compétences en milieu scolaire et en milieu professionnel*, 179-189.

Fiske, E. B. (1997, May 1). America's test mania. *New York Times (Education Supplement).*

Fontana, D., & Fernandes, M. (1994). Improvements in mathematics performance as a consequence of self-assessment in Portuguese primary school pupils. *British Journal of Educational Psychology*, 64(3), 407-417.

Frary, R. B., Cross, L. H., & Weber, L. J. (1993). Testing and grading practices and opinions of secondary teachers of academic subjects: Implications for instruction in measurement. *Educational Measurement: Issues and Practice*, 12(3), 23-30.

Freudenthal, H. (2012). *Mathematics as an educational task*. Springer Science & Business Media.

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 7.

Grenier, D. (1998). Milieu et contrat dans l'étude de l'enseignant et des interactions didactiques. *Actes des secondes journées didactiques de La Fouly,* 123-146.

Grugeon B. (1995). Etude des rapports institutionnels et des rapports personnels des élèves à l'algèbre élémentaire dans la transition entre deux cycles d'enseignement : BEP et Première G. Thèse de Doctorat, Université Paris 7.

Grugeon B. (1997). Conception et exploitation d'une structure d'analyse multidimensionnelle en algèbre élémentaire. *Recherches en didactique des mathématiques,* 17(2), 167-210.

Grugeon, B., Pilet, J., Chenevotot-Quentin, F., & Delozanne, E. (2012). Diagnostic et parcours différenciés d'enseignement en algèbre élémentaire. *Recherches en*

Didactique des Mathématiques, Numéro spécial hors-série, Enseignement de l'algèbre élémentaire: bilan et perspectives, 137-162.

Grugeon-Allys, B. (2015). Réguler l'enseignement en algèbre élémentaire : une approche multidimensionnelle. *Actes du séminaire national de didactique des mathématiques*, 40.

Grugeon-Allys, B. & Grapin, N. (2016). *Validité des évaluations : une approche épistémo-didactique*. Communication présentée au colloque Evaluation en mathématiques, Université Paris-Est Créteil.

Gueudet, G., Trouche, L. (2008). Du travail documentaire des enseignants : genèses, collectifs, communautés. Le cas des mathématiques. *Education et didactique*, 2(3), 7-33.

Gueudet, G., & Trouche, L. (2010). Ressources en ligne et travail collectif enseignant : accompagner les évolutions de pratique. In *Congrès Actualité de la Recherche en Education* (pp. 1-10). Université de Genève.

Gueudet, G., Pepin, B., & Trouche, L. (2012). *From text to Lived ressources: Mathematics Curriculum Material and Teacher Development*. New York: Springer

Gueudet, G. (2013). Les professeurs de mathématiques et leurs ressources professionnelles. *Actes du 20ème colloque de la CORFEM 13-14 Juin 2013*, Université Joseph Fourier, Grenoble.

Haberman, M. (1991). The pedagogy of poverty. *Phi Delta Kappan,* 73(4), 290-94.

Hadji, C. (2012). *Faut-il avoir peur de l'évaluation ?* (pp. 283-290). De Boeck Supérieur.

Hadji, C. (1989). Eléments pour un modèle de l'articulation formation/évaluation. *Revue française de pédagogie*, 49-59.

Haspékian, M., Kiwan, M. & Roditi, E. (2016). *Feedback des enseignants, évaluation formative et régulation des apprentissages : conséquences pour une approche didactique des interactions professeurs-élèves en classe de mathématiques*. (Rapport de recherche transmis à l'ANR pour le projet NeoPraEval). Paris, France.

Heuvel-Panhuizen, M., & Becker, J. (2003). Towards a didactic model for assessment design in mathematics education. *Second international handbook of mathematics education*, 689-716.

Heritage, M., Kim, J., Vendlinski, T., & Herman, J. (2009). From evidence to action: A seamless process in formative assessment? *Educational Measurement: Issues and Practice*, *28*(3), 24-31.

Hill, H. C., & Ball, D. L. (2004). Learning mathematics for teaching: Results from California's mathematics professional development institutes. *Journal for research in mathematics education*, 330-351.

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American educational research journal*, 42(2), 371-406.

Hodgen, J., & Marshall, B. (2005). Assessment for learning in English and mathematics: A comparison. *Curriculum journal*, 16(2), 153-176.

Hodgen, J., & Van den Heuvel-Panhuizen, M. (2013). Improving assessment in school mathematics. *MasterClass in Mathematics Education: International Perspectives on Teaching and Learning*, 27-38.

Horoks, J. (2006). *Les triangles semblables en classe de 2nde: Des enseignements aux apprentissages Etude de cas,* Thèse de Doctorat, Université Paris-Diderot, Paris 7.

Horoks, J., Pilet, J., & Haspekian, M. (2015). Quelles pratiques d'évaluation en algèbre au collège?. In *XXIIe Colloque de la CORFEM (COmmission de Recherche sur la Formation des Enseignants de Mathématiques)*.

Horoks, J., & Pilet, J. (2015). Etudier et faire évoluer les pratiques d'évaluation des enseignants de mathématiques en algèbre au collège dans le cadre d'un Léa. *L. Theis, Actes EMF2015, Pluralités culturelles et universalité des mathématiques: enjeux et perspectives pour leur enseignement et leur apprentissage, Alger, GT9*, 791-804.

Horoks, J., Pilet, J. (2016). Analyser les pratiques d'évaluation des enseignants à travers la prise en compte des élèves. *In Actes du XXVIIIème Colloque de l'ADMEE-Europe-Lisbonne-13-15 janvier 2016*.

Houdement, C., & Kuzniak, A. (2006). Paradigmes géométriques et enseignement de la géométrie. *Annales de didactique et de sciences cognitives*, 11, 175-193.

Issaieva, É., Pini, G., & Crahay, M. (2011). Positionnements des enseignants et des élèves du primaire face à l'évaluation: une convergence existe-t-elle?. *Revue française de pédagogie*, (3), 5-26.

Kahn, S. (2010). Pédagogie différenciée. *Le point sur la recherche en éducation*.

Kane, M. (2006). Content-related validity evidence in test development. *Handbook of test development*, 131-153.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement, (*4th ed.; p. 17-64). Westport, CT: American Council on Education and Praeger.

Kilpatrick, J. (1993). The chain and the arrow: From the history of mathematics assessment. In *Investigations into assessment in mathematics education* (pp. 31-46). Springer Netherlands.

Krummheuer, G. (1988). Structures microscopiques des situations d'enseignement des mathématiques'. In *Actes du premier colloque franco-allemand de didactique des mathématiques, La Pensée Sauvage, Grenoble* (pp. 41-51).

Lafortune, L., & Allal, L. K. (2008). *Jugement Professionnel en Evaluation: Pratiques Enseignantes Au Québec Et à Genève* (Vol. 21). PUQ.

Landsheere, D. (1971). *Evaluation continue et examens, Précis de docimologie*, Bruxelles, Ed.

Laveault, D., & Grégoire, J. (1997). *Introduction aux théories des tests en sciences humaines*. Bruxelles: De Boeck Université.

Laveault, D., & Grégoire, J. (2002). *Introduction aux théories des tests: en psychologie et en sciences de l'éducation*. De Boeck Supérieur.

Laveault, D. (1999). Autoévaluation et régulation des apprentissages. *C. Depover & B. Noël (Éds.). L'évaluation des compétences et des processus cognitifs*. Bruxelles: DeBoeck-Université.

Laveault, D. (2007). Quelles compétences, quels types de validité pour l'évaluation. *L. Bélair; D. Laveault; C. Lebel (Éds). Les compétences professionnelles en enseignement et leur evaluation*, 27-50.

Laveault, D. (2008). Le jugement professionnel d'évaluation scolaire: Enjeux, tensions et synergies nouvelles. *Revue suisse des sciences de l'éducation*, 30(3), 483-500.

Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment: Minute by minute, day by day. *Educational Leadership*, 63(3), 19–24.

Leder, G. C. (1992). Curriculum planning+ assessment= learning. *Assessment and learning of mathematics*, 330-344.

Leplat, J. (1997). *Regards sur l'activité en situation de travail: contribution à la psychologie ergonomique*. Presses universitaires de France.

Ligozat, F., & Leutenegger, F. (2008). Construction de la référence et milieux différentiels dans l'action conjointe du professeur et des élèves. Le cas d'un problème d'agrandissement de distances. *Recherches en didactique des mathématiques*, 28(3), 319-378.

Lin, P. J. (2006). Conceptualizing teachers' understanding of students' mathematical learning by using assessment tasks. *International Journal of Science and Mathematics Education*, *4*(3), 545-580.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational researcher*, 20(8), 15-21.

Loewenberg Ball, D., Thames, M. H., & Phelps, G. (2008). Content knowledge for teaching: What makes it special? *Journal of teacher education*, 59(5), 389-407.

Looney, J., & Poskitt, J. (2005). New Zealand: Embedding formative assessment in multiple policy initiatives. *Formative assessment: Improving learning in secondary classrooms*, 177-184.

Loye, N. (2010). 2010, odyssée des modèles de classification diagnostique (MCD). *Mesure et évaluation en éducation*, 33(3), 75-98.

Martin, J. (2002). Aux origines de la « sciences des examens » 1920-1940. *Histoire de l'éducation*, 177-199.

Margolinas C., & Wozniak, F. (2009). Usage des manuels dans le travail de l'enseignant : l'enseignement des mathématiques à l'école primaire. *Revue des sciences de l'éducation*, 35 (2), 59-82.

Margolinas C., & Wozniak, F. (2010). Rôle de la documentation scolaire dans la situation du professeur : le cas de l'enseignement des mathématiques à l'école élémentaire. In Gueudet G., Trouche L. (Eds.) (pp. 233-249), *Ressources vives, le travail documentaire des professeurs, le cas des mathématiques*. Rennes: Presses Universitaires de Rennes et INRP.

Marro, C. (1995). Réussite scolaire en mathématiques et physique, en passage en 1ère S: Quelles relations du point de vue des élèves et des enseignants? Étude différentielle suivant le sexe des élèves. *Revue française de pédagogie*, 27-35.

Maulini, O. (2012). Le chiffre et la lettre? entre culte du résultat et culture de la règle: comment l'enseignement change, et comment ce changement peut évoluer. In *Écoles en mouvements et réformes* (Vol. 1, pp. 53-63). De Boeck Supérieur.

Maury, H. (2006). *Cohérence dans le discours oral en français langue étrangère: démarches et procédés d'identification et d'évaluation*, Thèse de doctorat, Université de Nantes.

Maxwell, G. S. (2002). *Moderation of teacher judgments in student assessment*. Discussion Paper on Assessment and Reporting, Queensland School Curriculum Council.

McGatha, M., & Bush, W. (2013). Classroom assessment in mathematics. *SAGE handbook of research on classroom assessment*, 448-461.

McLean, L. D. (1982). Educational assessment in the Canadian provinces. *Educational Analysis*, 4(3), 79-96.

McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice*, *20*(1), 20-32.

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational measurement: Issues and practice*, *22*(4), 34-43.

McMillan, J. H. (2013). Why we need research on classroom assessment. *SAGE handbook of research on classroom assessment*, 3-16.

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research*, 95(4), 203-213.

McMillan, J. H. (2013). *Classroom Assessment: Pearson New International Edition: Principles and Practice for Effective Standards-Based Instruction*. Pearson Higher Ed.

Mercier, A. (1992). *L'élève et les contraintes temporelles de l'enseignement, un cas en calcul algébrique,* Thèse de Doctorat, Université Sciences et Technologies-Bordeaux I.

Mercier, A. (1995), Approche biographique de l'élève et des contraintes temporelles de l'enseignement : un cas en calcul algébrique. *Recherches en didactique des mathématiques*, 15(1)1, 97-142.

Merle, P. (1996). *L'évaluation des élèves: enquête sur le jugement professoral*. Presses universitaires de France.

Merle, P. (2007). *Les notes. Secrets de fabrication*. Paris. PUF.

Merle, P. (2015). *Les notes. Secrets de fabrication*. Presses universitaires de France.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, 18(2), 5-11.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9), 741.

Morgan, C. (1999). *Assessment in mathematics education: a critical social research perspective*. In J. Portela (Ed.), Actas do IX Seminario de Investigaçao em Educaçao Matematica (pp. 5-23). Guimaraes: Associaçao de Professores de Matematica.

Morgan, C., & Watson, A. (2002). The interpretative nature of teachers' assessment of students' mathematics: Issues for equity. *Journal for research in mathematics education*, 78-110.

Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practice*, *22*(4), 13-25.

Mosconi, N. (2001). Comment les pratiques enseignantes fabriquent de l'inégalité entre les sexes, *Les dossiers des sciences de l'éducation*, (5), 97-109.

Mottier Lopez, L., & Allal, L. (2008). Le jugement professionnel en évaluation: un acte cognitif et une pratique sociale située. *Revue suisse de sciences de l'éducation*, *30*(3), 465-482.

Mottier Lopez, L. (2012). *La régulation des apprentissages en classe*. De Boeck.

Mottier Lopez, L. (2014). L'évaluation pédagogique va-t-elle enfin marcher sur ses deux pieds? Les enseignements de l'histoire récente de l'école primaire genevoise. *Éducation et francophonie*, 42(3), 85-101.

Mottier Lopez, L. (2015). L'évaluation formative des apprentissages des élèves: Entre innovations, échecs et possibles renouveaux par des recherches participatives. *Questions vives*, (23).

Mottier Lopez, L. & Tessaro, W. (2016). Introduction : pourquoi interroger le jugement professionnel dans l'évaluation et la régulation des apprentissages des élèves ? In L. Mottier Lopez & W. Tessaro (Ed.), *Le jugement professionnel, au coeur de l'évaluation et de la régulation des apprentissages* (pp. 1-24). Bern: Peter Lang (collection Exploration).

Niss, M. (1993). Assessment of mathematical applications and modelling in mathematics teaching. *Innovation in mathematics education by modelling and applications*, 41-51.

Niss, M. (1993). *Investigations in to assessment in mathematics education*, An ICMI Study, Springer Netherlands.

Noirfalise, R. (1986). Attitudes du maître et résultats scolaires en mathématiques. *Recherches en didactique des mathématiques*, *7*(3), 75-112.

Nortvedt, G. A., Santos, L., & Pinto, J. (2016). Assessment for learning in Norway and Portugal: the case of primary school mathematics teaching. *Assessment in Education: Principles, Policy & Practice*, *23*(3), 377-395.

Nunziati, G. (1990). Pour construire un dispositif d'évaluation formatrice. *Cahiers Pédagogiques*, 280, 48-64.

Osterlind, S. J. (1998). *What Is Constructing Test Items?* Springer Netherlands.

Pelletier, G. (1997). L'évaluation des programmes universitaires québécois: Le cas des formations à l'enseignement. *Gestion de l'enseignement supérieur*, *9*(2), 75-90.

Perrenoud, P. (1989). *Pour une approche pragmatique de l'évaluation formative*. Conseil de l'Europe.

Perrenoud, P. (1995). *La pédagogie à l'école des différences*. Paris: ESF.

Perrenoud, P. (1998). L'évaluation des élèves. *De la fabrication de l'excellence à la régulation des apprentissages. Entre deux logiques, Paris-Bruxelles,* De Boeck & Larcier.

Perrin-Glorian, M. J., & Hersant, M. (2003). Milieu et contrat didactique, outils pour l'analyse de séquences ordinaires. *Recherches en didactique des mathématiques*, *23*(2), 217-276.

Pézard, M., Butlen, D., & Masselot, P. (2012). *Professeurs des écoles débutants en ZEP: quelles pratiques? Quelle formation?*. la Pensée sauvage.

Piéron, H. (1963). *Examens et docimologie* (Vol. 15). Presses Universitaires de France.

Philipp, R. A. (2007). Mathematics teachers' beliefs and affect. In F. K. Lester, Jr. (Ed.) *Second handbook of research on mathematics teaching and learning* (pp. 257-315). Charlotte, NC: Information Age.

Pilet, J. (2015). Réguler l'enseignement en algèbre élémentaire par des parcours d'enseignement différencié. *Recherches en didactique des mathématiques*, *35*(3), 273-312.

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, Université Paris-Diderot-Paris VII.

Pluvinage, F. (1979). Loto-questionnaires (pour l'evaluation et l'auto-controle en mathématiques). *Educational Studies in Mathematics*, *10*(4), 443-485.

Quiquempois, G. (2016). *Un dispositif de formation basé sur l'EPCC et ses différentes variantes*. Communication présentée au colloque Evaluation en mathématiques, UPEC Université Paris-Est Créteil.

Raymond, A. M. (1997). Inconsistency between a beginning elementary school teacher's mathe- matics beliefs and teaching practice. *Journal for Research in Mathematics Education*, 28(5), 550–576.

Raynal, F. & Rieunier, A. (1997, 4e éd. 2003). *Pédagogie : dictionnaire des concepts clés*, Paris : ESF éditeur.

Régner, I., & Huguet, P. (2007). Stereotype Threat Among School Girls in Quasi-Ordinary Classroom Circumstances. *Journal of Educational Psychology*, *99*(3), 561-574.

Remesal, A. (2007). Educational reform and primary and secondary teachers' conceptions of assessment: The Spanish instance, building upon Black and Wiliam (2005). *The Curriculum Journal*, *18*(1), 27-38.

Rémond, M. (2006). Éclairages des évaluations internationales PIRLS et PISA sur les élèves français. *Revue française de pédagogie. Recherches en éducation*, (157), 71-84

Rey, B. (1996). *Les compétences transversales en question*. Paris : ESF.

Rey, B. (2014). Compétence et évaluation en milieu scolaire: une relation complexe. *L'évaluation des compétences en milieu scolaire et en milieu professionnel*, 8-30.

Rey, O., & Feyfant, A. (2014). Évaluer pour (mieux) faire apprendre. *Dossier de veille de l'Ifé*, *94*.

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

Robert, A. (1998). Outils d'analyse des contenus mathématiques à enseigner au lycée et à l'université. *Recherches en didactique des mathématiques*, *18*(2), 139-189.

Robert, A. (2008). Problématique et méthodologie communes aux analyses des activités mathématiques des élèves en classe et des pratiques des enseignants de mathématiques. In Vandebrouck F. (Ed). *La classe de mathématiques: activités des élèves et pratiques des enseignant* (pp. 33–57). Toulouse: Octares.

Robert, A., & Rogalski, J. (2002). Le système complexe et cohérent des pratiques des enseignants de mathématiques: une double approche. *Canadian Journal of Math, Science & Technology Education*, *2*(4), 505-528.

Robert, A., & Vivier, L. (2013). Analyser des vidéos sur les pratiques des enseignants du second degré en mathématiques: des utilisations contrastées en recherche en didactique et en formation de formateurs–quelle transposition? *Education & didactique*, *7*(2), 115-144.

Roditi, E. (2011). *Recherches sur les pratiques enseignantes en mathématiques: apports d'une intégration de diverses approches et perspectives*, Note de synthèse pour l'Habilitation à Diriger des Recherches, Université René Descartes-Paris V.

Roditi, E., & Salles, F. (2015). Nouvelles analyses de l'enquête PISA 2012 en mathématiques. *Éducation et formations*, *86*, 236-267.

Roditi, E. (2013). Le métier d'enseignant et la recherche collaborative. *In* N. Bednarz (Dir.). *Recherche collaborative et pratique enseignante. Regarder ensemble autrement* (pp. 351-364). Paris: L'Harmattan.

Rogalski, J., & Robert, A. (2015). De l'analyse de l'activité de l'enseignant à la formation des formateurs. Le cas de l'enseignement des mathématiques dans le secondaire. Raisons Éducatives, 19, 95-114. (In V. Lussi-Borer, M. Durand, & F. Yvon (Eds.), *Analyse du travail et formation dans les métiers du supérieur* (pp. 95-114). Paris, De Boeck Supérieur.)

Romainville, M. (2002). *L'évaluation des acquis des étudiants dans l'enseignement universitaire*. Rapport établi à la demande du Haut Conseil de l'Evaluation de l'école.

Rogalski, J. (2007). *Approche de psychologie ergonomique de l'activité de l'enseignant*. Communication présentée au Séminaire International Professionnalisation des enseignants de l'éducation de base : les recrutements sans formation initiale. Centre International d'études pédagogiques, Sèvres.

Ross, J. A., Hogaboam-Gray, A., & Rolheiser, C. (2002). Student self-evaluation in grade 5-6 mathematics effects on problem-solving achievement. *Educational Assessment*, *8*(1), 43-58.

Ruminot Vergara, C. (2014). *Effets d'un système national d'évaluation sur l'enseignement des mathématiques: le cas de SIMCE au Chili*. Thèse de Doctorat, Université Paris-Diderot, Paris 7.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional science*, *18*(2), 119-144.

Samuelowicz, K., & Bain, J. D. (2002). Identifying academics' orientations to assessment practice. *Higher education*, *43*(2), 173-201.

Sarrazy, B. (1995). Note de synthèse [Le contrat didactique]. *Revue française de pédagogie*, *112*(1), 85-118.

Sayac, N. (2008). Former à enseigner les mathématiques à l'école primaire, en France : quelles pratiques pour quelle formation ?. *Colloque AMSE 15e Congrès international*, Marrakech (Maroc).

Sayac, N. (2010). Appréhender la formation des professeurs des écoles en France à travers les pratiques des formateurs en mathématiques. *Colloque AREF 2010*, Genève (Suisse).

Sayac, N. (2012). Analyser les pratiques de formateurs en mathématiques pour questionner la formation des enseignants dans le 1er degré, *Les Cahiers du Cerfee (30)*, 183-200.

Sayac, N. (2013). Pratiques de formateurs en mathématiques dans le 1er degré : les savoirs de formation, *Recherche et Formation,* (71), 115-130.

Sayac, N., & Grapin, N. (2013). Facteurs de compétence et de complexité en mathématique: un outil au service de la formation des enseignants. *Actes du 25ème colloque de l'ADMEE-Europe, Fribourg*.

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

Sayac, N., & Grapin, N. (2015). Évaluation externe et didactique des mathématiques: un regard croisé nécessaire et constructif, *Recherche en didactique des mathématiques,* 35(1), 101-126.

Scallon, G. (1997). L'autoévaluation: une tendance lourde en évaluation. *Vie pédagogique*, *103*, 27-31.

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

Schneider-Gilot, M. (2006). Quand le courant pédagogique «des compétences» empêche une structuration des enseignements autour de l'étude et de la classification de questions parentes. *Revue française de pédagogie. Recherches en éducation*, (154), 85-96.

Schubauer-Leoni, M. L. (1986). Le contrat didactique: un cadre interprétatif pour comprendre les savoirs manifestés par les élèves en mathématique. *European Journal of Psychology of Education*, *1*(2), 139-153.

Schubauer-Leoni, M. L. (1991). L'évaluation didactique: une affaire contractuelle. *L'évaluation: problème de communication*, 79-95.

Sensevy, G., & Mercier, A. (Eds.). (2007). *Agir ensemble: l'action didactique conjointe du professeur et des élèves*. Presses universitaires de Rennes.

Sensevy, G. (2008). Le travail du professeur pour la théorie de l'action conjointe en didactique. *Recherche & formation*, (1).

Sensevy, G. (2011). *Le sens du savoir. Éléments pour une théorie de l'action conjointe en didactique* (p. 800). De Boeck.

Schoenfeld, A. H. (1998). Toward a theory of teaching-in-context. *Issues in education*, *4*(1), 1-94.

Schoenfeld, A. H. (2007). *Assessing mathematical proficiency* (Vol. 53). Cambridge university press.

Shavelson, R.J. 2008. Guest editor's introduction. *Applied Measurement in Education* 21, no. 4: 293–4

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational researcher*, *29*(7), 4-14.

Shepard, L. (2001). The role of classroom assessment in teaching and learning.

Shepard, L. A. (2005). Linking Formative Assessment to Scaffolding. *Educational leadership*, *63*(3), 66-70.

Shepard, L. A. (2006). Classroom assessment. *Educational measurement*, *4*, 623-646.

Shepard, L. A. (2008). Formative assessment: Caveat emptor. *The future of assessment: Shaping teaching and learning*, 279-303.

Shepard, L. A. (2009). Commentary: Evaluating the validity of formative and interim assessment. *Educational Measurement: Issues and Practice*, *28*(3), 32-37.

Shirley, M. L., Irving, K. E., Sanalan, V. A., Pape, S. J., & Owens, D. T. (2011). The practicality of implementing connected classroom technology in secondary mathematics and science classrooms. *International Journal of Science and Mathematics Education*, *9*(2), 459-481.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational researcher*, *15*(2), 4-14.

Skott, J. (2015). Towards a Participatory Approach to "Beliefs" in Mathematics Education. In B. Pepin & B. Roesken-Winter (Eds.), *From beliefs to dynamic affect systems in mathematics education*, 3–23.

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of experimental social psychology*, *35*(1), 4-28.

Stiggins, J. R., Arter, J. A., Chappuis, J., & Chappuis, S. (2005). *Classroom Assessment for Student Learning. Doing it right-using it well*. Portland, Oregón, USA: Assessment Training Institute.

Suurtamm, C., Koch, M., & Arden, A. (2010). Teachers' assessment practices in mathematics: Classrooms in the context of reform. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 399-417.

Suurtamm, C., & Koch, M. J. (2014). Navigating dilemmas in transforming assessment practices: experiences of mathematics teachers in Ontario, Canada. *Educational Assessment, Evaluation and Accountability*, *26*(3), 263-287.

Suurtamm, C., Thompson, D. R., Kim, R. Y., Moreno, L. D., Sayac, N., Schukajlow, S., ... & Vos, P. (2016). Assessment in Mathematics Education. In *Assessment in Mathematics Education* (pp. 1-38). Springer International Publishing.

Taras, M. (2005). Assessment–summative and formative–some theoretical reflections. *British journal of educational studies*, *53*(4), 466-478.

Terrier, C. (2014). Un coup de pouce pour les filles? Les biais de genre dans les notes des enseignants et leur effet sur le progress des élèves. *Les notes de l'Institut des Politiques Publiques*, 14.

Tessaro, W. (2013). Améliorer la qualité des pratiques évaluatives des enseignants: une articulation entre formation initiale et formation continue. *Bulletin de la Haute école pédagogique de Bern, du Jura et de Neuchâtel*, 21 (février), 8-9.

Thompson, D. R. (1992). *An evaluation of a new course in precalculus and discrete mathematics*, Doctoral dissertation, University of Chicago, Department of Education.

Thompson, D. R., & Kaur, B. (2011). Using a Multi-Dimensional Approach to Understanding to Assess Students' Mathematical Knowledge. In *Assessment In The Mathematics Classroom: Yearbook 2011, Association of Mathematics Educators* (pp. 17-31).

Thouin, M. (1993). L'évaluation des apprentissages en mathématiques: une perspective constructiviste. *Mesure et évaluation en éducation*, *16*(1-2), 47-64.

Torrance, H. (1993). Formative assessment: some theoretical problems and empirical questions. *Cambridge journal of education*, *23*(3), 333-343.

Tourmen, C. (2009). L'activité évaluative et la construction progressive du jugement. *Les dossiers des sciences de l'éducation*, *22*, 101-119.

Tourmen, C. (2014). Quand la didactique professionnelle s'intéresse aux apprentissages culturels. *Travail et Apprentissages*.

Trouilloud, D., & Sarrazin, P. (2002). L'effet Pygmalion existe-t-il en éducation physique et sportive? Influence des attentes des enseignants sur la motivation et la performance des élèves. *Science et Motricité: revue scientifique de l'ACAPS/ACAPS*, *46*(2), 69-94.

Tunstall, P., & Gipps, C. (1996).'How does your teacher help you to make your work better?' Children's understanding of formative assessment. *The Curriculum Journal*, *7*(2), 185-203.

Van den Heuvel-Panhuizen, M. H. A. M. (1996). *Assessment and realistic mathematics education* (Vol. 19). Utrecht University.

Van den Heuvel-Panhuizen, M., & Becker, J. (2003). Towards a didactic model for assessment design in mathematics education. *Second international handbook of mathematics education*, 689-716.

Van de Walle, B., & Rutkowski, A. F. (2006). A fuzzy decision support system for IT service continuity threat assessment. *Decision support systems*, *42*(3), 1931-1943.

Veldhuis, M., Van den Heuvel-Panhuizen, M., Vermeulen, J. A., & Eggen, T. J. (2013). Teachers' use of classroom assessment in primary school mathematics education in the Netherlands. *Cadmo*.

Vantourout, M. (2004). *Etude de l'activité et des compétences de professeurs des écoles et de professeurs de mathématiques dans des situations "simulées" d'évaluation à visée formative en mathématiques,* Thèse de Doctorat, Université Paris 5.

Vantourout, M., & Maury, S. (2006). Quelques résultats relatifs aux connaissances disciplinaires de professeurs stagiaires dans des situations simulées d'évaluation de productions d'élèves en mathématiques. *Revue des sciences de l'éducation*, *32*(3), 759-782.

Vantourout, M., Goasdoué, R., Maury, S., & Nabbout, M. (2012). À la frontière entre l'écologique et l'expérimental: des situations aménagées pour étudier l'activité évaluative en mathématiques. In M. Altet, M. Bru, C. Blanchard-Laville (Eds), *Observer les pratiques enseignantes* (pp.191-204). Paris: l'Harmattan.

Vantourout, M., & Goasdoué, R. (2014). Approches et validité psycho-didactiques des évaluations. *Education et Formation*, (e-302).

Vergnaud, G. (1989). Difficultés conceptuelles, erreurs didactiques et vrais obstacles épistémologiques dans l'apprentissage des mathématiques. *Construction des savoirs. Obstacles et conflits*, 33-44.

Vergnaud, G. (1997). The nature of mathematical concepts. *Learning and teaching mathematics: An international perspective*, 5-28.

Vergnaud, G. (2001). Psychologie du développement cognitif et évaluation des compétences. In G. Figari & M. Achouche (Eds). *L'activité évaluative réinterrogée : regards scolaires et socioprofessionnels* (pp. 43-51). Bruxelles : De Boeck Université.

Vial, M. (1995). Nature et fonction de l'auto-évaluation dans le dispositif de formation. *Revue française de pédagogie*, 69-76.

Vial, M. (1996). *La gestion du projet professionnel: guide de l'apprentissage à l'usage des étudiants élèves-professeurs en formation d'enseignants*, M. Vial, J. Ravestein, G. Sensevy, C. Eymard-Simonian. J.-J. Bonniol (Eds), Paris: Harmattan.

Vygotsky, L. (1978). Interaction between learning and development. *Readings on the development of children*, *23*(3), 34-41.

Vygotsky, L. (1987). Zone of proximal development. *Mind in society: The development of higher psychological processes*, *5291*, 157.

Weil-Barais, A. (1996). La médiation des apprentissages par l'enseignant. *Educations, juin-octobre*, 37-39.

Wiliam, D. (1993). *Assessing authentic tasks: norms, criteria, and other referents*. A paper presented at the Nordic Symposium Research on Assessment in Mathematics Education, University of Göteborg, November 5, 1993.

William, D. (2000). Integrating summative and formative functions of assessment. Keynote address. In *First Annual Conference of the European Association for Educational Assessment. Prague, Czech Republic*.

Zhao, X., Van den Heuvel-Panhuizen, M., & Veldhuis, M. (2016). Teachers' use of classroom assessment techniques in primary mathematics education—an explorative study with six Chinese teachers. *International Journal of STEM Education*, *3*(1), 19.

Zhao, X., Van den Heuvel-Panhuizen, M., & Veldhuis, M. (2017). Classroom assessment in the eyes of Chinese primary mathematics teachers: A review of teacher-written papers. *Studies in Educational Evaluation*, *52*, 42-54.

Résumé :
Le cadre didactique de l'évaluation proposé dans cette note de synthèse vise à étudier et analyser l'évaluation et ses pratiques, en mathématiques. Il émane du constat que l'approche scientifique développée en sciences de l'éducation pour étudier les questions d'évaluation (externe, de classe, de pratiques, etc.) n'est pas suffisante quand des contenus mathématiques sont en jeu, mais que l'approche didactique ne l'est pas non plus quand elle ne prend pas en compte des résultats cruciaux développés en sciences de l'éducation. Il a été conçu pour penser et analyser les "faits évaluatifs" (Chevallard, 1986) en conjuguant savoirs scientifiques en évaluation (dans la diversité des champs scientifiques concernés) et savoirs didactiques. Il prend en compte à la fois des contenus disciplinaires et des réalités professionnelles permettant de définir une nouvelle approche scientifique de l'évaluation, plus didactique et volontairement ancrée dans la réalité des pratiques en classe. Ce cadre didactique ne retient pas une entrée par les différentes fonctions de l'évaluation pour étudier les pratiques évaluatives des enseignant·e·s en mathématiques. La notion d'*épisode évaluatif* y est centrale et permet d'appréhender l'évaluation en mathématiques sous toutes ses formes, au delà de ses fonctions usuelles (diagnostique, formative, sommative, etc.). Dans ce cadre c'est l'étude des différents épisodes évaluatifs proposés au cours d'une séquence d'enseignement qui permet, en fonction du moment où ils sont proposés, du contenu des tâches évaluatives en jeu, de leur gestion et du contrat didactique s'y rattachant (*contrat didactique en évaluation*) qui permet de caractériser la pratique d'évaluation d'un·e enseignant·e en mathématiques. Cette pratique est pilotée par une "logique évaluative" appréhendée à partir d'indicateurs retenus pour la décrire, notamment la façon dont l'enseignant·e conçoit et élabore les différents épisodes évaluatifs qui ponctuent la séquence (ressources utilisées, *documents évaluatifs*, méthodes), le *jugement professionnel et didactique en évaluation* et la notation qu'il ou elle adopte.
Ce cadre articule deux axes d'analyse des pratiques d'évaluation en mathématiques, l'un focalisé sur les épisodes évaluatifs proposés aux élèves et l'autre structuré par la "logique évaluative" des professeur·e·s. Il permet d'ouvrir des perspectives à la fois en termes de problématiques scientifiques et à la fois en termes de méthodologies de recherche et de formation, favorisant le rapprochement entre chercheur·e·s et praticien·ne·s et la production de savoirs scientifiques nouveaux au service des différentes communautés.