

ÉVALUATIONS NATIONALES AU CP EN MATHÉMATIQUES EN FRANCE : BIAIS ET IMPACTS SUR LE TRAITEMENT DES DONNÉES ISSUES DU DISPOSITIF ÉVALAIDE

| GUENAI* MÉLANIE

Résumé | Des résultats récents issus des évaluations nationales de CP et CE1 en France font état de différences de réussite genrées en mathématiques dès la mi-CP (Martinot *et al.*, 2021), semblant montrer des ancrages précoces des stéréotypes, jamais encore observés en France. Nous montrons, à partir d'observations en classe de CP, que ces différences pourraient être dues à des biais liés aux conditions de passation des tests cadrés par le dispositif ÉvalAide, dont les buts mal définis interrogent la pertinence des données produites et de leur exploitation.

Mots-clés : évaluations standardisées, biais de mesures, performance, pilotage éducatif, compétences mathématiques, inégalités de genre

Abstract | Recent results from national assessments of CP and CE1 in France report gender differences in achievement in mathematics from the middle of CP (Martinot *et al.*, 2021), appearing to show early anchoring of stereotypes, never observed before. We show, based on observations in CP classes, that these differences could be due to biases linked to the conditions of administration of the tests framed by the ÉvalAide system, whose poorly defined goals call into question the relevance of the data produced and their use.

Keywords: Standard assessments, measurement bias, performance, educational management, mathematical skills, gender inequalities

I. INTRODUCTION

Cette communication a pour but d'analyser l'adéquation des mesures réalisées lors des évaluations nationales de CP-CE1 cadrées par le dispositif ÉvalAide au regard des compétences mathématiques des élèves qu'elles prétendent décrire. En prenant appui sur l'analyse du dispositif, on s'intéresse à l'ensemble du processus ayant abouti à ces résultats, depuis la conception de l'outil de mesure jusqu'aux publications. À partir d'observations réalisées en CP au cours du premier semestre 2024-2025, on examine notamment les conditions dans lesquelles se réalisent effectivement ces évaluations pendant la classe.

Nous mettons au jour des décalages entre les observations et les préconisations du dispositif ÉvalAide, tant du point de vue de la pratique de classe que de l'analyse des données et du traitement qui en est fait. En particulier, l'analyse didactique *a priori* et des productions des élèves observés montre que certaines des mesures sont sans rapport avec les compétences annoncées ou qu'elles les dévalorisent en raison de biais externes aux mathématiques. Les observations des ressentis des élèves fournissent également des pistes explicatives des différences de performances genrées récemment médiatisées (Martinot *et al.*, 2021), trop hâtivement assimilées à des performances mathématiques. Elles ouvrent des perspectives d'études complémentaires relevant de la sociologie de l'éducation et de la psychologie de l'enfant, dont les enjeux sont essentiels : il s'agit d'évaluer la balance coûts bénéfices

* Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay – France – melanie.guenais@universite-paris-saclay.fr

d'un dispositif imposé à tous les élèves dans le cadre scolaire et aux effets potentiellement discriminatoires dans la construction de leur rapport aux savoirs, notamment en mathématiques.

II. DESCRIPTION DU DISPOSITIF ÉVALAIDE

La France a mis en place, depuis 2018, des évaluations nationales pour tous les élèves à l'entrée au CP, à la mi-CP et à l'entrée en CE1¹. Le cadre est fixé par le dispositif ÉvalAide élaboré par un groupe de travail du Conseil scientifique de l'Éducation nationale (CSEN) dirigé par le neuroscientifique S. Dehaene. Nous en décrivons la démarche et les objectifs à partir de l'analyse du document d'accompagnement (Sprenger-Charolles *et al.*, 2019).

1. *Un objectif principal concernant à mesurer des compétences cognitives*

Le dispositif vise à évaluer

« les compétences cognitives de tous les élèves français en début et en milieu de CP ainsi qu'en début de CE1 [pour] identifier les besoins spécifiques de chaque élève [...]. Le nouveau dispositif utilise des épreuves issues de la recherche qui permettent de bien repérer la nature des difficultés (qui peuvent concerner les traitements visuels, phonologiques ou sémantiques, par exemple) » (p. 3).

Pour les mathématiques, les items sont regroupés selon 5 thèmes définis par les auteurs : (1) la connaissance des symboles des nombres ; (2) le passage rapide des symboles à leur sens et vice-versa (dénombrement, comparaison des quantités) ; (3) la compréhension de la position des nombres sur une ligne numérique ; (4) la résolution de petits problèmes arithmétiques adaptés au niveau de l'élève ; (5) la maîtrise de certains concepts de base dans le domaine de la géométrie

2. *Un objectif secondaire de dépistage précoce de difficultés par des tests « prédictifs »*

Les choix des tests sont développés dans la partie B.2 du document. Les tests portant sur les points 1, 2, 3 et 5 constituent des prédictifs de réussite ou difficulté supposés² pour les compétences ultérieures en mathématiques, bien que la causalité ne soit pas prouvée³. Seul le test du point 4 est déterminé pour rendre compte de la capacité à résoudre des problèmes arithmétiques, faiblesse des élèves français selon l'enquête internationale TIMSS.

3. *Des modalités d'organisation des passations éloignées des sources citées.*

Les auteurs déclarent s'appuyer sur des dispositifs étrangers portant pour la plupart sur le dépistage précoce de troubles des apprentissages (surtout le langage). Bien que le protocole de test de performance de lecture en Angleterre soit présenté en modèle, ils s'en démarquent cependant délibérément (multiplication des items, aspect longitudinal des tests, dates de passations), insistant sur l'objectif d'une mise en place de tests prédictifs, pour permettre un traitement précoce des élèves en

¹ Élèves ayant entre 6 ans (en CP) et 7 ans (en CE1) dans l'année.

² « Pour chaque test, nous indiquons ses fondements scientifiques et son intérêt pour détecter ou anticiper d'éventuelles difficultés scolaires ». p. 26 : « indicateurs de capacités ultérieures », p. 27 : « L'épreuve Symp de comparaison numérique [...] corrélée à la réussite scolaire en mathématiques, a été intégrée au dispositif ÉvalAide [et] permet de détecter les élèves en difficulté. »

³ p. 24, les auteurs précisent : « Les méta-analyses récentes signalent des relations modestes, mais significatives, entre la capacité précoce de discrimination approximative des nombres et la réussite ultérieure en mathématiques. Cependant, la direction de la causalité que reflète cette corrélation reste débattue ».

difficulté potentielle⁴. Ils généralisent les items en temps court et contraint, pour « faciliter les comparaisons entre élèves », « limiter la fatigue », ou rendre compte de l'automatisation. Ils introduisent des distracteurs destinés à « mieux cerner les compétences acquises par un élève, ainsi que l'origine des erreurs qu'il commet ».

4. *Exploitation des données : un outil multitâche, à destination d'abord pédagogique ?*

Alors qu'il s'agit, page 35, de « fournir aux équipes pédagogiques des repères sur les compétences de chaque élève », les tests sont aussi destinés aux statistiques nationales, comme « indicateurs précieux sur les progrès et les difficultés des élèves [...] même si ce n'est pas l'objectif primaire du dispositif », ou bien pour « mesurer si les gestes pédagogiques sont efficaces pour réduire les inégalités et faire progresser tous les élèves ». Côté recherche, ces données devraient permettre de confirmer la nature prédictive de ces tests, car les preuves scientifiques manquent dans le domaine des nombres et surtout de la géométrie.

Les mesures des résultats sont fondées sur un comptage des items réussis, qui fournit un score global sur une passation⁵. Deux types d'exploitations sont ensuite diffusées : des valeurs seuils définissant 3 groupes : « à besoin », « fragile » ou « satisfaisant », et des scores de performance, analogues aux évaluations bilans internationales.

III. PROBLÉMATIQUE

Le dispositif ÉvalAide est présenté comme une aide pédagogique au service de la progression des élèves. Il est constitué de séries de tests « prédicteurs » des difficultés futures, sur la base de corrélations supposées avec des compétences générales dont la valeur reste faible, ne fait pas consensus et doit être contrôlée. Alors qu'il n'est pas fondé sur des compétences mathématiques visées par des programmes scolaires, les tests sont présentés comme « acquis fondamentaux » dans le guide du professeur⁶ dont les parents reçoivent un graphe étoilé des scores obtenus. De même, l'exploitation publique des données produites assimile ces tests à des mesures de performances en mathématiques, interrogeant la pertinence de cette exploitation :

1. Ces tests permettent-ils effectivement de mesurer et comparer des compétences en mathématiques ? Comment prendre en compte des entraînements ciblés sur les tests ?
2. Quels sont les facteurs impactant les scores calculés ? En particulier, comment analyser les différences filles-garçons apparaissant au cours du CP ?
3. En place depuis 6 ans, quels bénéfices pour les enfants en difficulté met-on en évidence ?
4. Quels sont les coûts d'un tel dispositif ? En particulier, que peut-on objectiver sur l'impact de tels dispositifs pour la vie de l'élève et son rapport aux savoirs, aux mathématiques ?

À partir d'observations en classe de CP, nous apportons ici de premières pistes d'analyse pour éclairer ces questions. En confrontant la présentation théorique des résultats diffusés nationalement à la réalité de leur production, nous mettons en évidence des biais qui les rendent finalement peu représentatifs des compétences en mathématiques des élèves.

⁴ p. 27 « Il ne s'agit donc pas de vérifier les acquis à la fin de l'année, mais d'alerter et d'accompagner les enseignants sur les difficultés potentielles afin de pouvoir aider l'élève bien avant qu'il se trouve en échec ».

⁵ Guide de scores, notice technique pour enseignants

⁶ Guide pour le professeur - Repère 2024, MEN, en ligne

IV. CONFRONTATIONS AUX OBSERVATIONS EN CLASSES EN CP

Les passations nationales de début CP comprennent 5 séquences (3 de français et 2 de mathématiques, séquences 2 et 4) qui se déroulent chaque jour de la semaine pendant la 2^e ou la 3^e semaine suivant l'entrée en primaire début septembre. Une seconde série de la même nature se déroule vers la mi-janvier. Nous restituons une synthèse des observations réalisées au cours du premier semestre 2024-2025 dans 2 classes de CP (CP1 : classe de 14 élèves en Réseau d'Éducation Prioritaire, IPS⁷ 86 ; CP2 : 22 élèves, milieu social mixte IPS 110). Ces observations se sont déroulées en 9 séances au total, incluant les évaluations nationales de début de CP et de mi-CP, ainsi que des observations en pratique ordinaire, dont l'une en évaluation de classe (commune à l'école en CP1 ou individuelle à la classe en CP2).

1. *Modalités : divergences entre consignes d'ÉvalAide et l'organisation habituelle*

Observations des organisations des tests nationaux en septembre (T1) et janvier (T2) :

Les enseignantes suivent les consignes dictées par le guide du professeur. Les enfants sont séparés ou déplacés pour leur interdire de copier, ne doivent ni parler ni se déplacer pendant la durée de la passation. La durée totale des passations en T1 est d'environ 40 minutes pour un temps de travail affiché de l'ordre de 10 minutes. Elle est d'environ 20 minutes en T2⁸. Le support est un livret blanc de taille A4 d'une soixantaine de pages numérotées. La plupart des items sont segmentés et contraints par un temps limité, parfois très court, que les enseignantes font respecter. Leur aide éventuelle doit se limiter à la compréhension des modalités de report du résultat. En pratique, l'un des tests est explicite. La contrainte de temps permet peu de venir en aide aux élèves qui en ont besoin, pour des raisons variées (crayon cassé, page mal tournée, dictée de nombres non entendue, enfant démobilité, etc.). Les réactions notées en T1 montrent de l'inquiétude et de la fatigue chez les élèves, du stress chez les enseignantes dans les deux classes, conduisant à des paroles ou comportements parfois brusques. Lors de la passation à T2, malgré l'assouplissement des modalités, l'anxiété est globalement perceptible en CP2 à l'annonce de l'utilisation du livret. En CP1 après l'évaluation, sont exprimés : du stress pour la majorité des garçons, du désintérêt pour deux ou trois filles et un garçon, de l'intérêt pour trois filles. La contradiction entre le discours préconisé pour présenter la passation de manière rassurante et le climat de classe résultant des prescriptions est perçue par les élèves⁹.

Observation des évaluations ordinaires de classe observées en décembre :

En CP1, les élèves se déplacent à une place dédiée aux évaluations individuelles pour éviter les copies, ils sont autonomes, contrairement au début d'année. En CP2, il n'y a pas de séparation ni de déplacement. La durée des évaluations est d'environ 15 à 20 minutes dans les 2 classes. L'évaluation se déroule sur une (CP2) ou plusieurs séances (CP1). L'enseignante présente aux enfants un ensemble d'activités à compléter sur une page A4 recto-verso pliée en deux qu'ils gèrent en autonomie dans une durée adaptée aux besoins des élèves les plus lents, les plus rapides pouvant choisir une activité libre lorsqu'ils ont fini. Les activités proposées à l'intérieur de chaque domaine évalué sont progressives. Dans une séance, elles peuvent être réalisées dans l'ordre souhaité au rythme de l'élève. L'enseignante circule et remobilise les élèves qui semblent perdus, réexplique la consigne individuellement, encourage, vérifie que le matériel est bien disponible (crayon cassé, couleurs, etc.), réajuste les

⁷ Indice de Position Sociale des établissements scolaires, accessibles [en open data](https://data.education.gouv.fr/) sur le site data.education.gouv.fr.

⁸ En T1, la séquence 2 observée en CP1, séquence 4 observée en CP2 ; en T2, seule la séquence 2 est observée.

⁹ Une élève restitue oralement en présentation en début de CP qu'« on a le droit de se tromper... un peu ».

consignes au besoin. Il n'a pas été noté lors de ces évaluations des réactions d'élèves permettant d'identifier des changements des habitudes de classe ni pour les élèves, ni pour l'enseignante.

2. *Analyse didactique des tâches mathématiques et des productions d'élèves*

Parmi les élèves en grande difficulté on identifie des enfants allophones (en T1) ou à besoins particuliers déjà identifiés, certains sont dispensés des tests en T2, contrairement aux consignes. Nous décrivons les résultats et la nature des difficultés observées lors des tests d'ÉvalAide¹⁰.

Numération

Ce domaine comporte un test de lecture et un test d'écriture en chiffres de 10 nombres dits à l'oral entre 0 et 10 en T1. En T2, seul un test d'écriture de 10 nombres dits entre 0 et 31 est proposé. Environ 20 % des élèves sont identifiés comme « fragiles » en T1. Les erreurs résultent surtout de problèmes externes aux mathématiques¹¹. Une élève signalée comme « à besoin » se révèle très performante aux évaluations de fin de trimestre. Une autre, non repérée en T1, est en grande difficulté. Elle présente des inversions dans l'écriture des nombres à 2 chiffres. Si les contenus mathématiques proposés sont adaptés pour des élèves de CP, leur mise en œuvre provoque des difficultés. Les modalités de passation tendent à déstabiliser les élèves et à sous-évaluer leurs compétences dans ce domaine.

Dénombrement – Comparaison

Ce domaine comporte un test de dénombrement et un test de comparaison en T1. Seule la comparaison est reproduite en T2. Le dénombrement contient 8 items qui ne posent globalement aucun problème de réalisation. Seuls quelques élèves en difficulté dans l'ensemble des tests sont aussi en difficulté pour dénombrer les collections désorganisées au-dessus de 5. La longueur des items impacte la durée de la séquence de tests, augmentant la fatigue des élèves. Le test de comparaison comporte en effet 40 paires de nombres à 1 puis 2 chiffres dont il faut barrer le plus grand. La consigne donnée aux élèves précise : « C'est un exercice de vitesse. Il dure seulement 1 minute. Ce qui compte, c'est d'en faire le plus possible. »¹². En T1, on relève des réactions d'anxiété chez les élèves dès la découverte de la page. Le chronométrage conduit à des situations de stress chez les élèves comme chez les enseignantes. La frustration et le sentiment d'échec sont perceptibles aussi bien en T1 qu'en T2¹³. Très peu d'erreurs de comparaison ont été observées malgré un taux de remplissage inférieur en moyenne à la moitié. Le sens de lecture de haut en bas nuit à la réalisation de la tâche pour de nombreux élèves, dont ceux identifiés comme « fragiles ». La consigne « barrer la bonne réponse » est source d'échec : les réponses des élèves qui utilisent à bon escient les signes < ou > ou qui entourent le plus grand sont comptées fausses ce qui conduit à des productions « fragiles ». Au contraire, une production « satisfaisante » comporte un taux d'erreur important dû à des inversions dans le sens de la lecture des nombres à 2 chiffres. Les modalités d'évaluation mettent la plupart des élèves en difficulté alors que la compétence semble acquise. Les mesures réalisées ne rendent donc pas compte des acquis en mathématiques et signalent à tort des élèves en difficulté.

¹⁰ Livret élève et Guide pour le professeur - Repère 2024, MEN, [en ligne](#).

¹¹ Nombre non entendu, report dans la mauvaise case, rythme trop rigide, manque de temps, élève en détresse liée aux conditions de passation.

¹² L'objectif de ce test issu de l'expérimentation « SYMP » est le dépistage précoce d'élèves présentant des déficiences en mathématiques (Brankaer *et al.*, 2017). L'article montre une corrélation positive qui tend à diminuer durant la scolarité avec les compétences mathématiques, en précisant l'absence de causalité.

¹³ Agitation, souffle bruyant, commentaires entre élèves, protestation, gestes d'énervement chez plusieurs garçons.

Ligne numérique

Il s'agit d'un test de repérage d'étiquette sur un segment gradué présenté sous forme de QCM présentant 6 nombres dans le désordre. 6 items sont proposés en T1 avec des graduations unitaires entre 0 et 10, puis 12 items en T2, avec des graduations unitaires sur des segments de longueur 4 ou 10, avec des bornes variant jusqu'à 35. En T1, l'annonce de devoir tourner les pages seul provoque des réactions d'inquiétudes et du découragement chez les élèves. Les erreurs proviennent souvent d'un manque de compréhension de la consigne. Le format QCM accentue l'échec¹⁴. Ce test ne relève pas de la comparaison en T1, tous les nombres proposés étant inclus dans le segment représenté. En T2, les productions n'ont pas été observées. Les résultats relevés montrent que les élèves en difficulté sur ce test ont des difficultés généralisées.

Opérations arithmétiques

Il s'agit de deux tests proposés en T2 sous forme de QCM comportant, dans les réponses proposées, des nombres dans le désordre et des distracteurs. L'un porte sur les additions, l'autre sur les soustractions, ils comportent chacun 10 items à réaliser en 3 minutes. Le problème de contrainte de temps est signalé dans le bilan de l'inspection générale lors du déploiement national du dispositif (Bruschini *et al.*, 2019¹⁵). Les additions portent sur deux nombres dont le premier est inférieur à 20 et le deuxième est compris entre 2 et 10. On note 3 passages de dizaine (avec des additions de 5 ou 6) et 2 de vingtaine. La progressivité est irrégulière. Les 2/3 des élèves sont en difficulté en CP2 alors qu'ils sont très peu à l'être en CP1. Les CP2 n'avaient pas été évalués à l'écrit sur les opérations symboliques en décembre, au contraire des CP1. Les attendus relèvent d'un niveau de CP très avancé à cette date de l'année. Les modalités augmentent l'échec des élèves. Les tests ne rendent pas compte des compétences visant les décompositions décimales et les compléments des petits nombres à la dizaine. Ils encouragent à modifier la progression habituelle en privilégiant le travail symbolique sur les opérations au détriment de la notion de dizaine, indispensable pour comprendre l'écriture décimale.

Résolution de problèmes additifs :

En T1, les 2 tests comportent chacun 3 problèmes de comparaison de collections sous forme de QCM. Certains énoncés ambigus peuvent avoir plusieurs réponses. Le premier test est figuratif et bien réussi par les élèves de CP1 qui ont été entraînés au protocole à suivre pour comparer des collections dessinées. La réussite d'un élève allophone suggère que le protocole est appliqué indépendamment de l'énoncé oral. Le second test, seulement oral, fait échouer les 2/3 des élèves dans les 2 classes. La présentation sous forme de QCM engendre des réponses comptées fausses qui ne sont pas forcément des erreurs : un certain nombre d'élèves entoure les données qui figurent dans les réponses proposées et écrit la réponse dans le cadre réservé au brouillon. En T2, 2 tests sont proposés comportant 2 puis 3 problèmes, le premier additif, les autres soustractifs¹⁶. L'échec est massif en CP2 contrairement au CP1 qui a été entraîné. Le format QCM est encore une fois source d'échec ou d'ambiguïté sur l'interprétation des réussites en raison du choix des variables numériques (les bonnes réponses peuvent être trouvées avec une démarche erronée). L'hétérogénéité des structures de problèmes présentés en T2 dénote de la répétition des comparaisons des problèmes en T1. Les attendus demandés du point de vue de la modélisation semblent très ambitieux en T2 et prématurés en T1. Le format et

¹⁴ Un élève choisit l'étiquette placée au même endroit que la graduation, plusieurs écrivent directement sur l'emplacement du label.

¹⁵ p. 12 « il reste difficilement acceptable pour les équipes de voir proposer une série de calculs que les élèves ne pourront pas tous traiter pendant le temps accordé afin de mesurer leur dextérité en calcul. Un travail d'explicitation sur ce point est encore nécessaire pour les enseignants, mais aussi pour les inspecteurs et les formateurs. »

¹⁶ Transformation ou composition, puis 2 transformations avec recherche de l'état final, puis de la transformation, et une comparaison

l'hétérogénéité des énoncés nuisent à la réussite des élèves, surtout les moins confiants. Cette analyse est cohérente avec la très faible réussite nationale à ces items (Le Breton, 2024).

Domaine de la géométrie

Le test de géométrie, présenté en début de CP et en début de CE1, consiste à entourer parmi plusieurs formes proposées celle qui correspond au « collage » de deux formes symétriques disposées de manière variée. Nous restituons 3 exemples parmi les 9 items proposés :

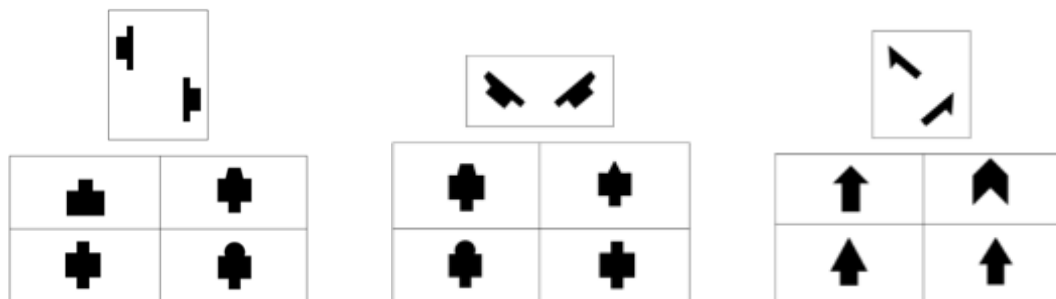


Figure 1 – 3 items de « géométrie » de l'exercice 8, séquence 2, CP (T1), reconduits en CE1 (T3). Consigne : « entoure la forme qu'on obtient en assemblant les 2 formes »

D'abord, la consigne est mathématiquement incorrecte : l'assemblage n'étant pas unique, les formes possibles sont nombreuses. Les compétences mathématiques relèvent de la comparaison des longueurs ou angles proches dans des configurations complexes. Cet exercice ne correspond à aucun attendu mathématique à ce niveau de classe et ne repose que sur la capacité visuelle à discriminer des petites différences. Il met en difficulté des élèves pour des raisons externes aux mathématiques (troubles de la vision, troubles dys, attention, compréhension de la consigne, etc.). Dans la classe observée, aucun élève ne réussit la totalité des items, deux élèves sont signalés en difficulté uniquement pour ce test. Ce test est sans rapport avec des compétences en géométrie en CP ou en CE1, il déroute les enseignantes qui trouvent sa présentation étonnante et croient dépister ainsi des élèves en difficulté qu'elles n'avaient pas identifiés.

V. ANALYSE

La littérature scientifique a étudié les protocoles des tests standardisés (e.g. Heulot *et al.*, 2012 ; Trosseilles et Rocher, 2015). Dans les rapports issus des analyses des tests antérieurs à 2015 en France, le Haut Conseil à l'Éducation identifie le manque de clarté sur la finalité, diagnostique ou bilan, comme une entrave à la valeur informative des données recueillies, susceptible de transmettre des informations non pertinentes (e.g. Haut Conseil de l'Éducation, 2011). Le Haut Conseil à l'Éducation (Salines et Vrignaud, 2001, annexe 6) recommande des protocoles distincts pour chacun de ces objectifs, les estimant non compatibles. En particulier, la passation d'un test bilan doit être uniforme avec des items non publics balayant un socle d'attendus et proposer un codage binaire des réussites laissant en retrait l'analyse d'erreur. Au contraire, une évaluation diagnostique peut ne porter que sur quelques items et doit pouvoir s'adapter aux besoins individuels. Si le modèle suédois cité dans Martinot (2021) s'inscrit bien dans ce cadre diagnostique¹⁷, les modalités présentées par le dispositif ÉvalAide s'y opposent.

¹⁷ Pas de visée comparative entre les élèves, de notes standardisées ni de remontée nationale, liberté aux enseignants de concevoir leur dispositif d'évaluation sur la base de propositions, évaluations individualisées, attention au bien être et à valoriser les compétences de l'enfant. Extraits : « Selon [...] l'Agence nationale suédoise pour l'éducation [...] il n'est pas approprié d'utiliser l'échelle nationale de notation dans le plan de développement. Les formulaires

3. *Peut-on parler de tests de compétences en mathématiques ?*

Des contenus non représentatifs des compétences en mathématiques en CP

L'analyse des contenus des tests d'ÉvalAide montre de nombreuses faiblesses qui diminuent leur validité. Certains tests sont sans lien clair avec des compétences mathématiques de CP (géométrie, ligne numérique), ou mal adaptés au niveau de classe. Des thèmes sont absents (système décimal) tandis que d'autres sont surreprésentés (comparaison), donnant à l'ensemble une répartition déséquilibrée et peu diversifiée. La progressivité est irrégulière, rendant peu fiable la mesure de la compétence visée (résolution de problèmes). Même par domaine, les contenus semblent mal adaptés pour rendre compte des compétences mathématiques de CP.

Des conditions de recueil des productions d'élèves inadaptées au cadre scolaire ordinaire

La présentation des contenus se démarque des évaluations usuelles conçues par les équipes éducatives : le support, la longueur, la répétitivité, le format ou le manque de progressivité sont autant de décalages de nature à perturber les conditions de travail des élèves. Les modalités de passation créent un environnement qui inquiète les élèves. En début d'année, il déstabilise les élèves qui ne connaissent encore ni les lieux ni les personnes et met l'enseignante face à des injonctions contradictoires. Ce cadre imposé nuit à la mise en confiance des élèves, favorise la mise en place d'un climat de compétition, et le découragement. Nos observations permettent d'identifier des biais dus aux items ou au contexte qui nuisent à la réalisation des tâches.

Des critères de mesure des productions qui sous-estiment les compétences mathématiques

Du point de vue strict des compétences mathématiques, le respect des modalités de calculs de score entraîne la non prise en compte des productions montrant des compétences acquises, mais non conformes au cadre attendu : entourer plutôt que barrer, écriture libre plutôt qu'entourer, absence d'erreur mais manque de temps pour réaliser les tâches dans un cadre chronométré. Ces biais d'items, dévalorisant les réussites, favorisent le traitement non uniforme des passations et des remontées des résultats par les enseignants.

Un pilotage national avec des items libres : quid des effets d'un entraînement au test ?

Enfin, l'exploitation des données pour des comparaisons nationales de scores présente un risque de surentraînement aux tests dont les contenus sont ouverts, sans contrôle sur les bénéfices réels apportés pour les apprentissages d'un tel entraînement. L'enjeu diagnostic des élèves en difficulté est donc en contradiction avec l'enjeu de classement, national ou entre élèves.

Finalement, ces tests présentent des déséquilibres des contenus, entre et à l'intérieur des domaines considérés. Le dispositif introduit des biais induits par la présentation, les conditions de réalisation et les critères de mesures de réussite qui déforment ou sous-estiment les compétences acquises. Il occulte ou déforme la réalité de la pratique des mathématiques, qui nécessite temps de réflexion et coopération pour travailler le raisonnement, indispensable à la progression des élèves et leur autonomie, point faible signalé régulièrement dans les études internationales. La conception des contenus comme les modalités d'évaluation nécessiteraient donc d'être reprises de manière critique pour donner une mesure fidèle des compétences en mathématiques, si telle est leur destination. Par conséquent, assimiler ces scores de performances à des compétences en mathématiques est scientifiquement peu fondé.

d'évaluation [de] l'Agence nationale suédoise pour l'éducation ne sont que des exemples de la manière dont de tels formulaires peuvent être conçus. »

4. *Facteurs impactant les calculs de score et analyses possibles des résultats*

Nos observations permettent de montrer que le dispositif introduit de nombreux biais relevant d'aptitudes sociales ou psychologiques impactant les performances : mise en compétition individuelle, rythme contraint et vitesse, items inhabituels, QCM, tâches monotones et qui font peu de sens. Des différences de scores filles/garçons dans les évaluations nationales en France apparaissent dans toutes les classes à partir de 2018, date de la mise en place des tests d'ÉvalAide (Eteve *et al.*, 2025). Elles sont faibles, persistantes et nouvelles¹⁸. Ces observations montrent la présence d'un avantage masculin cohérent avec les impacts de ces biais identifiés par la littérature (e.g. Stoevenbelt *et al.*, 2023 ; Montolio *et al.*, 2021). Interpréter des petites différences de score comme des compétences mathématiques qui seraient inégales entre filles et garçons dès quelques mois de scolarisation au CP est donc non pertinent.

5. *Quels apports du dispositif pour les élèves en difficulté et le niveau général ?*

Ce dispositif est administré à l'ensemble des classes de CP et CE1 depuis 6 ans, aucun bilan n'est publié sur son intérêt pour les élèves. Une légère augmentation des résultats de la mi-CP peut être liée à un entraînement spécifique ou un changement dans les contenus présentés, sans rapport établi avec de meilleures compétences. Les résultats des tests mettent en évidence des inégalités filles-garçons nouvelles dès le milieu de CP qui s'accroissent entre 2018 et 2021 puis se stabilisent (Eteve *et al.*, 2025). Il n'y a pas de progression significative aux résultats de TIMSS 2023 en CM1, avec des scores pour la France parmi les plus faibles des pays participants. Ce bilan ne permet donc pas d'établir l'existence d'un bénéfice apporté par ce dispositif, ni global pour l'ensemble des élèves, ni pour ceux en difficulté.

6. *Les risques d'un dispositif intrusif et d'informations biaisées sur les apprentissages*

L'exploitation médiatisée de ces scores à des fins de classement national ne correspond pas à l'intention d'aide au diagnostic de difficultés précoces présentée aux enseignants. Elle risque d'encourager l'entraînement aux tests dans un but de performance, au détriment de son objectif diagnostic. L'omission des biais introduits par le dispositif entraîne la transmission d'informations déformées sur les résultats publiés, aussi bien pour les familles qu'au national. D'une part, en focalisant sur de faibles différences de performances genrées, attribuées sans discernement aux compétences en mathématiques, on contribue à renforcer des idées reçues sur des moindres compétences supposées des filles en mathématiques. D'autre part, en imposant des modalités rigides de passation qui créent de l'inquiétude chez les enfants dès leur entrée à l'école et les sous-évaluent, on risque une dégradation de la relation aux savoirs en mathématiques, dont on donne une image dénaturée de la pratique.

VI. CONCLUSION

Les observations de terrain permettent de montrer des décalages importants du vécu des évaluations par les élèves et leurs enseignants avec les intentions affichées du dispositif pour l'intérêt de l'enfant. Ces décalages sont renforcés par une ambiguïté dont les évaluations sont porteuses entre des tests prédictifs de difficulté, d'ordre cognitif, et des tests de mesure de compétences mathématiques, d'ordre didactique. Si cette confusion n'est pas nouvelle, elle introduit des biais qui impactent le milieu scolaire et peuvent fausser les mesures et leurs interprétations. Dans ces conditions, le dispositif pourrait s'avérer inefficace à la fois dans son rôle diagnostic et comme bilan de performances en mathématiques à l'échelle du pays.

¹⁸ Aucune différence significative lors des évaluations de : CP panels entre 1997 et 2011, fin de CE1 en 2014, fin de CE2 et début de 6e jusqu'en 2017 ; fin de 3e en 2015 [MENESR-DEPP, 2015, 2016, 2017, 2018, 8.1].

Les modalités de passation collectives et contraintes, motivées par une exploitation nationale, conduisent à la création d'un climat de classe insécurisant et hors cadre scolaire, contraire à une mise en confiance des élèves et la possibilité d'une passation diagnostique sereine adapté aux besoins. Les contenus des tests sont peu représentatifs des compétences mathématiques visées selon les programmes scolaires. La présentation et les modalités de passation de ces tests dénaturent et tendent à sous-évaluer les compétences acquises, mettant parfois de très bons élèves en difficulté. Par conséquent, les mesures de scores transmettent une information peu pertinente des compétences en mathématiques au niveau national.

Notre analyse permet d'identifier également dans le dispositif des biais qui induisent des conditions psychologiques ou sociales favorisant la réussite des garçons, sans rapport avec les mathématiques. Ils pourraient expliquer les petites différences de scores entre les filles et les garçons. Alors qu'on ne peut établir de bénéfices de ce dispositif pour diminuer les inégalités scolaires, ses effets pervers questionnent. En négligeant les biais introduits par l'administration de ces tests cognitifs, on transmet une interprétation infondée en attribuant à des petites différences de scores des inégalités de compétences en mathématiques, renforçant des stéréotypes négatifs sur les femmes. Plus largement, l'installation répétée de ce dispositif, au-delà de son coût financier et humain pour l'État, pourrait nuire aux apprentissages et à l'image renvoyée de la pratique des mathématiques, effets pervers déjà identifiés par la Corée.

RÉFÉRENCES

- Brankaer, C., Ghesquière, P. et De Smedt, B. (2017). Symbolic magnitude processing in elementary school children: A group administered paper-and-pencil measure (SYMP Test). *Behavior Research Methods*, 49(4), 1361-1373. <https://doi.org/10.3758/s13428-016-0792-3>
- Bruschini, B., Hunault, O., Yebbou, J. (pilotes), Galais, P., Richet, B. et Turin-Bartier, M. (2019). L'organisation et la passation des évaluations nationales dans le premier degré [Rapport de recherche] (n° 2019-096). <https://www.education.gouv.fr/l-organisation-et-la-passation-des-evaluations-nationales-dans-le-premier-degre-41684>
- Eteve, Y., Garnerio, M. et Paillet, V. (2025, février). Évolution des écarts de performances entre filles et garçons en mathématiques, au fil du temps et de la scolarité. *Note d'Information*, (25.04). <https://doi.org/10.48464/ni-25-04>
- Heulot, F., Robert, A., Rousse, S., Vancauwenberghe, E., Roditi, E. et Chesné, J. (2012). Un point de vue didactique sur les questions d'évaluation en éducation. Les compétences et le socle commun. Dans M. Lattuati, J. Penninckx et A. Robert (dir.), *Une caméra au fond de la classe de mathématiques* (p. 279-319). Presses universitaires de Franche-Comté.
- Haut Conseil de l'Éducation. (2011). *Les indicateurs relatifs aux acquis des élèves. Bilan des résultats de l'École*. <https://www.vie-publique.fr/files/rapport/pdf/114000565.pdf>
- Le Breton, S. et Neirac, L. (2024, juillet). Évaluations point d'étape à mi-CP 2023-2024 : amélioration des résultats par rapport à 2023 et réduction des écarts entre secteurs de scolarisation. *Note d'Information*, (24.24).
- Martinot, P., Deheane, S., Bressoux, P., Huguet, P., Potier-Watkins, C., Sprenger-Charolles, L. et Ziegler, Z. (2021). Qu'apprend-on des évaluations de CP-CE1 ? *Note du CSEN*, (03.24). https://www.reseau-canope.fr/fileadmin/user_upload/Projets/conseil_scientifique_education_nationale/Note_CSEN_2021_03.pdf

- MENESR-DEPP. (2015). *Repères et références statistiques. Enseignements, formation, recherche*.
https://www.enseignementsup-recherche.gouv.fr/sites/default/files/content_migration/document/rers_2015_457907.pdf
- Montolio, D. et Taberner, P. (2021). Gender differences under test pressure and their impact on academic performance: A quasi-experimental design, *Journal of Economic Behavior & Organization*, 191, 1065-1090. <https://doi.org/10.1016/j.jebo.2021.09.021>
- Salines, M. et Vrignaud, P. (2001, juin). *Apprécier et certifier les acquis des élèves en fin de collège : brevet et évaluations-bilans* [Rapport]. Haut Conseil de l'évaluation de l'école. <https://www.vie-publique.fr/files/rapport/pdf/024000205.pdf>
- Sprenger-Charolles, L., Dehaene, S. et Ziegler, J. (dir.). (2019). *Évaluer pour mieux aider. ÉvalAide, un dispositif scientifique de prévention des difficultés en lecture et en mathématiques en CP et en CE1*.
<https://eduscol.education.fr/document/7736/download>
- Stoevenbelt, A. H., Wicherts, J. M., Flore, P. C., Phillips, L. A. T., Pietschnig, J., Verschuere, B., Voracek, M. et Schwabe, I. (2023). Are speeded tests unfair? Modeling the impact of time limits on the gender gap in mathematics. *Educational and Psychological Measurement*, 83(4), 684-709.
<https://doi.org/10.1177/00131644221111076>
- Trosseille, B. et Rocher, T. (2015). Les évaluations standardisées des élèves. *Éducation et formations*, (86-87), 15-35. <https://shs.hal.science/halshs-03572761/file/EF-86-87-2015-article-01-evaluations-standardisees-eleves-perspective-historique-pdf.pdf>