Les évaluations nationales en France : outils pédagogiques et outils de pilotage, intérêts, limites et nouvelles perspectives

Bruno Suchaut

Les évaluations nationales en France : outils pédagogiques et outils de pilotage, intérêts, limites et nouvelles perspectives.

Bruno Suchaut
Université de Bourgogne et Irédu C.N.R.S., Dijon, France
19ème colloque de l'ADMEE EUROPE (11-13 septembre 2006, Luxembourg)

Depuis 1989, le système éducatif français s'est doté d'outils d'évaluation des élèves à différents niveaux de la scolarité primaire et secondaire[1]. Ces évaluations de masse, distribuées et utilisées à l'échelon national ont, d'une part contribué de manière progressive à la diffusion de la culture de l'évaluation à l'ensemble de la communauté éducative et, d'autre part, permis aux enseignants d'approfondir leur réflexion pédagogique (Joutard, Thélot, 1999 ; M.E.N., 2000). L'objectif initial de ces évaluations nationales était de «*permettre l'observation des compétences et apprécier les réussites et les difficultés éventuelles de chaque élève considéré individuellement, à un moment précis de la scolarité. Elles fournissent aux enseignants des repères exploitables quant aux types d'erreurs fréquemment produites par les élèves au cours de leur apprentissage…* »[2]. Même si la dimension diagnostique reste aujourd'hui prioritaire, ces évaluations peuvent être utilisées comme des instruments de pilotage car elles rendent objectivement compte du niveau d'acquisition des élèves à plusieurs niveaux : national, régional, départemental, établissement scolaire.

Pour le chercheur en éducation, les évaluations présentent également un intérêt dans le sens où celles-ci balaient un vaste ensemble de compétences et de connaissances alors que les tests couramment mobilisés dans les recherches françaises sont beaucoup plus restrictifs sur ce plan. Ce texte expose une analyse originale des épreuves des évaluations nationales de CE2 de français et de mathématiques sur la base d'une recherche récente qui étudie les résultats des évaluations de CE2 (de septembre 1999) et de sixième (de septembre 2002) d'une même cohorte d'élèves (Morlaix, Suchaut, 2006). Nous essaierons de montrer en quoi les outils d'évaluation livrent des informations capitales sur les apprentissages réalisés par les élèves, à condition de ne pas se limiter pas à la seule approche institutionnelle, notamment à la classification des compétences établies par les concepteurs des épreuves.

Nous rappellerons dans un premier temps les précautions à prendre dans l'utilisation des résultats des évaluations nationales, puis, dans un second temps, nous exposerons une méthodologie originale qui permet de mieux appréhender la structure des acquisitions des élèves. Nous énoncerons pour conclure les perspectives nouvelles qu'offre cette nouvelle approche dans différents domaines.

[1] Les évaluations nationales mis en place par la D.E.P. (Direction de l'Evaluation et de la Prospective) ont concerné à l'origine trois niveaux de la scolarité : CE2 (3ème année de l'enseignement élémentaire), sixième (1ère année de l'enseignement secondaire) et seconde (5ème année de l'enseignement secondaire).

[2] Note de service 2002-105 du 30.04.2002, B.O. N°19 du 9 mai 2002.

L'utilisation des résultats des évaluations nationales : précautions et limites

Les exercices proposés dans les épreuves d'évaluation reflètent les programmes scolaires officiels en vigueur et constituent à ce titre une référence commune pour l'ensemble des enseignants et des élèves[3]. Sur la base des simple taux de réussite aux items, il est possible d'avoir une vision du niveau d'acquisition des élèves à un moment donné de leur scolarité et d'identifier les lacunes de chacun en vue d'y remédier au cours de l'année scolaire. Outre cette perspective diagnostique et pédagogique, les résultats des évaluations fournissent des indicateurs objectifs de la réussite des élèves, au sein de la classe, de l'école, mais aussi à des niveaux supérieurs. Les comparaisons spatiales et temporelles peuvent ainsi servir d'éléments utiles au pilotage d'une circonscription primaire ou d'un département. Le Ministère insiste toutefois sur l'utilisation inadaptée qui pourrait être faite des résultats des évaluations, ceux-ci ne pouvant, en aucun cas, être considérés comme des normes à atteindre, mais seulement comme un diagnostic, en début de cycle[4], des réussites et des difficultés des élèves. Il est également clairement énoncé dans le discours officiel que les épreuves d'évaluation ne rendent que partiellement compte des compétences et des connaissances des élèves puisque la passation est collective et limitée à des exercices écrits. La prudence est aussi conseillée quant à l'interprétation de l'évolution au cours du temps des taux de réussite, les épreuves n'étant pas la plupart du temps identiques d'une année sur l'autre.

L'institution éducative annonce donc nettement les limites à une utilisation des évaluations nationales à des fins de pilotage par les résultats. Il est vrai que les taux de réussite en eux-mêmes n'ont pas de signification bien précise et il pourrait en effet être hasardeux de les interpréter de manière directe. A titre d'illustration, le graphique suivant présente les pourcentages de réussite en français à l'entrée en CE2 entre 1996 et 2005. On voit très nettement que les résultats varient sensiblement au cours de ces dix années, la moyenne des scores globaux oscille en effet entre 60,5% (en 2001) et 73,5% (en 2004) avec un chiffre moyen de 67,3%. Des écarts sensibles apparaissent aussi pour un même domaine ; par exemple, les résultats en production d'écrits chutent fortement entre 1996 et 1997, alors que pour la même période les scores dans la dimension des outils de la langue connaissent une augmentation[5]. Les différences constatées d'une année sur l'autre renvoient principalement à

[3] Différents personnels de l'Education nationale participent à la conception et à la sélection des exercices : enseignants, inspecteurs de l'Education nationale, Inspecteurs généraux…

[4] L'enseignement primaire français est organisé en 3 cycles. Le cycle I (cycle des apprentissages premiers) concerne l'école maternelle ; le cycle II (cycle des apprentissages fondamentaux) regroupe la grande section de maternelle, le cours préparatoire et le CE1 ; le cycle III (cycle des approfondissements) intègre le CE2, le CM1 et le CM2.

[5] Pour l'année 2005, le domaine habituellement nommé « outils de la langue » est scindé en deux champs « reconnaissance des mots » et « écriture et orthographe ». Nous avons reporté sur le graphique les scores correspondant à l'ensemble des items des deux champs.

la composition des épreuves, les chiffres étant proches quant les exercices sont identiques et des écarts marqués apparaissent quand les épreuves sont totalement renouvelées. Le poids variable accordé à chaque domaine d'acquisition et le degré de difficulté des exercices composant ces domaines expliquent aussi pourquoi, selon les années, c'est telle ou telle dimension des acquis qui semble la moins (ou la mieux) maîtrisée. Dans tous les cas, les allures des courbes du graphique témoignent du caractère fluctuant des taux de réussite et les variations temporelles constatées ne renvoient pas nécessairement à une baisse ou une hausse du niveau des élèves.

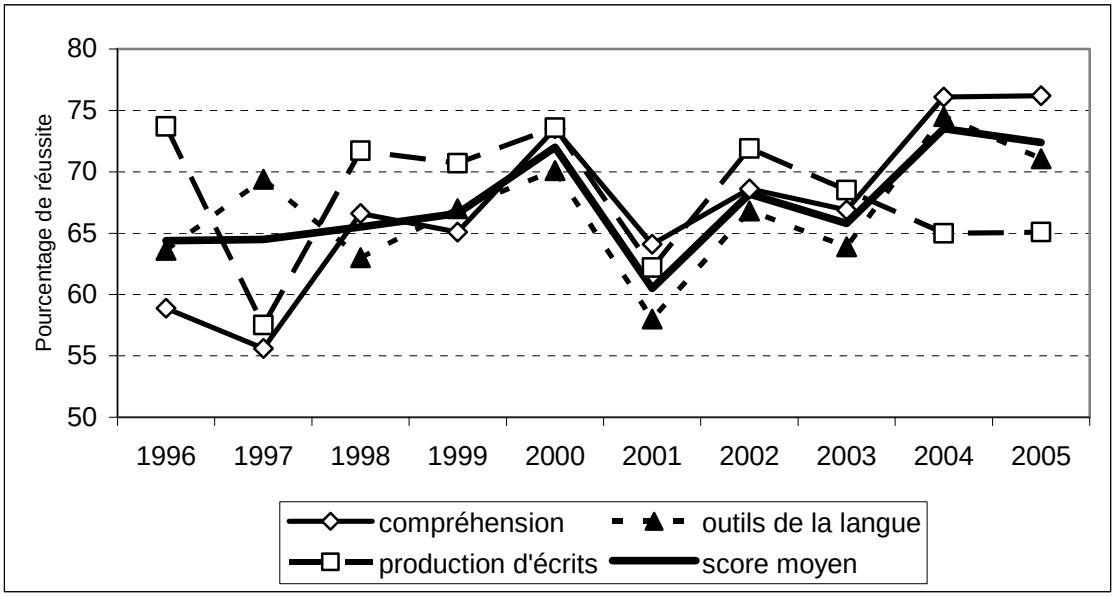

Figure1 : Evolution temporelle des résultats aux évaluations nationales de français au CE2

On peut également s'interroger sur l'échelle de mesure utilisée qui varie dans l'absolu entre 0% et 100% de réussite ; cette échelle ne fait pas référence à des critères de mesure précis et ne peut être considérée comme une mesure étalonnée du niveau des élèves. Les concepteurs des épreuves s'attendent-ils à ce qu'un élève qui aurait assimilé la moitié du programme du cycle II obtienne 50% de réussite au test ? La réponse est incertaine dans la mesure où un nombre infime d'élèves réussit la totalité des exercices présents dans les cahiers d'évaluation, ce qui ne veut évidemment pas dire qu'aucun élève français ne maîtrise correctement totalement le programme du cycle II. Ces limites ne signifient pas que les résultats des évaluations nationales ne peuvent pas être utilisés à des fins de comparaisons spatiales et temporelles, mais ces comparaisons impliquent l'usage d'une mesure relative et non absolue du niveau des élèves ; ceci est d'ailleurs aisé à mettre en œuvre par les standardisation des scores dans une échelle arbitraire.

Au-delà de ce premier problème lié à l'interprétation des taux de réussite, on peut s'interroger plus en profondeur sur la mesure des acquisitions des élèves. Les évaluations nationales sont

composées d'items regroupés dans des exercices, chaque exercice étant censé évaluer la maîtrise d'une seule compétence. Les exercices appartiennent à des champs (ou domaines) différents, eux-mêmes appartenant à une discipline (mathématiques ou français). A titre d'exemple, les épreuves CE2 de 1999 comportaient 91 items en français et 80 items en mathématiques ; le tableau 1 fournit la répartition de ce nombre d'items ainsi que les compétences qui y sont associées dans les différents domaines.

Tableau 1 : Nombre d'items et de compétences dans les différents champs
Evaluations de CE2 de 1999

		Nombre de compétences	Nombre d'items
Français	Compréhension	6	41
	Outils de la langue	7	40
	Production d'écrits	2	10
Mathématiques	Travaux géométriques	8	17
	Mesure	8	22
	Travaux numériques	6	32
	Résolution de problèmes	5	9

On observe un double déséquilibre dans cette répartition. En premier lieu, on note que les différents champs de l'évaluation ne sont pas représentés avec la même intensité dans les épreuves. Une quarantaine d'items évaluent la compréhension et les outils de la langue alors que seulement 9 items mesurent la résolution de problèmes et 10 items ont trait à la production d'écrits. Cette disproportion pourrait se justifier par la place plus ou moins importante accordée dans les programmes scolaires à ces différents domaines d'acquisition, mais aucun document émanent du Ministère indique que les épreuves ont été élaborées en choisissant le nombre d'items selon une technique de stratification des contenus d'enseignement (Laveault, Grégoire, 1997).

En second lieu, on relève un fort déséquilibre entre le nombre d'items et le nombre de compétences. Ainsi, les 6 compétences de compréhension sont mesurées par 41 items (soit en moyenne un peu moins de 7 items pour une compétence) alors que les 5 compétences en résolution de problèmes sont évaluées par seulement 9 items (soit en moyenne moins de 2 items pour une compétence). L'échelle utilisée pour apprécier la réussite des élèves est donc très variable d'une compétence à l'autre selon le nombre d'items retenus. Pour prendre des exemples extrêmes se rapportant toujours à l'évaluation CE2 de 1999, la compétence «*comprendre un texte…*» est évaluée dans 12 items alors que la compétence «*construire un figure simple sur un quadrillage…*» ne concerne qu'un seul item. Selon le cas, la graduation de la réussite est très fluctuante, allant de la simple dichotomie « échec / réussite » (score de 0 ou 1) à une échelle en 12 scores. Certaines compétences sont donc mesurées dans des situations plus nombreuses et plus variées que d'autres et on peut douter de la précision de la

mesure pour les compétences qui sont évaluées à travers un très petit nombre d'items. L'institution s'interroge elle aussi sur la pertinence des protocoles ne faisant intervenir que très peu d'items pour mesurer une compétence (M.E.N., 2000). Ces remarques nous interrogent sur la validité des épreuves au sens psychométrique du terme et le constat de la maîtrise d'un bon nombre de compétences sera davantage le fait d'une inférence à partir de situations d'évaluation forcément limitées. Cela pose d'autant plus problème s'il s'agit de compétences qui se manifestent principalement dans des situations complexes (Gérard, 2005).

La question du seuil de réussite des compétences se pose également. Il serait possible de fixer un seuil de réussite commun à toutes les compétences en adoptant les principes en usage (75% est généralement le seuil retenu), mais les dissymétries entre les différentes échelles rendent cette solution trop imparfaite, notamment pour les compétences qui comportent très peu d'items. Une solution consisterait alors à avoir recours à des juges ou à des experts (des professionnels de terrain compétents dans le domaine) qui définiraient eux-mêmes et individuellement les critères d'acceptation de la performance. Plusieurs techniques statistiques peuvent ensuite être mobilisées pour exploiter les différents jugements et définir un score de réussite correspond à chaque compétence (Laveault, Grégoire, 1997). Cette méthode, même si elle présente des intérêts évidents, sans doute du fait de sa lourdeur de mise en œuvre, n'a pas été retenue par les concepteurs des évaluations nationales. Cette question du seuil de réussite reste donc entière quand on souhaite savoir si une compétence est maîtrisée ou non par un élève.

Indépendamment des points relatifs à la mesure, les évaluations nationales peuvent donner lieu à des interrogations en ce qui concerne la définition même des compétences évaluées. Ces compétences sont définies par les commissions chargées de l'élaboration des épreuves et leur nombre peut varier d'une année sur l'autre. Les évaluations CE2 auxquelles nous nous référons dans cette recherche comportent 15 compétences en français et 27 en mathématiques (tableau 2). Les résultats des évaluations nationales sont exploités généralement à deux niveaux dans deux perspectives différentes. Le premier niveau, utilisé dans une perspective de pilotage, concerne une analyse globale des scores totaux, par discipline ou par champ. Le second niveau, utilisé dans une optique pédagogique, se centre sur la réussite ou l'échec aux compétences. Ces exploitations classiques des évaluations nationales fournissent une image incomplète des apprentissages effectivement réalisés par les élèves, soit en se situant à un niveau trop global qui ne permet pas d'identifier quelles dimensions des apprentissages sont les plus centraux, soit en dissociant les compétences les unes des autres sans examiner les liens qu'elles entretiennent entre elles. Quand on adopte au contraire une démarche basée sur la recherche des liens entre les multiples facettes des acquisitions des élèves, on soulève un certain nombre d'interrogations sur la pertinence de la définition institutionnelle des compétences et de leur rattachement à des domaines d'acquisition[6].

[6] Les analyses qui vont suivre sont basées sur l'étude d'un échantillon de 671 élèves pour lesquels nous disposons des résultats détaillés aux évaluations de CE2 (1999) ainsi que d'informations socio-démographiques et scolaires.

Tableau 2 : Liste des compétences des évaluations CE2 de 1999

Compétences de français
Savoir utiliser une bibliothèque ou une BCD, repérer et identifier des ouvrages appartenant à des domaines textuels différents.
Distinguer des textes différents (récits, documents,…) en utilisant des indices extérieurs au texte.
Comprendre et savoir appliquer les consignes courantes du travail scolaire.
Comprendre un texte et montrer qu'on l'a compris.
Reconstituer la chronologie des évènements dans des textes de statuts variés
Se représenter les relations spatiales et les lieux évoqués dans des textes de statuts variés.
Savoir produire un écrit bref appartenant à un type de texte défini et dans une situation de communication déterminée en s'appuyant sur une consigne, une sollicitation, des documents de référence.
Savoir produire, de manière plus autonome, un texte cohérent dans le cadre d'une situation de communication déterminée.
Repérer les usages typographiques courants et se situer dans l'espace graphique d'une page imprimée.
Reconnaître les mots écrits.
Mieux comprendre les mots d'un texte.
Copier un mot, une phrase, un texte, en respectant les exigences de présentation et en écrivant lisiblement.
Ecrire sous la dictée des mots courants, de petites phrases ou de petits textes.
Transformer un texte en appliquant des règles simples.
Identifier certains aspects d'un texte.
Compétences de mathématiques
Se repérer et se déplacer dans un quadrillage.
Utiliser les instruments de dessin pour achever un tracé.
Construire une figure simple sur un quadrillage en utilisant des propriétés de la figure.
Compléter par pliage (symétrie) une figure dessinée sur un quadrillage.
Associer une figure à une description.
Compléter un plan à partir de consignes.
Se repérer dans l'espace.
Tracer une figure à partir de consignes.
Se repérer dans la journée
Mesurer ou tracer un segment de longueur donnée.
Ranger des longueurs.
Associer une unité usuelle à une grandeur
Utiliser le calendrier.
Comparer des distances.
Résoudre un problème faisant intervenir une grandeur.
Choisir l'unité la mieux adaptée à un mesurage.
Effectuer des additions, posées, en ligne ou à poser.
Calculer des produits et des différences (calcul exact ou approché)
Calculer mentalement (calcul exact ou approché)
Transcrire en lettres des nombres écrits en chiffres et inversement.
Ranger des nombres
Comparer des nombres donnés sous formes diverses.
Lire et/ou remplir un tableau à double entrée.
Exploiter un document « brut »
Résoudre un problème à une opération.
Résoudre une situation de partage ou de groupement.
Effectuer un choix et en formuler la justification.

Une première remarque concerne la distinction entre les deux disciplines. Si en français, toutes les compétences sont effectivement reliées positivement les unes aux autres (les coefficients de corrélation sont tous significatifs au seuil de 1%), on relève une plus bien

grande hétérogénéité des situations en mathématiques puisque certaines compétences n'entretiennent aucune liaison statistique entre elles. Par exemple, les compétences «*associer une unité usuelle à une grandeur*» et « *lire et/ou remplir un tableau à double entrée* » ne sont pas corrélées (r = - 0,01, non significatif). On note également que les liens statistiques entre certaines compétences issues de disciplines différentes peuvent être plus intenses que ceux existants entre des compétences originaires d'une même discipline. Par exemple, *« calculer mentalement »* (mathématiques) et « *écrire sous la dictée des mots courants, de petites phrases ou de petits texte »* affichent un coefficient de corrélation élevé (r = +0,41, significatif à .01). En outre, les corrélations les plus fortes ne concernent pas obligatoirement des compétences appartenant à un même champ. Ainsi, en français les compétences : *« comprendre un texte et montrer qu'on l'a compris»* et «*identifier certains aspects d'un texte* » (corrélation de +0,45) appartiennent respectivement aux champs de la compréhension et des outils de la langue. En mathématiques les compétences «*calculer mentalement* » et « *résoudre un problème à une opération* » (corrélation de +0,50) se rattachent respectivement aux travaux numériques et aux problèmes numériques. La distinction entre les différents domaines d'apprentissage ne va donc pas de soi et laisse la place à un certain arbitraire quand on observe les relations statistiques entre les scores obtenus par les élèves aux différentes compétences.

Une seconde remarque, plus fondamentale, s'interroge sur la pertinence de l'identification de certaines compétences[7]. On observe globalement que les items appartenant à un même exercice, donc dans la grande majorité des items étant censés mesurer la même compétence, ne sont pas plus corrélés entre eux que les items appartenant à des exercices différents. Certains exercices présentent même des corrélations nulles entre les items les composant, c'est le cas pour l'exercice 7 de français qui évalue la compétence « *comprendre et savoir appliquer les consignes courantes du travail scolaire* » à l'aide de 5 items (tableau 3).

Tableau 3 : Corrélations entre les items de l'exercice 7 de français
Evaluations de CE2 de 1999

	Item 28	Item 29	Item 30	Item 31
Item 29	+0,02 n.s.			
Item 30	+0,16***	+0,00 n.s.		
Item 31	-0,01 n.s.	+0,04 n.s.	+0,00 n.s.	
Item 32	+0,17***	+0,09**	+0,14***	+0,03 n.s.

n.s. : non significatif, ** : significatif au seuil de 5%, *** : significatif au seuil de 1%

Il n'existe ainsi aucune relation statistique entre les items 28 et 29, 28 et 31, 29 et 30, 29 et 31, 30 et 31, 31 et 32 alors que les items 30 et 32, 30 et 28, 28 et 32, 29 et 32 sont positivement corrélés. Cette situation peut s'expliquer quand on se reporte au contenu des différents items proposés dans cet exercice. Si ces tâches mesurent bien toutes une aptitude ayant trait à un

[7] Les analyses portent sur les corrélations entre l'ensemble des items présents dans les épreuves.

comportement scolaire, en l'occurrence «*savoir appliquer des consignes*», elles sont en revanche d'une nature différente d'un item à l'autre. Les items 28, 30 et 32 mobilisent une activité de discrimination visuelle semblable (identifier une ou deux lettres dans un mot simple) qui explique la corrélation relevée entre ces trois items. La réussite de l'item 29 nécessite que l'élève sache se repérer dans l'espace (il s'agit de mettre une croix entre un carré et au-dessus d'une ligne) ; cette activité est donc assez éloignée de la précédente ce qui peut expliquer l'absence de corrélation entre l'item 29 et les items 28 et 30.

De façon complémentaire, les analyses suggèrent que certains items proposés dans les épreuves pourraient s'associer différemment pour mesurer la maîtrise des compétences. Par exemple, la compétence visée par l'exercice 24 de mathématiques «*exploiter un document brut* » et mesurée par les items (73 et 74) pourraient aussi bien être intégrée dans la mesure d'une autre compétence, c'est en tout cas ce que nous préconise notre approche empirique. Ainsi, l'analyse des corrélations entre items indique une corrélation positive et significative entre l'item 73 et 35 ($r = + 0,10$, significatif au seuil de 1%), ce dernier item mesurant à lui seul la compétence *«résoudre un problème faisant intervenir une grandeur* », compétence intuitivement proche des habilités requises pour répondre à la question de l'item 73 qui implique une comparaison de grandeurs numériques.

Nous mettons ici l'accent sur la difficulté à saisir concrètement la notion de compétence dans le domaine de l'éducation et surtout à la définir de façon univoque. Il est toutefois possible aujourd'hui de retenir une définition relativement consensuelle de la compétence, celle-ci pouvant être considérée comme une aptitude à la mise en œuvre d'un ensemble organisé de savoirs, de savoirs faire d'attitudes permettant d'accomplir un certain nombre de tâches (Crahay, 2006). Une compétence intègre une mise en relation de connaissances, qu'elle s'applique à une famille de situations et qu'elle est orientée vers une finalité (Gillet, 1991). Plus généralement, la notion de compétence renvoie à un ensemble intégré de connaissances susceptibles d'être mobilisées pour accomplir des tâches (Crahay, 2006). En ce qui nous concerne, nous nous situons à la marge du débat théorique relatif à cette question en nous contenant de faire émerger les diverses compétences mise en œuvre par les élèves à partir de situations d'évaluation. Le terme de compétence intègre ici dans sa définition les différentes capacités, aptitudes, connaissances et attitudes qui mobilisent chez les individus des ressources cognitives pour réaliser une tâche.

Le choix de situations d'évaluation des compétences relève d'un certain arbitraire auquel les épreuves nationales n'échappent pas. Les remarques précédentes montrent bien que les apprentissages des élèves s'inscrivent dans une logique d'interdépendance dont l'approche disciplinaire ne peut rendre compte, mais aussi qu'une définition a priori des compétences, n'est pas adéquate. Pour progresser dans ce domaine et permettre une identification empirique des compétences, nous avons adopté une méthode basée sur l'analyse des corrélations entre les résultats aux items. Ce seront les corrélations les plus fortes entre items qui nous permettrons d'identifier des compétences en regroupant les items qui mesurent des aptitudes

proches chez les élèves. Notre démarche s'inspire globalement des procédures utilisées pour les validations empiriques a posteriori des épreuves d'évaluation et plus précisément les validations empiriques internes (De Ketele, Gérard, 2005).

L'identification des compétences selon une approche empirique

Cette approche au niveau des items permet tout d'abord d'utiliser une échelle de mesure commune puisque tous les items présentent le même barème de cotation : 0 pour une réponse erronée, 1 pour la réponse attendue. Le second avantage, découlant nécessairement du précédent, est que la question du seuil de réussite ne se pose plus et que celui-ci est défini objectivement par la réussite ou l'échec, sans possibilité de situations intermédiaires[8]. L'analyse des relations entre items passe en premier lieu par le calcul de corrélations. Compte tenu du nombre d'items (171 items pour le CE2 en 1999), la matrice de corrélations produit 14535 coefficients. L'examen des corrélations permet d'identifier une grande variété des situations, certains exercices pouvant être perçus comme très cohérents dans la mesure où les items qui les composent tendent à mesurer la même chose (corrélations fortes) alors que d'autres exercices sont composés d'items beaucoup plus indépendants les uns des autres (corrélations faibles ou nulles). Cela confirme l'observation précédente selon laquelle certaines compétences seraient mesurées de façon imparfaite.

Une difficulté liée à notre approche tient à la sélection des corrélations inter-items en vue de l'étude approfondie des relations statistiques. Le nombre très élevé de corrélations nous contraint à effectuer des choix et nous avons sélectionné uniquement les relations les plus fortes parmi l'ensemble. Le seuil sur lequel nous nous basons pour effectuer ce choix revêt forcément un caractère arbitraire. Nous avons décidé de ne retenir que les corrélations supérieures à +0,20 : elles sont au nombre de 317 ; ce seuil est en grande partie déterminé par les possibilités d'analyse qui tiennent d'une part aux capacités de traitement du logiciel (il existe une contrainte d'arbitrage entre le nombre de corrélations et le nombre d'observations) et d'autre part aux possibilités d'interprétation des résultats ; celles-ci seraient peu opérationnelles si tous les liens entre items, notamment les plus faibles étaient examinés en détail. On signalera que les 317 corrélations représentent la grande majorité des items (141 items sur les 171).

Afin d'organiser l'analyse des relations entre items, une phase préparatoire est nécessaire pour étudier individuellement chaque corrélation et dresser ainsi une cartographie de l'ensemble des situations. Le principe de cette étape préalable est d'identifier des blocs de relations au sein desquels on retrouve le plus souvent les mêmes items. Cette procédure est systématique puisque pour chaque item, on identifie tous les autres items qui lui sont associés dans les

[8] On mentionnera que cette dichotomie (échec / réussite) n'est pas adaptée à une interprétation diagnostique des évaluations nationales et que les barèmes de cotation de certains items comportent à l'origine d'autres paliers qui permettent une analyse des erreurs des élèves.

corrélations. Au terme de cette phase, on aboutit à un ensemble de 29 regroupements comportant chacun un nombre d'items varié. La phase suivante consiste à étudier chacun de ces regroupements de corrélations entre items et d'isoler les compétences à l'aide d'analyses en variables latentes.

L'objectif est ici d'analyser les liaisons statistiques entre items de façon à mettre à jour des variables latentes pouvant être interprétées comme des compétences, des aptitudes ou des capacités mobilisées par les élèves dans les évaluations nationales. Les modèles ont été estimés avec le logiciel LISREL[9] ; ils permettent de rendre explicite une dimension latente des phénomènes étudiés en postulant l'existence de variables inobservées (ou inobservables) qui rendent compte des relations ente variables observées, en l'occurrence ici les résultats aux items. L'adoption de cette démarche contraint à la formulation d'un modèle théorique postulé « a priori ». Elle vise en fait *« à vérifier la validité d'une théorie causale préalablement formulée en testant l'ajustement d'un modèle mathématique à des données observées»* (Aish-Van Vaerenbergh, 1997). Le logiciel LISREL permet d'estimer des modèles de mesure en établissant des relations entre les variables latentes (compétences supposées) et leurs indicateurs (les items). Pour illustrer cette démarche, nous présentons les estimations d'un modèle de mesure faisant intervenir un regroupement de 8 items de français.

Sur ces 8 items, 5 items (items 62, 63, 64, 66, 67) appartiennent au même exercice et sont censés mesurer la même compétence «*écrire sous la dictée des mots courants, de petites phrases ou de petits textes*»[10]. La réussite à la dictée dépend de plusieurs éléments : de compétences orthographiques, de la capacité à se remémorer la graphie des mots préalablement observés et, dans une certaine mesure, d'une capacité d'attention. L'item 60 est supposé rendre compte de la compétence « *copier un mot, une phrase, un texte…* » mais cet item n'évalue qu'un aspect de la production des élèves : le respect de la ponctuation (majuscules et points). L'item 12 fait partie d'un exercice dans lequel on demande aux élèves de repérer des types d'écrits à partir d'un extrait de livres différents ; la compétence visée est *«distinguer des textes différents en utilisant des indices extérieurs au texte »*. Enfin, l'item 16 vise la compétence : « *comprendre un texte et montrer qu'on l'a compris* ». Un texte est proposé aux élèves et ceux-ci doivent répondre à des questions concernant ce texte.

Une première étape a consisté à définir un variable latente qui détermine les résultats des élèves à tous ces items. Les statistiques fournies avec ce premier modèle de mesure indiquent que celui-ci peut-être amélioré et le logiciel LISREL suggère qu'une nouvelle variable latente peut être introduite dans l'analyse en isolant les items 64 et 66. Un second modèle a donc été

[9] LISREL :LInear Structurel RELationship.

[10] Dans cet exercice, il est demandé aux élèves d'écrire la phrase suivante, dictée par l'enseignant : «*pendant la récréation, les garçons et les filles jouent aux billes.* ». Cette phrase a été auparavant écrite au tableau, les élèves étant invités à mémoriser l'orthographe des mots. La phrase a été ensuite effacée et un exercice différent a été proposé aux élèves ; ce n'est qu'après cet exercice que la phrase a fait l'objet d'une dictée.

estimé en définissant deux variables latentes La première variable « *comp1* » est déterminée par les items 12, 16, 60, 62, 63, la seconde « *comp2* » par les items 64 et 66. Le graphique suivant présente le diagramme causal associé au modèle de mesure tel que le produit le logiciel. Il est d'usage de représenter le modèle de mesure par un diagramme dans lequel les variables latentes sont symbolisées par des ellipses et les variables observées (ou indicateurs de ces variables latentes) sont matérialisées par des rectangles. Les flèches en traits pleins matérialisent l'intensité des relations qui lient chacun des indicateurs (items) à la variable latente (compétence), un coefficient de régression (et son degré de significativité) pour chaque indicateur fournit une indication sur la validité du modèle[11] (Morlaix, 2002).

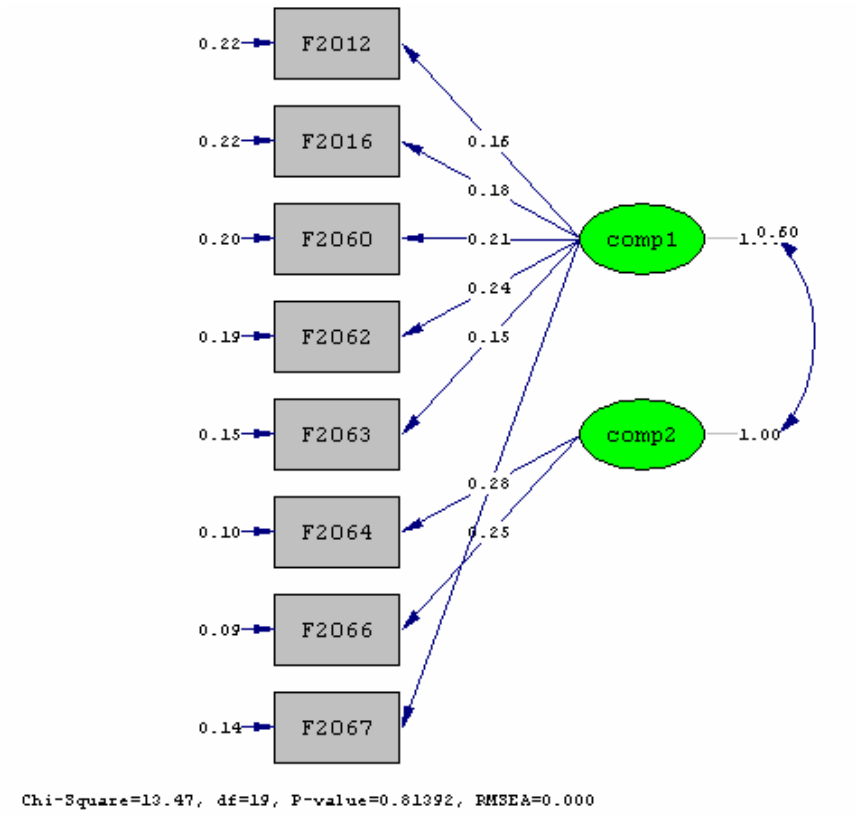

Figure 2 : Modèle de mesure établi à partir d'un regroupement de 8 items
Evaluations CE2, 1999

On voit apparaître sur le schéma la corrélation entre les deux variables latentes (coefficient de +0,60), cela témoigne de la proximité des deux compétences. Toutes les relations matérialisées sur le schéma sont significatives. Il semble donc bien que les items 64 et 66 mesurent une compétence spécifique qui rende compte des aptitudes des élèves à

[11] Le logiciel LISREL fournit des informations sur la qualité du modèle grâce à plusieurs indicateurs statistiques : χ^2 (khi carré) mesure la distance séparant les deux matrices de corrélations, celle théorique et celle portant sur les corrélations estimées, la π value (P value) mesure la probabilité d'obtenir la valeur du χ^2 correcte, l'indice RMSEA (Root Mean Square Error of Approximation) se rapporte à la moyenne des résidus du modèle et à leur significativité.

orthographier correctement des noms écrits au pluriel (« *garçons* » et «*filles* »). Les 6 autres items (12, 16, 60, 62, 63 et 67) mesureraient pour leur part une compétence différente. Il est difficile d'interpréter rigoureusement ces relations sans avoir recours à des éléments théoriques. En effet, les résultats des élèves aux items de ce premier regroupement semblent dépendre de plusieurs aptitudes : mémorielles, attentionnelles, orthographiques…. En tout cas, nous sommes sans doute loin de la seule compétence orthographique formulée dans les évaluations nationales. Les modèles de mesure estimés sur la totalité des regroupements d'items ont ainsi donné lieu à l'identification de 63 variables latentes de composition variée (Morlaix, Suchaut, 2006) qui forment une recomposition des compétences de l'évaluation CE2. Parmi les 63 variables mises à jour, 27 d'entre elles (soit 43%) correspondent, souvent de façon partielle, à des regroupements d'items déjà présents dans les épreuves nationales. Le recouvrement avec les compétences des épreuves est très imparfait puisque seules 5 variables correspondent exactement à des compétences figurant dans les évaluations (tableau 2).

L'étape suivante a consisté à identifier des ensembles de compétences plus consistants sachant que certaines variables latentes peuvent, du fait de la méthode employée, être très proches les unes des autres[12]. Nous avons donc conduit de nouvelles analyses qui s'appuient sur la matrice de corrélations entre les différentes variables latentes. Les coefficients de corrélation présentent des valeurs allant de 0 à +0,80 (ce qui est un chiffre très élevé). Pour isoler les compétences majeures parmi l'ensemble des variables, nous avons sélectionné les corrélations les plus importantes (supérieures à +0,70). Cela permet de dégager trois grands groupes de compétences qui sont représentées sur le graphique suivant (les chiffres correspondent aux numéros associés aux variables latentes).

[12] En effet, les modèles de mesure ont été estimés séparément sur chacun des regroupements d'items.

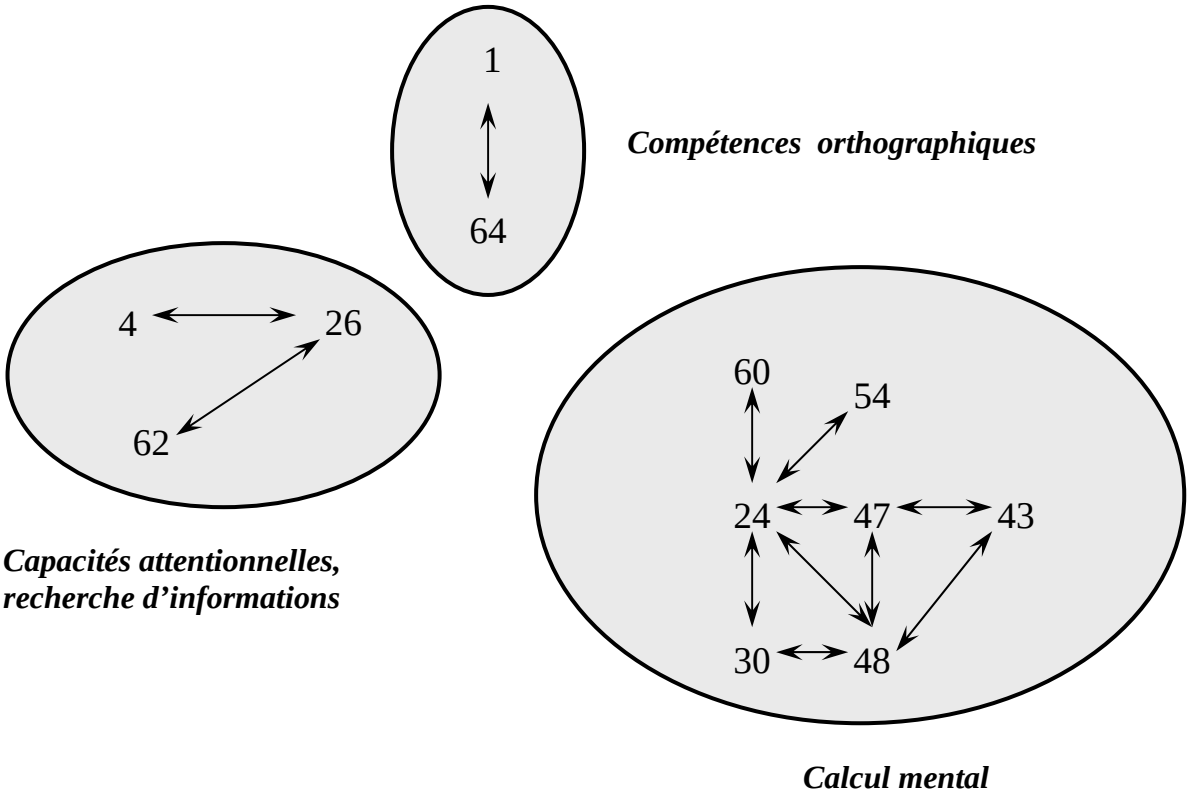

Figure 3 : Ensembles de compétences des évaluations CE2 (1999)

Un premier ensemble est composé des variables « *comp1* » et « *comp64* ». Il s'agit ici clairement de compétences orthographiques qui sont mesurées avec ces deux variables. En effet, la presque totalité des items rendant compte de ces variables latentes concernent deux exercices de dictée. Le deuxième ensemble regroupe les variables « *comp4* », « *comp26* » et « *comp62* ». Cet ensemble d'apparence disparate prend sens quand on l'examine sous l'angle de la psychologie cognitive[13]. Autant les items de français que ceux de mathématiques présents dans ce regroupement évaluent la capacité que peuvent avoir les élèves à rechercher de l'information plus ou moins complexe à partir de supports divers (texte, mots, calendrier, emploi du temps, plan, énoncé de problème). Ce sont donc les capacités attentionnelles des élèves qui sont mises à contribution pour la maîtrise de cette compétence globale. Le troisième ensemble établi sur les corrélations les plus fortes, regroupe 7 variables latentes. Le lien commun entre ces variables est également clair puisque les items de calcul mental interviennent systématiquement pour chacune d'entre elles et 21 fois au total dans ce regroupement.

[13] Des collègues psychologues du L.EAD (Université de Bourgogne), Pierre Barrouillet et Valérie Camos, ont été associés à cette partie de la recherche, ils ont notamment contribué à donner une signification à certaines variables latentes identifiées par les modèles LISREL.

Les acquisitions des élèves à l'entrée au CE2 s'organisent principalement autour de ces trois compétences qui ne sont pas de même nature. Si l'acquisition des compétences orthographiques dépend principalement d'un enseignement systématique, les deux autres compétences sont davantage associées à des processus plus complexes qui interviennent de façon transversale dans de nombreuses situations d'apprentissage. On devrait s'attendre à ce que ces compétences majeures contribuent fortement à l'explication des différences de réussite entre élèves à l'entrée au CE2. Pour vérifier cela, une régression « pas à pas » a été estimée avec comme variable dépendante le score global de CE2 et comme variables explicatives l'ensemble des compétences mises à jour précédemment. On constate alors que les trois compétences les plus prédictives (« *comp48* », « *comp4* », « *comp64* ») appartiennent chacune à un des groupes identifiés auparavant. Ces trois compétences expliquent à elles seules 82% de la variance du score global, elles sont bien au centre des acquisitions des élèves à l'entrée au CE2. Les habiletés en calcul mental, les capacités attentionnelles (et à trouver rapidement des informations dans des supports variés), la maîtrise de l'orthographe structurent ainsi fortement les résultats des élèves au début du cycle III.

Nouvelles perspectives d'utilisation des évaluations nationales

Nous insisterons pour conclure sur les pistes qu'ouvre notre démarche d'analyse. Nous avons mis l'accent dans ce texte sur le fait que la définition des compétences des évaluations nationales française était contestable dans la mesure où un même exercice peut mobiliser des compétences parfois nombreuses et diverses alors que la catégorisation « officielle » se limite à l'identification d'une seule compétence pour cet exercice. Notre méthode, basée sur une analyse statistique des relations entre les résultats obtenus aux items, permet au contraire d'associer des situations d'évaluation mobilisées dans des exercices distincts et de faire émerger des compétences capables de mieux cerner la structure des apprentissages des élèves.

Parmi la soixantaine de compétences identifiées, certaines ont un statut particulier dans le sens où elles rendent compte, à elles seules des différences d'acquisition des élèves à l'entrée au cycle III. Ces compétences concernent trois domaines : l'orthographe, l'attention et la recherche d'information, le calcul mental[14]. On rejoint ici un constat réalisé par la D.E.P. sur la base du panel 1997[15] selon lequel l'épreuve d'attention partagée administrée à l'entrée au CP s'est révélée fortement corrélée avec la réussite générale à l'évaluation des CE2, deux années plus tard ; « l*a capacité d'attention pourrait donc bien constituer un facteur majeur de la réussite* » (M.E.N., 2000). Un premier intérêt à notre travail est donc de pouvoir proposer des épreuves d'évaluation réduites, concentrées sur les items mesurant les compétences

[14] . Ces deux derniers éléments pouvant intervenir dans des situations d'apprentissage très variées qui dépassent largement l'approche disciplinaire.

[15] Elèves entrés au CP en 1997.

dégagées avec notre approche. Cela pourrait être très utile pour identifier les élèves en difficulté à l'entrée en CE2 ; cette identification pouvant être réalisée au niveau de la classe dans une perspective diagnostique, au niveau de l'école dans une perspective de pilotage.

Notre recherche a aussi permis d'approfondir les relations entre les compétences en mettant en évidence une structure hiérarchique (Morlaix, Suchaut, 2006), certaines compétences étant indispensables à l'acquisition d'autres compétences. Sans entrer ici dans le détail de cette structure, signalons que la compétence de CE2 qui se trouve au sommet de la hiérarchie correspond à la maîtrise de la soustraction qui apparaît être un processus nécessitant de la part des élèves diverses capacités et habiletés préalables. La compétence se situant au plus bas niveau de la structure hiérarchique concerne des exercices qui visent la compréhension de consignes simples (alors que les évaluations nationales ont répertorié ces exercices dans le domaine de la mesure). En prenant en compte la structure hiérarchique d'apprentissage effective des élèves, c'est-à-dire en respectant le fait, trivial à présent, que certains apprentissages ne peuvent se réaliser que si d'autres déjà maîtrisés (Bloom, 1979), les résultats évoqués précédemment débouchent naturellement sur un questionnement lié à la planification des activités d'enseignement, et plus largement au contenu des programmes scolaires. Il est aussi essentiel de s'interroger aussi sur le temps que l'on doit consacré à l'acquisition des différentes compétences, sachant que le temps alloué est très lié aux progrès des élèves et que celui-ci s'inscrit, pour l'enseignant, dans une logique d'arbitrage entre des activités diverses et concurrentielles (Morlaix, 2000 ; Suchaut, 1996).

Les évaluations nationales ont sans aucun doute contribué à faire évoluer positivement le système éducatif français en dotant l'institution d'outils pertinents utilisables à différents échelons. Ces outils, malgré leurs qualités initiales, demandent une certaine prudence quand on souhaite interpréter correctement les résultats des élèves, surtout dans une perspective d'évaluation diagnostique. Plus de 15 ans après leur mise en place, on peut souhaiter une évolution significative au niveau de la définition des compétences censées être évaluées à travers les situations proposées dans les épreuves. Les évaluations nationales sont dans un sens une bonne illustration de la vision simplificatrice et réductionniste que peut avoir de nos jours la notion de compétence dans le domaine de l'éducation. Peut-être que notre approche empirique contribuera modestement à éclairer « *la caverne d'Ali Baba conceptuelle*» (Crahay, 2006) que représente la notion de compétence, c'est en tous cas ce que l'on a essayé de faire apparaître dans notre démonstration. Il n'en reste pas moins que des recherches complémentaires doivent être conduites pour tenter de mieux cerner ce que les situations d'évaluation mesurent réellement chez les élèves, quels concepts sont manipulés, quels processus d'apprentissage sont mobilisés.

Pour conclure, cette recherche a montré que les évaluations nationales pouvaient constituer une base empirique d'analyse très riche, à partir du moment où l'on ne se contente pas des regroupements d'items déjà proposés dans les documents d'accompagnement. L'utilisation d'analyses statistiques en variables latentes a permis d'identifier des compétences et des

habiletés différentes de celles visées par les évaluations diagnostiques qui peuvent donner lieu à une réflexion nouvelle de la part des acteurs. La hiérarchisation des compétences peut quant à elle fournir des indications quant à la définition des curricula, en priorisant certaines compétences considérées comme centrales et se développant en amont du cycle III.

Bibliographie

Aish-Van Vaerenbergh, AM.(1997), « Modèles statistiques et inférences causales : analyse de structures de covariances avec LISREL ». In *Faut-il chercher aux causes une raison ? L'explication causale en sciences humaines.* Aish-Van Vaerenbergh, AM. et al. Librairie philosophique Vrin, pp106-130.

Bloom, B.S. (1979) *Caractéristiques individuelles et apprentissages scolaires*, Bruxelles, Labor, Paris, Nathan.

Crahay M. (2006), Dangers, incertitudes et incomplétude de la logique de la compétence en éducation. *Revue française de pédagogie*, N°154, pp. 97-110.

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*, À paraître.

Gerard F.M. (2005), *L'évaluation des compétences à travers des situations complexes*. Actes du Colloque de l'Admee-Europe, IUFM Champagne-Ardenne, Reims, 24-26 octobre 2005.

Gillet P. [éd.] (1991), *Construire la formation : outils pour les enseignants et les formateurs*. Paris, ESF.

Laveault D., Grégoire J. (1997), *Introduction aux théories des tests en sciences humaines*. De Boeck Université.336 p.

M.E.N. / Direction de l'Enseignement Scolaire (2000). *L'exploitation de l'évaluation nationale en CE2 : la lecture. Programme national de pilotage : actes du séminaire national*. C.R.D.P. de l'Académie de Versailles.

Morlaix S. (2000). Rechercher une meilleure répartition du temps scolaire en primaire pour favoriser la réussite au collège. *Revue française de pédagogie.* N°130. pp.121-131

Morlaix S. (2002). Intérêts et apports de l'analyse des variables latentes pour le chercheur en sciences sociales : exemple d'application à l'économie de l'éducation. *Orientation scolaire et professionnelle*, vol 31., N°1. pp.117-138.

Morlaix S., Suchaut B. (2006). *Evolution et structure des compétences des élèves à l'école élémentaire et au collège. Une analyse empirique des évaluations nationales*. Rapport pour l'UNSA-Education. 196 p. et annexes. Dijon, IREDU.

Suchaut B. (1996), « La gestion du temps à l'école maternelle et primaire : diversité des pratiques et effets sur les acquisitions des élèves ». *L'année de la recherche en sciences de l'éducation*, pp. 123-153.