DÉTERMINATION DES STANDARDS MINIMAUX POUR ÉVALUER LES COMPÉTENCES DU SOCLE COMMUN

Nicolas Miconnet
MENESR-DEPP, bureau des études statistiques sur les élèves

Ronan Vourc'h
MENESR-DEPP, bureau de l'évaluation des élèves

Depuis 2012, la direction de l'évaluation, de la prospective et de la performance (DEPP) du ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche est en charge de la production d'indicateurs relatifs à la maîtrise des compétences du socle commun. Pour ce faire, elle a progressivement mis en place des évaluations standardisées auprès d'échantillons représentatifs d'élèves en fin de CM2 et en fin de troisième. Celles-ci permettent de recueillir des informations fiables et comparables dans le temps, alors que celles obtenues à partir de l'attribution des compétences du socle commun par les enseignants peuvent varier en fonction des caractéristiques individuelles des élèves, mais aussi de facteurs liés à leur établissement.
La mise au point de tels indicateurs impose d'établir des scores seuils permettant de distinguer ceux qui atteignent la compétence évaluée et ceux qui ne l'atteignent pas. Pour cela, on a recours à des méthodes qui confrontent les résultats issus des évaluations standardisées avec le jugement d'enseignants et d'experts sur le niveau des élèves et le contenu des évaluations.
Parmi les méthodes utilisées, celle dite « des marque-pages » se révèle la mieux adaptée à ce contexte d'évaluation. Elle permet, à l'exception des langues vivantes étrangères au collège, d'aboutir à des pourcentages de validation qui varient d'environ 70 % à 80 % selon les niveaux scolaires et les disciplines. Ces pourcentages ainsi déterminés diffèrent selon le secteur de scolarisation, le sexe et l'âge. Enfin, l'analyse du devenir d'un échantillon d'élèves de troisième vient conforter la démarche mise en œuvre pour déterminer les seuils de maîtrise.

Le socle commun a été inscrit dans la loi en 2005. Celle-ci arrête que *« la scolarité obligatoire doit au moins garantir à chaque élève les moyens nécessaires à l'acquisition d'un socle commun constitué d'un ensemble de connaissances et de compétences qu'il est indispensable de maîtriser pour accomplir avec succès sa scolarité, poursuivre sa formation, construire son avenir personnel et professionnel et réussir sa vie en société*[1]*»*.
Le socle commun est aujourd'hui détaillé en sept compétences – elles-mêmes définies au travers de plusieurs domaines –, qui sont évaluées par les enseignants à trois étapes de la scolarité : le palier 1 en fin de CE1 (uniquement les compétences 1, 3 et 6) ; le palier 2 en fin de CM2 et le palier 3 en fin de troisième ▸ **Encadré p. 143**.
Pour ce faire, ils disposent d'un livret définissant les différents domaines et compétences à valider, ainsi que de grilles d'aide à la décision.
Le décret relatif au socle commun[2] paru en 2006 a instauré la règle de non-compensation des compétences, considérant notamment que la maîtrise du socle commun ne pouvait être *« que globale, car les compétences qui le constituent, avec leur liste principale de connaissances, de capacités et d'attitudes, sont complémentaires et également nécessaires »*. À ce principe de non-compensation s'ajoute une règle de collégialité : la décision d'attribution des compétences pouvant, par exemple, intervenir à l'occasion d'un conseil de classe.

Ces modalités d'attribution des compétences du socle commun par les enseignants ne sont pas sans introduire une certaine variabilité en fonction des caractéristiques individuelles des élèves, mais aussi de facteurs liés à leur établissement. En effet, on sait que le jugement porté par les enseignants ne s'explique pas uniquement par les performances effectives des élèves, mais qu'il peut aussi être influencé par des éléments d'ordre contextuel, tel que le niveau de la classe par exemple [Bressoux et Pansu, 2003]. D'une façon générale, les facteurs susceptibles d'introduire des biais dans la notation sont bien connus et documentés par les études entreprises dès le début du XX^{e} siècle par l'école d'Henri Piéron. Concernant le socle commun, de récents travaux menés par la DEPP ont montré combien, pour des résultats identiques aux tests, les attestations des enseignants pouvaient varier selon des variables sociodémographiques et scolaires [Daussin, Rocher, Troseille, 2010]. C'est, par exemple, le cas des élèves « en retard », qui, à score et caractéristiques fixés, ont moins de chances de recevoir une attestation de compétences en français et en mathématiques.

Ces analyses rappellent toute la légitimité des évaluations standardisées[3] réalisées auprès d'échantillons représentatifs d'élèves pour répondre à la demande d'indicateurs comparables dans le temps tels que ceux exigés par la loi organique relative aux lois de finances (LOLF).

De 2007 à 2012, la DEPP a ainsi calculé une série d'indicateurs mesurant les proportions d'élèves maîtrisant les compétences « de base » en français et en mathématiques en fin d'école et en fin de collège, déclinées selon le secteur de l'établissement (public hors éducation prioritaire, public relevant de l'éducation prioritaire[4], privé). Ces épreuves

1. Loi d'orientation et de programme pour l'avenir de l'École (article 9) – Loi n° 2005-380 du 23 avril 2005.
2. Décret relatif au socle commun de connaissances et de compétences et annexe – Décret n° 2006-830 du 11 juillet 2006.
3. On définit ici les évaluations standardisées comme des dispositifs qui *« visent à mesurer les acquis cognitifs des élèves sur la base d'épreuves dont les conception, administration et correction sont uniformisées »* [Mons, 2009].
4. L'éducation prioritaire comprend les établissements qui relèvent du dispositif Éclair (Écoles, collèges et lycées pour l'ambition, l'innovation et la réussite) et du réseau de réussite scolaire (RRS).

LES SEPT COMPÉTENCES DU SOCLE COMMUN

Le socle commun s'organise actuellement en sept compétences : la maîtrise de la langue française (compétence 1), la pratique d'une langue vivante étrangère (compétence 2), les principaux éléments de mathématiques et la culture scientifique et technologique (compétence 3), la maîtrise des techniques usuelles de l'information et de la communication (compétence 4), la culture humaniste (compétence 5), les compétences sociales et civiques (compétence 6), l'autonomie et l'initiative (compétence 7).

Il convient de préciser que la présente étude se fonde sur la précédente version du socle. Le nouveau socle a été publié en avril 2015, après une consultation nationale lancée en 2014.

avaient été élaborées en 2005 et testées en 2006 [Rocher, Chesné, Fumel, 2008]. Au moment de leur conception, la notion de socle commun existait bien, mais aucun texte n'en définissait précisément le contenu. Les tests avaient donc été établis à partir d'éléments issus des programmes scolaires en relation avec le socle commun de connaissances et de compétences. Les résultats obtenus indiquaient une stabilité dans le temps en CM2 alors qu'en troisième ils mettaient en évidence une baisse des taux de maîtrise dans les établissements relevant de l'éducation prioritaire [*L'état de l'École,* 2012].

Les indicateurs mesurant les proportions d'élèves maîtrisant les compétences « de base » ont aujourd'hui laissé la place aux indicateurs de maîtrise des compétences du socle commun pour lesquels des expérimentations ont été entreprises dès 2009. La mise au point de tels indicateurs impose d'établir des scores seuils permettant de distinguer ceux qui atteignent le niveau souhaité de ceux qui ne l'atteignent pas. Mais comment déterminer ces seuils dont la définition n'est pas univoque ? C'est ce que cet article se propose d'étudier. Pour cela, il décrit trois méthodes de détermination des seuils proposées dans la littérature psychométrique. Il compare ensuite leur mise en œuvre respective dans le cadre de la détermination des seuils de maîtrise des compétences du socle commun évaluées par des tests standardisés en fin de CM2 et en fin de troisième. Il analyse enfin la pertinence des résultats obtenus au regard des caractéristiques des élèves, de leur environnement et de leur parcours scolaire.

ÉVALUATION DE LA MAÎTRISE DU SOCLE AUX PALIERS 2 ET 3 : QUELS OUTILS ?

Les données recueillies

C'est en 2011 que des tests standardisés ont été utilisés pour la première fois pour renseigner les indicateurs de maîtrise des compétences du socle commun de connaissances et de compétences en fin de CM2 et en fin de troisième. Cette année-là, les tests ont permis de fournir des indicateurs pour les compétences 1 et 3 en fin d'école et pour la compétence 1 en fin de collège. En 2012, cette démarche a été élargie aux compétences 2 et 5 à l'école et aux compétences 2, 3 et 5 au collège. En moyenne, les échantillons se composent d'environ 7 000 élèves par évaluation et les taux de réponse dépassent les 90 %. À l'école comme au collège, il s'agit

d'un sondage stratifié selon le secteur (public hors éducation prioritaire, public relevant de l'éducation prioritaire, privé). À l'école, les évaluations concernent tous les élèves de CM2 des établissements sélectionnés. Au collège, elles concernent tous les élèves d'une même classe de troisième. Pour les compétences 1 et 3, la dimension des échantillons a été augmentée en 2013 pour améliorer la précision des estimations. En particulier, le secteur de l'éducation prioritaire a été surreprésenté. En effet, pour ces compétences, les indicateurs doivent être déclinés pour le secteur public et le secteur privé, mais aussi, au sein du secteur public, pour les établissements relevant de l'éducation prioritaire (Éclair et RRS). En outre, les indicateurs de la LOLF demandent de renseigner les écarts observés entre les résultats des élèves de l'éducation prioritaire et ceux du secteur public hors éducation prioritaire.

Construction des épreuves

L'élaboration de ce dispositif d'évaluations standardisées tient compte de deux contraintes principales : le temps (les indicateurs doivent être mis à jour chaque année au mois de janvier) et le coût. Pour répondre à ces contraintes, la construction d'un test sous forme de QCM (questions à choix multiples) a été retenue. Les épreuves ont été élaborées spécifiquement pour chaque niveau évalué par des groupes de concepteurs composés d'enseignants et de conseillers pédagogiques en collaboration avec l'inspection générale. Ce format de questions assure une correction rapide, fiable et économique. En contrepartie, une interrogation sous cette forme exclut d'emblée l'évaluation de certains domaines de compétences. Par exemple, le domaine « dire » pour la compétence 1 aux paliers 2 et 3 et le domaine « écrire » pour la compétence 1 au palier 3.

Pour la compétence 5, à l'école et au collège, l'indicateur a été construit à partir d'items issus des évaluations Cedre (Cycle des évaluations disciplinaires réalisées sur échantillon)[5] pour l'histoire-géographie et l'éducation civique. Il convient donc de noter que cette évaluation ne concerne qu'une partie de la culture humaniste couverte par la compétence 5.

Les modèles de réponse à l'item

Une fois les données recueillies, une échelle de performances a été élaborée pour chaque évaluation en utilisant les modèles de réponse à l'item [voir notamment Grégoire et Laveault, 2002]. Ces derniers, développés dans la seconde moitié du XX^{e} siècle, modélisent la probabilité de réussite à un item en fonction de certaines de leurs caractéristiques, dont la difficulté, et en fonction du niveau de compétence des élèves.

Le modèle de réponse à l'item le plus simple a été développé en 1960 par Georg Rasch, le seul paramètre d'item à estimer étant la difficulté de l'item. Dans le cadre des évaluations des élèves conduites par la DEPP, un modèle de réponse à l'item à deux paramètres est utilisé. Ces deux paramètres d'items sont d'une part la difficulté et d'autre part la discrimination. Formellement, un modèle de réponse à l'item à deux paramètres s'écrit :

5. Les évaluations Cedre ont pour finalité de mesurer les atteintes des objectifs fixés par les programmes dans une discipline donnée et de comparer les performances des élèves dans le temps. Elles portent donc sur des contenus plus élargis que ceux évalués par les épreuves du socle.

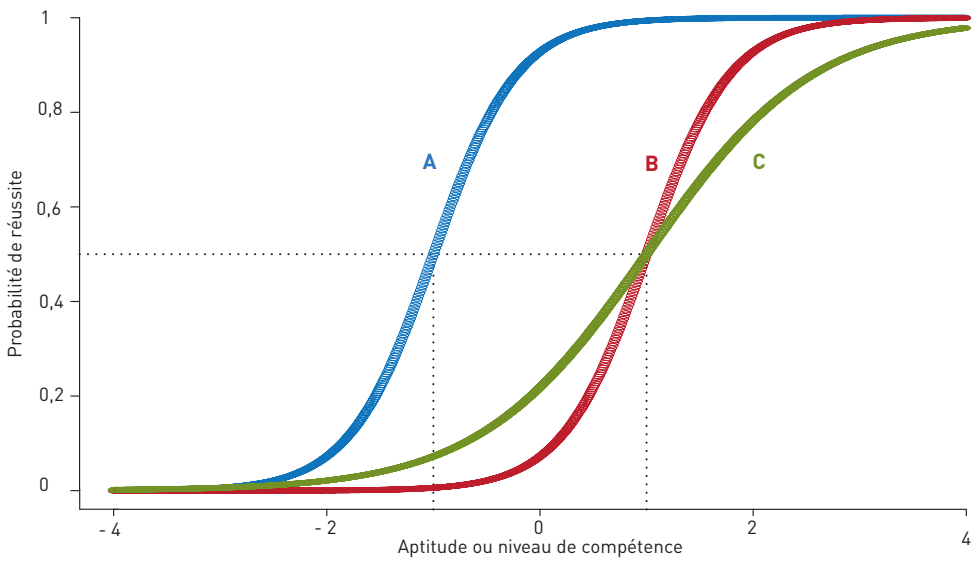
$$P_j(\theta_i) = \frac{e^{Da_j(\theta_i - b_j)}}{1 + e^{Da_j(\theta_i - b_j)}}$$

où $P_j(\theta_i)$ est la probabilité qu'un élève *i*, possédant une aptitude θ_i, réponde correctement à l'item *j*, b_j et a_j sont respectivement le paramètre de difficulté et le paramètre de discrimination de l'item *j*, et D un facteur d'échelonnement, constante fixée à 1,7 permettant ainsi de se rapprocher de la loi normale.

Cette dernière fonction peut être représentée par une courbe appelée courbe caractéristique de l'item qui met en relation l'aptitude ou la compétence de l'élève avec la probabilité de réussir un item donné. Généralement, l'aptitude θ des élèves est comprise entre – 4 et + 4. La **figure 1** illustre trois courbes caractéristiques pour des items dont le niveau de difficulté ou de discrimination varie. Par convention, la valeur qui représente la difficulté d'un item est la valeur de θ pour laquelle la probabilité de donner une réponse correcte est de 0,5, c'est-à-dire une chance sur deux. Ainsi, l'aptitude des élèves et les difficultés des items sont représentées sur une même échelle. Il s'agit là d'un avantage indéniable des modèles de réponses à l'item par rapport à l'approche usuelle (nombre de bonnes réponses), notamment lors de la détermination de seuils de maîtrise.

En effet, cette échelle permet de décrire les compétences d'un niveau donné en se servant des items correspondants comme descripteurs de ces compétences quand l'approche classique ne permet que de situer les élèves au sein de la distribution des scores sans renseigner sur les compétences attribuables aux niveaux où ils se situent. Sur la figure 1, la difficulté de l'item A est de – 1 et celle des items B et C de 1. Quant à la discrimination de l'item, elle est représentée par la pente de la courbe

Figure 1 Différentes courbes caractéristiques des items (A, B, C) selon la valeur du paramètre de difficulté et de discrimination



Source : MENESR-DEPP.

au point d'inflexion (plus elle est forte, plus l'item est discriminant). Ici, les items A et B présentent la même discrimination alors que l'item C est moins discriminant (la pente de la courbe étant plus faible).

Une fois le modèle théorique défini, il s'agit d'estimer les paramètres du modèle : la difficulté et la discrimination de chaque item et l'aptitude de chaque élève. L'estimation simultanée de tous ces paramètres s'avère complexe et dépasse le cadre de cet article (le lecteur intéressé par cette problématique pouvant se reporter à Rocher, dans ce numéro, p. 37).
Notons que l'aptitude des élèves, telle que définie par le modèle de réponse à l'item est très fortement corrélée avec le score (défini par le nombre de bonnes réponses à l'évaluation). Ainsi, le modèle de réponse à l'item donne des résultats proches de la théorie classique consistant à sommer le nombre de bonnes réponses. L'aptitude estimée par le modèle de réponse à l'item le plus simple (un seul paramètre, la difficulté) est d'ailleurs en bijection avec le nombre de bonnes réponses. L'ajout du paramètre de discrimination permet de casser ce lien univoque entre nombre de bonnes réponses et aptitude estimée par le modèle. En quelque sorte, ce modèle à deux paramètres revient à pondérer les items selon leurs discriminations (à titre d'illustration, si deux élèves ont le même nombre de bonnes réponses, mais obtenues sur des items différents, celui ayant réussi les items les plus discriminants aura une aptitude plus élevée).

Les modèles de réponse à l'item présentent aussi l'avantage de pouvoir situer sur une même échelle de compétences des élèves qui n'ont pas nécessairement été soumis aux mêmes items. C'est notamment le cas lorsque l'on utilise la méthode des « cahiers tournants » pour évaluer un nombre important d'items sans allonger le temps de passation. Cette méthode consiste à répartir les items dans des cahiers différents qui comportent des items communs.

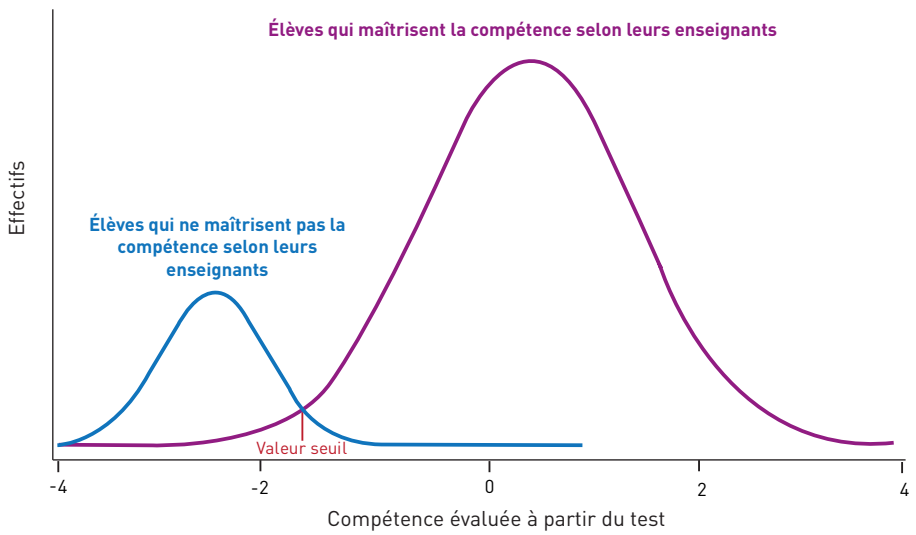
TROIS MÉTHODES POUR DÉTERMINER LE SEUIL DE MAÎTRISE DES COMPÉTENCES

Le calcul des taux de réponses correctes aux items ne permet donc pas d'estimer directement la proportion des élèves qui maîtrisent ces compétences, car le « degré de maîtrise » n'a pas encore été défini à ce stade. En effet, les items peuvent être de difficulté très variable, quand bien même ils portent sur une compétence dite du « socle ». De ce fait, pour être considéré comme un élève qui maîtrise les compétences du socle, 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 on considérera qu'il maîtrise les compétences du socle. La détermination de ce seuil ne s'impose pas d'elle-même.
Différentes méthodes ont été proposées dans la littérature pour déterminer ce point de bascule entre maîtrise et non-maîtrise. Quelle que soit la méthode utilisée, ce point de césure est défini par la confrontation entre les attentes des enseignants ou d'un groupe d'experts et les résultats statistiques. Trois de ces méthodes ont été retenues dans le cadre de la détermination des seuils de maîtrise des compétences du socle commun. La première repose sur le jugement des enseignants sur leurs élèves et les deux autres sur le jugement des items par des enseignants et des experts.

La première méthode possible pour déterminer le point de césure est celle dite des contrastes décrite par Grégoire et Laveault [2002]. Cette méthode confronte le jugement des enseignants sur les élèves avec les résultats des élèves aux tests standardisés. Il est demandé aux enseignants d'indiquer pour chacun de leurs élèves si celui-ci maîtrise ou non la compétence évaluée. On répartit alors les élèves en deux groupes, selon leur maîtrise supposée de la compétence évaluée, et reporte alors sur un même graphique la distribution de l'aptitude des élèves, estimée à partir du modèle de réponse à l'item pour chacun des deux groupes.
Pour que cette méthode puisse être utilisée, les deux distributions doivent être suffisamment disjointes. Un cas d'école, où les deux distributions se croisent, est reporté sur la **figure 2**. Le point d'intersection correspond alors au point de césure entre maîtrise et non-maîtrise (valeur seuil sur la figure 2). Cette méthode présente notamment l'avantage de pouvoir comparer un dispositif de classement des élèves par les enseignants – dont nous avons déjà présenté les limites – avec les résultats d'évaluations standardisées.

Une deuxième méthode peut aussi être utilisée pour déterminer des seuils de maîtrise. Il s'agit d'une adaptation de la méthode dite des « zones de jugement », ou Hofstee, décrite par Bunch et Cizek [2007]. En théorie, cette méthode utilise le score, c'est-à-dire le nombre de bonnes réponses au test, mais elle peut également être adaptée en utilisant l'aptitude estimée par le modèle de réponse à l'item. C'est notamment le cas lorsque l'évaluation utilise un nombre important d'items sans que les élèves les passent nécessairement tous (méthode des « cahiers tournants » par exemple).
La présentation ci-dessous fait référence à l'utilisation de la méthode Hofstee dans le cas où le score est défini en termes de réponses correctes.
Comme avec la méthode des contrastes, l'information apportée par les enseignants des élèves est utilisée, mais également celle tirée des experts ayant participé à la

Figure 2 **Exemple théorique de la méthode des contrastes**



Source : MENESR-DEPP.

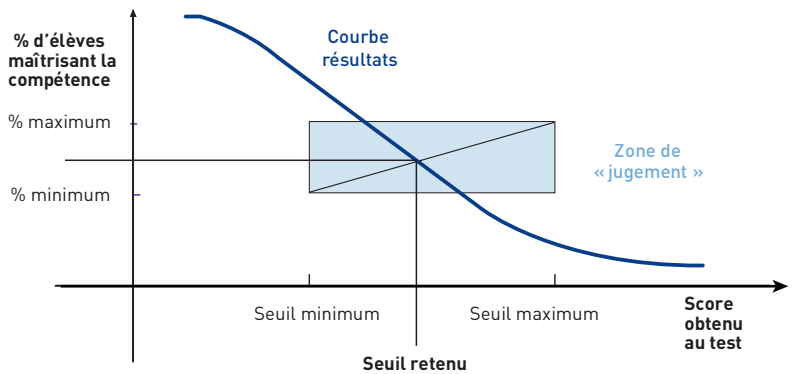
conception du test. Deux questions leur sont posées. La première leur demande de se prononcer, en se basant sur leur expérience professionnelle, sur le pourcentage minimum et le pourcentage maximum d'élèves maîtrisant, selon eux, la compétence en question. Il leur est ensuite demandé de déterminer à partir de quel score les élèves maîtrisent cette compétence. Puisqu'il est difficile de se déterminer sur le score, les enseignants et les experts travaillent sur les items. Plus précisément, ils doivent classer chaque item du test en trois catégories (A, B, C) selon le degré d'importance accordé à la réussite de l'item pour la validation de la compétence ▸ **Tableau 1**. Le nombre minimum de bonnes réponses pour maîtriser la compétence, correspondant à une notation souple, est obtenu en sommant les items classés « A » et le score maximum, correspondant à une notation plus sévère, est obtenu en sommant les items classés « A » ou « B ».
À chaque expert correspond alors une zone de jugement ▸ **Figure 3**. Cette zone témoigne à la fois des attentes et du niveau d'exigence de chaque expert. À partir des données issues de l'expérimentation, il est possible de calculer pour chaque score le pourcentage d'élèves situés au-delà 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.

▸ Tableau 1 Classification des items en trois catégories (méthode Hofstee)

Selon vous, les élèves maîtrisant la compétence…	… doivent absolument réussir cet item (A)	… devraient pouvoir réussir cet item (B)	… ne doivent pas forcément réussir cet item (C)
Item 1			
Item 2			
…			
Item N			

Source : MENESR-DEPP.

▸ Figure 3 Méthode des « zones de jugement »

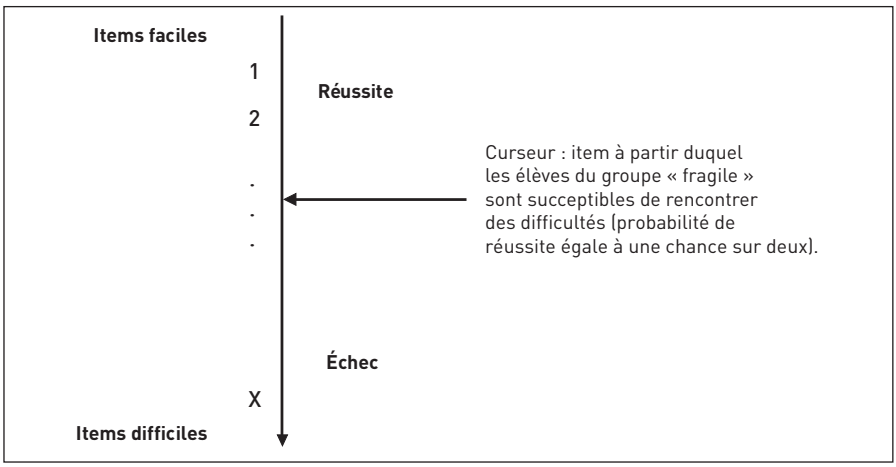


Source : MENESR-DEPP.

Enfin, une troisième méthode est utilisée pour déterminer le point de césure entre maîtrise et non-maîtrise des compétences du socle. Il s'agit de la méthode des marque-pages (bookmarks). Simple à mettre en œuvre et plus couramment utilisée que les précédentes, elle est aussi décrite par Bunch et Cizek [2007]. À partir des résultats de l'estimation du modèle de réponse à l'item, les items sont classés par ordre croissant selon la valeur de leur paramètre de difficulté. Les items du début de la liste correspondent à des items faciles, c'est-à-dire très réussis, et ceux de la fin sont plus difficiles ▸ **Figure 4**.

Comme nous l'avons vu, grâce aux modèles de réponse à l'item, les paramètres de difficulté des items et les niveaux de compétences des élèves sont positionnés sur une même échelle. Plus précisément, chaque item est positionné à un niveau tel que les jeunes situés à ce niveau ont une chance sur deux de réussir cet item et ceux qui se situent en dessous ont une probabilité de réussite plus faible. Il est alors demandé à chaque expert d'imaginer un groupe d'élèves fragile (élèves situés à la frontière entre le groupe « maîtrise » et le groupe « non-maîtrise ») et d'indiquer l'item (ou une zone réduite d'items) à partir duquel (de laquelle) ces élèves sont susceptibles de rencontrer des difficultés et d'avoir une chance sur deux de réussir. Ainsi, pour chaque expert, un pourcentage (s'il a fourni un seul item comme « zone de bascule ») ou un intervalle (si plusieurs items fournis) de maîtrise de la compétence est déterminé.
La seconde phase consiste à faire converger les attentes des différents experts pour aboutir à un consensus autour de la définition d'un seuil de maîtrise au regard des résultats obtenus à partir de la mise en œuvre de ces trois méthodes.

▸ **Figure 4** **Méthode des marque-pages**



Source : MENESR-DEPP.

DÉTERMINATION DES POINTS DE CÉSURE POUR L'ÉVALUATION DES COMPÉTENCES DU SOCLE COMMUN

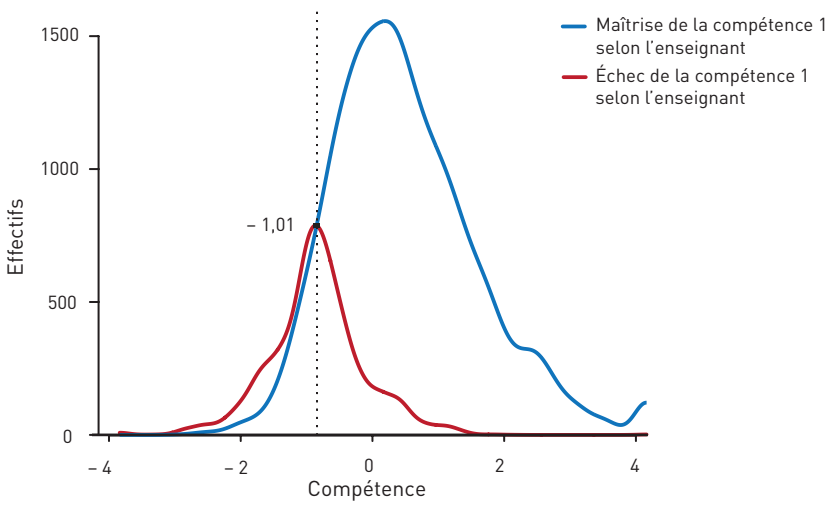
Selon les évaluations, le point de césure entre maîtrise et non-maîtrise a été défini en 2011 ou en 2012 en mettant en application les trois méthodes présentées précédemment. Selon les évaluations, le point de bascule a aussi été défini en 2011 ou en 2012. Il a été repris à l'identique les années suivantes, les évaluations étant composées exclusivement d'items communs dont les paramètres sont supposés invariants entre les différentes années.

Exemple pour la compétence 1 à l'école

Le point de césure pour la compétence 1 à l'école, c'est-à-dire le niveau d'aptitude à partir duquel un élève de CM2 maîtrise la compétence, a été déterminé en 2011 sur la base d'un échantillon d'environ 7 500 élèves.

La méthode des contrastes a été utilisée dans un premier temps. On a donc demandé aux enseignants de l'échantillon d'indiquer pour chacun de ses élèves si celui-ci maîtrise ou non la compétence 1, sur la base du travail effectué tout au long de l'année scolaire. La distribution des scores des élèves pour chacun des deux groupes (maîtrise / non maîtrise) a ensuite été représentée en fonction de cette information fournie par chaque enseignant ▸ **Figure 5**. Le point d'intersection entre ces deux distributions correspond à un niveau de compétence de – 1,01. Il conduit à un niveau de validation de 83 %. Cependant, le recouvrement des deux distributions ne permet pas de classer avec confiance un élève dans l'une ou l'autre des deux catégories à partir de son score au test. On ne saurait donc déterminer un seuil pertinent à partir de ces seuls résultats.
Quant à la méthode Hofstee (zone de jugement), utilisée sur la base du score estimé par le modèle de réponses à l'item, elle conduit à des pourcentages de validation

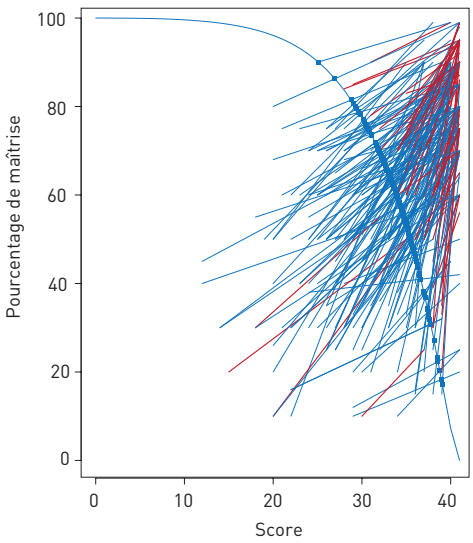
▸ **Figure 5** **Méthode des contrastes pour la compétence 1 à l'école en 2011**



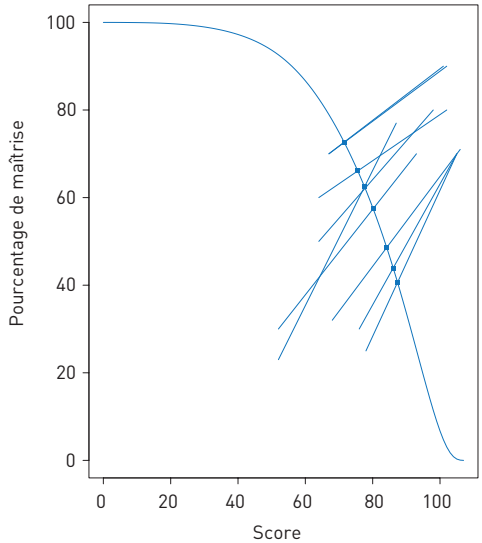
Source : MENESR-DEPP.

de la compétence 1 très variables à partir des données recueillies auprès des 263 professeurs des classes de l'échantillon ▸ **Figure 6**. La médiane est de 57 %, mais pour certains enseignants la proportion d'élèves maîtrisant la compétence ne dépasse pas 20 %, alors que pour d'autres, elle se situe au-delà de 80 %. Une des explications à cette dispersion, validée par des analyses secondaires entreprises autour de la méthode Hofstee, réside dans les différences de jugement des enseignants selon le secteur de leur établissement. Les résultats obtenus à partir des travaux effectués par neuf experts indiquent eux aussi des disparités de jugement comparables à celles observées parmi les enseignants de l'échantillon ▸ **Figure 7**. D'une manière générale, la mise en application de cette méthode ne permet donc pas d'obtenir un consensus compte tenu de la dispersion des taux de maîtrise qui en résultent.

▸ Figure 6 Méthode des « zones de jugement » pour la compétence 1 à l'école à partir des données recueillies auprès des professeurs de l'échantillon en 2011



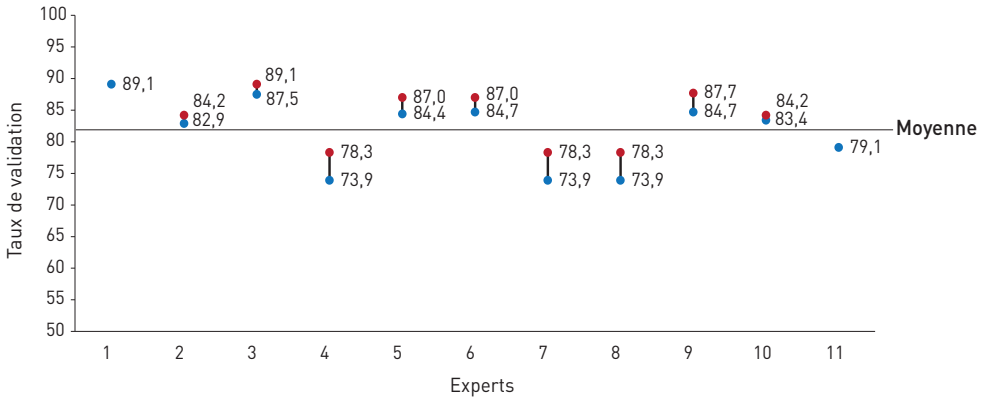
▸ Figure 7 Méthode des « zones de jugement » pour la compétence 1 à l'école à partir des données recueillies auprès des experts en 2011



Note : l'axe des abscisses est le score à l'évaluation, l'axe des ordonnées est le pourcentage d'élèves maîtrisant la compétence 1. La courbe représente, pour chaque score, le pourcentage d'élèves situés au-delà de ce score. Le point d'intersection entre cette courbe et la diagonale de la zone de jugement fournit le score-seuil retenu. Seules les diagonales représentées en bleu coupent la courbe.
Source : MENESR-DEPP.

Le seuil de coupure entre maîtrise et non-maîtrise a été obtenu par la méthode des marque-pages appliquée avec onze experts, dont les neuf ayant appliqué la méthode des « zones de jugement ». Les items de l'évaluation ont été classés par ordre de difficulté croissante (tri sur les paramètres de difficulté estimés par le modèle de réponse à l'item) et chaque expert a déterminé un item (ou un intervalle d'items) comme point de césure. La variabilité inter-experts est marquée, mais est tout de même sensiblement plus faible que celle observée lors de l'utilisation de la méthode des « zones de jugement ». Le pourcentage de maîtrise de la compétence sur les données 2011 oscille entre 73,9 % et 89,1 %. La moyenne est à 82,6 % ▸ **Figure 8**.

Figure 8 **Méthode des marque-pages pour la compétence 1 à l'école en 2011**



Note de lecture : l'intervalle d'items choisi par l'expert numéro 4 conduit à un taux de validation de la compétence 1 variant de 73,9 % à 78,3 %.
Source : MENESR-DEPP.

Conformément à la méthode, une discussion s'est alors engagée pour dégager un éventuel consensus. Finalement, l'item retenu comme point de césure présente une difficulté de - 0,9 conduisant à une maîtrise de la compétence 1 par 79 % des élèves. Les experts ont considéré que les élèves fragiles commençaient à échouer sur des items ayant trait au traitement de l'implicite. Ces élèves parviennent à mettre en relation les informations explicites d'un texte, mais peinent à s'engager dans un processus d'interprétation.

Les résultats de 2012 et de 2013 s'appuient sur les enseignements tirés lors de la session précédente. En effet, les items de ces évaluations ont été extraits de l'évaluation 2011. Dans le cadre de la théorie du modèle de réponse à l'item, les paramètres des items sont supposés invariants entre ces deux sessions [Rocher, dans ce numéro, p. 37]. Plus précisément, le paramètre de difficulté de chaque item est supposé être constant. Le point de césure déterminé en 2011 est de nouveau appliqué les années suivantes et conduit, en 2013, à un taux de validation de 79,8 % pour la compétence 1 à l'école **Tableau 2 p. 155**.

Pour les autres compétences, une démarche similaire à celle qui vient d'être exposée a permis de déterminer un item-seuil défini comme point de bascule entre maîtrise et non-maîtrise. Les trois méthodes présentées (contrastes, zone de jugement, marque-pages) ont aussi été mises en œuvre, mais ce sont les résultats fournis par la méthode des marque-pages qui se sont révélés les plus probants. En effet, la méthode des contrastes n'a pas toujours permis d'aboutir à des résultats utilisables. Pour la majorité des compétences, l'aptitude des élèves ne différait pas suffisamment selon que les enseignants des élèves validaient ou non la compétence. Cependant, même lorsque cette méthode ne peut pas être utilisée, elle n'en est pas moins utile puisque son échec tend une nouvelle fois à prouver que la validation des compétences par les enseignants n'est pas sans poser question et justifie *a posteriori* le recours à une évaluation standardisée pour déterminer la proportion de maîtrise du socle. Quant aux résultats obtenus à partir de la méthode des « zones de

jugement », ils ont le plus souvent permis d'éclairer les discussions des experts lors de la phase de recherche de consensus de la méthode des marque-pages.
Concernant la pratique d'une langue vivante étrangère au collège (compétence 2), dont le taux de maîtrise est plus faible, la méthode des marque-pages a aussi été appliquée à d'autres données d'évaluation ▸ **Encadré**. De manière générale, il est notable que, malgré la variabilité dans les attentes des enseignants et des experts, l'item retenu comme point de césure dans toutes les compétences correspond au passage de l'explicite à l'implicite.

DÉTERMINATION DU SEUIL DE MAÎTRISE POUR LA COMPÉTENCE 2 AU COLLÈGE

Pour la compétence 2 au collège, l'indicateur a été établi pour la première fois en 2012 à partir des réponses apportées par un échantillon d'élèves à des épreuves standardisées portant sur leurs compétences en anglais (compréhension orale et compréhension écrite).
Pour la compréhension orale, le pourcentage de maîtrise retenu (26,9 %) est très proche de celui observé dans les évaluations de l'Étude européenne sur les compétences en langues (ESLC) en fin de scolarité obligatoire (jeunes de 14 à 16 ans) [Bessonneau et Verlet, 2012]. Les résultats de cette enquête montrent en effet que, pour la compréhension orale, seuls 26 % des élèves français maîtrisent au moins le niveau A2[6]. En revanche, pour la compréhension écrite, on observe un écart significatif : 40,4 % à partir des épreuves du socle contre 22,8 % dans ESLC.
Au regard de ces résultats, il a été demandé au groupe d'experts de participer à une nouvelle séance de mise en application de la méthode des marque-pages aux items de compréhension écrite de l'épreuve ESLC. Même si cette évaluation ne repose pas sur les mêmes protocoles et méthodologies, elle comporte un plus grand nombre d'items que celle du socle, permettant ainsi d'affiner la détermination du point à partir duquel s'opère le basculement entre maîtrise et non-maîtrise. Ce nouveau seuil a permis d'aboutir à un taux de validation de la compétence 2 assez proche de celui observé à partir de l'épreuve du socle (43 %), ce qui témoigne d'une cohérence dans le jugement apporté par les experts. Ce résultat demeure nettement supérieur à celui observé dans ESLC où la définition des seuils de maîtrise a été effectuée par des experts de différents pays[7], méthode susceptible d'introduire des biais liés aux limites de la comparabilité des niveaux de difficultés des items entre pays.
L'indicateur de maîtrise de la compétence a ensuite été construit en faisant la moyenne entre la compréhension écrite et la compréhension orale. Ainsi, en 2012, 35,1 % des élèves maîtrisent la compétence 2 en anglais en fin de collège. En 2013, en appliquant le même seuil, ils sont 36,6 % ▸ **Tableau 2 p.155**.

6. Sur l'échelle européenne (CECRL), le niveau A2 – exigé pour la validation du socle commun de connaissances et de compétences – correspond à la mention « utilisateur élémentaire de l'anglais ». À ce niveau, l'utilisateur élémentaire peut comprendre des phrases isolées et des expressions fréquemment utilisées en relation avec des domaines immédiats de priorité (par exemple, informations personnelles et familiales simples, achats, etc.) ; peut communiquer lors de tâches simples et habituelles ne demandant qu'un échange d'informations simple et direct sur des sujets familiers et habituels ; peut décrire avec des moyens simples son environnement immédiat et évoquer des sujets qui correspondent à des besoins immédiats.

7. Les méthodologies utilisées dans le cadre de ESLC sont décrites dans le rapport technique du consortium : https://crell.jrc.ec.europa.eu

ANALYSE DES RÉSULTATS ET PERSPECTIVES

Les démarches entreprises lors de la détermination des seuils de maîtrise avaient pour objectif d'obtenir des résultats fiables et destinés à être comparés dans le temps, mais aussi de contribuer à la réflexion méthodologique sur le sujet.
En 2013, selon la procédure présentée ci-dessus, en fin de CM2, 79,8 % des élèves maîtrisent la compétence 1 et 70,9 % la compétence 3 ▸ **Tableau 2**. En fin de troisième, ils sont respectivement 79,2 % et 78,3 %. À l'école, les garçons sont moins nombreux à maîtriser la compétence 1 que les filles (77,1 % contre 82,6 %, l'écart étant significatif à un seuil inférieur à 1/1000). La différence s'accentue au collège (72,3 % contre 85,9 %).
Pour la compétence 3, la différence selon le sexe (là aussi, statistiquement significative) s'inverse légèrement à l'école (72,5 % des garçons contre 69,3 % des filles), mais les filles devancent les garçons au collège (80,5 % des filles contre 76,2 % des garçons). Il peut être souligné que les disparités entre filles et garçons avaient déjà été obtenues lors des épreuves réalisées en 2012, mais la dimension plus faible des échantillons ne permettait pas de conclure systématiquement à des différences significatives. La maîtrise plus fréquente des compétences du socle dans les disciplines scientifiques (mathématiques, physique-chimie, sciences de la vie et de la Terre, technologie) en fin de collège par les filles se retrouve également sur les notes au contrôle continu du diplôme national du brevet (DNB). En 2013, la moyenne des filles était supérieure à celle des garçons de 0,2 point en physique-chimie, de 0,4 point en mathématiques, de 0,7 point en technologie et de 0,8 point en sciences de la vie et de la Terre. De tels écarts étaient également constatés à la session 2012 du DNB. Sauf à considérer qu'il existe un biais de notation important en faveur des filles, les résultats de maîtrise du socle selon le sexe des élèves sont corroborés par les résultats exhaustifs au contrôle continu du DNB. Pour la compétence 2 (pratique d'une langue vivante étrangère), les performances des filles sont aussi supérieures à celles des garçons, que ce soit à l'école ou au collège. Ces résultats confirment les observations issues des évaluations Cedre [Bessonneau, Beuzon, Boucé *et alii*, 2012 ; Bessonneau, Beuzon, Daussin *et alii*, 2012]. En revanche, elles sont proportionnellement moins nombreuses à maîtriser la compétence 5 (culture humaniste) à l'école et au collège. Ici aussi, les résultats sont à rapprocher des analyses effectuées à partir des évaluations Cedre en histoire-géographie et en éducation civique. Ceux-ci indiquent que les garçons sont plus nombreux parmi les élèves de haut niveau. En revanche, les proportions d'élèves dans les groupes de bas niveau sont pratiquement les mêmes chez les filles et les garçons [Garcia et Pastor, 2013 ; Garcia et Krop, 2013]. Les résultats des deux évaluations ne se recoupent donc que partiellement.

Les élèves en retard représentent en moyenne 12 % des élèves de fin de CM2 interrogés. En fin de troisième, ils sont un peu plus d'un quart. Que ce soit en fin d'école ou en fin de collège, la proportion d'élèves qui maîtrisent les compétences évaluées est nettement moins importante parmi les élèves « en retard » que parmi les élèves « à l'heure ». La différence entre les deux groupes d'élèves est particulièrement marquée à l'école où elle se situe autour de 40 points de pourcentage pour les compétences 1 et 3 et autour de 30 points pour la compétence 2. Au collège, les différences sont un peu moins élevées, mais l'écart entre les deux groupes reste tout de même important.

Tableau 2 Proportion d'élèves qui maîtrisent les compétences du socle commun en 2013 (en %)

Compétence	Public hors EP	RRS	Éclair	Privé	Public	Garçons	Filles	« En retard »	« À l'heure »	Ensemble
C1 école	81,8	69,8	62,5	87,4	78,6	77,1	82,6	46,2	84,7	79,8
C3 école	74,2	56,5	47,3	79,2	69,6	72,5	69,3	33,1	76,3	70,9
C2 école	81,6	61		85,9	78,1	75,0	83,8	54,3	82,8	79,3
C5 école	70,9	59,1	53,5	78,9	70,6	74,9	69,1	35,5	76,6	72,1
C1 collège	80,6	70,1	56,7	87,9	78,7	72,3	85,9	55,6	86,5	79,2
C3 collège	80,4	67,7	51,5	88,1	77,9	76,2	80,5	52,7	86,6	78,3
C2 collège	36,1	20,9		48,3	33,4	33,2	39,9	15,4	41,9	36,6
C5 collège	67,0	55,1	39,3	79,5	66,6	71,1	68,4	40,7	77,9	69,8

Lecture : en fin de CM2, 81,8 % des élèves des établissements public hors éducation prioritaire maîtrisent la compétence 1.
Note : comme toutes les estimations réalisées sur échantillons, les résultats du tableau présentent des intervalles de confiance. Selon les compétences et la variable étudiée, les intervalles de confiance sont compris dans une fourchette allant de +/- 1,2 à +/- 4,1.
Champ : France métropolitaine + DOM.
Source : MENESR-DEPP.

Les données permettent aussi de comparer les niveaux de maîtrise selon les secteurs d'enseignement et, plus particulièrement, entre les établissements publics relevant de l'éducation prioritaire et les autres. On constate ainsi qu'en fin de CM2, comme en fin de troisième, les élèves scolarisés en Éclair maîtrisent moins bien que les autres élèves les compétences 1, 3 et 5 du socle commun. Par exemple, si 62,5 % des élèves de CM2 des écoles du programme Éclair maîtrisent la compétence 1 du socle, ils sont 69,8 % dans les écoles RRS et 81,8 % dans les écoles publiques hors éducation prioritaire.

L'origine sociale de l'élève influence également la maîtrise des compétences du socle. Les enfants de cadres ou de personnes exerçant une profession intellectuelle supérieure maîtrisent très souvent les compétences 1 et 3 du socle (respectivement 91,3 % et 92,3 % pour les compétences 1 et 3 au collège). Cette proportion de maîtrise est plus faible pour les enfants ayant une origine sociale défavorisée[8] (75,8 % maîtrisent la compétence 1 ; 73,9 % maîtrisent la compétence 3).

La situation de la grande majorité des élèves de troisième à la rentrée scolaire suivant l'évaluation a pu être déterminée à partir du système d'information du ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche : 70,5 % d'entre eux étaient en seconde générale et technologique, 22,3 % en seconde professionnelle, 4,4 % en CAP et 2,7 % de nouveau en classe de troisième.
Il s'avère que les résultats aux évaluations présentent un caractère prédictif intéressant. En effet, 9 élèves sur 10 ayant poursuivi en seconde générale et technologique ont été évalués positivement à l'évaluation en troisième (90,7 % pour la compétence 1 ; 91 % pour la compétence 3). Les élèves qui sont toujours en classe de troisième à la rentrée suivant l'évaluation maîtrisent moins fréquemment la compétence 1 ou la compétence 3 (72,8 % pour la compétence 1 ; 71,8 % pour la compétence 3). La maîtrise de ces compétences est encore moins fréquente

8. Ouvriers, retraités ouvriers et employés, inactifs (chômeurs et inactifs n'ayant jamais travaillé).

pour les élèves orientés dans l'enseignement professionnel (66,5 % en seconde professionnelle et 53 % en CAP pour la compétence 1 ; 65 % en seconde professionnelle et 44,3 % en CAP pour la compétence 3). Ces résultats sur le devenir des élèves confèrent une forme de validité aux évaluations standardisées administrées par la DEPP.

Les résultats présentés dans cet article rappellent aussi tout l'intérêt de la confrontation des méthodes de définition des seuils de maîtrise. Parmi elles, c'est la méthode des marque-pages qui s'est révélée la plus adaptée. De ce fait, elle sera privilégiée dans les prochains travaux portant sur la détermination des seuils de maîtrise qui vont de nouveau être entrepris par la DEPP sur les compétences 1 et 3. Ils porteront tout d'abord sur le palier 1 pour lequel des évaluations se sont tenues en fin de CE1 en mai 2014. Ensuite, ils s'appuieront sur des évaluations qui seront effectuées respectivement en début de sixième (2015) et en fin de troisième (2016).

BIBLIOGRAPHIE

Bessonneau P., Verlet I., 2012, « Les compétences en langues étrangères des élèves en fin de scolarité obligatoire. Premiers résultats de l'Étude européenne sur les compétences en langues 2011 », *Note d'information*, n° 12.11, MEN-DEPP.

Bessonneau P., Beuzon S., Boucé S., Daussin J.-M., Garcia É., Levy M., Marchois C., Trosseille B., 2012, « L'évolution des compétences en langues des élèves en fin de collège de 2004 à 2010 », *Note d'information*, n° 12.05, MENJVA-DEPP.

Bessonneau P., Beuzon S., Daussin J.-M., Garcia É., Levy M., Marchois C., Trosseille B., 2012, « L'évolution des compétences en langues des élèves en fin d'école de 2004 à 2010 », *Note d'information*, n° 12.04, MENJVA-DEPP.

Brennan R., Kolen M., 2004, *Test Equating, Scaling, and Linking. Methods and Practices,* 2nd edition, New York, Springer, 548 p.

Bressoux P., Pansu P., 2003, *Quand les enseignants jugent leurs élèves*, Paris, Presses universitaires de France, 190 p.

Bunch M., Cizek G., 2007, *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests,* London, Thousand Oaks, Sage Publications, 352 p.

Daussin J.-M., Rocher T., Trosseille B., 2010, « L'attestation de la maîtrise du socle commun est-elle soluble dans le jugement des enseignants ? » *Éducation & formations*, n° 79, MENJVA-DEPP, p. 45-58.

Garcia É., Pastor J.-M., 2013, « CEDRE 2012 histoire-géographie et éducation civique en fin d'école primaire : grande stabilité des acquis depuis six ans », *Note d'information*, n° 13.10, MEN-DEPP.

Garcia É., Krop J., 2013, « CEDRE 2012 histoire-géographie et éducation civique : baisse des acquis des élèves de fin de collège depuis six ans », *Note d'information*, n° 13.11, MEN-DEPP.

Grégoire J., Laveault D., 2002, *Introduction aux théories des tests en psychologie et en sciences de l'éducation,* 2e édition, Bruxelles, De Boeck, 377 p.

L'état de l'École, 2012, Paris, MEN-DEPP.

Le Donné N., Rocher T., 2010, « Une meilleure mesure du contexte socio-éducatif des élèves et des écoles. Construction d'un indice de position sociale à partir des professions des parents », *Éducation & formations,* n° 79, MENJVA-DEPP, p. 103-115.

Mons N., 2009, « Effets théoriques et réels des politiques d'évaluation standardisée », *Revue française de pédagogie,* n° 169, p. 99-140.

Rocher T., Chesné J.-F., Fumel S., 2008, « Méthodologie de l'évaluation des compétences de base en français et en mathématiques en fin d'école et en fin de collège », *Note d'information,* n° 08.37, MEN-DEPP.