

Pour citer cet article : Grapin, N., & Sayac, N. (2026, prépublication). Éclairage didactique et psychosocial sur les évaluations nationales en France en mathématiques en début d'école élémentaire. *La Revue L'EeE*, 11. <https://revue.leeeonline/index.php/info/article/view/274>

ÉCLAIRAGE DIDACTIQUE ET PSYCHO-SOCIAL SUR LES ÉVALUATIONS NATIONALES EN FRANCE EN MATHÉMATIQUES EN DÉBUT D'ÉCOLE ÉLÉMENTAIRE

Nadine GRAPIN, Nathalie SAYAC

Version de la prépublication : juillet 2026
Évaluation ouverte et collaborative

Résumé

En France, les évaluations menées sur l'ensemble des élèves à l'entrée de chacun des niveaux du primaire révèlent des écarts de performance entre filles et garçons en mathématiques et ce, dès la première année du primaire. Pour éclairer ce résultat, nous avons étudié la validité des inférences produites à partir du modèle statistique utilisé en nous concentrant sur les deux facettes de la validité considérées en didactique des mathématiques, les facettes épistémologique et psychodidactiques. Notre recherche repose sur deux types d'étude. Une analyse a priori des tâches du test, fondée sur une praxéologie de référence, permet d'apporter des preuves de validité du côté du savoir. Elle est complétée, au niveau de l'exercice du test proposé aux élèves de CE1 (grade 2) qui présentait en 2023 l'écart de performance le plus important entre garçons et filles, par une analyse des réponses de 260 élèves et d'entretiens menés avec une trentaine d'élèves.

Nous montrons d'abord que le contenu du test, tel que présenté, ne permet pas de répondre précisément aux objectifs de l'évaluation, puis que le format de l'exercice analysé et les consignes données à l'élève influent fortement sur leurs réponses et conduisent à des interprétations qui peuvent être erronées, en particulier si on se réfère au diagnostic produit pour chacun des élèves à la suite de cette évaluation. Nos analyses apportent également des éléments explicatifs quant aux écarts de score entre garçons et filles et permettent de relativiser les constats récurrents de supériorité des garçons en mathématiques à ce niveau scolaire.

Mots-clés : validité, didactique des mathématiques, genre, évaluation diagnostique

Abstract

In France, assessments carried out on all pupils at the beginning of each year of primary school reveal performance gaps between girls and boys in mathematics, starting as early as the first year of primary school. To better understand this result, we studied the validity of the inferences generated from the statistical model used, focusing on the two aspects of validity considered in mathematics education: epistemological and psycho-didactic validity. Our research is based on two types of study. An a priori analysis of the test tasks, based in a reference praxeology, provides evidence of validity from the knowledge perspective. This is completed, in relation to the test exercise set for Grade 2 pupils (CE1), which in 2023 showed the largest performance gap between boys and girls, by an analysis of the responses of 260 pupils and interviews conducted with around thirty pupils.

We first demonstrate that the content of the test, as presented, does not allow for a precise fulfilment of the assessment objectives; we show that the format of the exercise analysed and the instructions given to pupils strongly influence their responses and lead to interpretations that may be incorrect, particularly when referring to the diagnostic report produced for each pupil after this assessment. Our analyses also provide explanations for the differences in scores between boys and girls and help to put into perspective the recurring findings of boys' superiority in mathematics at this school level.

Keywords: validity, mathematics education, gender, diagnostic assessment

Depuis la rentrée 2024, en France, tous les élèves de l'école élémentaire sont soumis à des évaluations standardisées en mathématiques et en français, au tout début de l'entrée dans chacun des niveaux scolaires, et même également quelques mois plus tard pour la première année d'école élémentaire¹. Ces évaluations appelées « Repères » visent un triple objectif : dresser un diagnostic individuel des connaissances de chaque élève pour informer les parents et outiller l'enseignant en début d'année scolaire, aider au pilotage à un niveau local par un ajustement des plans de formation au niveau académique et disposer d'indicateurs au niveau national pour mesurer les performances des élèves (DEPP, 2023). Au-delà des bilans individuels qui donnent lieu à des rendus aux parents et à une répartition des élèves dans des groupes (fragiles, à besoin, satisfaisant)², une synthèse des résultats à l'échelle nationale est produite par la Direction de l'évaluation, de la prospective et de la performance (DEPP). Un élément revient quasi systématiquement chaque année, il concerne les écarts de résultats entre les filles et les garçons et notamment les meilleures performances des garçons en mathématiques. Celles-ci se manifestent très tôt et conduisent au constat : en deuxième année d'école élémentaire (CE1 – Cours élémentaire 1^{ère} année), contrairement au début de l'année précédente (CP – Cours préparatoire), « les écarts de performances selon le sexe sont toujours à l'avantage des garçons à deux exceptions près : la géométrie [...] et le calcul mental³ » (DEPP, 2023, p. 40). Ce type de résultats récurrents au fil des années, et convergeant avec ceux des évaluations nationales et internationales à grande échelle (PISA, TIMSS, CEDRE) déclenche régulièrement des reprises fatalistes dans les journaux télévisés ou dans les médias, juste après la parution des notes d'informations les révélant. Des études complémentaires sont également produites par le Conseil supérieur de l'éducation nationale (CSEN, 2021) et l'Institut des politiques publiques (Breda et al., 2024 ; Terrier, 2024) pour tenter d'éclairer ce funeste constat.

¹ Ces évaluations existent depuis 2018 pour les deux premiers niveaux de l'école élémentaire.

² Trois groupes d'élèves sont identifiés : « ceux pour lesquels on peut considérer, dès le début de l'année, qu'ils sont en difficulté [...] ; ceux dont les acquis sont fragiles [...] ; ceux pour lesquels il n'y a pas de difficultés identifiées » (DEPP, 2023, p. 24).

³ Il est entendu ici par « calcul mental », la restitution de faits numériques (tables d'addition) et non du calcul mental réfléchi.



En tant que didacticiennes des mathématiques, ce résultat nous a également fortement interpellées, c'est pourquoi nous avons souhaité l'étudier de plus près. Notre intérêt s'est porté sur le test mené en CE1 pour deux raisons. D'une part, il nous semblait intéressant d'analyser la façon dont étaient évaluées les connaissances des élèves dans le domaine numérique, en particulier pour l'écriture et le nom des nombres puisqu'en CP, se construit la compréhension de la numération écrite chiffrée à travers ses deux aspects : positionnel (savoir que la valeur d'un chiffre dépend de sa position) et décimal (la position de chaque chiffre correspond à une unité qui est dix fois plus grande que celle qui est à sa droite). D'autre part, puisque les premiers écarts de performance entre garçons et filles apparaissent dès le milieu du CP, il nous paraissait pertinent d'étudier ce qui se joue au début de l'année suivante, au CE1.

Après avoir présenté les évaluations « Repères », nous rendons compte dans cet article de nos analyses portant sur la validité des inférences produites à partir des modèles utilisés dans ces évaluations à deux niveaux : sur l'entièreté du test⁴ et sur un exercice spécifique pour lequel l'écart de performance est le plus important entre les filles et les garçons (16% au niveau national). Nous présentons d'abord les éléments théoriques sur lesquels nous nous appuyons pour fonder notre problématique, puis nous explicitons notre méthodologie et présentons les résultats obtenus ; ces résultats interrogent le contenu du test et permettent de relativiser les constats récurrents de supériorité des garçons en mathématiques dès le début de l'école primaire.

1. Evaluations « Repères »

Les épreuves des évaluations Repères sont déployées en début d'année scolaire par chaque enseignant·e dans sa classe. En mathématiques, en CE1 (grade 2, élèves âgés de 7-8 ans), elles se déroulent collectivement, en papier-crayon, durant deux séquences d'une quinzaine de minutes chacune. Après la passation, chaque enseignant·e saisit, sur une application ministérielle dédiée, les réponses de ses élèves ; les résultats par élève et pour sa classe lui sont ensuite communiqués via cette même application. En compléments des résultats nationaux, des résultats par école sont également disponibles au niveau de la circonscription⁵ pour accompagner au pilotage local.

Le cadre de l'évaluation s'appuie sur les programmes de l'école primaire en vigueur ; il a été défini principalement par la DEPP et le CSEN (DEPP, 2023). Deux domaines sont évalués en mathématiques : nombres et calculs (7 exercices - 66 items) & espace et géométrie (1 exercice – 8 items) ; sept « compétences » sont identifiées pour le domaine numérique et une pour la géométrie (Tableau 1). Trois niveaux de maîtrise sont déterminés pour chacune de ces compétences (à besoins, fragile, satisfaisant) et les seuils pour définir ces niveaux sont déterminés à partir de la méthode des marque pages, décrite dans DEPP (2023, p. 24-25). Les seuils restent identiques d'une année sur l'autre (à condition qu'il n'y ait pas de modification dans le contenu des exercices) pour permettre des comparaisons annuelles. Par exemple, à partir du tableau 1, pour la compétence « additionner – calculer en ligne », un élève est considéré « à besoins », s'il réussit moins de trois calculs sur les huit proposés ; la maîtrise de cette compétence est considérée comme « fragile » s'il réussit entre 4 et 5 calculs ; au-delà, on considère qu'il a une maîtrise satisfaisante.

Tableau 1

Compétences évaluées et seuils retenus en 2023 – Extrait de DEPP (2023, p. 28)

⁴ Nous appelons test, le support qui sert à l'évaluation, c'est-à-dire l'entièreté des exercices composés d'items proposés aux élèves. Les tests des évaluations Repères sont adaptés aux niveaux scolaires correspondants.

⁵ Regroupements d'écoles à une échelle communale.

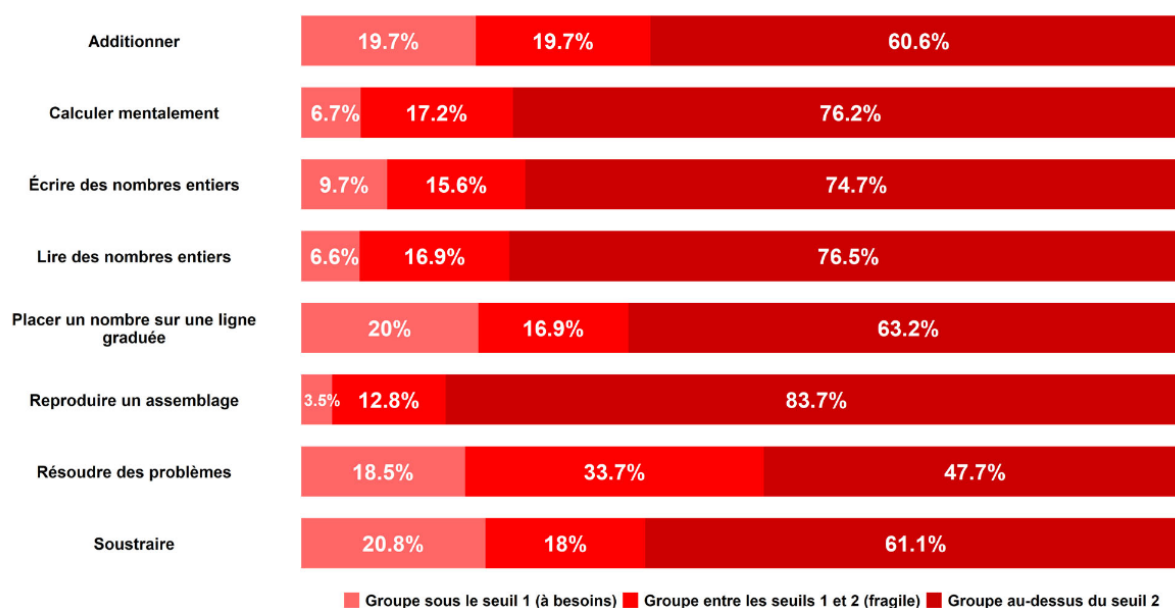


Compétences	Nombre d'items	Seuil 1	Seuil 2
Lire des nombres entiers (Reconnaître des nombres dictés)	10	4	7
Écrire des nombres entiers (sous la dictée)	10	4	7
Placer un nombre sur une ligne graduée	15	4	8
Résoudre des problèmes	6	2	4
Calculer mentalement	10	4	7
Additionner (Calculer en ligne)	8	3	5
Soustraire (Calculer en ligne)	7	1	3
Reproduire un assemblage	8	2	4

Les enseignants et les parents sont informés du groupe auquel appartient chaque élève pour chaque compétence. Au niveau national, les résultats peuvent prendre la forme de répartition par compétence et par groupe (Figure 1).

Figure 1

Répartition des élèves dans les groupes selon les compétences à la rentrée 2023 – Extrait de DEPP (2023, p.32)



2. Éléments théoriques et problématique

Pour les concepteurs d'un test, s'assurer qu'il mesure effectivement ce qu'il prétend mesurer et uniquement cela est une préoccupation essentielle qui relève de sa validité. Recueillir des preuves de validité (Laveault & Grégoire, 2023), des inférences réalisées à partir des scores au test est pour ces auteurs « un processus long et complexe ». Nous ne revenons pas ici sur la façon dont les psychométricien·nes appréhendent le concept de validité, nous expliquons d'abord comment les didacticien·nes des mathématiques s'en sont emparé, puis montrons l'intérêt d'une approche psycho-sociale prenant en compte les stéréotypes de genre pour



reconsidérer les écarts de performance entre garçons et filles et concluons par la problématique soulevée dans cet article.

2.1 Validité et didactique des mathématiques

En accordant une place importante au savoir évalué, les recherches menées en didactique vont s'intéresser à des preuves de validité basées « sur le contenu du test » et sur « les processus de réponse » (Laveault & Grégoire, 2023, p. 200) ; en didactique des mathématiques, Grapin & Grugeon-Allys (2018) ont ainsi identifié des preuves didactiques, du côté du savoir et du côté de l'élève, correspondant aux facettes épistémologique et psychodidactique de la validité (Goasdoué & Vantourout, 2014 ; Grapin et al., 2022a).

2.1.1 Preuves de validité du côté du savoir

Pour rechercher des preuves du côté du savoir, Grapin & Grugeon-Allys (2018) se placent dans le cadre d'une approche anthropologique (Chevallard, 1999) ; « les tâches de l'évaluation sont considérées comme relevant d'un échantillon de l'ensemble des types de tâches constitutives des praxéologies mathématiques évaluées » (Grugeon-Allys & Grapin, 2018). Dans ce cadre, une étude épistémologique d'un domaine mathématique fixé conduit à la définition d'une praxéologie de référence qui permet de décrire les praxéologies à enseigner, enseignées et apprises. Ainsi, chaque tâche proposée dans un test peut être décrite à partir du type de tâche duquel elle relève, de la ou des technique(s) qui permettent de la résoudre et les éléments qui justifient mathématiquement ces techniques. Par exemple, en fin de première année d'école élémentaire, la tâche « calculer mentalement $18+4$ » relève d'un type de tâche de calcul mental réfléchi d'une addition et peut être résolu par différentes techniques, comme :

- ajouter les unités entre elles ($8+4 = 12$), convertir le nombre d'unités en dizaines et unités (12 unités = 1 dizaine 2 unités), puis ajouter ensemble les dizaines pour obtenir 22. Au-delà des faits numériques relatifs à l'addition ($8+4 = 12$), cette technique s'appuie sur des savoirs relatifs aux aspects décimal et positionnel des écritures chiffrées, permettant notamment la conversion de 12 unités en 1 dizaine et 2 unités ;
- sur-compter de 4 à partir de dix-huit (dix-neuf ; vingt ; vingt-et-un ; vingt-deux) ; les savoirs en jeu ne relèvent dans ce cas que de la numération orale ;
- décomposer additivement 4 en $2 + 2$, puis calculer $18+2 = 20$ puis $20 + 2 = 22$; dans ce cas, des savoirs relatifs à la décomposition additive des nombres et à l'associativité de l'addition sont en jeu.

La définition d'une praxéologie de référence permet ainsi, au niveau de chaque tâche, d'identifier les savoirs qui peuvent être évalués et de s'assurer qu'il y a adéquation entre l'objectif d'évaluation assigné à la tâche par les concepteur·rices du test et ce qu'elle permet *a priori* d'évaluer. La praxéologie de référence liste aussi l'ensemble des praxéologies relevant d'un domaine mathématique donné. Au regard de cette référence, il sera ainsi possible d'étudier le recouvrement des types de tâches présents dans le test, au regard de ceux de la praxéologie de référence (couverture du domaine), mais aussi de s'assurer d'une certaine variété des tâches proposées en termes de complexité par exemple. L'analyse menée au niveau de l'ensemble des tâches du test peut aussi permettre une mise en perspective des réponses des élèves à différentes tâches au niveau de la justification des techniques.



2.1.2 Preuves de validité du côté de l'élève

Rechercher des preuves de validité du côté de l'élève demande de s'intéresser à l'activité de l'élève lorsqu'il résout la tâche ; nous entendons ici la tâche comme « ce qui est à faire » et l'activité comme « ce que développe un sujet lors de la réalisation de la tâche » (Rogalski, 2003). Il s'agit alors d'étudier l'écart entre la tâche prescrite (id.) par le ou la conceptrice, celle que l'élève s'est redéfinie (id.) et son activité effective. Pour s'assurer alors de la qualité des inférences qui peuvent être faites à partir des résultats, et en particulier lorsque les évaluations ont une visée de diagnostic individuel, il est nécessaire de mener des observations cliniques permettant de déterminer les processus cognitifs qui conduisent à la réponse (Goasdoué & Vantourout, 2014). Par exemple comme nous l'avons montré dans une recherche antérieure sur l'emploi des QCM dans les évaluations en fin d'école élémentaire (Sayac & Grapin, 2014), il est difficile de déterminer ce qui conduit un élève à choisir sa réponse, même lorsque les distracteurs sont conçus à partir d'erreurs classiques, liées à des conceptions erronées identifiées et/ou à des procédures utilisées de façon incorrecte. Si l'analyse *a priori* de la tâche peut renseigner sur les techniques erronées conduisant à certains distracteurs, notre étude montre que cette analyse n'est pas suffisante pour évaluer les connaissances des élèves, puisqu'ils ou elles peuvent choisir leur réponse en utilisant des stratégies qui ne sont pas des stratégies de savoir (Sayac & Grapin, 2014).

Au-delà des variables didactiques considérées habituellement dans l'analyse *a priori*, d'autres éléments peuvent alors être pris en compte pour rechercher des types de preuve du côté de l'élève, comme l'impact du support (papier/crayon par exemple), le niveau de langue des énoncés ou encore le contexte de passation.

2.2 Stéréotypes de genre

Il est aujourd'hui reconnu que les filles sont plus sensibles à la pression évaluative (Marsh et al., 2008 ; Genoud et al., 2009) et plus anxieuses que les garçons vis-à-vis de leur travail scolaire (Devine et al., 2012 ; Chen et al., 2013). Même si les travaux ayant attesté de ces différences ont concerné des élèves plus âgés que les élèves de CP et CE1, on peut tout à fait concevoir que passer des évaluations en mathématiques et en français, la semaine qui suit la rentrée scolaire engendre une certaine pression et de l'anxiété chez des filles de 6-7 ans (Sayac, 2023). L'observation d'une quarantaine de passations des évaluations « Repères » en mathématiques dans des classes de CP en janvier 2023, nous a permis de relever des éléments susceptibles de provoquer cette pression évaluative, comme l'usage ostensible du chronomètre pour certains exercices ou le placement d'un grand livre pour séparer l'espace de deux élèves partageant la même table. Par ailleurs, les évaluations « Repères » ont un caractère très institutionnel, avec un protocole et des consignes très strictement précisés dans les livrets d'accompagnement. Les professeur·es que nous avons observé·es les respectent très majoritairement, mais des différences de ton, de discours et d'étayage dues à la façon dont ils/elles s'approprient ces consignes et à leur personnalité peuvent potentiellement générer de l'anxiété auprès de certains élèves et plus particulièrement des filles.

Marie-Christine Tocsek (2005) a montré combien les stéréotypes de sexe pouvaient influencer les performances scolaires des élèves, notamment à travers la nature des consignes données avant la réalisation d'exercices, lors des évaluations nationales, et qu'ils pouvaient avoir une incidence sur leurs performances en fonction de leur genre en mobilisant le phénomène de menace du stéréotype (Steele et Aronson, 1995, ; Régner et al., 2014). En croisant les données issues des évaluations « Repères » administrées à plus de 2,5 millions d'élèves scolarisés en France entre 2018 et 2022 avec des informations sur le contexte familial, les enseignant·es et les caractéristiques générales des écoles (publique, privée, en réseau



d'éducation prioritaire ou à pédagogie alternative), les chercheur·es de l'IPP qui souhaitent élucider la raison de l'émergence des écarts de performance en mathématiques entre les filles et les garçons observés en CP et en CE1 ont conclu qu'aucun de ces facteurs ne pouvait, à lui seul, expliquer ces écarts (Breda et al., 2024). Ils et elles ont donc conclu que ces écarts devaient être dus à « une dynamique diffuse dans la société », sous-entendant que les stéréotypes de sexe étaient le facteur explicatif le plus plausible.

2.3 Problématique

Comme le soulignaient Troseille & Rocher (2015, p.15), assigner à une même évaluation une fonction de diagnostic à destination de l'enseignant et de bilan pour le pilotage « est potentiellement source d'erreurs et de troubles, tant sur le plan scientifique que sociétal ». Or, les évaluations « Repères » visent à répondre à ce double objectif.

Afin de remplir la fonction revendiquée de diagnostic, ces évaluations doivent donc renseigner l'enseignant·e aussi précisément que possible sur l'état des connaissances de ses élèves afin qu'il ou elle puisse mettre en œuvre les régulations adaptées aux besoins de chacun. Or, les résultats produits (que ce soit au niveau individuel de l'élève ou de la classe, de l'école) renvoient à des performances relatives à des seuils de maîtrise de compétences et ne rendent pas compte du type d'erreur commise ; il semble alors difficile pour l'enseignant·e, à partir des résultats qu'il ou elle reçoit, de mettre en place une régulation adaptée sans savoir ce qui a conduit l'élève à produire ses réponses. Quelles inférences sur les connaissances des élèves le contenu du test permet-il de faire ? quelles preuves de validité pouvons-nous apporter après une étude des facettes épistémologique et psycho-didactique ?

Pour éclairer d'un point de vue complémentaire et didactique les écarts de performance apparaissant entre garçons et filles, des recherches montrent l'intérêt d'une analyse des procédures des élèves, en lien avec une analyse *a priori* des tâches. Ainsi, Che et al. (2012) ont pointé des différences dans les types de procédures utilisées par les garçons et les filles lors de la résolution de problème de proportionnalité au grade 6. En France, Roditi et Salles (2015), après une analyse des niveaux de mise en fonctionnement des connaissances mathématiques (Robert, 1998) requis dans les items du PISA 2012, ont montré que « les filles sont d'autant plus en difficulté par rapport aux garçons que le niveau requis de mise en fonctionnement des connaissances est un niveau exigeant » (Roditi & Salles, 2015, p. 252). Enfin, lors de nos recherches antérieures menées sur des tâches extraites d'évaluation mathématiques en fin d'école, nous avons observé que les garçons attribuaient, en moyenne, à leurs réponses un degré de certitude (Leclercq, 1982 ; Sayac & Grapin, 2014, 2016) plus élevé que les filles, quelle que soit la justesse de leurs réponses. Les garçons auraient donc des connaissances plus assurées que les filles, mais aussi, une ignorance ignorée (Gilles, 1996) plus fréquente, selon ces indicateurs. Même si les résultats que nous avons obtenus sont à relativiser (nombre réduit d'élèves et d'exercices, modalités de passation, etc.), la menace du stéréotype semble, dès ce jeune âge, impacter plus particulièrement les filles qui se confortent dans l'idée que « elles ne sont pas bonnes en mathématiques », mais aussi les garçons pour qui « l'assurance d'être bons en mathématiques » peut les leurrer sur leur déficit de compétences.

Si les études que nous citons ici montrent l'intérêt d'une approche didactique pour mieux comprendre les écarts de performance entre garçons et filles, nous faisons l'hypothèse que croiser les résultats obtenus à partir d'une analyse didactique avec ceux issus des recherches sur les stéréotypes de genre permettra de mieux comprendre et interpréter ces écarts. Même si les recherches en didactique des mathématiques auxquelles nous faisons ici référence ont été menées avec des élèves plus âgés, nous considérons que leurs résultats sont transposables aux élèves de notre étude. Par ailleurs, aucune d'entre elles ne prend en



compte le contexte dans lequel sont passées ces évaluations, ni les effets qu'elles peuvent produire sur les élèves. Or, ces deux éléments (contexte et effets) ne sont pas sans incidence sur les performances des élèves, comme de nombreux travaux le présument (Tocsek, 2005 ; Goudeau et al., 2017). Au-delà des questions de validité en lien avec la visée diagnostic du test, notre recherche vise donc également à répondre aux questions suivantes : en quoi une analyse des réponses et des procédures des élèves permet-elle de mieux comprendre les écarts de score observés lors des évaluations nationales ? peut-on éclairer ces résultats par des éléments liés à des considérations psycho-sociales (anxiété générée par les modalités de passation, auto-évaluation de la qualité de la réponse fournie) ?

3. Méthodologie

Notre recherche combine deux études complémentaires :

- une première menée sur l'ensemble des exercices du domaine « nombres et calculs » du test pour le niveau CE1 (grade 2) ; une analyse *a priori* de chaque tâche du test est menée au regard d'une organisation mathématique de référence, comme nous l'avons mentionné dans la partie 1, afin d'apporter des preuves de validité du côté du savoir ;
- une seconde centrée sur l'exercice qui présente l'écart de performance le plus important entre les garçons et les filles⁶. Pour cet exercice de calcul mental, afin de mieux comprendre les écarts de score qu'il engendre, mais aussi de prendre en compte d'autres éléments que ceux présents dans l'analyse *a priori*, en particulier les conditions de passation, nous avons étudié des réponses d'élèves au test et mener des observations cliniques ; il s'agit ici d'apporter des preuves de validité du côté de l'élève.

3.2 Présentation de l'exercice de calcul mental

Pour les concepteurs, cette tâche vise à évaluer les connaissances des élèves dans le calcul en ligne d'additions et de soustractions. Quinze calculs (additions et soustractions mélangées avec une alternance irrégulière) sont proposés les uns à la suite des autres sur une double page (cf. Figure 2). Le premier calcul $5 + 3$, donné à titre d'exemple, est destiné à être fait collectivement pour que l'élève comprenne la tâche qui lui est prescrite, la consigne étant : « il faut trouver le nombre qui est la bonne réponse. Pour répondre, entourez ce nombre » (MENJS, 2023a, p. 20). Les élèves disposent de 7 minutes pour traiter individuellement et à leur propre rythme l'ensemble des calculs. Il est explicitement indiqué qu'aucune aide, ni intervention de l'enseignant·e ne doivent être apportées par l'enseignant·e durant ce temps.

Figure 2

Reproduction des pages 21 et 22 du cahier d'évaluation de l'élève à l'entrée au CE1 (MENJS, 2023b)

⁶ Les résultats à l'échelle nationale pour cet exercice (DEPP, 2023, p.40) montrent qu'il y aurait 15% de garçons de plus de filles qui aurait une maîtrise satisfaisante de la compétence « additionner » (53% de filles au-dessus du seuil 2 vs 68% de garçons) et 8% de plus sur la compétence « soustraire » (57% de filles au-dessus du seuil 2 vs 65 % de garçons).



Exercice 7

$5 + 3 =$ 53 9 8 2 5 3

$10 + 8 =$ 19 108 10 8 18 2

$9 - 4 =$ 94 9 13 5 6 4

$9 + 9 =$ 0 99 17 18 90 9

$50 + 50 =$ 0 50 79 99 100 5050

$10 - 2 =$ 10 8 12 102 9 2

$15 - 5 =$ 11 155 10 15 20 5

Exercice 7 (suite)

$20 + 30 =$ 20 51 2030 10 50 30

$18 + 4 =$ 4 184 18 14 22 21

$19 - 2 =$ 19 21 192 18 17 2

$56 - 10 =$ 66 56 45 5610 46 10

$21 + 53 =$ 2153 32 74 21 53 75

$8 + 46 =$ 54 38 8 846 46 55

$50 + 10 + 8 =$ 50108 60 18 68 67 58

$35 - 15 =$ 15 3515 21 50 20 35

$33 - 5 =$ 28 33 38 5 335 29

23CE1E
Page 21

23CE1E
Page 22

3.3 Analyse de réponses d'élèves et observations cliniques

Un premier travail a été réalisé à partir des réponses données par 265 élèves de CE1 (138 filles et 127 garçons) d'une même circonscription parisienne lors de la passation des évaluations nationales en septembre 2023. Les réponses ont été traitées, élève par élève, en prenant en compte les distracteurs repérés par l'analyse *a priori*.

Comme pour tous les exercices de calcul mental, il est difficile d'accéder directement à l'activité de l'élève qui réalise ses calculs (Butlen, 2005, p.39), mais nous pouvons inférer des éléments de cette activité à partir de certaines traces (choix de réponse, traces écrites, explications orales de procédures, etc.). Or, peu de traces figurent habituellement sur les cahiers de passation des élèves. Il nous semblait par conséquent important de mener une expérimentation complémentaire permettant de déterminer les stratégies, notamment celles de savoir, qui ont conduit l'élève choisir une réponse plutôt qu'une autre tout en prenant en compte son ressenti sur ce type d'exercice.

Pour ce faire, des passations individuelles ont été menées en juin 2023 auprès de 37 élèves (20 filles / 17 garçons) de fin de CP⁷ résolvant cette tâche dans des conditions similaires à celles des évaluations nationales : l'exercice a été présenté avec le même format que celui du cahier de CE1 (cf. Figure 2 : mêmes calculs, même organisation sur une double page, etc.), les consignes ont été les mêmes que celles prescrites pour la passation en CE1. A la suite de la réalisation de cet exercice, nous avons demandé à chaque élève ce qu'il ou elle avait pensé

⁷ Ce choix correspondait aux élèves allant passer les épreuves de CE1, juste après les congés scolaires d'été.



de cet exercice ; il s'agissait pour nous, avec une question ouverte, de savoir quel était le ressenti de l'élève, s'il ou elle évoquait le caractère complexe ou non de l'exercice, certaines des difficultés rencontrées, etc. Nous lui avons également demandé la certitude qu'il ou elle accordait à l'ensemble de ses réponses au test. Pour ce faire, nous avons repris de nos travaux précédents (Sayac & Grapin, 2014) une échelle de Likert avec 4 degrés de certitude (pas sûr du tout, un peu sûr, sûr, sûr et certain), mais nous l'avons adaptée puisque d'après les recommandations de Houssa & Nader-Grosbois (2018, p. 16) : « lorsqu'on utilise une échelle de Likert avec des jeunes enfants, les options de réponse doivent être concrètes et créatives et doivent les aider à comprendre la tâche ». Les options de réponse ont été ainsi proposées avec des smileys et chacune d'elle a été formulée oralement, en montrant chacun des smileys.

Chaque élève a été ensuite interrogé-e sur la façon dont il ou elle avait procédé pour fournir sa réponse à quatre calculs spécifiquement retenus pour cette étude : $9 - 4$; $20 + 30$; $18 + 4$; $19 - 2$. Cette étude ayant été réalisée en amont de la première, le choix des calculs ne repose pas sur les observations faites à la suite de l'étude quantitative. Nous les avons choisis *a priori*, parce qu'ils combinent additions et soustractions, mais aussi en fonction de leur place dans la série de calculs proposés et des procédures peuvent être mobilisées pour donner la bonne réponse ; cela permettra ainsi d'étudier si des différences de procédures et-ou d'explications des procédures apparaissent entre garçons et filles.

Plus précisément :

- $9 - 4$ est la première soustraction qui apparaît dans le test ; nous supposons que le risque de confusion avec une addition est plus important sur ce calcul et souhaitons étudier si ce type de confusion est plus fréquent chez les garçons ou les filles. Par ailleurs, nous souhaitons également déterminer si les élèves qui donnent pour réponse la somme des 2 nombres sont capables de modifier leur réponse lorsqu'ils constataient que l'opération à réaliser était une différence. Le résultat peut être connu (fait numérique) ou trouvé en passant par un décomptage à partir de 9 ou encore à partir de la somme $4 + 5 = 9$;
- $20 + 30$: le résultat peut être donné en appui sur le nombre de dizaines ;
- $18 + 4$: le résultat peut être issu d'un sur-comptage ou d'une décomposition additive de 4 en $2 + 2$ ($18 + 4 = 18 + 2 + 2 = 20 + 2 = 22$) ou encore d'une technique en appui sur les dizaines et unités (similaire à celle de l'algorithme de l'addition posée) ;
- $19 - 2$: deux techniques sont possibles, décompter ou utiliser le fait que $9 - 2 = 7$.

Durant la passation, nous avons particulièrement observé si l'élève utilisait ou non ses doigts de façon visible pour répondre et nous avons pris en note les explications qu'il ou elle nous donnait quant aux techniques utilisées pour donner sa réponse ; des enregistrements audio et vidéo de certaines passations ont également été réalisés. Ces observations cliniques remplissent pour nous un double objectif : apporter des preuves de validité du côté de l'élève (indépendamment de la différence de performance garçons- filles), mais aussi des éléments permettant d'éclairer didactiquement la différence de performance entre garçons-filles.

4. Résultats

Nous organisons cette partie de la même façon que celle dédiée à la méthodologie. Nous revenons d'abord sur les preuves de validité du côté du savoir (en lien avec les tâches du test), puis sur celles du côté de l'élève.



4.1 Analyse *a priori* des tâches du test : preuve de validité du côté du savoir

Nous présentons dans le tableau 2 la liste des types de tâches présents dans le test, accompagnés d'exemples de tâches ; l'exercice de géométrie n'a pas été conservé pour notre analyse.

Tableau 2

Répartition des tâches du test par sous domaines et types de tâches avec exemples

	Type de tâches et nombre de tâches	Exemples de tâches avec précisions sur le temps de passation
Résolution de problèmes arithmétiques	Résoudre des problèmes additifs (5)	<i>Consigne orale : pour répondre, entourez votre réponse en dessous du cadre.</i> <i>L'énoncé est lu deux fois. Les élèves disposent de 1minute 30 secondes pour chaque problème, sauf pour un problème à deux étapes (2 minutes).</i> Pierre avait 10 billes. Il en gagne 4 à la récréation. Combien de billes a-t-il maintenant ? 14 10 13 6 15 4
	Résoudre des problèmes multiplicatifs (1)	
Numération	Reconnaitre l'écriture chiffrée d'un nombre dont le nom est donné oralement (10)	<i>Consigne orale : entourez le nombre vingt-neuf. Les élèves disposent de 5 secondes après une première lecture et de 5 secondes supplémentaires après une seconde lecture.</i> 92 89 29 30 20 9
	Ecrire un nombre en chiffres à partir de son nom donné oralement (8)	<i>Consigne : Ecrivez soixante-dix-neuf (pas de proposition de réponse).</i>
	Placer des nombres sur une droite graduée (15)	<i>Consigne : Pour répondre, entourez le bon nombre sur la ligne en dessous. Les élèves disposent de 5 minutes pour traiter l'ensemble des questions.</i> 35 45 34 38 39 46 75 48



Calcul	Restituer des faits numériques additifs (10 résultats de tables d'addition)	<i>Consigne : pour répondre, entourez la bonne réponse parmi les 6 nombres proposés sur la ligne. Mettez votre doigt sur le rond blanc, le calcul est $3 + 2$, faites $3 + 2$ dans votre tête mais ne dites pas le résultat. Entourez le résultat sur la ligne du rond blanc. Chaque calcul est dit deux fois ; les élèves ont 10 secondes pour répondre par calcul.</i> 
	Calculer mentalement (calcul réfléchi) des additions et soustractions (15)	<i>Consigne orale : il faut trouver le nombre qui est la bonne réponse. Pour répondre, entourez ce nombre. Les élèves disposent de 7 minutes pour traiter l'ensemble des calculs.</i> Cf. Figure 2

La plupart des tâches proposées sont sous format QCM, ce qui n'est pas sans conséquence sur l'activité mathématique de l'élève, peu habitué·e à ce format de question.

4.1.1 Exemple d'analyse a priori

Nous présentons ici, à titre d'exemple, l'analyse *a priori* faite pour la tâche de calcul mental réfléchi ; elle s'appuie notamment sur les travaux de Butlen (2015) et de Grapin & al. (2022b). Nous observons que les résultats de 10 calculs sur 15 peuvent être donnés par comptage (décomptage / surcomptage) puisqu'un des deux nombres en jeu est inférieur ou égal à 10 ; les élèves peuvent évidemment utiliser de telles techniques pour des nombres plus grands, le temps par calcul n'étant pas limité, mais avec un risque d'erreur plus important. Cinq calculs font intervenir un nombre entier de dizaines, leurs résultats peuvent donc être déterminés à partir des propriétés de la numération écrite chiffrée (par exemple $20 + 30 = 2$ dizaines + 3 dizaines = 5 dizaines = 50). Enfin les cinq premiers calculs peuvent être associés à des faits numériques que les élèves doivent connaître à ce stade de la scolarité selon les attendus de fin de CP (MENJS, 2020).

Pour chaque calcul, nous observons que les mêmes types de distracteurs⁸ sont proposés, à savoir :

- la bonne réponse à plus ou moins 1 près selon les exemples ;
- le nombre formé par l'écriture accolée des termes du calcul (108 pour $10+8$ ou 5610 pour $56-10$) ;
- le résultat de l'opération inverse (la différence lorsque le calcul est une addition et réciproquement, la somme lorsque le calcul est une soustraction) ;
- chacun des deux termes de la somme ou de la différence (ou à un des deux dans le cas où les termes sont identiques, comme pour $50+50$).

La position de la bonne réponse, comme celle des distracteurs, est variable d'un calcul à l'autre. Il est aisé de décrire la façon dont les distracteurs ont été conçus ; en revanche, il est plus difficile d'inférer la procédure qui a pu conduire l'élève à choisir sa réponse à partir de cette dernière uniquement, sauf pour le choix de la somme lorsque c'est une soustraction qui est proposée (ou l'inverse).

⁸ Dans un QCM, un distracteur est une possibilité de réponse autre que celle qui est correcte.



4.1.2 Validité du côté du savoir

Nous revenons ici sur la façon dont sont évaluées les connaissances des élèves selon les trois sous-domaines identifiés : la résolution de problèmes, la numération et le calcul.

Les problèmes proposés sont basiques (Houdement, 2017) et seul un demande deux étapes de résolution ; ils appartiennent à des classes de problèmes différentes (Vergnaud, 1991) et mettent en jeu des nombres inférieurs à 30. Chaque énoncé est lu par l'enseignant·e deux fois, ce qui permet de ne pas pénaliser des élèves qui présenteraient des difficultés en lecture. Le cadre prévu pour la recherche laisse la possibilité à l'élève de procéder par calcul ou comptage en dessinant des collections. Cependant, comme il est demandé à l'élève de répondre à chaque problème en moins d'1 minute 30 secondes (2 minutes pour le problème avec deux étapes), recourir aux dessins de collections peut prendre trop de temps et ne pas lui permettre de répondre au problème.

Les distracteurs sont construits systématiquement de la même façon : chacun des nombres présents dans le problème, le résultat de l'opération inverse, la bonne réponse à + ou - 1 près. Ils permettraient ainsi dans une analyse plus approfondie des réponses et de déterminer où se situe l'erreur de l'élève, en particulier de repérer si le modèle mathématique choisi n'est pas adapté (résultat de l'opération inverse) ou si l'élève n'a pas compris le problème (choix d'un seul des nombres présents dans l'énoncé). Le choix du format de réponse interroge également. Il est demandé à l'élève d'entourer un nombre, la qualification (Houdement, 2011) de ce nombre n'étant pas demandé. Une telle pratique peut laisser penser à l'élève que résoudre à un problème, c'est trouver un nombre, résultat d'une opération ou d'un comptage et ainsi favoriser les démarches superficielles au détriment des démarches expertes, qui le conduiraient à revenir à la situation évoquée dans le problème (Mounier & al., 2024).

Seulement trois types de tâches évaluant les numérations orales et écrites chiffrées sont proposées alors qu'il en existe de nombreux autres. Il n'y a pas de tâche de comparaison ni de dénombrement de collections par exemple, mais surtout aucune tâche évaluant la connaissance des aspects décimal et positionnel des écritures chiffrées n'est proposée (par exemple, écrire en chiffres 5 unités et 7 dizaines), alors que ces connaissances sont particulièrement travaillées au CP et sont mises en jeu à la fois dans les techniques opératoires de l'addition et de la soustraction, mais aussi dans l'écriture des nombres entiers plus grands ; il est donc indispensable que l'enseignant, au début du CE1, puisse connaître le niveau de maîtrise de ses élèves pour ces connaissances. Les nombres en jeu dans les exercices sont tous inférieurs à 100, ce qui est conforme aux attendus des programmes de CP. Cinq mêmes nombres (67-76-83-98-90) sont proposés dans les deux exercices de reconnaissance et d'écriture en chiffres des nombres ; il pourrait être intéressant de mettre en perspective les réponses d'un·e même élève pour s'assurer de leur cohérence, mais aucun résultat n'est produit en ce sens. Enfin, le nombre de tâches correspondant à chacun des types de tâches nous interroge : pourquoi demander 15 fois à un·e élève de reconnaître la position d'un nombre sur une droite graduée alors que des types de tâches tout aussi importants ne sont pas proposés ?

Concernant le calcul, seules des tâches de calcul mental sont présentes dans le test ; aucun calcul posé d'addition n'est demandé aux élèves alors que la technique est enseignée au CP (grade 1). Au-delà du nombre de calculs proposés (10 calculs pour évaluer la connaissance des faits numériques et 15 pour le calcul mental réfléchi), nous nous interrogeons également sur le format de passation de l'exercice de calcul mental réfléchi. En effet, l'élève dispose de 7 minutes pour traiter l'ensemble des calculs ; il peut donc procéder par comptage, poser certaines additions et ainsi donner la bonne réponse sans mobiliser de stratégie de calcul mental, ce qui interroge l'interprétation qui pourra être faite des résultats obtenus.



Pour conclure, même si nous sommes conscientes qu'il n'est pas possible en un temps limité et dans une telle évaluation standardisée de proposer à l'élève l'ensemble des types de tâches de chaque sous-domaine, le choix d'écarter toute évaluation de savoirs liés la numération décimale et de proposer un nombre important de tâches correspondant au même type de compétences (cf. exercice sur la droite graduée) nous semble devoir être repensé afin que le test réponde aux fonctions qui lui sont assignées, notamment celle de diagnostic. En effet, alors qu'au CE1 l'élève devra être capable de mobiliser ses connaissances sur les unités de numération (dizaines / unités) pour écrire des nombre plus grands (à trois chiffres) et pour comprendre les algorithmes de calcul posé, aucune information n'est fournie à l'enseignant-e en début d'année sur ces connaissances.

4.2 Preuves de validité du côté de l'élève

Nous avons choisi de présenter conjointement les résultats de l'ensemble des échantillons (260 cahiers d'élèves, 37 entretiens) sur la tâche de calcul mental réfléchi tout en faisant apparaître les résultats concernant les différences de performance entre garçons et filles.

4.2.1 Analyse des réponses sur les cahiers d'élèves

Les résultats que nous obtenons à l'échelle de la circonscription (tableau 3) rejoignent ceux obtenus à l'échelle nationale ; la performance des garçons sur l'ensemble des calculs est meilleure que celle des filles puisque le score moyen des garçons est de 10,7 réponses justes (sur 15) alors qu'il est de 9,3 réponses justes (sur 15) pour les filles ; l'écart est encore plus marqué en faveur des garçons concernant les soustractions (Tableau 3).

Tableau 3

Nombre moyen de bonnes réponses par type de calcul (addition – soustraction) et selon le sexe

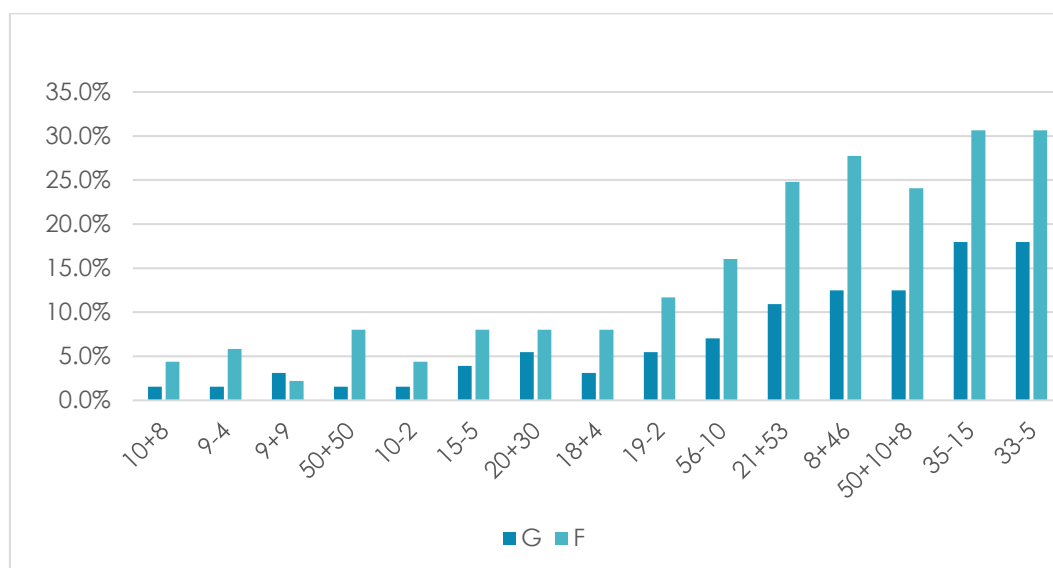
	Additions et soustractions /15	Additions / 8	Soustractions / 7
Totalité des élèves de l'échantillon	10,7	6,2	4,5
Garçons	11,3	6,3	4,8
Filles	9,3	5,6	3,7
Ecart (garçons – filles)	2	0,7	1,1

Si nous nous intéressons à présent aux réponses produites pour chaque calcul par les élèves, nous constatons d'abord une augmentation du taux de non-réponse au fur et à mesure de l'avancée dans les calculs (Figure 3) avec systématiquement un nombre plus important de non-réponse chez les filles (1,1 non-réponse en moyenne chez les garçons vs 2,1 chez les filles) et un écart entre le pourcentage de non-réponse par calcul entre garçons et filles qui se stabilise autour de 13 % à partir du onzième calcul (21 + 53). Différentes hypothèses peuvent être formulées pour expliquer l'augmentation de la non-réponse : un effet de lassitude qui conduirait certains élèves à se déconcentrer et ne plus répondre, un manque de temps dû à une utilisation trop fréquente du comptage sur les doigts, un manque de connaissances qui ne permet pas de répondre aux derniers calculs, plus complexes que les premiers.

Figure 3

Pourcentage de non-réponse selon le calcul en fonction du sexe





Intéressons-nous désormais aux choix de réponses et aux erreurs commises. Le nombre composé des deux termes accolés est une erreur peu fréquente (moins de 5% des réponses) sauf pour trois calculs : $21 + 53$ (8,3% des réponses) ; $50 + 10 + 8$ (13,6%) ; $35-15$ (5,7%) ; l'écart entre garçons et filles étant particulièrement marqué pour $50 + 10 + 8$ (10,2% des garçons et 17,5% des filles commettent cette erreur).

Les bonnes réponses à plus ou moins 1 près représentent plus de 10% des réponses sur trois calculs : $56-10$ (10,2% des réponses) ; $8+46$ (11,3%) ; $33-5$ (10,6%) ; l'écart entre garçons et filles étant particulièrement marqué pour $8+46$ (14,8% des garçons et 8% des filles commettent cette erreur).

Enfin, la confusion addition / soustraction représente environ 10 % des réponses pour chacune des soustractions (elle est inférieure à 4% pour les additions). La différence entre garçons et filles est particulièrement marquée pour ce type d'erreur sur les trois premières soustractions où la proportion des filles commettant cette erreur est supérieure de plus de 7% à celle des garçons (cf. Tableau 4). L'écart semble se réduire sur les dernières soustractions, voire s'inverser, mais si on ne prend en compte que les calculs renseignés, c'est-à-dire en écartant les non-réponses, il perdure.

Tableau 4

Pourcentage de réponses correspondant à la confusion addition / soustraction selon le calcul et selon le sexe

	9-4	10-2	15-5	19-2	56-10	35-15	33-5
Totalité des élèves							
Garçons	8,6%	10,2%	6,3%	7,8%	15,6%	10,9%	8,6%
Filles	16,1%	18,2%	16,8%	10,9%	9,5%	9,5%	14,6%
Ecart (F-G)	7,5%	8%	10,5 %	3,1%	-6,1%	-1,5%	6%

Ces résultats interrogent les inférences produites à partir des scores. En effet, nous considérons, en tant que didacticiennes des mathématiques, que le choix d'une réponse correspondant à l'opération inverse et l'absence de réponse ne signifient pas pour autant que



l'élève ne maîtrise pas le savoir attendu ; différentes raisons peuvent être apportées pour expliquer ce constat (manque de temps, déconcentration, attractivité d'un distracteur, etc.). Pour autant, dans le calcul des seuils de maîtrise et donc la caractérisation du niveau global de compétences des élèves, seule la réponse correcte est prise en compte et il est considéré qu'un élève qui ne répond pas ou se trompe d'opération ne maîtrise pas le savoir. Par ailleurs, il est difficile d'interpréter les mauvais choix des élèves par la mobilisation de techniques erronées. Par exemple, il est difficile de savoir si les réponses à plus ou moins 1 près proviennent d'un surcomptage / décomptage inadapté ou d'une connaissance approximative d'un fait numérique. Par conséquent, l'évaluation répond difficilement à sa fonction de diagnostic puisqu'il n'est pas possible pour l'enseignant·e, à partir des seules réponses de l'élève, de savoir si ses erreurs proviennent de difficultés émanant du comptage ou d'un manque de connaissance des faits numériques.

Notre analyse didactique permet partiellement de comprendre en partie l'origine des écarts de score entre les filles et les garçons en fonction du type de réponse qu'ils ou elles font, car d'autres interrogations surviennent : pourquoi les filles sont-elles plus nombreuses que les garçons à produire des non-réponses ? pourquoi choisissent-elles de manière plus marquée l'opération inverse de celle attendue ? Ces différences constatées entre les réponses des filles et celles des garçons ne peuvent s'expliquer du seul point de vue cognitif, mais la piste d'une pression évaluative plus grandement ressentie par les filles ou celle de l'effet de la menace du stéréotype sont des pistes à explorer. L'étude qualitative devrait nous permettre de mieux comprendre la façon dont un·e élève choisit une réponse et ainsi de caractériser plus justement ses performances.

4.2.2 Analyse des entretiens

Précisons d'abord que l'échantillon de 37 élèves que nous avons interrogé·es n'est pas représentatif de la population de la circonscription et que les résultats que nous produisons sont à relativiser puisque l'échantillon est de petite taille.

Le score moyen de cet échantillon est de 7,9 sur 15 ; non seulement l'écart de score entre garçons et filles est plus réduit que celui à l'échelle de la circonscription, mais il est à la faveur des filles, leur score moyen étant de 8,3 bonnes réponses et celui des garçons de 7,5. Comme à l'échelle de la circonscription, nous retrouvons un nombre important de non-réponses. La non-réponse apparaît de façon importante (plus de 10% des réponses) dès le calcul $20+30$ et par la suite, pour tous les calculs qui mobilisent des nombres à deux chiffres, le pourcentage moyen de non-réponse sur l'ensemble du test étant de 12%. Nous avons pu observer que tous les élèves ont terminé l'exercice durant le temps imparti, l'augmentation du nombre de non-réponses lors des derniers calculs n'est donc pas liée à un manque de temps. Comme pour l'étude précédente, le nombre de non-réponse moyen reste plus élevé chez les filles que chez les garçons et ce dans les mêmes proportions (2,2 vs 1,2). Parmi les quatre calculs que nous avons abordés durant l'entretien, c'est surtout sur $20 + 30$ que l'écart de non-réponses est le plus marqué : un seul garçon n'a pas répondu lors de la passation alors qu'elles sont cinq filles à être dans ce cas. Lors de l'entretien, elles sont cependant capables, pour quatre d'entre elles, de donner une réponse accompagnée d'un début d'explication de procédure, mais tout laisse à penser que durant la passation, elles ont préféré ne pas répondre plutôt que de choisir une réponse dont elles n'étaient pas sûres du tout. La dernière des cinq filles, tout comme le garçon qui n'a pas répondu à $20+30$, explique que le calcul « *est trop dur* » ; ils ne proposeront ni réponse ni début de procédure. On peut alors supposer que la difficulté perçue par les élèves sur les calculs avec des nombres à deux chiffres, et en particulier pour les filles, les conduit à ne pas répondre.

Nous remarquons ensuite que les élèves utilisent régulièrement leurs doigts de façon visible pour répondre : les filles les utilisent davantage, en moyenne pour 4 calculs et les garçons



pour 3. Si l'objectif d'évaluation assigné à cette tâche est d'évaluer des connaissances en calcul en ligne des élèves, comme notifié dans le guide d'évaluation, nous en concluons que les modalités de passations ne sont pas adaptées. Par ailleurs, les explications données par les élèves lors des échanges avec les chercheuses montrent que si les filles et les garçons utilisent dans une proportion similaire une procédure de comptage, les filles semblent davantage mentionner explicitement l'utilisation des doigts pour expliquer la façon dont elles procèdent. La différence est particulièrement marquée pour $18 + 4$ où aucune fille ne propose une décomposition d'un des deux nombres pour trouver le résultat, alors qu'ils sont trois garçons à le faire.

Les entretiens menés à la suite de la passation ne permettent pas de montrer des différences notoires dans le ressenti des élèves face à la difficulté de l'exercice. Signalons cependant que spontanément la durée de l'exercice n'a été évoquée que par deux garçons (« *c'était dur, mais quand même, j'ai arrivé à faire le minuteur* », « *mon record dans tout ça c'était 3'50 si je ne l'avais pas battu, j'aurais pas aimé* ») et seules trois filles ont fait référence à ce qu'elles ont travaillé au préalable avec leur enseignante pour expliquer leur ressenti face à la difficulté de certains calculs (« *On n'a pas appris avec la maitresse* » ; « *y'en a qu'on avait appris avec la maitresse* »).

5. Discussion et conclusion

L'analyse *a priori* des tâches du test ainsi que les analyses quantitatives et qualitatives que nous avons menées à partir de réponses d'élèves montrent que l'évaluation ne permet guère de fournir à l'enseignant un diagnostic des connaissances mathématiques de chacun de ses élèves. D'une part, certains types de tâches autour de la numération écrite chiffrée sont absents du test alors qu'ils évaluent des savoirs fondamentaux pour construire les connaissances de l'année à venir ; d'autre part, pour l'exercice de calcul mental « en ligne », les modalités de passation ne garantissent pas que les élèves calculent mentalement (nos observations montrent une utilisation importante du comptage). Pour autant, la méthodologie statistique mise en œuvre par la DEPP permet de réaliser des comparaisons temporelles et d'apporter des éléments d'évaluation à des fins de pilotage. Nos conclusions sur le contenu du test rejoignent ainsi celles de Mercier-Brunel (2024) sur l'évaluation nationale en français à l'entrée en Sixième et nous pouvons reprendre en partie la conclusion qu'il apporte : « Sans doute faudrait-il [...] clarifier la priorité de ce test, entre le recueil de données à des fins de pilotage et l'outillage des enseignants à des fins de régulation, deux objectifs difficiles à maintenir simultanément », ce que soulignaient déjà en 2015 Troseille et Rocher.

L'observation des scores obtenus par les garçons et les filles dans les deux études tendrait à montrer que les conditions de passation choisies pour l'étude qualitative semblent réduire l'écart de performance, voire l'inverser, entre garçons et filles. Rappelons que les consignes qui ont été données lors de cette expérimentation étaient les mêmes que celles de l'évaluation nationale, mais que les conditions dans lesquelles les élèves étaient placés étaient différentes : l'exercice a été réalisé individuellement en présence d'une chercheuse et nous avons notamment veillé à ce que les élèves soient rassurés sur le fait qu'il n'y avait aucun enjeu en termes d'évaluation à l'exercice qu'ils et elles faisaient. Ainsi, l'anxiété générée chez les filles de 6-7 ans par les conditions de passation des évaluations nationales constatée par Sayac (2023) pourrait se retrouver chez des filles un peu plus âgées, mais cela demanderait à être confirmé. On pourrait également invoquer cette anxiété plus grande des filles pour expliquer le fait de leur plus grande confusion des opérations (additions, soustractions).

Au-delà des conditions de passation, les formats et les modalités de questionnement pourraient également avoir des répercussions sur les performances des élèves. Réaliser une



tâche en temps limité permet certes, de renseigner sur un certain type de connaissances, mais cela peut aussi avoir des répercussions sur l'activité de l'élève et sur ses performances. Le score plus élevé de non-réponses des filles pourrait, en partie, être une conséquence de ce choix de contrainte temporelle. Ajoutons à cela le fait que ce type de contrainte n'est pas très répandu dans les classes d'école primaire, en dehors des séances spécifiquement dédiées à la mémorisation de faits numériques ou de calcul mental.

On peut également émettre l'hypothèse que l'utilisation plus importante des doigts par les filles pour répondre à des problèmes numériques ou des calculs est une procédure qui pourrait davantage les rassurer, ou bien correspondre à des procédures que les enseignant·es accepteraient, voire encourageraient, davantage pour les filles que pour les garçons. En effet, on sait que le stéréotype de supériorité des garçons en mathématiques circule, même dans les salles de classe, et qu'il peut amener les enseignant·es à avoir des pratiques différenciées selon le genre des élèves (Mosconi, 2001).

Références

- Breda, T., Sultan Parraud, J. & Touitou, L. (2024). Le décrochage des filles en mathématiques dès le CP : une dynamique diffuse dans la société. *Notes de l'Institut des Politiques Publiques*, 101. <https://www.ipp.eu/publication/le-decrochage-des-filles-en-mathematiques-des-le-cp-une-dynamique-diffuse-dans-la-societe/> (consulté le 09/08/2024).
- Butlen, D. (2005). *Le calcul mental entre sens et technique*. Presses Universitaires de Franche Comté.
- Che, M., Wiegert, E., & Threlkeld, K. (2012). Problem solving strategies of girls and boys in single-sex mathematics classrooms. *Educational Studies in Mathematics*, 79(2), 311-326.
- Chen, H. Y., Chen, M. F., Lee, Y. S., Chen, H. P., & Keith, T. Z. (2013). Gender reality regarding mathematic outcomes of students aged 9 to 15 years in Taiwan. *Learning and Individual Differences*, 26, 55-63.
- Chevallard, Y. (1999). L'analyse des pratiques enseignantes en théorie anthropologique du didactique. *Recherches en didactique des mathématiques*, 19(2), 221-266. <http://rdm.penseesauvage.com/L-analyse-des-pratiques.html>
- CSEN (2021) *Qu'apprend-t-on des évaluations de CP-CE1 ?* Notes du CSEN 2021-03. <https://www.reseau-canope.fr/conseil-scientifique-de-leducation-nationale-site-officiel/publications-et-ressources/notes-du-csen.html>
- DEPP (2023). Repères CP, CE1 – Premiers résultats. *Série Etudes*. Document de travail n°2023-E05. Novembre 2023. <https://www.education.gouv.fr/sites/default/files/document/%C3%89valuations%202023%20Rep%C3%A8res%20CP%2C%20CE1%20-%20Premiers%20r%C3%A9sultats-366546.pdf>
- Devine, A., Fawcett, K., Szűcs, D., & Dowker, A. (2012). Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and brain functions*, 8, 1-9.
- Genoud, P. A., Ruiz, G., & Gurtner, J. L. (2009). Evolution de la motivation scolaire des adolescents. Différences selon la filière et le genre. *Schweizerische Zeitschrift für Bildungswissenschaften*, 31(2), 377-395.
- Gilles, J.-L. (1996). Utilisation des degrés de certitude et normes de réalisme en situation d'examen et d'auto-estimation à FA.P.S.E. - ULG. *Communication présentée à ADMEE-EUROPE "Dix années de travaux de recherche en évaluation"*, Grenoble, France. <http://hdl.handle.net/20.500.12162/4809>
- Goasdoué, R., Vantourout, M. (2014). Approches et validité psycho-didactiques des évaluations. *Education et Formation*. e-302, décembre.
- Goudeau, S., Autin, F., & Croizet, J. C. (2017). Etudier, mesurer et manipuler la classe sociale en psychologie sociale: Approches économiques, symboliques et culturelles [Studying, measuring



- and manipulating social class in social psychology: Economic, symbolic and cultural approaches]. *International Review of Social Psychology*, 30(1), 1-19.
- Grapin, N., Chenevotot, F., Ledan, L., Beylot, D., Mounier, E., Blanchouin, A. (2022) Etude exploratoire de procédures d'élèves de 7-8 ans en calcul mental aditif. *Revue Math-Ecole*, 238, 29-40.
- Grapin N., Grugeon-Allys B. (2018). Approches psychométrique et didactique de la validité d'une évaluation externe en mathématiques : quelles complémentarités ? Quelles divergences ? *Mesure et évaluation en éducation*. 41-2, 37-66.
- Grapin N., Vantourout, M., Grugeon-Allys B. (2022). S'assurer de la validité des évaluations en milieu scolaire. *Recherches en didactiques*, 33, p. 59-83. DOI : 10.3917/rdid1.033.0059.
- Houdement, C. (2017). Résolution de problèmes arithmétiques à l'école. *Grand N*, 100, 59-78.
- Houdement, C. (2011). Connaissances cachées en résolution de problèmes. *Annales de didactique et de sciences cognitives*, 16, 67 – 96.
- Houssa & Nader-Grosbois (2018) Réalisation d'enquêtes quantitatives par questionnaire auprès de jeunes enfants (5-8ans). Etape 1 : Rapport final de recherche. Institut des sciences psychologiques (IPSY, UCL). Université catholique de Louvain. https://oejaj.cfwb.be/fileadmin/sites/oejaj/uploads/Hors_PublicationsTravaux/Documents/CID E/Participation des enfants en FWB/outils pratiques/rapport_final_interroge_enfant.pdf
- Laveault, D., Grégoire, J. (2023). *Introduction aux théories des test en psychologie et en sciences de l'éducation*. De Boeck supérieur : Louvain la Neuve.
- Leclercq, D. (1982). Confidence marking: its use in testing. *Evaluation in Education*, 6(2), 163-287.
- Marsh, H. W., Martin, A. J., & Cheng, J. H. (2008). A multilevel perspective on gender in classroom motivation and climate: Potential benefits of male teachers for boys?. *Journal of educational psychology*, 100(1), 78.
- MENJS (2020a) Pour enseigner les nombres, le calcul et la résolution de problèmes au CP.
- MENJS (2020b). Attendus de fin d'année CP – mathématiques. <https://eduscol.education.fr/document/13936/download>
- MENJS (2023a). Repères CE1 2022. Guide pour le professeur.
- MENJS (2023b). Repères CE1 2022. Cahier de l'élève.
- Mercier-Brunel, Y. (2024) Le test national de compréhension d'un texte littéraire écrit à l'entrée en sixième peut-il constituer un outil pour l'apprentissage ? *Repères*, 69, 147-164.
- Mosconi, N. (2001). Comment les pratiques enseignantes fabriquent-elles de l'inégalité entre les sexes?. *Les dossiers des sciences de l'éducation*, 5(1), 97-109.
- Régner, I., Steele, J. R., & Huguet, P. (2014). Stereotype threat in children: Past and present. *Revue internationale de psychologie sociale*, 27(3), 5-12.
- Robert A. (1998). Outil d'analyse des contenus mathématiques à enseigner au lycée et à l'université. *Recherches en didactique des mathématiques*, 18(2), 139-190.
- Rogalski, J. (2003). Y a-t-il un pilote dans la classe ? Une analyse de l'activité de l'enseignant comme gestion d'un environnement dynamique ouvert. *Recherches En Didactique Des Mathématiques*, 23(3), 343–388.
- Sayac, N. et Grapin, N. (2016). Stratégies et degrés de certitude des filles et des garçons en mathématiques. *Repères IREM*, 104, 43-57.
- Sayac N. et Grapin N. (2014). Évaluer par QCM en fin d'école : stratégies et degrés de certitude. *Annales de Didactique et de Sciences Cognitives*, 19, 169-199.
- Sayac, N. (2023). Comment expliquer les écarts de performance entre les filles et les garçons en mathématiques aux évaluations nationales de CP ? *Évaluer. Journal international de recherche en éducation et formation*, 9(3).
- Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of personality and social psychology*, 69(5), 797.



- Terrier, C. (2024). Un coup de pouce pour les filles ? Les biais de genre dans les notes des enseignants et leur effet sur le progrès des élèves. *Notes de l'Institut des Politiques Publiques*, 14. <https://www.ipp.eu/actualites/un-coup-de-pouce-pour-les-filles-les-biais-de-genre-dans-les-notes-des-enseignants-et-leur-effet-sur-le-progres-des-eleves/> (consulté le 27/07/2024)
- Toczek, M. C. (2005). Réduire les différences de performances selon le genre lors des évaluations institutionnelles, est-ce possible ? Une première étude expérimentale... *L'orientation scolaire et professionnelle*, (34/4), 439-460.
- Vergnaud, G. (1991). La théorie des champs conceptuels. *Recherches en didactique des mathématiques*, 10 (2/3), 133 – 170.

Remerciements

Nous remercions celles et ceux qui ont permis de rendre possible l'enquête qualitative auprès des élèves : Madame Samut, Inspectrice de l'éducation nationale de la circonscription de Paris, Monsieur Lasnet, conseiller pédagogique ainsi que les enseignant·es des classes de CP des écoles Primo Levi, Jean Simon et des Grands Moulins

Nadine Grapin

Maitresse de conférences en didactique des mathématiques à l'Université Paris-Est-Créteil, elle appartient au Laboratoire de Didactique André Revuz et mène des recherches sur la validité des évaluations. Ses intérêts actuels portent notamment sur les apprentissages en résolution de problèmes et sur les gestes évaluatifs des enseignant·es pouvant soutenir l'autorégulation de ces apprentissages.

Nathalie Sayac

Nathalie Sayac est professeure des universités en didactique des mathématiques à l'université de Rouen Normandie. Ses travaux portent sur l'évaluation et ses pratiques en mathématiques, mais également sur les inégalités de genre en mathématiques. Elle est également inspectrice générale de l'éducation, du sport et de la recherche depuis septembre 2024.

