L'ÉVALUATION DES COMPÉTENCES DES ÉLÈVES : UN PROCESSUS DE MESURE SINGULIER

*Thierry Rocher**

Les évaluations standardisées des compétences élèves se sont aujourd'hui imposées dans le débat public sur l'éducation. Leurs résultats sont utilisés à la fois pour éclairer les politiques éducatives mais également pour construire des indicateurs statistiques qui alimentent des outils de suivi mobilisés par les acteurs du système.

La mise au point de ces instruments suit une démarche assez singulière, du fait que l'objet de mesure – la compétence – est lui-même une construction, qui ne pré-existe pas à l'opération de mesure elle-même. Il en découle des procédures et des modélisations spécifiques, relevant de la psychométrie, domaine assez méconnu en France, alors qu'il y trouve une part importante de ses origines.

Cet article vise à donner un tour d'horizon des méthodes utilisées pour mesurer les compétences des élèves, en insistant sur leurs spécificités et en pointant quelques perspectives enrichissantes.

Standardised assessments of students' competences have now become an integral part of the public debate on education. Their results are used both to inform educational policies and to construct statistical indicators that feed into the monitoring tools used by educational stakeholders.

Building these instruments follows a rather singular approach, in that the object of measurement – the competence – is itself a construction, which does not pre-exist the measurement operation itself. This results in specific procedures and modelling, which are part of psychometrics, a field that is relatively unknown in France, even though it has a large part of its origins there.

This article aims to provide an overview of the methods used to measure students' competences, emphasising their specific features and pointing out some enriching perspectives.

* Adjoint au sous-directeur des Évaluations et de la performance scolaire, Depp, *thierry.rocher@education.gouv.fr*

MESURE DES COMPÉTENCES : QUAND L'INSTRUMENT FAIT NAÎTRE SON OBJET

Quel est le niveau des élèves ? Comment évolue-t-il ? Comment les élèves français se situent-ils par rapport à leurs camarades d'autres pays ? Au-delà de l'école, quel est le niveau d'adéquation entre compétence et emploi ? Ces questions font, et c'est bien naturel, l'objet de beaucoup d'attentions. Elles engendrent commentaires et débats, tant médiatiques qu'entre professionnels et spécialistes, pas toujours exempts de préjugés[1]. Elles appellent donc des réponses objectives, statistiquement fondées, et constituent ainsi un enjeu pour la statistique publique.

Mais qu'est-ce qu'un niveau de compétence, comment le mesurer ?

De multiples aspects du processus de mesure sont relativement classiques dans le champ des enquêtes statistiques. Cependant, la nature même de la variable mesurée distingue de façon très singulière ces évaluations, car les compétences ne s'observent pas directement. Seules les manifestations des compétences sont observables. Ce seront par exemple les résultats obtenus à un test standardisé : l'existence supposée de la compétence visée est alors matérialisée dans la réussite au test – plus précisément dans l'agrégation des résultats de chaque *item* du test.

> *« Les compétences ne s'observent pas directement. Seules les manifestations des compétences sont observables. »*

D'une certaine manière, on pourrait avancer que c'est l'opération de mesure elle-même qui définit concrètement l'objet de la mesure. D'ailleurs, dans le domaine de la psychométrie, qui s'intéresse à la mesure de dimensions psychologiques en général, c'est le terme de *« construit »* qui est employé pour désigner l'objet de la mesure : l'intelligence logique, la lecture, la mémoire de travail, etc. D'où le célèbre pied de nez attribué à Alfred Binet[2], l'inventeur de la discipline en France au début du XX^{e} siècle : à la question « *Qu'est-ce que l'intelligence ?* », il aurait répondu « *C'est ce que mesure mon test.* ».

Bien entendu, une majorité de statistiques peut être considérée comme une construction, basée sur des conventions[3]. Cependant, des distinctions avec celles ayant trait à l'évaluation de compétences peuvent être opérées, en lien avec le caractère tangible de la variable visée.

Par exemple, la réussite scolaire peut être appréhendée par la « réussite au baccalauréat » qui est mesurable directement, car elle est sanctionnée par un diplôme, donnant lieu à un acte administratif que l'on peut comptabiliser (Evain, 2020). Le « décrochage scolaire », quant à lui, est un concept qui doit reposer sur une définition précise, choisie parmi un ensemble de définitions possibles : ce choix conventionnel fait acte de construction. Une fois la définition établie, le calcul repose le plus souvent sur l'observation de variables administratives existantes, telles que la non ré-inscription dans un établissement scolaire.

1. Les débats suscités par la couverture médiatique des résultats des enquêtes PISA (*Programme international pour le suivi des acquis des élèves*) de l'OCDE en sont l'illustration la plus flagrante.
2. Alfred Binet (1857-1911) est un pédagogue et psychologue français. Il est connu pour sa contribution essentielle à la psychométrie.
3. *« On peut présenter la sociologie de la quantification comme perpétuellement tendue entre deux conceptions des opérations statistiques, l'une « réaliste métrologique » (l'objet existe antérieurement à sa mesure), et l'autre « conventionnaliste » (l'objet est créé par les conventions de la quantification : exemples du taux de pauvreté, du chômage, du quotient intellectuel ou de l'opinion publique). »* (Desrosières, 2008).

En comparaison, la mesure des compétences des élèves se présente comme une démarche de construction assez particulière. L'idée sous-jacente de la psychométrie consiste à postuler qu'un test mesure des performances qui sont la manifestation concrète d'un niveau de compétence, non observable directement. Ainsi, l'objet de la mesure est une **variable latente**. Cette approche n'est pas propre au domaine de la cognition. On retrouve ce type de variable en économie avec la notion de propension, en sciences politiques avec la notion d'opinion ou encore en médecine avec la notion de qualité de vie (Falissard, 2008).

Cette singularité n'est pas simplement d'ordre conceptuel, elle a des implications concrètes quand il s'agit de répondre objectivement, avec des indicateurs statistiques, à des questions du débat public, comme celle du niveau des élèves.

LE NIVEAU DES ÉLÈVES : QUESTION ANCIENNE, RÉPONSES RÉCENTES

Il y a plus de trente ans, le débat sur la baisse supposée du niveau scolaire faisait rage, attisé par des interrogations sur la nécessaire transformation du système éducatif (un système de masse peut-il être performant ?) ou sur le lien entre éducation et économie (les compétences des élèves comme levier dans la compétition économique internationale ? Voir (Goldberg et Harvey, 1983) aux États-Unis). En France, (Thélot, 1992), tout comme (Baudelot et Establet, 1989) soulignaient alors le manque criant de mesures directes et objectives des acquis des élèves.

Si d'aucuns étaient tentés de mobiliser les statistiques sur les examens, elles ne permettaient cependant pas de se prononcer sur l'évolution du niveau des élèves. En effet, chaque année les sujets d'examen changent, sans qu'il n'ait été établi de comparaison rigoureuse de leur difficulté. Si bien que, par exemple, la comparaison de deux taux de réussite au baccalauréat n'est pas pertinente pour mesurer une évolution dans le temps : si le taux de réussite augmente, est-ce parce que le niveau d'exigence est moins élevé ou bien parce que le niveau des élèves est meilleur ?

Jusque dans les années quatre-vingt-dix, les seules données disponibles permettant une comparaison temporelle rigoureuse étaient celles issues des tests « psychotechniques » passés pendant les « *trois jours* » organisés par le ministère de la Défense. Elles ne concernaient cependant pas tous les élèves.

Le recours à des tests standardisés est alors apparu comme la solution adaptée. Ce type de dispositif de mesure trouve ses origines en France, dans les travaux d'Alfred Binet et de ses collaborateurs au début du XX^{e} siècle : pourtant, la psychométrie y reste une discipline très méconnue aujourd'hui encore. Paradoxalement, les évaluations des élèves sont très présentes dans le système scolaire français, à travers les contrôles continus fréquents conduits par les enseignants. Des études docimologiques, menées depuis près d'un siècle, avec notamment les travaux de la commission Carnegie sur le baccalauréat en 1936, montrent pourtant que le jugement des élèves par les enseignants est en partie empreint de subjectivité et peut dépendre de facteurs étrangers au niveau de compétence des élèves (Piéron, 1963). La notation des élèves est ainsi susceptible de varier sensiblement selon les caractéristiques des enseignants, des contextes scolaires, ainsi que des élèves eux-mêmes.

En revanche, dans d'autres pays, la psychométrie s'est considérablement développée, notamment aux États-Unis, à travers des thématiques telle que la méritocratie scolaire (assurer un traitement équitable des élèves) ou bien l'intelligence, sujet ayant d'ailleurs conduit à certaines dérives idéologiques (Gould, 1997).

Ainsi, malgré une demande sociale forte et récurrente, la question du niveau des élèves et de son évolution a longtemps souffert d'un manque de cadrage conceptuel et méthodologique. Le recours à des dispositifs d'évaluations standardisées est relativement récent dans le paysage des enquêtes statistiques françaises.

AUJOURD'HUI, IL EXISTE UN VASTE SYSTÈME D'ÉVALUATIONS STANDARDISÉES…

Forte de ce constat, dans les années quatre-vingt-dix, la direction de l'Évaluation et de la Prospective du ministère de l'Éducation nationale[4] a conduit plusieurs études visant à mesurer l'évolution des acquis des élèves. Ces travaux avaient clairement pour objectif de répondre aux tenants de la faillite du système éducatif, souvent nostalgiques d'un modèle scolaire révolu. Cependant, ces premières enquêtes comparatives montraient quelques faiblesses méthodologiques.

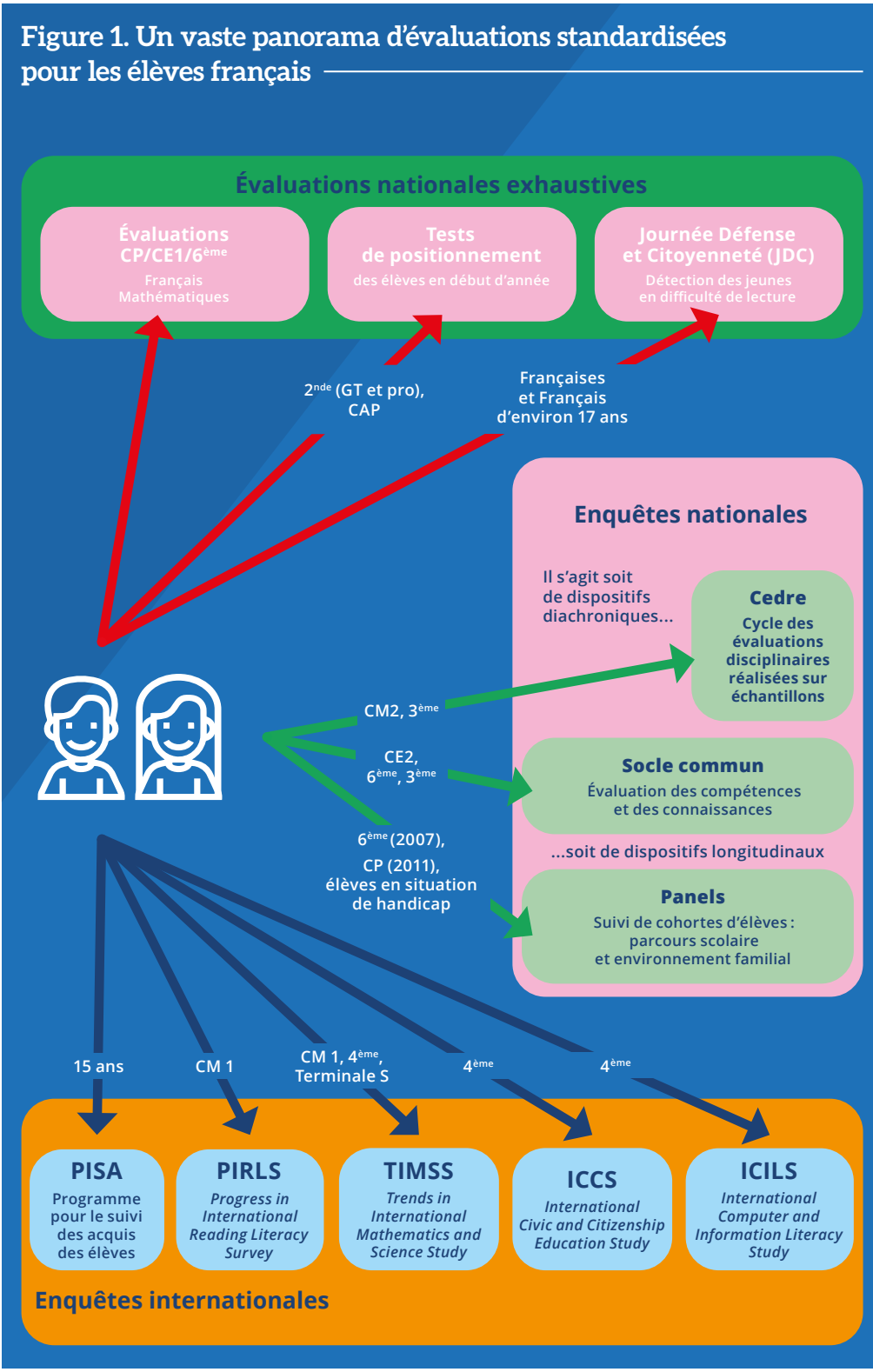
La France avait pourtant une longue expérience de campagnes d'évaluations des élèves, notamment à travers les évaluations nationales diagnostiques passées par tous les élèves de CE2 et de 6ème, à chaque rentrée scolaire, entre 1989 et 2007. Mais ces évaluations ne permettaient pas d'établir des comparaisons temporelles statistiquement robustes. D'une part, leur objectif premier, tout comme celui des examens, n'était pas de rendre compte de l'évolution du niveau d'acquisition des élèves dans le temps, mais de servir d'outils de repérage individuel des difficultés pour les enseignants. D'autre part, les connaissances dans le champ de la mesure en éducation et plus largement en psychométrie étaient très peu diffusées et vulgarisées ; le constat est encore actuel, bien que l'expérience de la Depp dans ce domaine se soit considérablement améliorée depuis une vingtaine d'années.

Progressivement, d'autres dispositifs d'évaluations construits pour permettre des comparaisons diachroniques se sont développés en France ***(figure 1)***. Plusieurs phénomènes relativement récents expliquent cet essor et cette multiplicité. Tout d'abord, la volonté de construire des indicateurs de suivi, pour le pilotage du système, est devenue de plus en plus prégnante, notamment sous l'impulsion de la LOLF[5], qui implique la construction d'indicateurs annuels de résultats, tels que le pourcentage d'élèves qui maîtrisent les compétences attendues, à différents niveaux scolaires.

Parallèlement, les dispositifs internationaux, tels que PISA[6] (OCDE, 2020), PIRLS[7] ou TIMSS[8] (Rocher et Hastedt, 2020) ont largement contribué à rendre incontournables les programmes d'évaluations standardisées à grande échelle (*Large-scale assessments*) dans le débat public sur l'École et dans les décisions en matière de politiques éducatives. Aujourd'hui, rares sont les papiers ou les discours sur le système éducatif qui ne renvoient pas aux évaluations internationales.

4. La DEP a été plus récemment transformée en direction de l'évaluation, de la prospective et de la performance (Depp), qui est le service statistique du ministère.
5. La loi organique relative aux lois de finances fixe le cadre des lois de finances en France. Promulguée en 2001, s'applique à toute l'administration depuis 2006 et vise à moderniser la gestion de l'État. Elle a favorisé le développement d'outils d'évaluation des résultats, sur l'ensemble de l'action administrative.
6. Programme international pour le suivi des acquis des élèves (*Programme for International Student Assessment*), mené par l'Organisation de coopération et de développement économiques (OCDE).
7. Programme international de recherche en lecture scolaire (*Progress in International Reading Literacy*) de l'IEA (*International Association for the Evaluation of Educational Achievement*) basée aux Pays-Bas.
8. *Trends in International Mathematics and Science Study* (TIMSS) est une enquête internationale de l'IEA sur les acquis scolaires en mathématiques et en sciences.

Figure 1. Un vaste panorama d'évaluations standardisées pour les élèves français



Enfin, depuis 2017, de nouvelles évaluations exhaustives ont été mises en place : elles concernent aujourd'hui tous les élèves de CP, de CE1, de 6ème et de 2nde, ainsi que de CAP[9] de lycée. Le déploiement de ces évaluations, qui concernent plus de trois millions d'élèves à chaque rentrée scolaire, a évidemment favorisé l'essor et la visibilité de ce type de dispositifs.

Ainsi, du point de vue des producteurs d'indicateurs dans le domaine des compétences des élèves, il est devenu primordial d'adopter un corpus méthodologique adapté permettant d'établir des mesures fiables dans le temps et dans l'espace. Or les indicateurs produits alimentent différents types d'usages, rendant plus complexe la configuration optimale d'un système d'évaluation de compétences.

… ADAPTÉ À DIFFÉRENTS USAGES

En effet, les utilisations des résultats d'évaluations standardisées des compétences des élèves ont connu différentes phases dans l'histoire, conduisant à une succession de dispositifs différents depuis quarante ans (Trosseille et Rocher, 2015). En particulier, les débats portent sur la configuration des évaluations nationales exhaustives, qui concernent tous les élèves d'un ou de plusieurs niveaux scolaires. La clarification précise des objectifs qui leur sont assignés est indispensable à l'efficacité de leur mise en œuvre. En combinant l'appréciation individuelle (diagnostic de difficulté) et le compte-rendu collectif (construction d'indicateurs statistiques), elles répondent à différents besoins. Ce *hiatus* était d'ailleurs déjà pointé par Alfred Binet (Rozencwajg, 2011) avec la distinction entre approche « clinique » et approche statistique.

Aujourd'hui, d'une façon générale, on identifie trois finalités différentes :

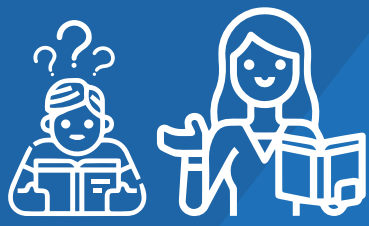
- fournir aux enseignants des repères sur les acquis de leurs élèves, compléter ainsi leurs constats et leur permettre d'enrichir leurs pratiques pédagogiques. Par exemple, en début de 6ème, les élèves passent un test de fluence (*i.e.* de rapidité de lecture), identique pour tous et préalablement calibré, permettant de repérer les élèves susceptibles d'être pénalisés lors de leur parcours au collège ;
- doter les « pilotes de proximité »[10] d'indicateurs leur permettant de mieux connaître les résultats des écoles et d'effectuer une vraie régulation. Par exemple, à partir des résultats obtenus aux évaluations nationales, un recteur peut situer son académie aux différents niveaux du parcours des élèves (CP, CE1, 6ème, 2nde), identifier des points de faiblesse et mettre en place des dispositifs d'actions pédagogiques ;
- disposer d'indicateurs permettant de mesurer, au niveau national, les performances du système éducatif, d'en apprécier les évolutions temporelles et d'en déduire des comparaisons internationales. Par exemple, le cycle des évaluations disciplinaires réalisées sur échantillon (Cedre) situe les élèves par rapport aux attendus des programmes scolaires, de manière fine, ce qui permet d'alimenter la réflexion sur d'éventuels ajustements de ces programmes.

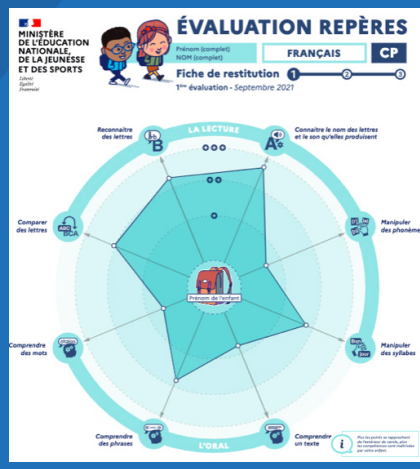
S'agissant des évaluations nationales exhaustives aujourd'hui, intentionnellement positionnées en début d'année scolaire, elles permettent à la fois d'aider à l'action pédagogique immédiate à partir des résultats individuels et à la fois d'alimenter des outils statistiques de suivi et de pilotage, notamment s'agissant des résultats du niveau scolaire précédent. À chaque niveau, les acteurs sont destinataires de résultats répondant à leurs besoins spécifiques ***(figure 2)***.

9. Le certificat d'aptitude professionnelle est un diplôme français d'études secondaires et d'enseignement professionnel.
10. Recteurs d'académie, directeurs académiques des services de l'Éducation nationale (DASEN, anciennement inspecteurs d'académie) ou inspecteurs de l'Éducation nationale (IEN).

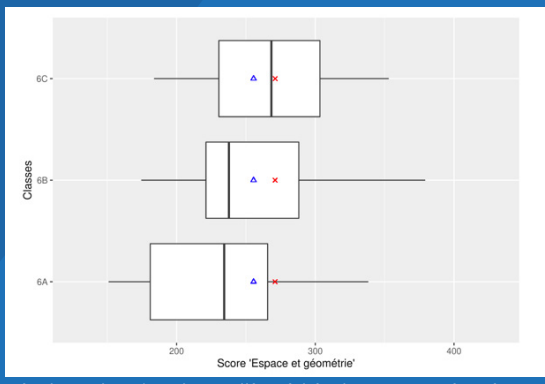
Figure 2. Les évaluations nationales exhaustives fournissent des indicateurs utiles aux différents acteurs éducatifs

Pour aider l'enseignant en CP à identifier les élèves en difficulté de lecture…





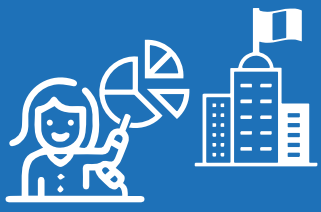
Résultat pour chaque élève de CP de l'évaluation en français de début d'année.

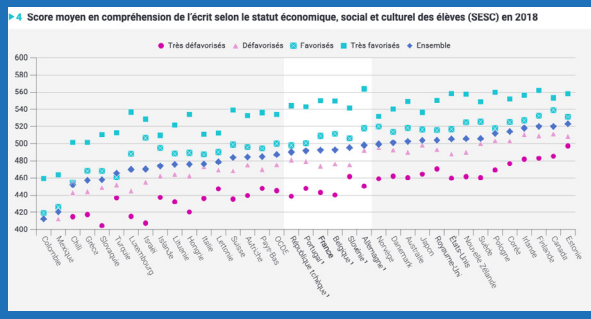


Résultats des 6èmes d'un collège à l'évaluation en géométrie.

… ou pour permettre au principal d'un collège et à son équipe pédagogique d'adapter leur action dans les classes de 6ème…

… ou pour que le Ministère adapte l'affectation des soutiens aux populations en difficulté scolaire…





Comparaisons internationales des scores en compréhension de l'écrit, selon le statut économique et social (PISA).

Depuis 2017, les évaluations nationales exhaustives concernent tous les élèves de CP, CE1, 6ème, 2nde, et de CAP.

… ET ANCRÉ DANS UN HÉRITAGE MÉTHODOLOGIQUE ROBUSTE

> *« Les programmes d'évaluations ont pour ambition d'établir une mesure objective, scientifique, des acquis des élèves, la plus indépendante possible des conditions d'observation, de passation, de correction. »*

Quel que soit le niveau d'usage, les programmes d'évaluations ont pour ambition d'établir une mesure objective, scientifique, des acquis des élèves, la plus indépendante possible des conditions d'observation, de passation, de correction. En ce sens, ces évaluations sont « standardisées ».

Ces opérations se situent au carrefour de deux traditions méthodologiques : celle de la psychométrie, pour ce qui relève de la mesure de dimensions psychologiques, en l'occurrence des acquis cognitifs ; et celle des enquêtes statistiques pour ce qui a trait aux procédures de recueil des données.

Si cette dernière tradition est largement partagée dans le champ de la statistique publique, celle relative à la psychométrie est nettement moins connue. Or, les modélisations statistiques proposées dans ce domaine sont très anciennes (cf. les analyses factorielles de Spearman en 1905) et donnent lieu encore aujourd'hui à une littérature très riche au niveau international. La diffusion de ce corpus méthodologique dans le cadre de dispositifs nationaux doit sans doute beaucoup à l'influence des grandes évaluations internationales (Rocher, 2015b). Les pays participants – et contributeurs actifs – ont bénéficié d'une acculturation à ces méthodes leur permettant d'opérer un transfert de technologie dans leurs systèmes nationaux d'évaluation.

Ces deux traditions – enquêtes statistiques et psychométrie – sont convoquées et combinées lors de la mise en œuvre d'une opération d'évaluation, qui suit un processus précis. Dans le cadre du Cedre (cf. *supra*), ce processus fait même l'objet d'une certification externe (voir sur le site de l'*Afnor*) basée sur un référentiel d'engagements de service pris à chaque étape, conférant ainsi au programme des qualités de reproductibilité et de transparence.

Le processus se déroule en six grandes étapes, depuis la définition du « construit » que l'on cherche à mesurer, jusqu'à la production de résultats contrôlés et redressés.

QUEL «CONSTRUIT» VISE-T-ON?

À la base du processus de mesure se situe le « construit », c'est-à-dire le concept visé par l'opération de mesure. Le construit est défini de manière précise dans un cadre conceptuel (*framework*). Ce cadre conceptuel peut être adossé à un plan de formation ou bien se référer à une théorie cognitive. Par exemple, le cycle Cedre mesure les acquis des élèves dans chaque discipline ; il repose ainsi sur ce qui est censé être enseigné à l'école, au regard de l'ensemble des programmes scolaires. Une illustration radicalement différente peut être donnée par l'identification de troubles d'apprentissage, comme la dyslexie. Les tests vont alors être élaborés pour détecter un déficit sur des dimensions très spécifiques.

Il est primordial que ce cadre soit le plus précis possible, au vu des définitions multiples que peut avoir un même objet. Par exemple, en quoi consistent les compétences en mathématiques ? Deux visions très différentes nous sont données par les évaluations internationales. D'un côté, l'évaluation TIMSS de l'IEA s'intéresse à la façon dont est structuré l'enseignement des mathématiques, pour élaborer l'évaluation selon des domaines assez partagés dans les différents *curricula* (nombres, géométrie, résolution de problèmes, etc.). L'évaluation va ainsi souvent porter sur des aspects « intra-mathématiques ». De son côté, l'enquête internationale PISA s'intéresse au concept de « *literacy* », c'est-à-dire la capacité des individus à appliquer leurs connaissances dans des situations de la vie réelle. La question est alors celle de pouvoir passer du monde réel au monde des mathématiques, et *vice et versa*. L'orientation est donc très structurante pour la suite, c'est-à-dire pour l'élaboration de l'instrument mais également pour l'interprétation des résultats.

Le cadre conceptuel décrit également la structure de l'objet, ou encore l' « univers des *items* », c'est-à-dire l'ensemble des *items* (plus petit élément de mesure) censés mesurer la dimension visée. C'est cette structuration qui définit au final l'objet mesuré.

LA CONCEPTION DES UNITÉS DE MESURE OU *ITEMS*

Une fois ce cadre posé, un groupe de concepteurs – composé d'enseignants, d'inspecteurs, possiblement de chercheurs – est donc chargé de construire un large ensemble d'*items*. Fruit d'un travail collectif, ils font l'objet de débats jusqu'à aboutir à un consensus.

Les *items* sont ensuite soumis à un « cobayage », c'est-à-dire à une passation auprès d'une ou plusieurs classes pour estimer leur difficulté, leur durée de passation minimale et recueillir les réactions éventuelles des élèves.

Les *items* passant avec succès l'étape du cobayage sont alors testés par un échantillon d'élèves représentatif du niveau visé (entre 500 et 2 000 élèves par *item*). Cette phase expérimentale a lieu un an avant la phase principale de l'évaluation, afin de respecter le positionnement de l'évaluation dans le calendrier de l'année scolaire.

L'ÉCHANTILLONNAGE DES ÉLÈVES

S'agissant des grandes enquêtes nationales ou internationales ***(figure 1)***, les échantillons sont composés de plusieurs milliers d'élèves (entre 4 000 et 30 000 selon les programmes et les cycles). Des problématiques classiques du domaine des sondages se posent alors, par exemple la définition du champ, les bases de sondage, les modalités de tirage, etc. Ces aspects sont documentés en détails dans les rapports techniques (Bret *et alii*, 2015 ; OCDE, 2020), en particulier concernant les grandes enquêtes internationales qui imposent le respect de nombreux standards dans ce domaine : il s'agit notamment de maximiser le taux de couverture de la population visée et de limiter les exclusions (par exemple, territoriales ou en raison de contraintes pratiques).

En général, les procédures de sondage procèdent par tirage à deux degrés, d'abord des établissements scolaires (ou directement des classes) et ensuite des élèves. Enfin, dans la mesure où plusieurs échantillons peuvent être tirés à partir des mêmes bases, la coordination de leur tirage est traitée avec précaution (Garcia *et alii*, 2015).

L'ADMINISTRATION DES ÉVALUATIONS

Dans le cadre de ces enquêtes sur échantillons, la passation des évaluations est le plus souvent assurée par des personnels extérieurs à l'établissement scolaire (par exemple : l'évaluation sur tablettes des compétences des élèves de l'école primaire ou les évaluations internationales TIMSS et PISA sur ordinateurs dans les collèges et lycées).

Pour les évaluations nationales exhaustives, en CP et CE1, ce sont les professeurs qui font passer les évaluations. Les consignes sont extrêmement précises, mais il est évident, dès lors que 45 000 professeurs sont concernés, qu'une certaine variabilité des conditions de passation, difficilement quantifiable, vient affecter l'erreur de mesure. Pour les évaluations 6ème et 2nde, les conditions sont en revanche plus favorables, grâce aux modalités de passation sur ordinateurs, qui garantissent une meilleure standardisation.

L'IMPLICATION DES ÉLÈVES

Comme dans toute enquête, il est important de s'interroger sur les dispositions des enquêtés. Dans le cas des évaluations standardisées, elles restent à faible enjeu (*low stakes)* pour les élèves y participant, même si elles renvoient à des enjeux politiques croissants.

Dans le système éducatif français, la notation tient une place prépondérante. Dès lors, face à une évaluation ne conduisant pas à une note, on peut s'interroger sur le degré de motivation des élèves. À partir des enquêtes du Cedre, une étude expérimentale a montré que les élèves s'investissaient davantage lorsqu'on leur annonce préalablement que le résultat obtenu conduira à une note.

LA «CORRECTION» DES RÉPONSES

Enfin, les réponses données par les élèves sont soit codées automatiquement (par exemple dans le cas de questions à choix multiples), soit codées par des correcteurs humains (par exemple dans le cas de productions écrites complexes). En cas de correction humaine, un processus de corrections multiples avec arbitrage est suivi. En effet, il s'agit de neutraliser les nombreux biais de correction qui peuvent apparaître et qui ont été documentés depuis plus d'un siècle par la docimologie, la science des examens (Piéron, 1963).

Enfin, l'analyse psychométrique permet d'identifier des *items* ayant un mauvais fonctionnement, par exemple les *items* n'étant pas corrélés à l'ensemble des *items* censés mesurer la même dimension. Ce processus suit une démarche très empirique, à façon, qui consiste à établir un ensemble d'*items* cohérent et le plus en phase avec le cadre conceptuel.

En guise d'illustration, l'opération Cedre Sciences a réalisé en 2017 l'expérimentation d'environ 400 *items*, pour en retenir 262 pour l'évaluation finale de 2018, dont 43 ont été repris à l'identique de l'enquête de 2007 et 31 de celle de 2013, afin d'assurer des comparaisons temporelles (Bret *et alii*, 2015).

QUELQUES CONCEPTS PSYCHOMÉTRIQUES AUTOUR DE LA NOTION DE VARIABLE LATENTE

Les éléments qui viennent d'être présentés sont potentiellement présents également dans d'autres domaines couverts par les enquêtes statistiques classiques. Comme nous l'avons indiqué en introduction, la spécificité des dispositifs d'évaluation de compétences tient plus particulièrement à l'objet de mesure, dont la matérialité se révèle uniquement à travers l'instrument de mesure. Il est ainsi convenu que l'instrument permet d'observer des performances, qui sont des manifestations concrètes de la compétence, variable qui ne nous est pas accessible directement : cette notion de **variable latente** est centrale en psychométrie.

Afin d'illustrer de façon pédagogique les grandes notions de psychométrie, un exemple classiquement utilisé porte sur la taille des individus ***(figure 3)***. La situation est la suivante : nous n'avons aucun moyen de mesurer directement la taille des individus d'un échantillon donné. Mais nous avons la possibilité de proposer un questionnaire, composé de questions appelant une réponse binaire (oui/non) et n'évoquant pas directement la taille. Nous nous plaçons ainsi artificiellement dans le cas de la mesure d'une variable latente que nous cherchons à approcher à l'aide d'un questionnaire, soit un dispositif de mesure apparemment comparable à celui d'une évaluation standardisée.

Ce questionnaire permet d'illustrer concrètement des concepts importants de psychométrie :

- la **validité** : le test mesure-t-il bien ce qu'il est censé mesurer ? En l'occurrence, il se trouve que la taille réelle des individus est fortement corrélée à un score calculé à partir des 24 *items*. Le score obtenu représente donc bien la variable latente visée ;
- la **dimensionnalité d'un ensemble d'*items*** : le calcul d'un score suppose que les *items* mesurent la même dimension, que le test est unidimensionnel. Cependant, il est clair que les *items* présentés ici ne mesurent pas purement la dimension taille, mais interrogent chacun une multiplicité de dimensions. L'idée est qu'un facteur commun prépondérant relie ces *items*, facteur lié à la taille. Différentes techniques existent pour déterminer si un test peut être considéré comme unidimensionnel.

PASSER DES UNITÉS À L'ÉCHELLE DE MESURE

Lorsqu'il s'agit de construire l'échelle de mesure, d'autres concepts sont mobilisés et peuvent également être illustrés à travers notre questionnaire sur la taille :

- les **fonctionnements différentiels** d'*items* : en guise d'illustration, à l'affirmation « À deux sous un parapluie, c'est souvent moi qui le tiens », 89 % des hommes répondent oui contre 52 % des femmes, soit un écart de 37 points, alors qu'en moyenne sur l'ensemble des *items*, la différence entre les hommes et les femmes est de 20 points seulement. La question est dite « biaisée » selon le sexe : la réponse donnée dépend d'une caractéristique de groupe et non pas seulement de la taille. L'étude des fonctionnements différentiels est fondamentale en matière de comparaison temporelle ou internationale, pour savoir si d'autres facteurs interviennent dans la réussite, au-delà du seul niveau de compétence ;
- le **pouvoir discriminant** des *items* ou la corrélation *item*-test permet de vérifier qu'un *item* mesure bien la dimension supposée. Par exemple, l'*item* « Dans un lit, j'ai souvent froid aux pieds. », repris d'un questionnaire similaire aux Pays-Bas, n'est pas corrélé avec les autres *items* sur l'échantillon français. Ainsi, cet *item* ne mesure pas la dimension taille en France mais plutôt une autre dimension décorrélée, telle que la frilosité… ;

Figure 3. Une illustration des grands concepts de psychométrie

Évaluer la taille des individus (la variable latente),
à l'aide d'un questionnaire de 24 *items*
auxquels il suffit de répondre par oui ou par non (extrait).

1 **Je dois souvent faire attention à ne pas me cogner la tête**

2 **Pour les photos de groupe, on me demande souvent d'être au premier rang**

3 **On me demande souvent si je fais du basket-ball**

4 **Dans la plupart des voitures, je suis mal assis(e)**

5 **Je dois souvent faire faire les ourlets quand j'achète un pantalon**

6 **Je dois souvent me baisser pour faire la bise**

Pour une description détaillée, consulter (Rocher, 2015a).

- l'**échelle de mesure** : le questionnaire ne permet pas de connaître la taille des individus, mais simplement de les classer selon une variable corrélée à la taille, en l'occurrence un score obtenu aux 24 *items*. Il est ainsi possible d'opérer des transformations linéaires sur ce score, ce qui ne modifie pas les rapports entre intervalles de scores entre individus. Typiquement les scores peuvent être standardisés de moyenne 0 et d'écart-type 1, mais le plus souvent ils sont transformés à des valeurs supérieures (moyenne 250 et écart-type 50 dans Cedre, ou moyenne 500 et écart-type 100 dans PISA) afin d'éviter des valeurs négatives.

UN BESOIN DE MODÉLISATION POUR RELIER LES OBSERVATIONS À LA VARIABLE LATENTE

Envisager les résultats à une évaluation comme résultant d'un processus de mesure d'une variable latente ne s'impose pas de lui-même. En effet, le calcul de scores à une évaluation peut sembler trivial : compter le nombre de bonnes réponses obtenues apparaît comme un indicateur adapté du niveau de compétences et il est tout à fait possible de considérer uniquement le nombre de points et de ne pas donner plus de significations à cette statistique qu'un score observé à un test.

> *« Distinguer ce qui relève de la difficulté du test de ce qui relève du niveau de compétence des élèves. »*

Mais cette démarche est très frustre d'un point de vue théorique et trouve vite des limites en pratique, car elle permet difficilement de distinguer ce qui relève de la difficulté du test de ce qui relève du niveau de compétence des élèves. En particulier, pour assurer la comparabilité entre différentes populations ou entre différentes épreuves, le recours à une modélisation plus adaptée, qui se situe au niveau des *items* eux-mêmes et non au niveau du score agrégé, est apparu nécessaire[11]. En particulier, les **modèles de réponse à l'*item*** (ou MRI), nés dans les années soixante, se sont imposés dans le champ des évaluations standardisées à grande échelle ***(encadré 1)***.

Ces modèles permettent de relier de manière probabiliste les réponses aux *items* et la variable latente visée. Ils sont très utiles dès lors qu'il s'agit de comparer les niveaux de compétence de différents groupes d'élèves. Cette problématique renvoie à la notion d'ajustement des métriques (*equating*). Il s'agit de positionner sur la même échelle de compétence les élèves de différentes cohortes, à partir de leurs résultats observés à des évaluations partiellement différentes. De nombreuses techniques existent et sont couramment employées dans les programmes d'évaluations standardisées (Kolen et Brennan, 2004). Typiquement, les comparaisons sont établies à partir d'*items* communs, repris à l'identique d'un moment de mesure à l'autre. Les modèles de réponse à l'*item* fournissent alors un cadre approprié, dans la mesure où ils distinguent les paramètres des *items*, qui sont considérés comme fixes, des paramètres des élèves, considérés comme variables.

11. Pour une lecture en français sur la théorie des tests, voir (Laveault et Grégoire, 2002).

Encadré 1. Des modèles probabilistes pour séparer niveau de compétence et difficulté de l'*item*

Pour un même objet de mesure, les questions (*items*) composant l'instrument de mesure peuvent être différentes. Dès lors, travailler sur un score agrégé trouve vite des limites et il est préférable de faire reposer l'analyse sur l'élément le plus élémentaire, c'est-à-dire l'*item*.

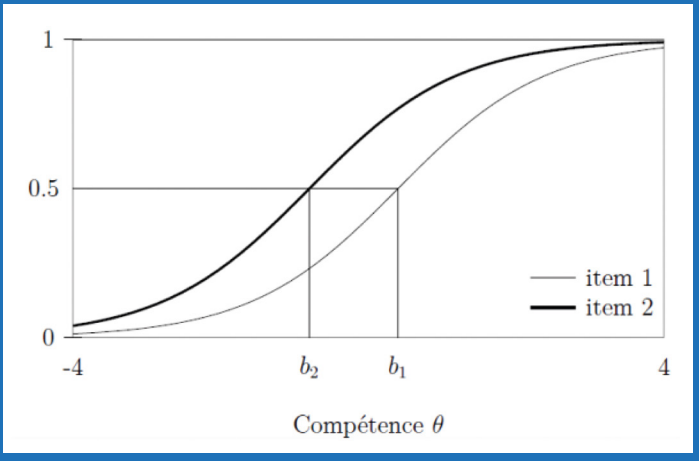
Les modèles de réponse à l'*item* (MRI), nés dans les années soixante, sont une classe de modèles probabilistes. Ils modélisent la probabilité qu'un élève donne une certaine réponse à un *item*, en fonction de paramètres concernant l'élève et l'*item*.

Dans le modèle le plus simple, proposé par le mathématicien danois George Rasch en 1960, la probabilité P_{ij} que l'élève i réussisse l'*item j* est une fonction sigmoïde du niveau de compétence θ_i de l'élève *i* et du niveau de difficulté b_j de l'*item j*. La fonction sigmoïde étant une fonction croissante (voir figure), il ressort que la probabilité de réussite augmente lorsque le niveau de compétence de l'élève augmente et diminue lorsque le niveau de difficulté de l'*item* augmente, ce qui traduit à l'évidence les relations attendues entre réussite, difficulté et niveau de compétence.

L'intérêt de ce type de modélisation, et ce qui explique son succès, c'est de séparer deux concepts-clés, à savoir la difficulté de l'*item* et le niveau de compétence de l'élève.

Ainsi, les **MRI ont un intérêt pratique pour la construction de tests :** si le modèle est bien spécifié sur un échantillon donné, les paramètres des *items* peuvent être considérés comme fixes et applicables à d'autres échantillons dont il sera alors possible de déduire les paramètres relatifs aux élèves, en l'occurrence leur niveau de compétence.

Autre avantage : le niveau de compétence des élèves et la difficulté des *items* sont placés sur la même échelle. Cette propriété permet d'interpréter le niveau de difficulté des *items* par rapprochement avec le continuum de compétence. Ainsi, les élèves situés à un niveau de compétence égal à b_j auront 50 % de chances de réussir l'*item*, ce que traduit visuellement la représentation des courbes caractéristiques des *items* selon ce modèle.



Note de lecture : la probabilité de réussir l'item (en ordonnées) dépend du niveau de compétence (en abscisse). Par définition, le paramètre de difficulté d'un item correspond au niveau de compétence ayant 50 % de chances de réussir l'item. Ainsi, l'item 1 en trait fin est plus difficile que l'item 2 en trait plein. La probabilité de le réussir est plus élevée quel que soit le niveau de compétence.

QUELLE UTILITÉ POUR LA STATISTIQUE PUBLIQUE?

Cet appareillage statistique et psychométrique permet d'élaborer des indicateurs robustes du niveau des élèves et surtout il permet d'établir des comparaisons temporelles et spatiales. Par exemple, la question de l'évolution du niveau scolaire dans le temps peut être abordée statistiquement grâce à la reprise à l'identique d'***items* dits d'ancrage**. En effet, la reprise identique de l'ensemble des *items* passés lors d'une précédente enquête n'est pas forcément pertinente, au regard de l'évolution des programmes scolaires, des pratiques, de l'environnement culturel, etc. Certains *items* doivent être retirés, d'autres ajoutés. Par conséquent, les élèves des deux cohortes passent une épreuve en partie différente. Dès lors, comment assurer la comparabilité des résultats ? Le simple calcul du nombre de bonnes réponses n'est plus pertinent et il faut recourir aux modélisations présentées ici. Ainsi, avec cette approche les enquêtes Cedre permettent d'établir des comparaisons du niveau en mathématiques des élèves depuis plus de dix ans et qui montrent d'ailleurs une dégradation préoccupante des résultats, à l'école primaire comme au collège (Ninnin et Pastor, 2020 ; Ninnin et Salles, 2020).

De leur côté, les évaluations internationales utilisent ces méthodologies pour assurer la comparabilité des difficultés des *items* d'un pays à l'autre. En effet, une hypothèse forte de ces enquêtes est que l'opération de traduction ne modifie pas la difficulté de l'*item*. Des procédures strictes de contrôle des traductions sont mises en œuvre. Cependant, des analyses montrent que la hiérarchie des paramètres de difficulté des questions posées est à peu près conservée pour des pays partageant la même langue, mais qu'elle peut être bouleversée entre deux pays ne parlant pas la même langue. Les modèles de réponse à l'*item* permettent de repérer ces sources de biais potentiels de comparabilité. Un exemple concret est donné par les enquêtes TIMSS auxquelles participent des dizaines de pays et qui ont récemment montré la place préoccupante de la France en mathématiques, de manière complémentaire et cohérente avec Cedre (Colmant et Le Cam, 2020 ; Le Cam et Salles, 2020).

EN PERSPECTIVE : ÉVALUER LES COMPÉTENCES TRANSVERSALES…

Les programmes d'évaluation font face à de nouvelles perspectives, liées à la demande d'évaluation de compétences plus complexes, ainsi qu'à l'essor du numérique.

Ainsi, aux difficultés méthodologiques évoquées précédemment viennent s'ajouter de nouveaux défis posés par une demande croissante (venant pour partie du monde économique) pour la mesure de dimensions beaucoup plus complexes que les compétences traditionnelles, académiques. On parle parfois de compétences transversales, de *soft skills*, compétences du XXIe siècle, de compétences socio-cognitives, etc. ***(encadré 2)***.

L'évaluation de ces dimensions constitue un vrai challenge. En effet, la définition de ces compétences (*framework*) n'est pas toujours très solide ou consensuelle. Ensuite, leur enseignement n'est pas toujours explicite, ce qui interroge sur la portée des résultats de l'évaluation. Enfin, leur structuration est complexe : elles impliquent le plus souvent des dimensions cognitives, mais également des attitudes, des dispositions, etc. Par exemple, évaluer l'esprit critique peut renvoyer à de multiples dimensions intriquées, telles que la compréhension, des composantes métacognitives, la curiosité, etc. Même si chaque composante est potentiellement évaluable, leur juxtaposition ne permet pas non plus de rendre compte d'un degré d'esprit critique. Des dispositifs plus « holistiques » doivent être imaginés.

Encadré 2. Les évaluations standardisées concernent des types de compétences très divers

	De quoi s'agit-il ?	Une illustration
Connaissances et compétences disciplinaires	La référence est celle des programmes scolaires qui définissent les attendus en matières d'apprentissage pour les différents cycles et niveaux scolaires.	**CEDRE Mathématiques 3ème** • Développer une expression algébrique • Appliquer le théorème de Thalès pour calculer des longueurs • etc.
Littératie et Numératie	Il s'agit de la capacité à mobiliser ses acquisitions scolaires pour agir dans la vie quotidienne.	**PISA Compréhension de l'écrit à 15 ans** • Localiser une information dans un texte • Relier des informations entre différentes sources • Évaluer la qualité et la crédibilité d'un texte • etc.
Compétences sociocognitives	Il s'agit d'un terme général pouvant regrouper de nombreuses dimensions spécifiques, parfois appelées aussi socio-comportementales, non cognitives ou conatives, provenant de la recherche en psychologie.	**Panels** Questionnaires dits « subjectifs » sur des thèmes tels que : • Motivation (ex : j'essaie de bien faire au collège parce que j'apprends des choses qui m'intéressent) • Sentiment de performance (ex : vous sentez-vous capable de réussir en mathématiques ?) • etc.
Compétences du XXIe siècle	Un ensemble de compétences censées être importantes face aux évolutions – notamment technologiques – de nos sociétés. Un des premiers cadres de référence est celui du « Partenariat pour les Compétences du 21e siècle » ou P21 qui définit les 4C : • Esprit Critique • Créativité • Collaboration • Communication. D'autres cadres se sont développés, intégrant par exemple deux autres C : • Citoyenneté • et *Character* (personnalité).	**Socle commun** Exemples de travaux en cours dans le champ des évaluations standardisées en France, *via* des conventions de recherche : • Esprit Critique : évaluation de la capacité des élèves à juger la véracité d'une information • Créativité : dans le langage, avec la création d'histoires ou en mathématiques, avec la recherche de solutions originales à des problèmes.

… ET INTÉGRER LA RÉVOLUTION NUMÉRIQUE

La révolution numérique entraîne des transformations profondes, y compris dans le domaine de l'évaluation des compétences des élèves. En 2015, les programmes d'évaluations standardisées ont entamé leur mue vers le format numérique. Aujourd'hui, dans le second degré, les évaluations sont toutes réalisées sur ordinateur et concernent chaque année près de deux millions d'élèves. Dans le premier degré, la situation est plus compliquée, en raison des équipements mal adaptés.

Le processus de mesure n'est pas bouleversé dans ses principes, mais l'évolution technologique apporte son lot de difficultés nouvelles. Ainsi :

- passer du papier/crayon au numérique pose des questions de comparabilité et d'éventuelles ruptures de séries ;
- l'aptitude des élèves à utiliser[12] ces nouveaux outils, voire la familiarité des élèves avec ces nouveaux environnements est mal connue ;
- utiliser le numérique amène des problématiques liées à la confidentialité ou à la sécurité.

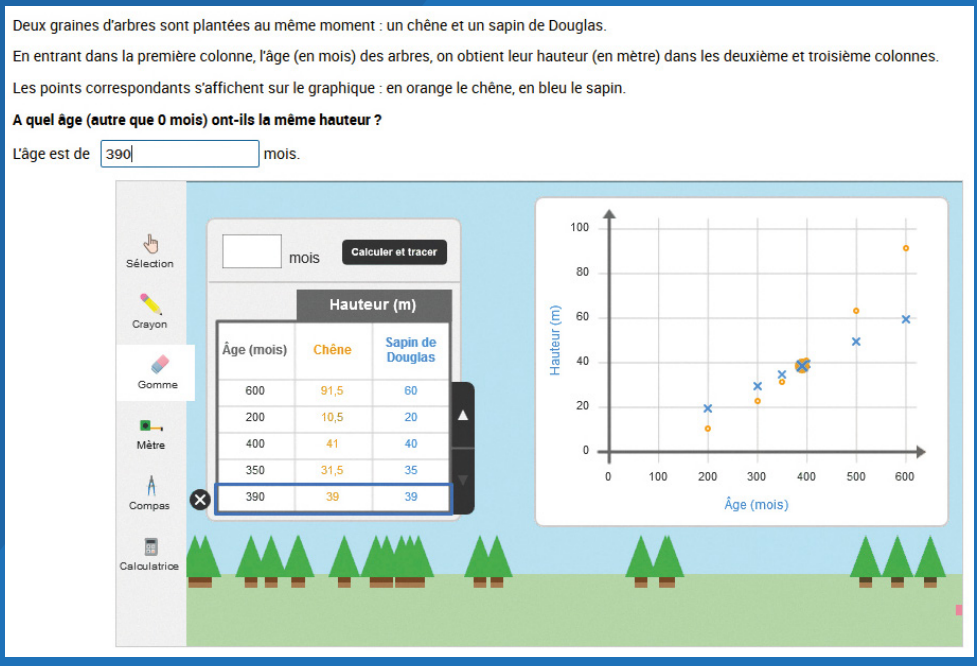
Mais *a contrario*, le numérique offre des fonctionnalités très intéressantes en matière d'évaluation (multimédia, accessibilité, etc.), des techniques plus sophistiquées (comme la possibilité d'introduire des processus adaptatifs), des situations interactives pour des expériences plus ludiques (*game-based*), etc.

Enfin, en matière d'analyse statistique, ces dispositifs permettent de recueillir beaucoup plus de données, à travers l'enregistrement des actions des élèves (« traces » des élèves) ***(encadré 3)***. Ces approches permettent déjà d'enrichir les analyses, et seront très utiles à la fois pour un retour individuel approfondi et pour des statistiques plus précises sur le niveau de compétences des élèves.

12. On parle alors d'« utilisabilité » en ergonomie informatique par exemple.

Encadré 3. L'analyse des traces numériques des élèves

Dans cet *item* interactif, l'élève doit réaliser une série d'essais pour déterminer le point de croisement entre deux fonctions : en entrant des valeurs dans un tableau, celles-ci sont positionnées automatiquement sur un graphique. L'élève peut utiliser différents outils numériques (crayon, gomme, etc.). À partir des traces laissées par ses différentes actions, une analyse basée sur les techniques de *data science* permet d'identifier des profils cognitifs pertinents (Salles *et alii*, 2020). Il est important de noter que cette étude n'est pas *data driven** – approche souvent vouée à l'échec dans ce domaine – mais s'est appuyée sur un cadre théorique didactique qui a guidé le processus d'analyse.



Source : Évaluation des compétences du socle commun en fin de 3ème.

** Le « pilotage par la donnée » supposerait de contextualiser ou de personnaliser l'outil à l'élève en fonction de ses caractéristiques.*

BIBLIOGRAPHIE

BAUDELOT, Christian et ESTABLET, Roger, 1989. *Le niveau monte*. Éditions du Seuil. ISBN 2-02-010385-0.

BRET, Anaïs, GARCIA, Émilie, ROCHER, Thierry, ROUSSEL, Léa et VOURC'H, Ronan, 2015. *Rapport technique de CEDRE, Cycle des Évaluations Disciplinaires Réalisées sur Échantillons. Sciences expérimentales 2013*, *Collège*. [en ligne]. Février 2015. MENESR-DEPP. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.education.gouv.fr/media/10742/download*.

COLMANT, Marc et LE CAM, Marion, 2020. *TIMSS 2019 – Évaluation internationale des élèves de CM1 en mathématiques et en sciences : les résultats de la France toujours en retrait*. [en ligne]. Décembre 2020. MENESR-DEPP. Note d'information n°20.46. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.education.gouv.fr/media/73349/download*.

DESROSIÈRES, Alain, 2008. *Gouverner par les nombres, L'argument statistique II*. [en ligne]. Presses des Mines. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://books.openedition.org/pressesmines/341*.

EVAIN, Franck, 2020. Indicateurs de valeur ajoutée des lycées : du pilotage interne à la diffusion grand public, In : *Courrier des statistiques*. [en ligne]. 31 décembre 2020. Insee. N° N5, pp. 74-94. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5008710/courstat-5.pdf*.

FALISSARD, Bruno, 2008. *Mesurer la subjectivité en santé – Perspective méthodologique et statistique*. 2e édition. Éditions Elsevier-Masson, Issy-les-Moulineaux. ISBN 978-2-294-70317-1.

GARCIA, Émilie, LE CAM, Marion, ROCHER, Thierry et alii, 2015. Méthodes de sondage utilisées dans les programmes d'évaluation des élèves. In : *Éducation & Formations*. [en ligne]. Mai 2015. MENESR-DEPP. N°86-87, pp. 101-117. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://cache.media.education.gouv.fr/file/revue_86-87/63/8/depp-2015-EF-86-87_424638.pdf*.

GOLDBERG, Milton et HARVEY, James, 1983. A Nation at Risk: The Report of the National Commission on Excellence in Education. In : *The Phi Delta Kappan*. [en ligne]. Vol. 65, n°1, pp. 14-18. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.jstor.org/stable/20386898*.

GOULD, Stephen Jay, 1997. *La mal-mesure de l'homme*. Éditions Odile Jacob. ISBN 978-2-7381-0508-0.

KOLEN, Michael J. et BRENNAN, Robert L., 2004. *Test Equating, Linking, and Scaling: Methods and practices*. 3e édition. Éditions Springer-Verlag, New York. ISBN 978-1-4939-0317-7.

LAVEAULT, Dany et GRÉGOIRE, Jacques, 2002. *Introduction aux théories des tests en psychologie et en sciences de l'éducation*. 3e édition, janvier 2014. Éditions De Boeck, Bruxelles. ISBN 978-2-804170752.

LE CAM, Marion et SALLES, Franck, 2020. *TIMSS 2019 – Mathématiques au niveau de la classe de quatrième : des résultats inquiétants en France.* [en ligne]. Décembre 2020. MENESR-DEPP. Note d'information, n°20.47. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.education.gouv.fr/timss-2019-mathematiques-au-niveau-de-la-classe-de-quatrieme-des-resultats-inquietants-en-france-307819*.

NINNIN, Louis-Marie et PASTOR, Jean-Marc, 2020. *Cedre 2008-2014-2019 Mathématiques en fin d'école : des résultats en baisse.* [en ligne]. Septembre 2020. MENESR-DEPP. Note d'information n°20.33. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.education.gouv.fr/cedre-2008-2014-2019-mathematiques-en-fin-d-ecole-des-resultats-en-baisse-306336*.

NINNIN, Louis-Marie et SALLES, Franck, 2020. *Cedre 2008-2014-2019 Mathématiques en fin de collège : des résultats en baisse*. [en ligne]. Septembre 2020. MENESR-DEPP. Note d'information n°20.34. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.education.gouv.fr/cedre-2008-2014-2019-mathematiques-en-fin-de-college-des-resultats-en-baisse-306338*.

OCDE, 2020. *PISA 2018 Technical Report.* [en ligne]. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.oecd.org/pisa/data/pisa2018technicalreport/*.

PIÉRON, Henry, 1963. *Examens et docimologi*e. 1er janvier 1963. Presses universitaires de France.

ROCHER, Thierry et HASTEDT, Dirk, 2020. *International large-scale assessments in education: a brief guide*. [en ligne]. Septembre 2020. IEA Compass: Briefs in Education, Amsterdam, n°10. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *http://files.eric.ed.gov/fulltext/ED608251.pdf*.

ROCHER, Thierry, 2015a. Mesure des compétences : méthodes psychométriques utilisées dans le cadre des évaluations des élèves. In : *Éducation et Formations*. [en ligne]. Mai 2015. MENESR-DEPP. N°86-87, pp. 37-60. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://cache.media.education.gouv.fr/file/revue_86-87/63/8/depp-2015-EF-86-87_424638.pdf*.

ROCHER, Thierry, 2015b. PISA, une belle enquête : lire attentivement la notice. In : *Administration et Éducation*. [en ligne]. Association Française des Acteurs de l'Éducation. N°145, pp 25-30. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.cairn.info/revue-administration-et-education-2015-1-page-25.htm*.

ROZENCWAJG, Paulette, 2011. La mesure du fonctionnement cognitif chez Binet. In : *Bulletin de psychologie*. [en ligne]. 2011/3, N°513, pp. 251-260. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.cairn.info/revue-bulletin-de-psychologie-2011-3-page-251.htm*.

SALLES, Franck, DOS SANTOS, Reinaldo et KESKPAIK, Saskia, 2020. *When didactics meet data science: process data analysis in large-scale mathematics assessment in France*. [en ligne]. 29 mai 2020. IEA-ETS Research Institute Journal, Large-scale Assessments in Education, 8:7. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://doi.org/10.1186/s40536-020-00085-y*.

THÉLOT, Claude, 1992. *Que sait-on des connaissances des élèves ?* Octobre 1992. Les Dossiers d'Éducation et formations. N°17.

TROSSEILLE, Bruno et ROCHER, Thierry, 2015. Les évaluations standardisées des élèves. Perspective historique. In : *Éducation et Formations*. [en ligne]. Mai 2015. MENESR-DEPP. N°86-87, pp. 15-35. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://cache.media.education.gouv.fr/file/revue_86-87/63/8/depp-2015-EF-86-87_424638.pdf*.