AVANT-PROPOS

Thierry Rocher

Le thème de l'évaluation est un sujet récurrent de débat dans le domaine de l'éducation. Encore est-il important de bien préciser ce dont il s'agit, tant les formes d'évaluations sont multiples : évaluation des élèves dans la classe par les enseignants, évaluation certificative, formative, évaluation diagnostique, auto-évaluation, etc. Très clairement, ce double numéro spécial de la revue *Éducation & formations* porte sur les « évaluations standardisées » des acquis. Organisées le plus souvent sous forme de « programmes », ces évaluations se distinguent par le fait qu'elles ambitionnent de fournir une mesure objective, scientifique, des acquis des élèves, la plus indépendante possible des conditions d'observation, de passation, de correction. En ce sens, elles sont « standardisées ». En outre, ces programmes d'évaluations apparaissent comme des évaluations externes, et non internes comme par exemple le contrôle continu ou l'évaluation formative. Si leurs objectifs peuvent sensiblement varier, ces évaluations ont en commun de pouvoir rendre compte des acquis des élèves au-delà du niveau individuel et, en particulier, d'apprécier les résultats du système éducatif pris dans sa globalité. Leurs modalités de conception sont relativement partagées, car tendues vers l'objectivité de la mesure et la comparabilité des résultats, entre groupes d'élèves, dans le temps, etc. Bien que leurs fondements méthodologiques restent assez méconnus en France, les résultats de ces évaluations sont largement diffusés et commentés dans le champ médiatique et politique. Ce numéro spécial est donc important, car il donne à voir les différentes facettes des évaluations standardisées des acquis des élèves, de leur conception à l'exploitation des résultats. Les cadres et les formateurs en particulier en verront tout l'intérêt.

Les treize articles de ce numéro sont organisés en trois parties. La première comprend quatre articles et livre un panorama général des problématiques, des usages et des concepts spécifiques aux évaluations standardisées. La seconde regroupe quatre articles portant plus spécifiquement sur des aspects méthodologiques. Enfin, cinq articles constituent la dernière partie qui rend compte de différents résultats issus de ces évaluations.

Pour ouvrir ce numéro, Bruno Trosseille et Thierry Rocher (*Les évaluations standardisées des élèves – perspective historique*) proposent une analyse historique du développement des programmes d'évaluations standardisées en France. Différentes périodes se sont succédé depuis leur naissance. Ces périodes ont correspondu à des objectifs spécifiques assignés à ces évaluations, objectifs qui doivent être clairement distingués sous peine de créer des confusions néfastes. Les fonctions de ces évaluations (servent-elles à observer, à sélectionner, à diagnostiquer, à piloter ?) ainsi que leurs usages (servent-elles aux élèves eux-mêmes, aux enseignants, aux responsables politiques ?) sont des questions fondamentales. Les auteurs

montrent à travers l'histoire récente que la poursuite d'objectifs différents pour une même évaluation peut conduire à une utilisation inefficace, voire dévoyée de leurs résultats.

La question des usages des résultats des évaluations est indissociable de la question de la mesure des compétences. Les problématiques de mesure relèvent du domaine de la psychométrie, domaine spécifique très peu investi en France, que ce soit dans le monde académique ou dans celui de la statistique publique. L'article de Thierry Rocher (*Mesure des compétences – Méthodes psychométriques utilisées dans le cadre des évaluations des élèves*) vise précisément à donner aux lecteurs de la revue un aperçu des procédures psychométriques, avec un souci pédagogique, sans toutefois sacrifier la rigueur statistique. L'article montre en particulier que tout instrument de mesure est un « construit », reposant sur un ensemble d'hypothèses et de méthodes spécifiques. Une illustration de ces méthodes est donnée à travers la présentation des procédures psychométriques employées dans le cadre du programme d'évaluations standardisées Cedre (Cycle des évaluations disciplinaires réalisées sur échantillons).

Ces méthodes sont globalement partagées au niveau international. L'ingénierie à mettre en place pour assurer la qualité et la comparabilité des données relève d'une forme de technicité bien décrite dans un autre contexte par Christophe Dierendonck, Amina Kafaï, Antoine Fischbach, Romain Martin et Sonja Ugen (*Les épreuves standardisées – Élément-clé du pilotage du système éducatif luxembourgeois*). Le programme d'évaluations standardisées mis en place au Luxembourg se distingue par ses objectifs, dans la mesure où il vise à aider au pilotage des établissements scolaires. Les établissements sont en effet accompagnés dans la rédaction d'un plan d'action reposant en partie sur l'analyse des données tirées des évaluations standardisées. Tout le défi est là, dépasser la méfiance envers un outil pouvant être perçu comme un contrôle, associer les acteurs aux constats et les accompagner dans la définition et la mise en place d'actions ciblées. La taille du Luxembourg favorise sans doute la réalisation de cet exercice difficile, mais précisément son expérience pourrait sans doute inspirer les acteurs du pilotage local en France, souvent demandeurs de dispositifs d'évaluations standardisées.

Pour conclure cette première partie de cadrage général, Fabrice Murat et Thierry Rocher (*L'évaluation des compétences des adultes – Quelles contraintes ? Quelles spécificités ?*) nous invitent à sortir du cadre scolaire et à considérer les évaluations d'adultes. Dans ce domaine, deux traditions statistiques, peu amenées à se côtoyer, doivent converger, à savoir celle des enquêtes-ménage, très maîtrisée par l'Insee, et celle de l'évaluation des compétences, naturellement plus développée à la DEPP. Une perspective de comparaison avec les évaluations des élèves montre les spécificités de l'évaluation des compétences des adultes, notamment le fait d'interroger des individus, chez eux, sur leurs compétences, ce qui peut réveiller de mauvais souvenirs scolaires chez certains d'entre eux. La question de leur implication dans

la situation d'évaluation est alors centrale. Les auteurs détaillent deux dispositifs d'évaluation – l'un national (IVQ), l'autre international (Piaac) – complémentaires, mais qui affichent des différences méthodologiques liées à différents objectifs ainsi qu'aux leçons que la France a pu tirer de précédentes enquêtes sur le sujet.

Les quatres articles suivants portent sur des questions méthodologiques précises : méthodes de sondage, questions de motivation, seuil de maîtrise et passage au numérique.

En France, les programmes d'évaluations standardisées ont pour objectif le pilotage d'ensemble du système éducatif ; ils sont donc principalement réalisés sur échantillons. Dès lors, des questions d'ordre technique se posent. Elles sont abordées par Émilie Garcia, Marion Le Cam et Thierry Rocher (*Méthodes de sondages utilisées dans les programmes d'évaluations des élèves*) : comment tirer un échantillon « représentatif » de la population ? Comment éviter de tirer deux fois la même école pour deux évaluations différentes ? Comment tenir compte des non-réponses ? Comment mesurer l'erreur d'échantillonnage inhérente à cette démarche ? Les auteurs proposent des réponses à travers une présentation pédagogique des méthodes et selon une démarche empirique, reposant sur des simulations. En filigrane, les auteurs montrent tout l'intérêt de tenir compte des informations disponibles en amont sur les élèves ou sur les établissements, pour améliorer la qualité des échantillons et des résultats. Notons avec eux au passage que ces informations sont aujourd'hui inégalement disponibles dans notre système : présentes dans le second degré, quasi-absentes dans le premier degré.

Une fois les élèves sélectionnés, se pose la question de leur participation et de leur implication dans l'activité d'évaluation. De nombreux travaux, notamment issus de la psychologie sociale, ont montré l'importance du contexte de l'évaluation sur les résultats eux-mêmes. En l'occurrence, les programmes d'évaluations standardisés en cours en France ne comportent aucun enjeu pour les élèves. Dans un système largement dominé par la notation, les élèves ne sont pas notés à l'issue de la passation d'une évaluation standardisée, n'ont pas connaissance de leurs résultats individuels et ne reçoivent pas d'incitations financières (à l'inverse de certains pays pour l'évaluation PISA). Dès lors, quel degré d'implication peut-on attendre des élèves ? Comment le mesurer ? Et quel lien peut-on faire entre leur niveau d'implication et leurs résultats ? C'est à ces questions que tentent de répondre Saskia Keskpaik et Thierry Rocher (*La motivation des élèves français face à des évaluations à faibles enjeux – Comment la mesurer ? Son impact sur les réponses*) en proposant un instrument de mesure de la motivation des élèves envers l'évaluation.

À supposer que les élèves aient témoigné d'un degré de motivation satisfaisant et que l'on puisse les classer selon un score en fonction de leurs réussites et de leurs échecs aux items qu'ils ont passés, certains programmes doivent alors produire un résultat « normatif ». C'est le cas des indicateurs calculés

dans le cadre de la LOLF (Loi organique relative aux lois de finances) qui reposent sur les pourcentages d'élèves qui maîtrisent les compétences du socle commun. Nicolas Miconnet et Ronan Vourc'h (*Détermination de standards minimaux pour évaluer les compétences du socle commun*) montrent que le statisticien ne peut pas répondre seul à la question de la détermination de seuils de réussite. En effet, il s'agit de se prononcer sur la « maîtrise » d'une compétence, au regard des objectifs visés et de ce qui est proposé dans l'évaluation. Les auteurs décrivent et appliquent des méthodes spécifiques développées pour répondre à cette problématique. Elles visent à croiser le jugement d'« experts » pédagogiques avec un corpus de données empiriques. Étant donné la valeur normative du résultat produit par l'évaluation, il est important que ces méthodes soient explicitées.

Le dernier article de cette partie axée sur la méthodologie porte sur un sujet d'actualité : la transition des évaluations, habituellement passées sous un format papier-crayon, vers un format numérique. Pascal Bessonneau, Philippe Arzoumanian et Jean-Marc Pastor (*Une évaluation sous forme numérique est-elle comparable à une évaluation de type « papier-crayon » ?*) présentent les résultats de deux expériences consistant à comparer les résultats obtenus aux mêmes épreuves, envisagées selon les deux formats différents. Si les auteurs parviennent à identifier certains paramètres pouvant expliquer les écarts obtenus, ils montrent qu'il n'est pas aisé de contrôler l'effet du passage à un environnement numérique sur la difficulté de la tâche proposée. C'est l'un des défis à venir : répondre à la préoccupation légitime de continuité des séries de résultats, indépendamment du format d'interrogation. Cependant, au-delà de cette attitude « conservatrice » visant à une simple dématérialisation des évaluations, notamment pour des raisons financières, les futures évaluations devraient grandement profiter du passage au numérique en termes de contenu (format des items, données de navigation des élèves, items interactifs, etc.).

Après ces considérations méthodologiques, les cinq articles suivants portent sur des analyses et des résultats issus de données d'évaluations.

Tout d'abord, l'article de Sylvie Beuzon, Émilie Garcia et Corinne Marchois (*Les compétences des élèves français en anglais en fin d'école et en fin de collège – Quelles évolutions de 2004 à 2010 ?*) rend compte des résultats des évaluations Cedre en anglais, la focale étant mise sur l'évolution du niveau de compétence des élèves de 2004 à 2010. La période envisagée se révèle intéressante. À l'école, les résultats positifs peuvent être rapprochés de la politique d'apprentissage précoce de la deuxième langue. Au collège, en revanche, le tableau est moins glorieux : malgré les efforts de l'institution sur la place de l'oral dans l'enseignement de l'anglais, le niveau baisse et les inégalités augmentent. L'autre apport important de l'article est de montrer l'importance du caractère standardisé des évaluations. Ainsi, en fin de troisième, Cedre, tout comme l'enquête européenne ESLC, montre qu'au plus 40 % des élèves maîtrisent le niveau A2 du cadre de référence européen, alors que la validation de ce même niveau, telle qu'enregistrée dans le livret personnel de compétences (LPC) dépasse 90 %...

Les comparaisons diachroniques ont l'avantage de révéler des points forts et des points faibles, en coupe, mais elles ne permettent pas de saisir la dynamique des évolutions de compétences. La DEPP conduit depuis les années 1960 des études longitudinales, appelées « panels ». Le panel d'élèves entrant en sixième à la rentrée 2007 est sans doute le panel le plus riche qu'ait conduit la DEPP : environ 35 000 élèves, évalués en fin de sixième, trois ans plus tard, en fin de troisième, deux prises d'informations sur les familles, des évaluations sur les aspects cognitifs mais également affectivo-motivationnels. Ce panel constituera très certainement une source majeure de connaissance pour la recherche en éducation dans les années futures. À partir de ces données, Linda Ben Ali et Ronan Vourc'h ont conduit une analyse sur l'évolution, de la sixième à la troisième, du niveau des élèves dans les différents domaines évalués, en fonction de leurs caractéristiques familiales. Cette étude est importante, car elle permet de renouveler les analyses sur les inégalités sociales à l'école. Elle montre que les inégalités sociales sont relativement figées de la sixième à la troisième, dans des disciplines telles que la lecture-compréhension et le raisonnement logique, mais qu'elles augmentent en mathématiques et en mémoire encyclopédique. C'est l'intérêt de cette étude de préciser la construction des inégalités sociales, en distinguant différentes dimensions des acquis, alors que les études sur les inégalités sociales portent très majoritairement sur les parcours ou les diplômes.

Il est difficile de réaliser un numéro spécial sur les évaluations des acquis sans faire référence aux évaluations internationales qui occupent une place importante dans ce domaine. Les nombreux rapports publiés rendent compte de résultats très variés, mais ils évoquent assez rarement les résultats sur le contenu même de l'évaluation. Rapidement, l'attention se concentre sur des palmarès globalisants des pays sur une échelle dont on perd de vue la façon dont elle a été construite. C'est précisément l'objectif de l'article d'Éric Roditi et de Franck Salles (*Nouvelles analyses de l'enquête PISA 2012 en mathématiques – Un autre regard sur les résultats*) que de revenir sur les principes de conception des items. Les auteurs sortent d'une analyse selon « le niveau en mathématiques » pour déconstruire le score global et montrer les leçons très utiles à tirer d'un point de vue pédagogique, dès lors que l'on adopte une grille de lecture pertinente et que l'on affine la granularité des analyses.

Toujours dans le domaine des mathématiques, l'article suivant offre néanmoins une toute autre perspective. Stéphane Herrero, Thomas Huguet et Ronan Vourc'h (*Évaluation des compétences des jeunes en numératie lors de la Journée défense et citoyenneté*) font état des résultats obtenus par les jeunes Français au test de numératie, introduit dans les tests de la Journée défense et citoyenneté (JDC) sur un large échantillon. La corrélation obtenue avec les résultats aux tests de lecture passés par tous les jeunes toute l'année montre la spécificité des difficultés des jeunes dans le domaine de l'usage des mathématiques. Cette évaluation se révèle riche d'enseignements, mais plus

généralement, le dispositif d'interrogation, adapté à l'interrogation massive de l'ensemble d'une génération, de façon relativement simple et efficace, renforce la JDC comme lieu d'observatoire de la jeunesse.

Pour clore ce numéro spécial, le dernier article porte sur l'évaluation standardisée au service de la mesure des effets d'une expérimentation. Marion Le Cam et Olivier Cosnefroy (*Évaluation des effets du dispositif expérimental d'enseignement intégré de science et technologie (EIST)*) présentent les résultats de l'évaluation du dispositif EIST. Le dispositif d'évaluation a consisté à suivre les élèves bénéficiant de l'EIST et de comparer leurs progressions à celles obtenues par un échantillon témoin. Le corpus de données est relativement unique : environ 4 000 élèves ont été suivis tout au long du collège, et évalués à cinq reprises, du début de la sixième à la fin de la troisième. Les effets de l'expérimentation ne sont pas concluants, pour des raisons dépassant certainement le seul contenu de l'expérimentation et certainement liées aux conditions de sa mise en œuvre. Si les résultats de l'expérimentation se révèlent décevants, les données recueillies constituent une source très riche d'informations sur les progressions des acquis en science et devraient renseigner sur le développement des compétences et des attitudes à l'égard des sciences au cours du collège.