Mesure et évaluation en éducation

La détermination de standards minimaux dans le cadre d'indicateurs de résultats
Méthodologie, intérêt, utilité

Thierry Rocher

Volume 31, numéro 2, 2008

URI : https://id.erudit.org/iderudit/1025008ar
DOI : https://doi.org/10.7202/1025008ar

Aller au sommaire du numéro

Éditeur(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (imprimé)
2368-2000 (numérique)

Découvrir la revue

Citer cet article

Rocher, T. (2008). La détermination de standards minimaux dans le cadre d'indicateurs de résultats : méthodologie, intérêt, utilité. *Mesure et évaluation en éducation*, *31* (2), 75–91. https://doi.org/10.7202/1025008ar

Résumé de l'article

Les questions méthodologiques que soulève la mise au point de standards minimaux sont discutées à travers l'exemple d'un dispositif français d'évaluation destiné à produire des indicateurs de résultats du système éducatif. Une attention particulière est portée sur la méthode employée pour fixer les seuils de performance.

La détermination de standards minimaux dans le cadre d'indicateurs de résultats : méthodologie, intérêt, utilité

Thierry Rocher
Ministère de l'Éducation nationale, France

Mots clé : Standards, score-seuil, compétences de base, indicateurs

Les questions méthodologiques que soulève la mise au point de standards minimaux sont discutées à travers l'exemple d'un dispositif français d'évaluation destiné à produire des indicateurs de résultats du système éducatif. Une attention particulière est portée sur la méthode employée pour fixer les seuils de performance.

Key words : Standards, threshold score, basic skills, indicators

The methodological issues raised by the development of minimum standards are discussed through the example of a French evaluation design intended to produce performance indicators of the education system. Special attention is given to the method used to set performance thresholds.

Palavras-chave : Standards, limiar de pontuação, competências básicas, indicadores

As questões metodológicas levantadas pelo desenvolvimento de standards mínimos são discutidas através do exemplo de um dispositivo francês de avaliação destinado a produzir indicadores de resultados do sistema educativo. É dada uma atenção particular sobre o método utilizado para fixar os limiares de desempenho.

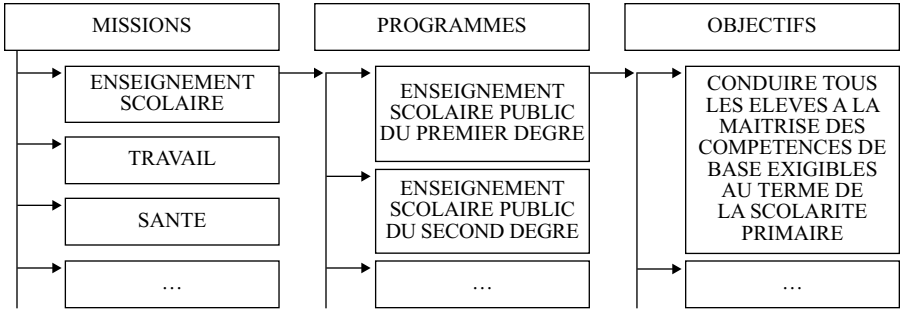
Note de l'auteur – Toute correspondance peut être adressée comme suit : Thierry Rocher, Direction de l'évaluation, de la prospective et de la performance, Ministère de l'Éducation nationale, DEPP B2, 61-65 rue Dutot, 75015 PARIS, France, ou par courriel à l'adresse suivante : [thierry.rocher@education.gouv.fr].

Introduction

Le développement de standards en matière d'éducation répond à une demande de plus en plus forte. Bien établi dans les systèmes éducatifs anglo-saxons, le concept protéiforme de standards se diffuse aujourd'hui dans plusieurs pays francophones (Behrens, 2005 ; Mons & Pons, 2006). D'un point de vue méthodologique, cette demande pose des questions intéressantes, notamment concernant l'aspect opérationnel du processus d'élaboration des standards, à savoir la fixation des seuils de performance. Ces questions sont discutées ici à travers l'exemple d'un dispositif français d'évaluation, conçu dans le but de produire des indicateurs de résultats pour rendre compte de l'efficacité des politiques publiques.

Le contexte

Depuis 2001, la France est engagée dans une refonte du cadre comptable du budget de l'État que définit la loi organique relative aux lois de finances (LOLF, cf. figure 1).



Note : Le nouveau cadre comptable défini par la LOLF comporte 34 missions, 133 programmes et près de 580 actions assorties d'indicateurs permettant de mesurer l'atteinte des objectifs que se fixent les lois. Les indicateurs « proportions d'élèves maîtrisant les compétences de base (définies en référence au socle commun) en français et en mathématiques » s'inscrivent dans l'objectif, pour le premier degré, de « conduire tous les élèves à la maîtrise des compétences de base exigibles au terme de la scolarité primaire » et pour le second degré, de « conduire le maximum d'élèves aux niveaux de compétences attendues en fin de scolarité et à l'obtention des diplômes correspondants », et ce, dans les programmes des secteurs public et privé. De plus, pour l'objectif « accroître la réussite scolaire des élèves en zones difficiles et des élèves à besoins éducatifs particuliers », ces indicateurs doivent également être calculés pour les élèves scolarisés dans les établissements d'éducation prioritaire – réseau de réussite scolaire (RRS) et réseau ambition réussite (RAR).

Figure 1. ***Les indicateurs « compétences de base » dans le cadre de la loi organique relative aux lois de finances (LOLF)***

Cette nouvelle architecture budgétaire s'inscrit dans une démarche de performance : pour tous les domaines de l'administration, des objectifs sont fixés et des indicateurs de résultats sont utilisés pour rendre compte de l'efficacité des politiques publiques.

Le domaine de l'éducation n'est pas en reste et, parmi les indicateurs de résultats retenus, figurent les proportions d'élèves qui maîtrisent les compétences de base en français et en mathématiques, en fin d'école primaire (CM2) et en fin de collège – *i.e.* secondaire inférieur – (troisième)[1]. Ces indicateurs doivent être calculés chaque année, à compter de 2007, sur la base des résultats obtenus par un échantillon d'élèves à des tests standardisés. Plus précisément, les quatre proportions sont à décliner selon les catégories suivantes : le secteur public, deux catégories du secteur de l'éducation prioritaire et le secteur privé. En outre, pour répondre complètement à la demande de la LOLF, ces seize taux devaient être produits dans chacune des trente académies qui correspondent aux circonscriptions éducatives françaises. Cependant, les contraintes logistiques et financières ont corrigé à la baisse ces ambitions de départ : en 2007, les résultats sont valables uniquement sur le plan national. De plus, ces mêmes contraintes impliquent la construction d'un test sous forme de QCM (questions à choix multiples). Ce format de questions assure une correction fiable, économique et rapide – les résultats doivent être disponibles chaque année avant la fin de l'été pour la préparation du débat budgétaire au Parlement.

Ces indicateurs répondent à une exigence forte sur le plan méthodologique : la comparabilité dans le temps. En effet, les indicateurs de résultats de la LOLF sont accompagnés de « cibles » fixées par le pouvoir politique, c'est-à-dire d'objectifs chiffrés à atteindre l'année suivante (*i.e.*, *benchmarks*). Par conséquent, les indicateurs doivent permettre de mesurer l'évolution des acquis des élèves sur des bases comparables.

En France, le programme de suivi des acquis des élèves, tel que mis en place par la Direction de l'évaluation de la prospective et de la performance (DEPP) du ministère de l'Éducation nationale, est organisé depuis 2003 en cycles de six ans, alternant chaque année la discipline évaluée (*cf.*, par exemple, Gibert, Levasseur & Pasteur, 2004). Cette organisation ne permet pas de répondre à la présente demande ; la DEPP a donc conçu un dispositif d'évaluation *ad hoc*, qui a été mis en place pour la première fois en 2007. Si chaque aspect de ce dispositif est décrit dans la suite de cet article, l'accent est plus particulièrement mis sur la méthode suivie pour déterminer le seuil à partir duquel un élève peut être considéré comme maîtrisant les compétences de base.

Le dispositif d'évaluation relatif aux indicateurs de résultats de la LOLF

Définition des compétences de base

Les compétences de base visées par l'indicateur font référence au Socle commun de connaissances et de compétences (Ministère de l'Éducation nationale, 2006) qui définit les compétences que doivent maîtriser les élèves à l'issue de la scolarité obligatoire. Dans ce cadre, les services en charge de l'élaboration des programmes scolaires ont mis au point, dans chaque discipline, à chaque niveau, des documents de travail dans lesquels ont uniquement été retenues les compétences dites « de base », c'est-à-dire considérées comme indispensables, en français et en mathématiques, en fin d'école primaire et en fin de collège.

Cependant, les compétences visées sont celles que la forme d'évaluation retenue – en l'occurrence, un test standardisé composé de QCM – permet de mesurer. De ce fait, en français, les compétences liées à l'expression orale et celles liées à l'expression écrite ne sont pas prises en compte. De la même manière, en mathématiques, l'écriture de nombres, la construction de figures géométriques ou encore la production de démonstrations ont été écartées.

Élaboration des épreuves

Afin d'élaborer les items destinés à évaluer les compétences retenues, des groupes de travail ont été réunis dans chaque discipline, à chaque niveau : ils étaient composés d'enseignants, de conseillers pédagogiques, d'inspecteurs de l'éducation nationale et de représentants des services ministériels chargés des programmes. Ces groupes ont mis au point un grand nombre d'items afin de permettre le renouvellement et l'évolution des tests dans les prochaines années.

Le tableau 1 décrit la composition des épreuves. En mathématiques, que ce soit à l'école ou au collège, la méthode des cahiers tournants (Weller & Romney, 1988) a permis d'expérimenter plus de trois heures de test en maintenant un temps de passation d'une heure. En français, la situation est différente car les items peuvent rarement être « décontextualisés ». Ils se rapportent à un stimulus, le plus souvent un texte écrit. De ce fait, il est moins aisé de construire, comme en mathématiques, des blocs équilibrés pour les cahiers tournants. La démarche envisagée en français repose sur la construction de versions « parallèles », c'est-à-dire proches dans leur architecture et leur questionnement.

Tableau 1
Description des épreuves expérimentées en 2006

Épreuves	*Items construits*	*Items testés*	*Cahiers*	*Temps de passation*	*Champs*
Maths CM2	244	221	13 blocs de 17 items pour 13 cahiers tournants de 4 blocs chacun	1 heure	Exploitation de données numériques (A) ; Connaissance des nombres entiers naturels (B) ; Connaissance des fractions simples et des décimaux (C) ; calcul (D) ; Espace et géométrie (E) ; Grandeurs et mesure (F)
Maths 3e	294	169	13 blocs de 13 items pour 13 cahiers tournants de 4 blocs chacun	1 heure	Organisation et gestion de données, fonctions (A) ; Nombres et calculs (B) ; Grandeurs et mesures (C) ; Géométrie (D)
Français CM2	142	142	2 versions « parallèles », respectivement de 67 et 75 items ; 2 cahiers pour tenir compte de l'ordre dans les versions	2 x 1 heure	Lecture ; maîtrise des outils de la langue
Français 3e	445	172	2 « corpus » (ensemble d'items avec un thème commun), respectivement de 95 et 77 items ; 4 cahiers tournants	1 1/2 heure	Compréhension de textes ; maîtrise des outils de la langue

Expérimentation 2006

En 2006, les échantillons ont été stratifiés selon l'académie et le secteur de scolarisation (secteur public, éducation prioritaire, secteur privé). Ce dispositif « allégé », qui concernait 15 000 élèves par niveaux scolaires (CM2 et troisième), a également permis d'évaluer la charge de travail et le coût financier qu'impliquerait la déclinaison de ces indicateurs sur le plan géographique, par académies. Ce projet a d'ailleurs été abandonné et, en 2007, l'échantillon est national.

L'analyse des items a montré que, selon les tests, de 10 % à 25 % des items sont faiblement discriminants, le plus souvent en raison d'ambiguïtés dans le questionnement ou dans les réponses proposées. Ces items ont été exclus par la suite. Par ailleurs, l'étude des réussites aux items selon leurs places dans les cahiers n'a pu déceler aucun effet de fatigue ou d'entraînement. Enfin, les

éventuels comportements de réponses au hasard, dont on pouvait craindre l'existence à cause du format des items, ne sont pas apparus à l'analyse. Dans la perspective de faciliter les comparaisons temporelles, le cadre d'analyse est celui des modèles de réponse à l'item (MRI à deux paramètres, unidimensionnel pour chaque discipline et chaque niveau, considérés séparément).

La détermination du seuil de maîtrise

Introduction

L'expérimentation a permis de recueillir les résultats des élèves à un ensemble d'items qui mesurent l'acquisition des compétences de base, telles qu'elles ont été définies en amont. Cependant, ces résultats ne permettent pas de calculer directement la proportion des élèves qui maîtrisent ces compétences de base, car le «degré de maîtrise» n'a pas encore été défini. En effet, les items peuvent être de difficulté très variable, quand bien même ils portent sur une compétence dite de «base». Par exemple, en mathématiques, il est possible de rendre très complexes des items dont la résolution ne fait appel qu'aux opérations les plus élémentaires. De ce fait, pour être considéré comme un élève qui maîtrise les compétences de base, l'élève doit-il réussir toutes les questions qui lui sont proposées? les trois quarts? la moitié? C'est ce seuil qui doit être fixé, seuil à partir duquel nous considérerons que les élèves maîtrisent les compétences de base.

Lors de précédentes évaluations menées par la DEPP, la fixation d'un tel seuil a été opérée de manière arbitraire, ce qui a pu nuire parfois au sens et à la légitimité des résultats. De ces expériences, il est ressorti que deux tentations sont à éviter: la tentation «relative», qui consiste à fixer ce seuil en fonction de la distribution des élèves, par exemple en considérant que les 15% les élèves plus faibles sont ceux qui ne maîtrisent pas les compétences de base; la tentation «absolue», qui consiste à fixer *a priori* une règle de calcul qui indique si l'élève maîtrise ou non les compétences de base. Le point de vue envisagé ici a consisté à mêler ces deux approches antagonistes en confrontant les observations issues de l'expérimentation avec les attentes et les exigences du système éducatif. Le choix a été fait d'un travail collaboratif entre statisticiens, pédagogues et responsables pédagogiques.

De nombreuses méthodes ont été proposées pour déterminer des seuils de performances (Blais, dans ce volume; Laveault & Grégoire, 2002). Elles s'appuient sur les avis d'un panel d'experts et sont généralement distinguées selon qu'elles se basent sur le contenu du test ou sur la performance des sujets (Jaeger, 1989). Par exemple, parmi les plus couramment utilisées, les métho-

des dites d'Angoff (1971) demandent un examen approfondi des items aux experts, qui doivent prédire pour chacun d'entre eux la probabilité de réussite des élèves. Elles ne nécessitent pas de disposer des résultats des élèves. En revanche, la célèbre méthode dite des « groupes contrastés » se base sur une classification des élèves *a priori*, qui est ensuite mise en regard de leurs performances réelles (Livingston & Zieky, 1982). L'approche retenue ici se fonde sur l'examen du contenu du test mais elle utilise également les résultats empiriques issus de l'expérimentation pour les confronter avec les jugements et les attentes des experts.

Du point de vue de la qualité de ces méthodes, des critères ont été mis au point (Berk, 1986). Par exemple, la simplicité de mise en œuvre, la facilité de communication, le nombre et les caractéristiques des experts sont autant de critères dont l'importance dépend principalement des objectifs du test. Hambleton (2001) préfère même poser des critères d'évaluation sous forme de questions plutôt que sous forme de consignes strictes, tant les situations peuvent être différentes.

Le choix et l'adaptation de ces méthodes à notre contexte ont tenu compte de contraintes spécifiques, notamment de la nécessité d'organiser le travail sur une seule journée, avec les concepteurs des tests qui tiennent lieu d'experts[2]. Le matin, les experts ont été conviés à un travail individuel sur les items, sans avoir accès aux résultats statistiques ; l'après-midi, la confrontation de leur travail avec les données empiriques leur a été présentée et, après discussion, un consensus a été dégagé pour établir les seuils. Deux méthodes différentes ont été élaborées, afin de croiser leurs résultats et de tirer parti des éventuelles contradictions. Ces méthodes sont complémentaires : la première est très générale, elle compare directement les jugements et les attentes des experts avec les données statistiques ; la seconde est plus précise et s'appuie sur les compétences attendues d'un groupe virtuel d'élèves, situés autour du seuil de maîtrise des compétences de base.

Première méthode

La première approche suit la méthode d'Hofstee (cf. Norcini, 2003). Elle consiste à établir la « zone de jugement » de chaque expert (*cf.* figure 2). Cette zone témoigne à la fois des attentes (axe des ordonnées) et du niveau d'exigence (axe des abscisses) des experts. Par ailleurs, à partir des données issues de l'expérimentation, il est possible de calculer, pour chaque score observé, le pourcentage d'élèves situés en deçà de ce score et de tracer la courbe correspondante. Le point d'intersection entre cette courbe et la diagonale de la zone de jugement fournit le score-seuil retenu.

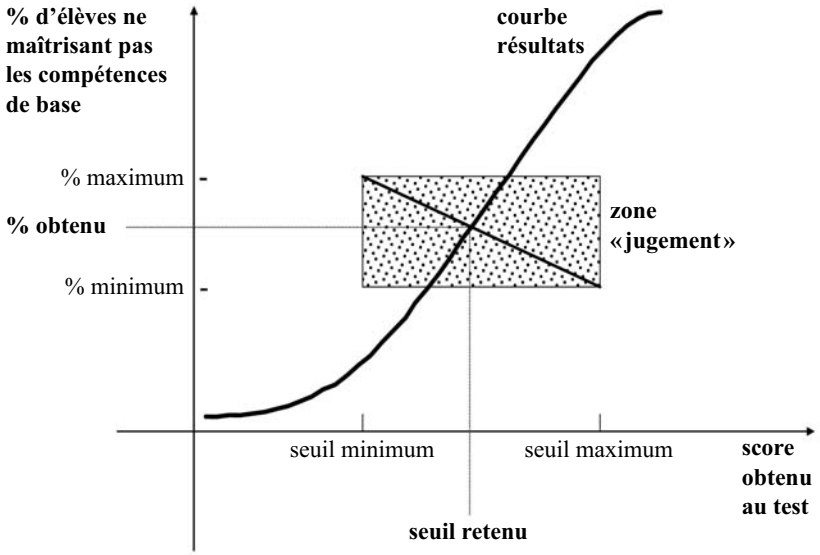


Figure 2. ***Première méthode de détermination du seuil***

Concrètement, pour chaque discipline – français et mathématiques – et pour chaque niveau – CM2 et troisième – les experts ont répondu à la question suivante : « Sur la base de votre expérience professionnelle, quels sont, selon vous, le pourcentage minimum et le pourcentage maximum d'élèves ne maîtrisant pas les compétences de base ? » Notons tout d'abord que le point de vue adopté est celui de la non-maîtrise des compétences de base, plutôt que de la maîtrise ; ce point de vue s'avère plus révélateur pour identifier le seuil.

Il leur a été demandé ensuite de déterminer quel est le score – quant au nombre de bonnes réponses – en deçà duquel les élèves peuvent être considérés comme ne maîtrisant pas les compétences de base. Plus précisément, il leur a été demandé de déterminer un score minimum « acceptable » correspondant à une notation large et un score maximum « acceptable » correspondant à une notation sévère. Un score « acceptable » leur a été présenté de la manière suivante : « Un score qui correspond à votre propre perception et qui vous semble légitime du point de vue de ce que l'institution scolaire peut attendre d'un élève en fin d'école. »

Les experts devaient travailler à partir d'un cahier de l'expérimentation. Cependant, la détermination de ces scores minimum et maximum n'est pas évidente, par la simple observation des items présents dans le cahier. Pour les aider dans cette tâche, les experts ont été invités à suivre la consigne suivante : « La borne inférieure relève d'un niveau d'exigence minimal. En

pratique, il s'agit de déterminer les questions fondamentales qui doivent absolument être réussies. *A contrario*, la borne supérieure renvoie à l'idée que les élèves devraient réussir toutes les questions posées, si l'on s'en tenait à une vision rigide des compétences de base En pratique, pour déterminer cette borne supérieure, il s'agit simplement d'éliminer certaines questions qui ne semblent pas relever *stricto sensu* des compétences de base. ». Ainsi, dans le même esprit que la méthode d'Ebel (1972), les experts devaient classer les items en fonction de leur importance : les items qu'un élève maîtrisant les compétences de base doit absolument réussir, devrait être capable de réussir ou ne doit pas forcément réussir. Le nombre d'items devant être absolument réussi fournit la borne inférieure et le nombre total d'items, exceptés les items ne devant pas forcément être réussis, fournit la borne supérieure.

Deuxième méthode

La deuxième approche s'inspire des méthodes dites d'Angoff, dans une version simplifiée. Elle consiste à caractériser la difficulté des items pour un groupe d'élèves dit « flottant » (*borderline*), situés au seuil de maîtrise des compétences de base. Selon les méthodes classiques d'Angoff, les experts doivent estimer, pour chaque item, le taux de réussite qu'obtiendraient ces élèves. Cette idée a été écartée pour des raisons d'efficacité – le travail devait durer une journée au maximum – et de recevabilité – les experts auraient difficilement accepté de chiffrer en détail chaque item. En fait, les items ont été triés par ordre de difficulté[3] et il a été demandé aux experts de situer dans la liste les items pour lesquels les élèves du groupe « flottant » obtiendraient 50 % de réussite.

En pratique, le travail leur a été présenté comme suit : « L'objectif final est de produire un pourcentage d'élèves qui maîtrisent les compétences de base. Ainsi, de fait, deux groupes d'élèves sont distingués : ceux qui maîtrisent et ceux qui ne maîtrisent pas ces compétences. Évidemment, la réalité n'est pas aussi tranchée et les groupes ne sont pas aussi clairement distincts l'un de l'autre. Certains élèves sont susceptibles d'être classés dans l'un ou l'autre des deux groupes selon des paramètres extérieurs comme par exemple la fatigue, l'environnement, les conditions de passation, *etc*. Ces élèves forment un groupe dit "flottant". Dès lors, il s'agit de repérer un niveau de difficulté d'items permettant de caractériser le mieux ce groupe d'élèves "flottant" ». C'est ce niveau de difficulté qui permet de fixer le seuil de maîtrise des compétences de base.

À partir des résultats issus de l'expérimentation, les items ont été classés par champ et par difficulté croissante. Les items du début de chaque liste ont été les mieux réussis lors de l'expérimentation et ceux de la fin les moins bien

réussis (*cf.* figure 3). Pour chaque champ, les élèves du groupe « flottant » sont susceptibles de réussir la majorité des items du début de la liste et d'échouer la majorité des items de la fin de la liste.

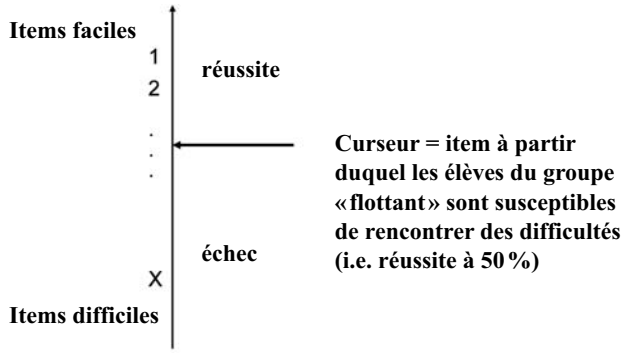


Figure 3. ***Deuxième méthode de détermination du seuil***

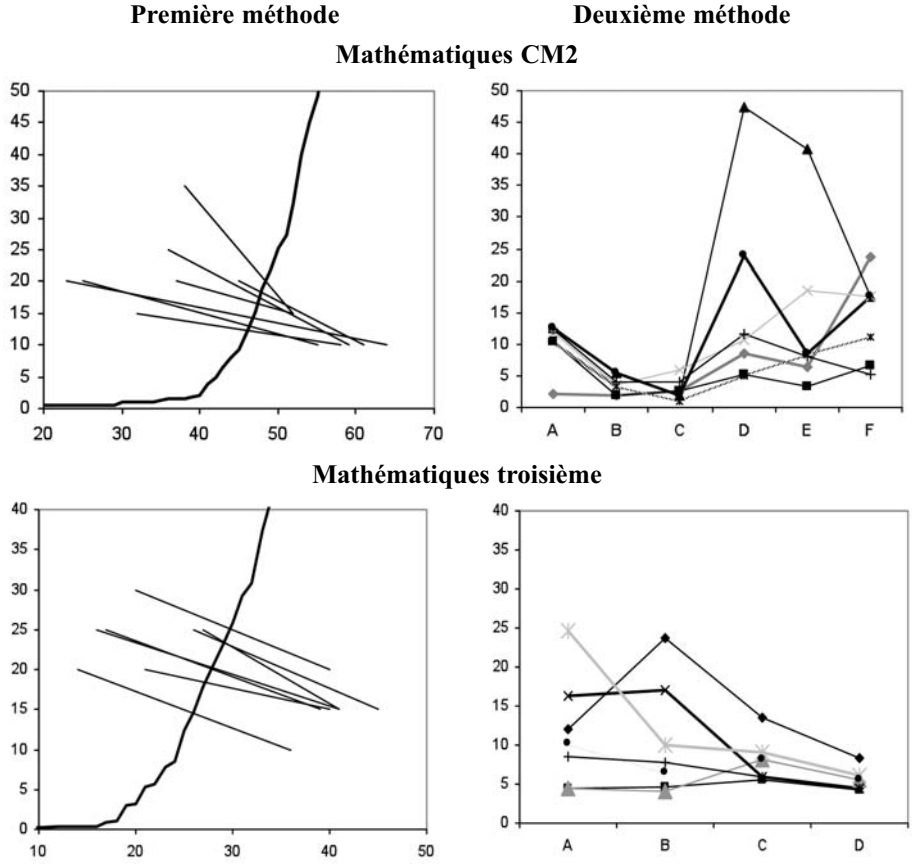
Il a été demandé aux experts de repérer dans la liste fournie la position de l'item à partir duquel peut s'opérer « la bascule », c'est-à-dire à partir duquel les élèves du groupe « flottant » vont commencer à rencontrer des difficultés. Pour permettre une certaine souplesse, il était possible d'indiquer trois items consécutifs plutôt qu'un seul. Il est important de noter que la liste des items n'était accompagnée d'aucun résultat complémentaire, d'aucun chiffre ou autre taux de réussite qui auraient pu influencer le choix des experts.

Résultats

La figure 4 donne les résultats de ce travail pour les mathématiques. Il apparaît que la première méthode conduit à des seuils – et donc à des pourcentages d'élèves ne maîtrisant pas les compétences de base – plus élevés que la seconde méthode. La variabilité entre les experts est limitée, surtout pour le CM2. Enfin, les niveaux d'exigence et d'attente diffèrent selon les champs (nombres, géométrie, *etc.*). La situation est différente en français (figure 5) : la première méthode laisse apparaître des niveaux d'exigence minimum (borne inférieure) trop élevés par rapport aux résultats des élèves, si bien que les résultats de la première méthode sont indéterminés pour la plupart des experts.

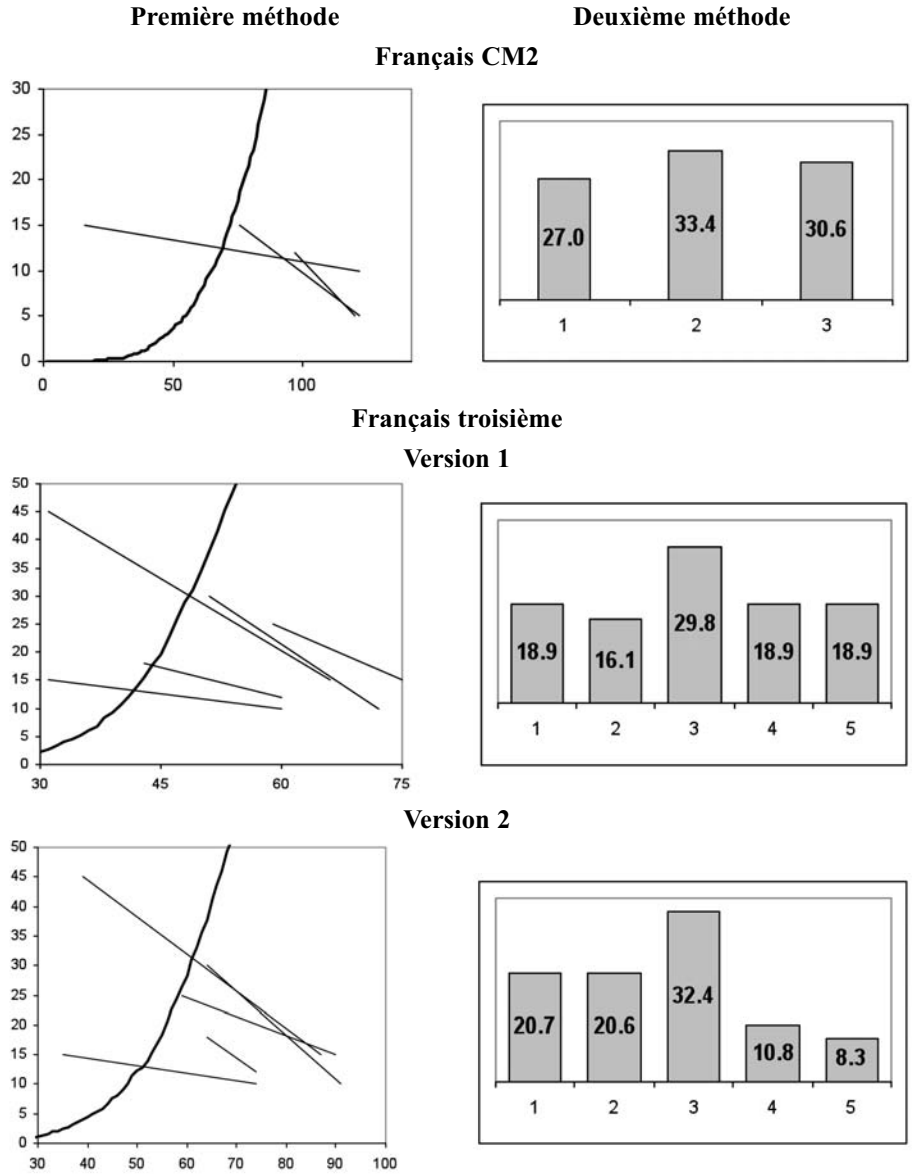
Les groupes d'experts ont discuté de ces résultats et des divergences constatées. Ils ont alors réanalysé les items, en s'appuyant sur les résultats statistiques et en prenant comme point de départ le seuil moyen déterminé par la seconde méthode. En effet, la première méthode fournit une bonne image des représentations des experts mais elle se révèle moins précise pour l'établissement

d'un seuil. Au final, les experts se sont mis d'accord sur l'établissement d'un seuil en commun. Plus précisément, ils ont fourni les items-seuil conformément à la deuxième méthode. Le niveau de difficulté de ces items permet d'obtenir, grâce aux modèles de réponse à l'item, le niveau de compétence à partir duquel un élève est considéré comme maîtrisant les compétences de base.



Note : Les graphiques ci-dessus présentent les résultats des deux méthodes de détermination des seuils pour les mathématiques, en CM2 et en troisième. Les deux groupes étaient composés chacun de sept experts. Les graphiques situés à gauche donnent les résultats de la première méthode : l'axe des abscisse est le score à l'évaluation (sur 68 pour le CM2, sur 52 pour la troisième), l'axe des ordonnées est le pourcentage d'élèves ne maîtrisant pas les compétences de base. Les graphiques situés à droite donnent les résultats de la deuxième méthode : en abscisse, il s'agit des différents champs évalués (cf. tableau 1), en ordonnées, il s'agit du pourcentage d'élèves ne maîtrisant pas les compétences correspondant au seuil retenu selon la deuxième méthode.

Figure 4. ***Résultats des deux méthodes de détermination pour les mathématiques***



Note : Comme pour la figure 4 qui concernaient les mathématiques, les figures de gauche concernent le français. En CM2, seuls trois experts étaient présents lors de la réunion. En troisième, cinq experts ont participé à ce travail et deux versions étaient évaluées (cf. tableau 1). Les graphiques de droite donne le pourcentage d'élèves ne maîtrisant pas les compétences correspondant au seuil retenu par la deuxième méthode, selon les experts. Mais contrairement aux mathématiques, il n'y a pas eu de distinction selon différents champs.

Figure 5. ***Résultats des deux méthodes de détermination pour le français***

Validation

En plaçant sur la même échelle les élèves et les items, il est possible de décrire les compétences maîtrisées par les élèves situés au-dessus du seuil établi, en CM2 et en troisième. Cette description a été présentée aux responsables de la politique pédagogique, qui ont pu valider les choix opérés en approuvant le sens précis donné à « la maîtrise des compétences de base ». À ce titre, il est intéressant de noter que les pourcentages sont sensiblement différents des mathématiques au français, surtout en troisième. La conception de la maîtrise des compétences de base n'est probablement pas la même. Si, en mathématiques, elles sont perçues comme un niveau de maitrise nécessaire pour la vie courante, en français, il s'agit plutôt de compétences minimales pour poursuivre les études dans le secondaire supérieur en filière générale (*i.e.* « académique »).

Test 2007 et indicateurs

Au final, le test doit fournir une mesure précise du niveau de compétences des élèves situés autour du seuil déterminé à l'étape précédente. Par exemple, il n'est pas nécessaire que le test soit précis pour les élèves les plus compétents. En conséquence, la sélection des items pour le test de 2007 a tenu compte de leur capacité à évaluer précisément le niveau de compétences correspondant au seuil retenu (dans le cadre des MRI, il s'agit de l'« information » des items). Cette sélection a également respecté l'équilibre entre les champs, les sous-compétences visées, les supports, *etc*. S'il n'a pas été nécessaire de procéder à un programme d'optimisation (*cf.* Van der Linden, 1998), la sélection des items s'est inscrite dans cette démarche.

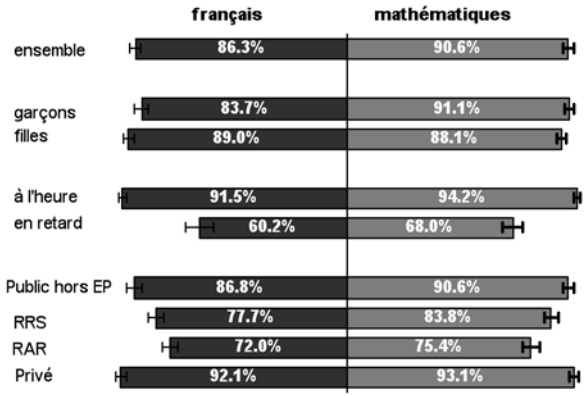
Au final, le test se présente sous la forme d'une seule version contenant des épreuves de français (75 items en CM2 et 127 items en troisième) et de mathématiques (68 items en CM2 et 52 en troisième). La durée de l'évaluation est d'une heure pour chaque discipline, ce qui permet aux élèves de répondre à tous les items.

Pour l'année 2007, des échantillons nationaux ont été tirés au sort : environ 2 000 élèves par zones (secteur public hors éducation prioritaire, le réseau de réussite scolaire, le réseau ambition réussite et le secteur privé) et par niveaux scolaires (CM2 et troisième). Au total, près de 8 000 élèves de CM2 et 8 000 élèves de troisième ont passé les tests. Comme dans toute enquête procédant par échantillonnage, les résultats doivent être accompagnés d'une marge d'incertitude, que l'on peut estimer en fonction des données[4]. En

l'occurrence, les intervalles de confiance varient de ±1,8 % à ±3,8 %, selon le niveau scolaire et la zone de scolarisation (figures 6 et 7). Ces précisions sont d'autant plus importantes à fournir que ces indicateurs, comme tous ceux de la LOLF, servent de base au vote du budget de l'État. Il s'agit ainsi d'éviter que des décisions soient engagées sur la base de variations provenant simplement d'erreurs d'échantillonnage.

	En **français**, 86,3 % élèves de fin de CM2 sont capables
Lecture	de chercher des informations en se référant à l'organisation d'un dictionnaire ; de comprendre globalement un texte littéraire ou documentaire court et d'y prélever des informations ponctuelles explicites
Maîtrise des outils de la langue	de maîtriser partiellement l'automatisation de la correspondance grapho-phonologique ; d'identifier les principaux temps de l'indicatif pour les verbes les plus fréquents ; de reconnaître les règles les plus simples d'orthographe lexicale et grammaticales

	En **mathématiques**, 90,6 % élèves de fin de CM2 sont capables
Exploitation de données numériques	de prélever une information dans un tableau ; de résoudre des problèmes simples relevant de l'addition et de la soustraction
Connaissance des nombres et calcul	de passer d'une écriture en lettres à une écriture en chiffres (ou le contraire) et de comparer, d'additionner et de soustraire des nombres entiers naturels ; de reconnaître le double ou la moitié d'un nombre entier « familier » ; de passer d'une écriture en lettres à une écriture sous forme fractionnaire (ou le contraire) de fractions simples
Espace et géométrie	de reconnaître visuellement un triangle, un triangle rectangle, un rectangle, un carré ; de reconnaître par une représentation en perspective un cube ou un parallélépipède rectangle
Grandeurs et mesure	de mesurer la longueur d'un segment ; d'utiliser les unités de mesure des durées (sans calculs)

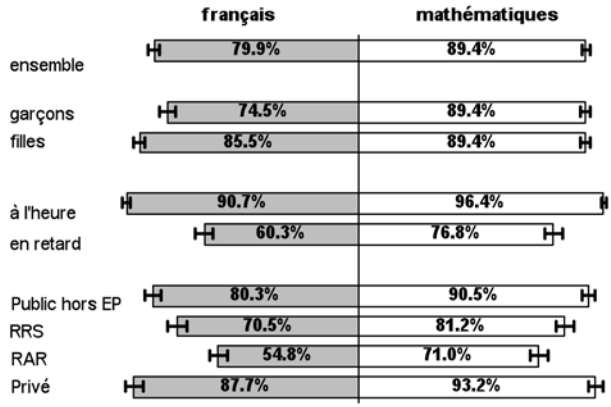


Lecture : 86,3 % des élèves de CM2 maîtrisent les compétences de base en français. L'intervalle de confiance de cet indicateur est de ± 2,1 %

Figure 6. ***Proportion d'élèves de fin de CM2 qui maîtrisent les compétences de base en français et en mathématiques***

	En **français**, environ 79,9% élèves de fin de troisième sont capables
Compréhension des textes	de reconnaître un texte explicatif; de distinguer les principaux genres de textes; de prélever des informations explicites; de faire des inférences simples; et de donner une interprétation d'un texte sans difficulté de compréhension, à partir d'éléments simples
Maîtrise des outils de la langue	d'identifier les structures syntaxiques fondamentales; d'analyser les principales formes verbales; de faire un emploi pertinent du vocabulaire courant; de repérer différents niveaux de langue; de reconnaître les règles d'orthographe et de ponctuation, d'usage courant

	En **mathématiques**, environ 89,4% élèves de fin de troisième sont capables
Organisation et gestion de données, fonctions	d'utiliser une représentation graphique dans des cas simples (lecture des coordonnées d'un point, lien avec un tableau numérique dans une situation de proportionnalité, détermination des données d'une série statistique); de calculer la moyenne d'une série statistique; de traiter des problèmes simples de pourcentages
Nombres et calculs	de comparer des nombres décimaux relatifs écrits sous forme décimale; d'utiliser les opérations élémentaires dans une situation concrète
Grandeurs et mesures	d'effectuer pour des grandeurs (durée, longueur, contenance) un changement d'unités de mesure (h min en min, km en m, L en cL); de calculer le périmètre d'un triangle dont les longueurs des côtés sont données; de calculer l'aire d'un carré, d'un rectangle dont les longueurs des côtés sont données dans la même unité
Géométrie	d'identifier des figures simples à partir d'une figure codée et d'en utiliser les éléments caractéristiques (triangle équilatéral, cercle, rectangle); d'écrire et d'utiliser le théorème de Thalès dans un cas simple; de reconnaître un patron de cube ou de parallélépipède rectangle



Lecture: 86,3% des élèves de CM2 maîtrisent les compétences de base en français.
L'intervalle de confiance de cet indicateur est de ± 2,1%

Figure 7. ***Proportion d'élèves de fin de troisième qui maîtrisent les compétences de base en français et en mathématiques***

Conclusion

Les indicateurs de résultats ont une importance croissante aujourd'hui dans le domaine de l'action publique. Utilisés pour évaluer les politiques engagées et rendre compte de leurs résultats, ils jouent un rôle prépondérant dans le processus de décision. Par conséquent, la méthode retenue pour construire de tels indicateurs nécessite une attention particulière.

Dans ce cadre, l'élaboration de standards minimaux en éducation implique un dispositif d'évaluation rigoureux – échantillonnage, passation, analyses statistiques, comparabilité temporelle, *etc*. Au-delà, il est également apparu important de donner une légitimité à ces indicateurs, en impliquant dans leur définition non seulement les spécialistes de la mesure mais aussi les différents acteurs du système éducatif : enseignants, inspecteurs, responsables politiques, *etc*. La confrontation des attentes et des exigences du système éducatif avec les résultats réels des élèves a révélé des contradictions et a permis de déterminer des seuils de maîtrise dans un esprit d'analyse et de concertation.

Les méthodes employées ici pour la détermination des seuils se sont avérées efficaces mais elles sont perfectibles, notamment par un travail sur le choix des experts – augmentation de leur effectif, élargissement aux professeurs des élèves de l'échantillon, *etc*. – ou sur le processus de confrontation entre les jugements et les résultats – par exemple avec la mise en place d'une procédure itérative. Ainsi, en tirant parti de cette expérience, il est envisagé d'étendre l'application de ces méthodes à d'autres évaluations d'élèves menées en France.

NOTES

1. Ces deux niveaux correspondent respectivement à la cinquième et à la neuvième années de scolarité obligatoire.
2. Le fait que les experts soient les personnes qui aient participé à l'élaboration du tests offre un avantage certain : l'appropriation du contenu de l'évaluation est plus rapide. À l'inverse, d'aucuns pourraient opposer une trop grande proximité des experts avec l'instrument. C'est pourquoi, en 2008, le panel d'experts a été élargi aux 250 enseignants des élèves de CM2 de l'échantillon. Les résultats apporteront des éléments éclairants sur la problématique du choix des experts.
3. Selon les paramètres de difficulté estimés par un MRI à deux paramètres.
4. La technique empirique du « jacknife » a été employée pour estimer la variance des estimateurs.

RÉFÉRENCES

Angoff, W.H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (éd.), *Educational measurement* (pp. 508-600). Washington, DC : American Council on Education.

Behrens, M. (2005). *Analyse de la littérature critique sur le développement, l'usage et l'implémentation de standards dans un système éducatif*. IRDP, Neuchâtel.

Berk, R.A. (1986). A Consumer's Guide to Setting Performance Standards on Criterion-Referenced Tests. *Review of Educational Research*, *56*(1), 137-172.

Ebel, R.L. (1972). *Essentials of educational measurement*. Englewood Cliffs, NJ : Prentice-Hall.

Gibert, F., Levasseur, J., & Pasteur, J.-M. (2004). La maîtrise du langage et de la langue française en fin d'école primaire. *Note d'évaluation*, 04.10. [http://educ-eval.education.fr/pdf/eva0410.pdf].

Hambleton, R.K. (2001). Setting Performance Standards on Educational Assessments and Criteria for Evaluating the Process. In G.J.Cizek (éd.), *Setting Performance Standards : Concepts, Methods, and Perspectives* (pp. 89-116). Mahwah, NJ : Lawrence Erlbaum Associates.

Jaeger, R.M. (1989). Certification of student competence. In R. Linn (éd.), *Educational Measurement* (3e éd. pp. 485-514). New York : American Council on Education/ MacMillan.

Laveault, D., & Grégoire, J. (2002). *Introduction aux théories des tests en éducation et en psychologie* (2e éd.). Bruxelles : DeBoeck-Université.

Livingston, S.A., & Zieky, M.J. (1982). *Passing scores : A manual for setting standards on educational and occupational tests*. Princeton : Educational Testing Service.

Ministère de l'Éducation nationale (2006). *Le socle commun des compétences et des connaissances. Décret du 11 juillet 2006*. MEN-DGESCO. [http://www.education.gouv.fr/cid2770/le-socle-commun-de-connaissances-et-de-competences.html].

Mons, N., & Pons, X. (2006). *Les standards en éducation dans le monde francophone : Une analyse comparative*. Neuchâtel : Institut de recherche et de documentation pédagogique.

Norcini, J.J. (2003). Setting standards on educational tests. *Medical Education*, 37, 464–469.

Van der Linden, W.J. (1998). Optimal Assembly of Psychological and Educational Tests. *Applied Psychological Measurement*, *22*(3), 195-211.

Weller, S.C., & Romney, A.K. (1988). *Systematic Data Collection* (Qualitative Research Methods, 10). Newbury Park : Sage.