UNE ÉVALUATION SOUS FORME NUMÉRIQUE EST-ELLE COMPARABLE À UNE ÉVALUATION DE TYPE « PAPIER-CRAYON » ?

Pascal Bessonneau
MENESR-DEPP, bureau de l'évaluation des actions éducatives et des expérimentations

Philippe Arzoumanian et Jean-Marc Pastor
MENESR-DEPP, bureau de l'évaluation des élèves

Aujourd'hui, la place prépondérante prise par l'informatique questionne l'école sur la transition d'un environnement dominé par le papier vers un environnement dominé par le support numérique. Cette transition est en marche dans le domaine des évaluations standardisées. Cependant, la question de la comparabilité de la mesure est posée. L'hypothèse sous-jacente d'une transition naturelle et sans contrainte d'un support à l'autre doit en effet être interrogée.
De nombreux articles tirés de la littérature scientifique comparent les performances des élèves à des évaluations proposées sur support papier et sur support électronique. Ces études indiquent des résultats divergents. Les tests sont parfois plus faciles, parfois plus difficiles, ou de même difficulté.
Une question se pose : un item peut-il être proposé aux élèves à l'identique dans les deux supports sans influencer sa difficulté et sans provoquer de modification des compétences mises en jeu ?
L'article présente les résultats de deux expériences menées sur ce thème dans le cadre des évaluations standardisées conduites par la DEPP. La première cherche à identifier les différences de difficulté des items entre le support papier et le support numérique, à partir d'une évaluation des compétences de base en français et en mathématiques conduite en fin de primaire. La seconde expérience tente de dégager les variables explicatives de ces différences sur la base d'une étude menée en mathématiques en fin d'école et en fin de collège dans le cadre de Cedre.

La DEPP a pour mission de concevoir et de développer des évaluations pour mettre à disposition des responsables du système éducatif des informations rigoureuses et objectives, tant sur l'évolution des connaissances et des

compétences des élèves que sur leur développement conatif [Trosseille et Rocher, dans ce numéro, p. 15]. De plus en plus, ces évaluations sont réalisées sur support numérique. Se pose donc la question de la comparabilité des épreuves sur les deux supports, notamment dans le cas de reprise d'épreuves afin d'établir des comparaisons temporelles. Les items peuvent-ils être proposés aux élèves à l'identique sans changer la difficulté et les compétences mises en jeu ?

De nombreux tests institutionnels et commerciaux sont proposés sous forme papier et/ou numérique en parallèle, ou bien exclusivement sous forme numérique. Pour citer des enquêtes internationales récentes, PISA (*Programme for International Student Assessment*) a interrogé en 2009 une partie des élèves sous forme numérique [OCDE, 2011] et pour la session 2015, les évaluations seront entièrement proposées sous forme numérique. L'enquête ESLC *(First European Survey on Language Competences)* a interrogé, en 2011, dans certains pays une partie des élèves sous format numérique (Commission européenne, 2012). De même, l'enquête Piaac (*Programme for the International Assessment of Adult Competencies*) a également évalué les individus selon les deux modalités, papier-crayon et numérique [voir Murat et Rocher, dans ce numéro, p. 83].

Ces enquêtes sont illustratives des problèmes de comparabilité rencontrés. En 2009, PISA a interrogé les élèves sur du matériel créé spécialement pour le support informatique. Il s'agissait d'une évaluation spécifique de la lecture dans un environnement électronique : ERA (*Electronic Reading Assessment*). En 2015, le mode d'interrogation sous forme numérique sera généralisé, à tous les domaines, précédemment évalués sous forme papier. Les responsables de PISA devront donc s'assurer de la comparabilité des résultats avec les vagues antérieures, notamment la session 2006, alors que les modalités d'interrogation sont différentes. C'était d'ailleurs l'objet d'une étude dans le cadre de l'expérimentation de PISA 2015, qui a eu lieu en 2014, et dont les résultats ne sont pas connus au moment où le présent article est rédigé. Les enquêtes ESLC et Piaac interrogent quant à elles les élèves ou les adultes d'un même pays, soit sur informatique soit sur papier, avec le même matériel d'évaluation, en supposant que les deux modalités sont comparables.

La première question, posée dans le cadre d'ERA, est celle de l'équivalence des compétences de lecture mises en jeu selon les deux modalités d'interrogation. Outre les questions ergonomiques, les concepteurs utilisent dans ce type de test la navigation hypertextuelle, les onglets, de nouvelles formes d'items, etc. Se pose alors la question de savoir si on ne mesure pas une nouvelle compétence distincte de la lecture sur support papier.

La comparaison réalisée dans le cadre de l'évaluation ERA repose sur l'analyse de la corrélation des scores. Le coefficient de corrélation observé entre l'épreuve papier de compréhension de l'écrit de PISA et l'épreuve numérique construite dans le cadre d'ERA est élevé, d'une valeur de 0,83 sur l'ensemble des 16 pays ayant participé à ERA. Il est cependant intéressant de comparer cette corrélation avec celle observée entre la compréhension de l'écrit sous forme papier et les mathématiques et les sciences, elles aussi évaluées sous forme papier ▸ **Tableau 1**.

Tableau 1 **Corrélations entre les épreuves de PISA**

	Littératie (papier)	Littératie (électronique)
Littératie (papier)	1	
Littératie (électronique)	0,83	1
Mathématiques	0,83	0,76
Sciences	0,88	0,79

Note de lecture : le tableau indique la corrélation entre les scores des différentes épreuves pour les 16 pays de l'OCDE participant à l'expérimentation.
Source : OCDE, 2011.

Il apparaît que la corrélation entre les épreuves papier et électronique est du même ordre de grandeur que la corrélation entre la compréhension de l'écrit sous forme papier et d'autres compétences telles que les mathématiques. La lecture sur support électronique apparaît donc bien comme un domaine distinct de la lecture sur papier, au même titre que les mathématiques ou les sciences. Enfin, il faut noter que la corrélation papier/électronique varie de manière importante selon les pays : de 0,71 à 0,89. Ce dernier point peut soulever un problème de comparabilité internationale, qui doit très certainement prendre en compte la familiarité des élèves avec le support numérique.

Pour l'enquête ESLC, le rapport n'indique pas, quant à lui, de différences majeures entre les difficultés des items issus du papier et du test électronique [Commission européenne, 2012]. Les données ne sont présentées que pour l'anglais en compréhension écrite et en compréhension orale : il semble ressortir que les niveaux de difficulté observés sur les deux supports, papier et électronique, sont moins comparables en compréhension de l'écrit qu'en compréhension orale.

Au niveau national, la DEPP a développé une application « Lecture sur support électronique » (LSE) afin d'évaluer spécifiquement la lecture sur support informatique. La comparaison entre les résultats à ce test et à un test de maîtrise de la langue issu du cycle Cedre [Colmant, Daussin, Bessonneau, 2011] a abouti pour des élèves de CM2 à des corrélations entre les tests relativement basses, de l'ordre de 0,6 [Bessonneau, 2012 ; Dierendonck, 2014]. En outre, elle mettait en avant de meilleurs résultats relatifs sur le format électronique pour les garçons et pour les élèves en zone d'éducation prioritaire notamment.

Au-delà des enquêtes d'évaluation, des revues de littérature comparant les performances sur papier et sur support électronique abondent. Par exemple, la revue de Wang et Shin [2009] fait état de résultats divergents. Les tests sont parfois plus faciles, plus difficiles ou de même difficulté. Ces études sont pour la plupart anglo-saxonnes et portent fréquemment sur des épreuves comportant un certain enjeu pour les élèves, car faisant partie intégrante de leur cursus scolaire. Or, un facteur important de différence pourrait avoir trait à la motivation des élèves face à la situation d'évaluation. Des enquêtes telles que PISA ou LSE sont des évaluations sans enjeu pour les élèves de l'échantillon, et cette absence d'enjeu pourrait expliquer certains écarts observés entre les différents supports. Une autre hypothèse sur la diversité des résultats peut porter sur le fait que les études sont menées à partir de logiciels spécifiques, qui adoptent des ergonomies différentes les uns par rapport aux autres.
Les résultats obtenus à la DEPP à travers les évaluations nationales et internationales, ainsi que les résultats contrastés rapportés par la littérature scientifique,

nous ont amené à conduire des expériences spécifiques pour apprécier la comparabilité des deux supports.

Ainsi, cet article présente les résultats de deux expériences :
– la première cherche à identifier les différences de difficulté des items entre le support papier et le support numérique, à partir d'une évaluation des compétences de base en français et en mathématiques conduite en fin de primaire, et selon un plan d'expérience (*design*) original ;
– la seconde expérience concerne les résultats d'un lot d'items identiques proposés sur support numérique et sur support « papier » dans le cadre des évaluations Cedre mathématiques.

Les résultats de ces expériences mettent en avant l'interaction entre la forme des items, la charge cognitive pour l'élève et les restrictions liées à l'un ou l'autre des supports.

EXPÉRIENCE 1 : ÉVALUATION DES COMPÉTENCES DE BASE EN FRANÇAIS ET EN MATHÉMATIQUES EN FIN DE PRIMAIRE

De 2007 à 2012, pour alimenter les indicateurs de performance du système éducatif attendus par la LOLF (loi organique des lois de finances), une évaluation annuelle de la DEPP au primaire (CM2) et au collège (troisième) réalisée sur échantillon a permis d'évaluer le niveau de maîtrise des compétences de base en français et en mathématiques. La description de la création de cette épreuve a fait l'objet d'une *Note d'information* [Rocher, Chesné, Fumel, 2008].

Concernant le CM2, l'épreuve finale était composée de 75 items de français et de 68 items de mathématiques. Ces items provenaient d'une large expérimentation d'items dont la création impliquait la DEPP, des enseignants, des conseillers pédagogiques, des inspecteurs de l'éducation nationale (IEN, IA-IPR, IGEN). La qualité des résultats de cette évaluation reposait sur sa conception, sur l'évaluation de larges échantillons d'élèves et sur l'utilisation d'outils psychométriques pour l'analyse des résultats [Laveault et Grégoire, 2002].

Les indicateurs attendus par la LOLF étaient les proportions d'élèves maîtrisant les compétences de base. Les scores des élèves devaient donc être utilisés pour différencier ces deux populations en créant un score seuil départageant ces deux populations d'élèves. La détermination de ce score seuil est établie selon des procédures dites de « *standard settings* » [Bunch et Cizek, 2007]. Ce travail permet de croiser les exigences et les attentes pédagogiques avec les résultats psychométriques en vue de déterminer un score seuil faisant consensus. Depuis 2011, de nouvelles épreuves ont été conçues par la DEPP pour évaluer les compétences du socle commun, avec les mêmes soucis méthodologiques que les épreuves évaluant les compétences de base [Miconnet et Vourc'h, dans ce numéro, p. 141]. Mais afin de s'assurer notamment de la consistance des résultats, les deux évaluations ont coexisté durant quelques années.

Les deux épreuves se caractérisent par la nécessité de produire des résultats fiables et comparables dans le temps. Or, pour des exigences matérielles, ces nouvelles épreuves du socle sont confrontées à terme à la possibilité d'être dématérialisée, c'est-à-dire que les cahiers d'évaluation pourraient être remplacés par une évaluation sur support numérique. C'est pour étudier les conséquences d'une éventuelle transition du papier vers le numérique qu'une expérimentation *ad hoc* a été proposée.

Description de l'expérimentation

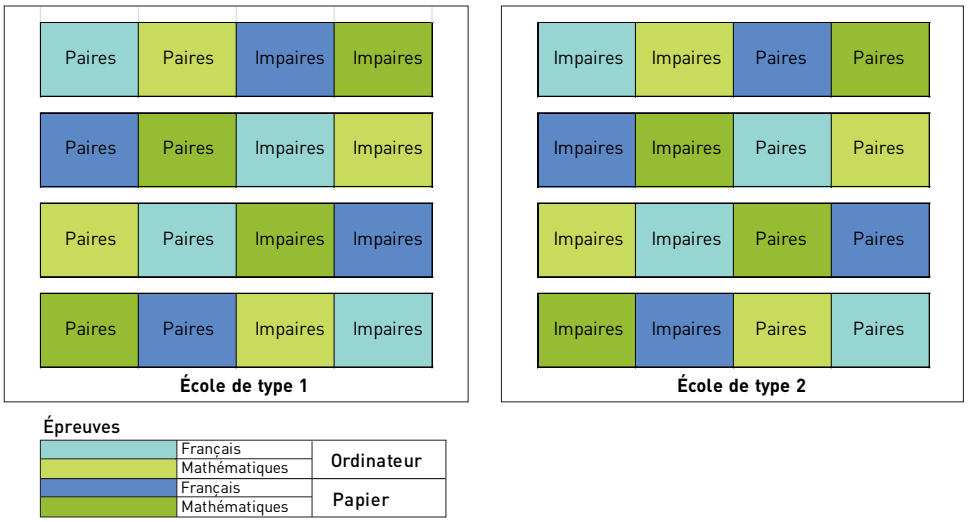
L'épreuve de compétences de base a été divisée en deux : une épreuve dite « paire » et une épreuve dite « impaire ». Les items pairs ont été inclus dans l'épreuve paire et les items impairs dans l'épreuve impaire. Cette alternance pair/impair permet de garder la position relative des items dans chaque épreuve.

Toutefois, les textes et supports longs présents dans l'épreuve n'ont pu être divisés. En effet, si on avait divisé les items en gardant les supports longs, l'épreuve aurait été :
– trop longue pour les élèves ;
– les élèves auraient retrouvé les mêmes textes en partie paire et impaire ;
– le décloisonnement des items pour chaque texte aurait nui à la comparabilité en cas de dépendance entre items.

L'élaboration du plan d'expérience (*design*) s'est appuyée sur plusieurs contraintes. Le but était de contrôler différents paramètres concernant l'ordre de passation des épreuves, à savoir le support de passation (informatique ou papier), l'épreuve (paire ou impaire) ainsi que la discipline (français ou mathématiques). Afin que les élèves ne puissent pas communiquer sur les items, la passation se fait en parallèle sur les mêmes items.
Ainsi, dans un premier type d'écoles dit « paire puis impaire » la moitié des élèves passait la première épreuve « paire » sur support informatique tandis que l'autre moitié des élèves passait l'épreuve « paire » sur papier. Chacun de ces deux groupes d'élèves était divisé en deux : l'un passant d'abord le français, l'autre passant d'abord les mathématiques.
La même organisation était appliquée dans les écoles du deuxième type dit « impaire puis paire », mais l'ordre de passation des épreuves était inversé (d'abord les épreuves impaires puis les paires).
Le *design* est donc complètement équilibré puisque chaque discipline, chaque série d'items et chaque support se retrouvent dans chaque position. Ce *design* est schématisé par la **figure 1 p. 164**.

Concernant le contenu de l'évaluation, toutes les questions sont des QCM avec 2, 3 ou 4 choix possibles. La présentation des items était presque identique : les items provenaient d'une même banque d'items et leur génération sur papier et sur électronique était automatique. L'épreuve impaire de français était composée de 36 items et celle de mathématiques de 33 items. L'épreuve paire est, quant à elle, composée de 33 items de français et de 32 items de mathématiques.
L'échantillon était composé d'écoles se répartissant en deux catégories :
– des écoles volontaires situées en province (un tiers de l'effectif) ;
– des écoles parisiennes tirées au hasard sur les écoles ne participant pas à une autre opération de la DEPP en CM2 (deux tiers de l'effectif).

Figure 1 Plan de rotation des blocs électronique et papier



Note de lecture : pour une école de type 1, quatre cas sont possibles. Dans les deux premiers cas, le français est vu avant les mathématiques alors que dans les deux derniers cas, les mathématiques sont passées en premier. Dans les cas 1 et 3, le test électronique est passé avant le test papier-crayon. Dans les cas 2 et 4, c'est l'inverse : le test papier-crayon est passé avant le test électronique.

Tous les élèves de CM2 des écoles sélectionnées participaient à l'expérience. Au total, nous avons collecté des résultats pour 44 écoles et environ 800 élèves.
Selon les résultats recueillis, il s'avère qu'une partie des écoles n'a pas respecté le plan de rotation. En cause notamment la difficulté pratique pour la mettre en place au sein d'une école. En outre, le respect des conditions de passation était parfois particulièrement difficile au regard du matériel informatique disponible. Ainsi, du point de vue du *design*, si les cahiers sont équilibrés, le respect de l'ordre de passation n'a pas été tout à fait respecté.

Des questions de « contrôle » étaient proposées sur la partie informatique telles que : « As-tu déjà passé l'épreuve papier ? », « Quel était l'identifiant de ton cahier ? », etc. En outre, il était demandé à l'élève d'indiquer son mois de naissance, son année de naissance et son sexe sur les deux supports. Le but de ces questions était de s'assurer lors de l'analyse que les résultats sur les deux supports correspondaient à un seul et même élève. Pour près de 200 élèves, les réponses à ces questions indiquaient qu'il s'agissait probablement d'un élève différent : les réponses étaient discordantes entre papier et support électronique. Ces élèves ont été supprimés pour l'analyse. L'échantillon final portait donc sur 632 élèves dans 39 écoles.

Des statistiques sur les établissements et les élèves ont été récoltées :
– nombre de filles et de garçons ;
– école de secteur privé ou école publique ;
– niveau de l'école aux épreuves nationales de CM2 en 2011 ;
– ordre de passation des épreuves.

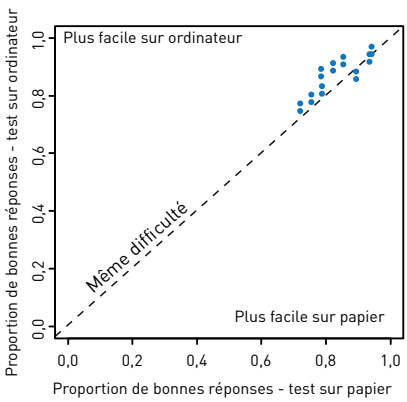
Ces statistiques ont été utilisées pour calculer des poids dans le but de redresser l'échantillon. En incluant les poids, chaque cahier présente la même répartition pour chacune des statistiques citées. Le calcul des poids a été réalisé par un calage sur marge [Sautory, 1993].

Résultats

En premier lieu, il s'agit d'identifier des différences entre la difficulté des items sur le support papier et le support numérique qui serait un indicateur important de comparabilité des deux supports. Dans un second temps, une analyse des scores selon les supports était envisagée. Toutefois étant donné la faible longueur de chacune des parties du test et la grande facilité de l'épreuve, une proportion importante d'élèves obtenait le score maximal à chaque partie du test. La distribution des scores pour plusieurs épreuves étant tronquée, l'analyse des scores n'était pas envisageable.

Comme les différents échantillons ont été équilibrés, les premières analyses portent sur la comparaison des proportions de bonnes réponses aux questions (items) entre leur version sur papier et leur version sur support électronique. En français, les proportions de réussite sont très voisines pour les items d'orthographe ▸ **Figure 2**. Inversement, des différences non négligeables apparaissent pour les items de compréhension de textes.

▸ **Figure 2** **Proportion comparée de bonnes réponses pour l'orthographe**

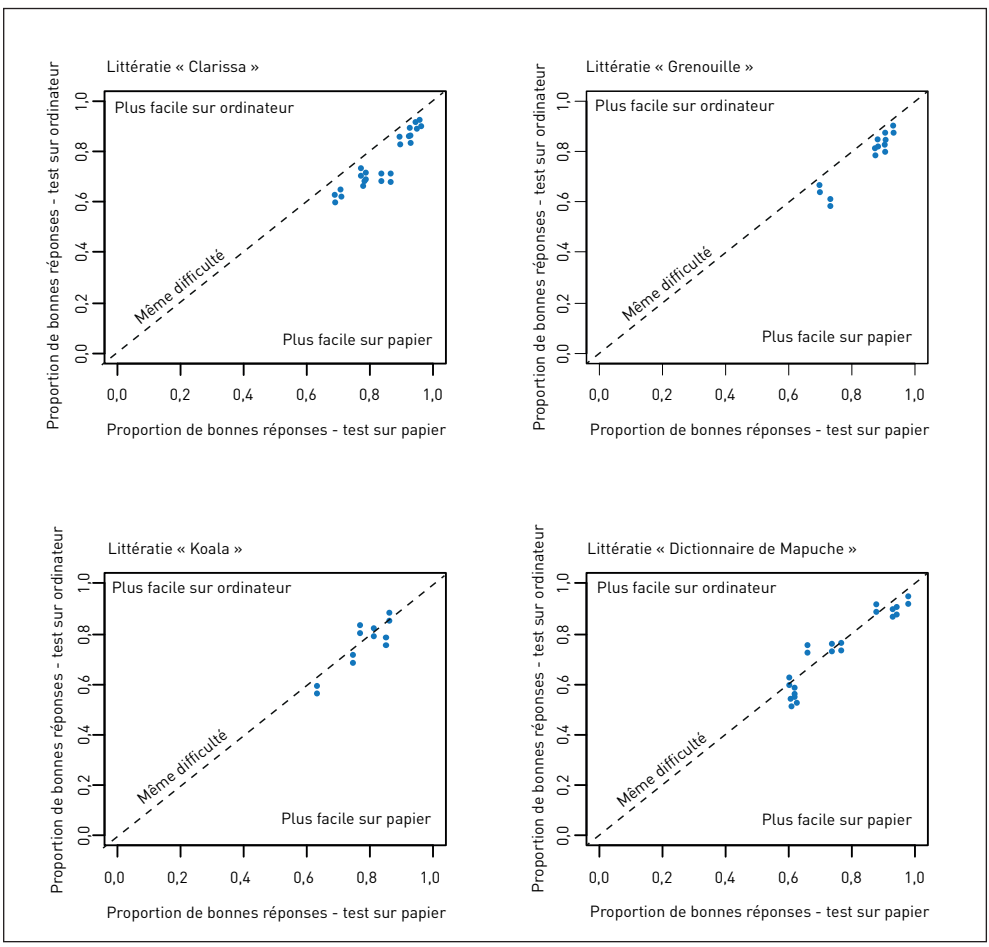


Note de lecture : chaque point représente un item. L'axe des abscisses représente le taux de réussite des items en version électronique, l'axe des ordonnées le taux de réussite des mêmes items dans leur version papier/crayon. La droite qui partage ce graphique indique un niveau de réussite identique sur les deux supports. Un éloignement de cette droite correspond soit à une réussite plus grande sur le papier (au-dessus de la droite) soit à une réussite plus grande sur le support numérique (au-dessous de la droite).

Globalement les items sont mieux réussis sur papier que sur ordinateur. Les différences sont notables pour les textes « Clarissa » et « Grenouille » tandis qu'elles sont moindres pour les textes « Koala » et « Dictionnaire de Mapuche » ▸ **Figure 3**.
Pour « Clarissa » et « Grenouille », on peut poser l'hypothèse que, les supports étant longs, le défilement du document sur ordinateur a rendu la tâche plus difficile. En outre pour « Grenouille », les questions sont en regard du texte sur la version papier, pas sur la version informatisée. Ce qui a conduit à diminuer la difficulté des items sur papier.
Pour « Koala », l'hypothèse inverse peut être formulée. Le texte assez court ne nécessite pas de manipulation supplémentaire (ascenseur vertical) pour lire le texte dans son intégralité, d'où une différence de difficulté moindre selon le support.

Pour le « dictionnaire de Mapuche », le document est long mais, comme pour « Grenouille », les questions se retrouvant en regard du support, elles sont plus faciles sur papier.
De manière générale, il apparaît une cohérence globale en termes de hiérarchie de difficulté des items, à l'exception de quelques items, notamment pour le texte « Clarissa », dont l'écart de difficultés entre les deux versions est plus prononcé que les autres items.

Figure 3 Proportions comparées de bonnes réponses pour les textes

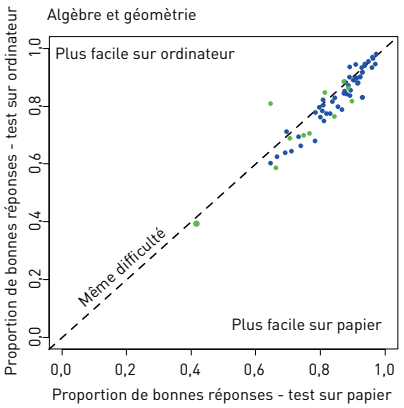


Note de lecture : chaque point représente un item. Un éloignement de la droite qui partage le graphique correspond soit à une réussite plus grande sur le papier (au-dessous de la droite) soit à une réussite plus grande sur le support numérique (au-dessus de la droite).

Pour les mathématiques, à part pour les questions les mieux réussies, l'épreuve apparaît plus facile sur papier que sur ordinateur ▸ Figure 4.
Nous notons une exception toutefois : un item est largement plus facile, car il s'agissait d'identifier un parallélépipède et le redimensionnement de l'image sur ordinateur rendait la tâche plus aisée.

Pour les mathématiques, une recherche a été menée sur certaines questions difficiles à réaliser sans brouillon. Nous avons recherché l'utilisation du cahier comme brouillon, mais sur les images des cahiers numérisés, aucune trace d'écriture n'était visible.

Figure 4 **Proportions comparées de bonnes réponses pour les items de mathématiques**



Note de lecture : chaque point représente un item. L'axe des abscisses représente le taux de réussite des items en version électronique, l'axe des ordonnées le taux de réussite des mêmes items dans leur version papier/crayon. La droite qui partage ce graphique indique un niveau de réussite identique sur les deux supports. Un éloignement de cette droite correspond soit à une réussite plus grande sur le papier (au-dessous de la droite) soit à une réussite plus grande sur le support numérique (au-dessus de la droite).

Conclusion de l'expérience 1

Les différences de difficulté des items semblent dépendre de la nature du support, du « contexte » de l'exercice sur le papier et de la nature des tâches.
Plus que des différences entre les supports, les différences entre numérique et papier mettent en lumière des différences qui pourraient exister entre différentes mises en page papier : la proximité du texte avec les questions, sa taille, etc. La version numérique y apparaît comme une mise en page de l'item parmi d'autres.
En mathématiques comme en orthographe, peu de différences entre les taux de réussite selon les supports sont observées. Toutefois cela est à relativiser car les supports de ces items sont tous de faible taille.

EXPÉRIENCE 2 : LES ACQUIS DES ÉLÈVES DANS LE CADRE DE L'ÉVALUATION CEDRE MATHÉMATIQUES

Les enquêtes du cycle d'évaluations disciplinaires réalisées sur échantillons (Cedre) sont réalisées tous les ans sur une discipline différente. En 2013, en vue de l'évaluation de 2014, des items de mathématiques ont été expérimentés en fin de collège et en fin de CM2 afin d'écarter ceux apportant peu d'informations sur la compétence des élèves. Cette sélection porte sur une analyse psychométrique et pédagogique [Rocher, dans ce numéro, p. 37].
Les techniques de l'information et de la communication viennent modifier aussi bien les pratiques pédagogiques que les compétences devant être mises en œuvre par les élèves.

Il est important d'observer ce que ces techniques numériques garantissent comme continuité par rapport aux techniques traditionnelles ; ce qu'elles proposent comme nouveautés et ce qu'elles ne peuvent pas remplacer.

Dans le cadre de l'évaluation des élèves et plus spécifiquement du protocole Cedre qui a pour but d'observer les évolutions temporelles, il est primordial de savoir si le passage d'une évaluation papier-crayon à une évaluation numérique permet d'assurer une mesure comparable. Comme pour la première expérience, nous avons établi un protocole de « dématérialisation », qui consiste à reproduire à l'identique un item sur le support « papier » et sur le support « numérique ». Les élèves de l'échantillon passent les items en partie sous une forme ou en partie sous l'autre.

Cette deuxième expérience a pour objectif d'analyser les différences entre les deux formats selon plusieurs critères liés au domaine des mathématiques :
– les champs mathématiques ;
– les compétences mises en jeu ;
– les facteurs de complexités liés à l'énoncé, à la connaissance mathématique et à la tâche demandée à l'élève.

L'analyse des résultats, à travers les taux de réussite observés selon le support, doit nous conduire à répondre aux questions suivantes :
– si des écarts sont constatés en fonction d'un support, peut-on quantifier ces écarts ?
– l'écart est-il toujours dans le même sens ?
– quelles sont les pistes explicatives ?

Notons tout d'abord qu'il existe des différences « intrinsèques » entre les deux modes :
– les écrans, même en haute résolution, n'offrent pas le confort de lecture du support écrit ; ils introduisent parfois des déformations. Ceci peut s'avérer un biais important pour les items de géométrie. L'écran impose une lecture verticale moins confortable que celle proposée par la lecture d'un texte posé sur un bureau ;
– la structure d'une page numérisée d'information n'utilise pas les mêmes référents que la page imprimée. Sur un support numérique, la segmentation de l'information est spécifique : les pages sont plus courtes, les repères graphiques multiples, les liens hypertextes renvoient à de nouveaux contenus. Tous les éléments ne sont donc pas accessibles d'un seul coup d'œil par le lecteur. Les pages ne sont pas numérotées, mais plutôt disposées en réseau ;
– des dispositifs particuliers permettent l'accès à ces pages (liens, fenêtres multiples, onglets, etc.). Le lecteur doit maîtriser ces structures spécifiques. Sur support numérique, l'accès aux contenus demande à l'élève de s'approprier le mode de navigation et les actions spécifiques de déplacements entre les pages numériques ;
– à la capacité de lecture-compréhension du message s'ajoute celle de l'accès à l'information. On ne déploiera pas les mêmes procédures pour accéder à un contenu donné dans un livre ou dans un site Web. Dans certains cas la transposition d'un support « papier » à un support « numérique » n'est pas simple.
Ces différences impliquent très certainement une variabilité en termes de procédures de résolution, de stratégies, de processus cognitifs. Cela peut expliquer que les taux de réussite des items soient dépendants du mode d'interrogation.
Dans le cadre de l'évaluation Cedre, nous trouvons cette différence. Chaque unité de

l'évaluation correspond à un ensemble d'exercices sur un même thème. Les exercices sont disposés « classiquement » sur le support papier. L'élève peut appréhender l'ensemble de la situation d'un seul coup d'œil. Sur le support numérique, nous avons procédé à une segmentation de l'unité qui donne à voir à l'élève un seul exercice à la fois.

Première partie : Cedre école

Lors de l'expérimentation de 2013, les élèves ont été évalués à la fois sur support papier et sur support numérique en mathématiques. L'échantillon comptait un peu plus de 4 500 élèves de CM2 répartis dans 172 écoles. Les réponses de 3 841 élèves répartis dans 149 écoles ont été analysées pour la partie papier et les réponses de 2 575 élèves répartis dans 118 écoles pour la partie numérique. Le faible pourcentage de retour de l'enquête numérique s'explique par la grande disparité et la quantité de matériel disponible dans les écoles, ce qui constitue aujourd'hui une première limite importante en matière de comparabilité des deux modes d'interrogation[1].

Le matériel d'évaluation était constitué de six cahiers pour les items en version papier-crayon et de douze modules pour les items en version numérique. Parmi l'ensemble des items expérimentés, 56 items ont été évalués sur les deux supports (papier et numérique). Ils recoupent les champs mathématiques de la connaissance des nombres entiers naturels, des nombres décimaux, de la géométrie, des grandeurs et mesures, de l'organisation et gestion de données.

Ces items étaient répartis dans les six cahiers et dans les douze modules numériques. Chaque élève se voyait attribuer, de manière aléatoire, un des six cahiers et un des douze modules. La correspondance entre cahiers et modules a été élaborée de manière à ce qu'un élève ayant répondu à un item sur le support « papier » ne retrouve pas le même item sur le support « numérique ».
L'évaluation est séquencée par une présentation de l'activité conduite par l'enseignant ; une phase d'entraînement et l'évaluation de mathématiques que l'élève réalise en complète autonomie.

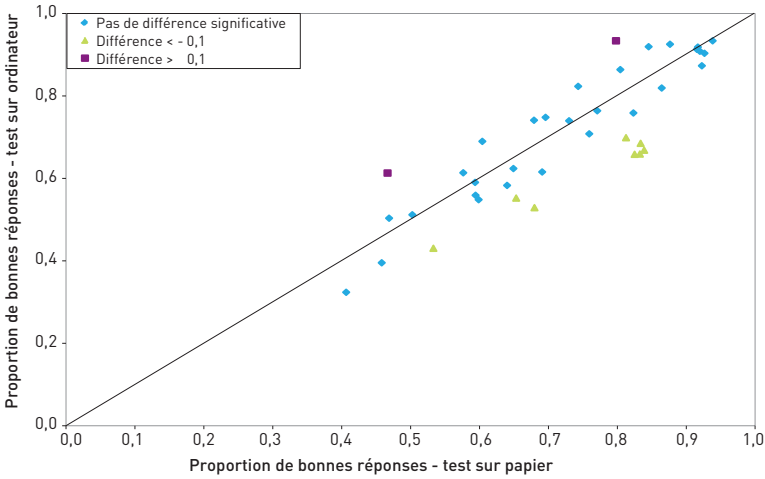
Sur le support papier, l'élève doit appréhender les formats de questions ; sur le support numérique s'ajoute le nécessaire apprentissage de la navigation.
À la fin de l'évaluation, les cahiers nous sont retournés pour le traitement des résultats ; pour la partie numérique, les résultats sont directement sauvegardés dans la base de données de la DEPP et immédiatement consultables.

Analyse globale

Trente-neuf items présentant des qualités psychométriques satisfaisantes ont été retenus pour l'analyse. Pour chaque item, le taux de réussite sur les deux supports a été calculé ▸ Figure 5 p.170.

1. Nous observons tous types de systèmes d'exploitation (Windows, MacOs, Linux) ou d'architectures (monoposte, réseau, terminaux, netbooks, tablettes). Les ordinateurs sont regroupés en salle informatique ou répartis dans les classes de l'école. Nous constatons l'impossibilité de faire une passation lorsqu'il y a trop peu d'ordinateurs (un ordinateur en fond de classe par exemple).

Figure 5 Taux de réussite aux items selon le support



Note de lecture : chaque point correspond à un item.
L'axe des abscisses représente le taux de réussite des items en version papier-crayon, l'axe des ordonnées le taux de réussite des mêmes items dans leur version numérique. La droite qui partage ce graphique indique un niveau de réussite identique sur les deux supports. Un éloignement de cette droite correspond soit à une réussite plus grande sur le papier (au-dessous de la droite) soit à une réussite plus grande sur le support numérique (au-dessus de la droite).

Une première approche montre que :
– 8 items sont mieux réussis sur le support papier. Ces items comportent des textes (consignes ou propositions de réponses) longs et nécessitant une lecture fine. Parfois, l'élève utilise la possibilité d'écrire sur le document pour marquer des repères, faire des annotations ou utiliser un outil de mesure lui permettant de vérifier une longueur, par exemple ;
– 2 items sont mieux réussis sur le support numérique. Ces items comportent des textes courts. La réponse portée est facilitée par un bouton radio (la réponse ne peut être vraie et fausse à la fois). Les graphiques semblent mieux mis en valeur sur ce support, mais les performances dépendent de la tâche demandée : lecture directe du graphique, sans de multiples prises d'indices ;
– 29 items ne présentant pas d'écart de réussite en fonction du support. Il est difficile de dégager des caractéristiques saillantes communes à ces items. Cependant, il nous a semblé pertinent de les regrouper selon quatre catégories liées au niveau de lecture ainsi qu'au graphisme :
T1 – item dont les consignes et/ou les propositions impliquent une lecture directe sans difficulté majeure ▸ **Figure 6** ;
T2 – item dont les consignes et/ou les propositions impliquent une lecture fine et attentive. La réponse n'est pas directe, il y a une nécessaire appropriation du message avant de porter la réponse ▸ **Figure 7** ;
T3 – item constitué d'un graphisme (ce peut-être une courbe, un organigramme, un tracé géométrique, etc.). La tâche de l'élève consiste à observer ce graphisme et d'en déduire directement la réponse ▸ **Figure 8** ;

Figure 6 Item dont les consignes et/ou les propositions impliquent une lecture directe sans difficulté majeure.

Pour chaque nombre de cette liste, indique s'il est multiple de deux :

	Vrai	Faux
5	☐ 1	☐ 2
14	☐ 1	☐ 2
25	☐ 1	☐ 2
40	☐ 1	☐ 2
33	☐ 1	☐ 2
124	☐ 1	☐ 2
250	☐ 1	☐ 2

Figure 7 Item dont les consignes et/ou les propositions impliquent une lecture fine et attentive. La réponse n'est pas directe, il y a une nécessaire appropriation du message avant de porter la réponse.

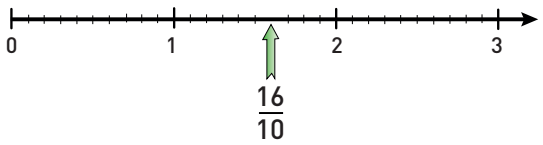
		Vrai	Faux
1	Le quotient est le résultat d'une addition	1 ☐	2 ☐
2	Le produit est le résultat d'une division	1 ☐	2 ☐
3	Le quotient est le résultat d'une division	1 ☐	2 ☐
4	Le produit est le résultat d'une multiplication	1 ☐	2 ☐

T4 – item constitué d'un graphisme. La tâche de l'élève consiste à relever plusieurs éléments avant de pouvoir porter sa réponse **Figure 9**.

Cette typologie montre que :
– les items proposant une lecture directe (présentés sous formes de losanges) sont systématiquement mieux réussis sur un support numérique **Figure 10**. Pour décrire la tâche demandée à l'élève, nous pourrions dire : l'élève est focalisé sur la lecture à l'écran, il clique avec sa souris, il n'a pas besoin d'étape intermédiaire[2] ;

2. Étape intermédiaire : l'élève doit utiliser un instrument, ou il doit prendre des notes, ou il doit mémoriser des éléments (faire des inférences).

Figure 8 **Item constitué d'un graphisme. La tâche de l'élève consiste à observer ce graphisme et d'en déduire directement la réponse**

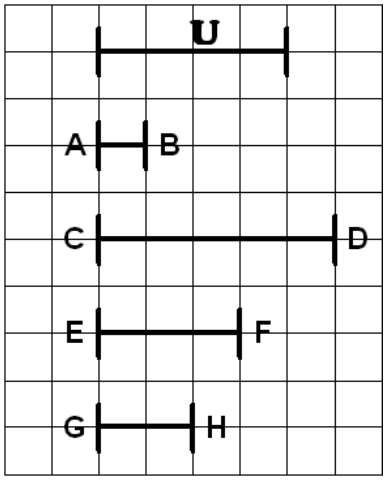


1 ☐ Oui

2 ☐ Non

Figure 9 **Item constitué d'un graphisme. La tâche de l'élève consiste à relever plusieurs éléments avant de pouvoir porter sa réponse**

Observe les segments.



Observation 1

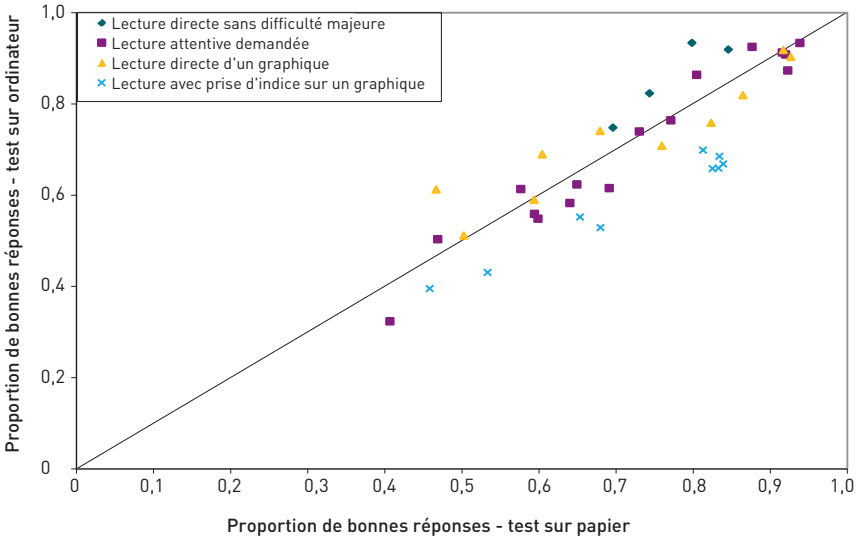
La longueur du segment [AB] est :

1	☐	$\frac{1}{2}U$
2	☐	$\frac{4}{1}U$
3	☐	$\frac{1}{8}U$
4	☐	$\frac{1}{4}U$

– à l'opposé, les items proposant une lecture avec prise d'indice sur un graphique (croix bleue) sont systématiquement mieux réussis sur un support papier. La tâche demandée à l'élève implique une étape intermédiaire ;
– entre ces deux pôles, nous trouvons les items réclamant une lecture attentive (carrés) et la lecture directe d'un graphique (triangles). Ces items se répartissent de part et d'autre de la droite indiquant une réussite identique entre les deux supports.

Il est sans doute difficile d'être catégorique après une étude ne reposant que sur 39 items. Néanmoins, nous pouvons dégager des pistes explicatives et répondre partiellement aux questions initialement posées. Concernant l'écart de difficulté entre les deux supports, il n'est pas possible d'y répondre simplement en faveur

Figure 10 Comparaison des niveaux de difficulté selon le type d'item



Note de lecture : chaque point correspond à un item.
L'axe des abscisses représente le taux de réussite des items en version papier-crayon, l'axe des ordonnées le taux de réussite des mêmes items dans leur version numérique. La droite qui partage ce graphique indique un niveau de réussite identique sur les deux supports. Un éloignement de cette droite correspond soit à une réussite plus grande sur le papier (au-dessous de la droite) soit à une réussite plus grande sur le support numérique (au-dessus de la droite).

de l'un ou de l'autre médium. En revanche, il est possible de caractériser les items en fonction de leur difficulté sur l'un ou l'autre de ces supports.
Pour rappel, la difficulté mesurée de l'item est la résultante de l'énoncé, de la connaissance mathématique mise en jeu et de la tâche à accomplir. Lorsque l'on propose des items qui réclament une lecture longue, de multiples inférences ou présentant une structure syntaxique riche, ils sont mieux réussis sur le support papier. Lorsque l'on propose un graphisme en tant que document, dans la mesure où la consigne concerne une prise d'indice directe, le résultat sur support numérique est meilleur. En d'autres termes, si l'élève peut agir directement et rapidement sans recourir à des outils ou à des étapes intermédiaires, le support numérique lui est favorable ; à l'opposé, s'il doit utiliser un instrument de mesure ou procéder à un brouillon, c'est le support papier qui est le plus efficace.

Seconde partie : Cedre collège

Lors de l'expérimentation de 2013, les élèves de troisième échantillonnés ont été évalués à la fois sur support papier et sur support numérique. L'échantillon comptait un peu plus de 5 000 élèves de troisième générale répartis dans 199 classes. Les réponses de 4 176 élèves répartis dans 194 collèges ont été analysées pour la partie papier ainsi que celles de 3 204 élèves répartis dans 148 classes pour la partie numérique.
Le matériel d'évaluation était constitué de huit cahiers pour les items en version papier-crayon et de huit modules pour les items en version numérique. Parmi l'ensemble des items expérimentés, cent trente-neuf items ont été évalués sur les

deux supports (papier et numérique). Chaque élève se voyait attribuer de manière aléatoire un des huit cahiers et un des huit modules. La correspondance entre cahiers et modules a été élaborée de manière à ce qu'un élève ayant répondu à un item sur le support «papier» ne retrouve pas le même item sur le support « numérique ».

Analyse globale

En fin de collège, l'analyse a porté sur 109 items dont les qualités psychométriques étaient satisfaisantes. Pour chaque item, le taux de réussite sur les deux supports a été calculé. La **figure 11** montre de manière évidente que les items sont nettement mieux réussis sur support « papier » que sur support « numérique ».

L'analyse détaillée des résultats par items tend à montrer que :

- l'item est moins bien réussi sur support numérique lorsque le type de tâche demandé relève d'une ou plusieurs des catégories suivantes :
  – il induit un raisonnement à plus d'une étape ;
  – il nécessite le recours à une schématisation de la situation ;
  – il nécessite le recours à des instruments de mesure ;
  – il nécessite des calculs intermédiaires ;
- l'item a des taux de réussite équivalents sur les deux supports lorsque le type de tâche demandé relève du calcul automatisé. Ce constat ouvre la voie à la création d'items d'activités mentales dans le cadre d'une éventuelle évaluation adaptative afin d'utiliser la plus-value apportée par l'environnement numérique ;
- l'item est mieux réussi sur support numérique lorsque le type de tâche demandé relève d'une méthode d'apprentissage.

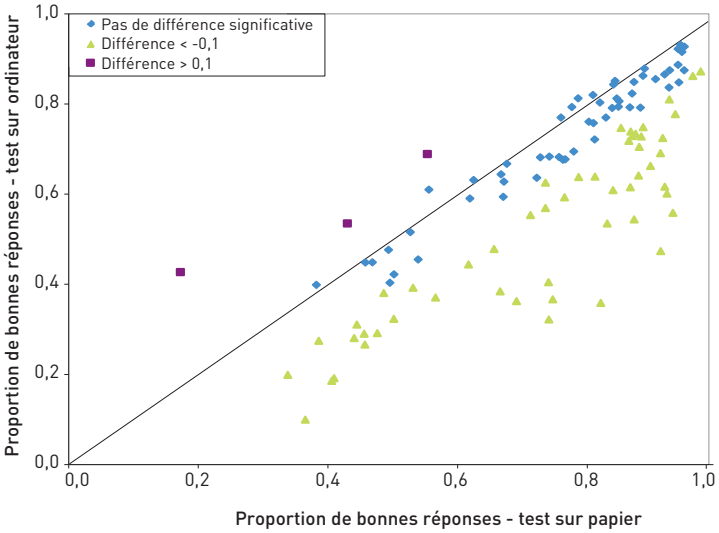
Nous illustrons des premiers constats en présentant un exemple d'item, proposé dans l'évaluation, par type de tâche ▸ **Figure 12**.

Sur support papier, l'élève a la possibilité d'exclure certaines figures lors de son raisonnement en les barrant par exemple, tandis que sur support « numérique », il est contraint de garder en mémoire l'ensemble des étapes de sa procédure de résolution. Ainsi pour cet item la différence de réussite est très importante : 74 % sur support papier et 32 % sur support numérique.

Dans l'exemple suivant, le constat est encore plus marqué ▸ **Figure 13**. Sur support « papier », l'élève peut colorier, barrer, cocher. Le support « numérique » contraint l'élève à repérer la case B3 pour chaque nouvelle question. Il amorce son raisonnement à partir des questions posées et teste les solutions proposées. Pour cet item également la différence de réussite est très importante : 91 % sur support papier et 47 % sur support numérique.

Sur support « papier », il est possible de faire une représentation imagée du problème (représenter les sacs puis chaque bille dans le sac par un point, etc.) ▸ **Figure 14**. Sur support « numérique », le raisonnement reste mental. On observe des différences de taux de réussite importantes entre les supports : 87 % sur papier et 54 % sur numérique.

Figure 11 Comparaison des taux de réussite selon le support (papier/ordinateur) et selon le type d'item



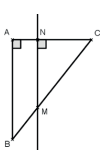
Note de lecture : chaque point correspond à un item. L'axe des abscisses présente le taux de réussite sur le support « papier », l'axe des ordonnées, le taux de réussite sur le support « numérique ». La diagonale tracée indique un taux identique sur les deux supports. Plus un point (un item) se rapproche de cette droite, plus les taux sont proches sur les deux supports. À l'inverse, un éloignement correspond à un taux plus important sur support « numérique » (au-dessus de la droite) ou à un taux plus important sur support « papier » (au-dessous de la droite).

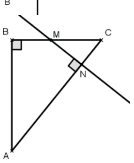
Figure 12 Item dont le type de tâche induit un raisonnement à plus d'une étape

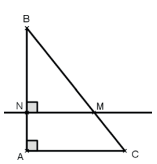
On donne le programme de construction suivant :

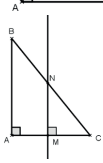
Tracer un triangle ABC rectangle en A.
Placer un point M sur le segment [BC]
Tracer la perpendiculaire à la droite (AC) passant par M.
Noter N son point d'intersection avec le segment [AC].

Parmi les constructions suivantes, cocher celle qui correspond à l'énoncé ci-dessus.

1 ☐ 

2 ☐ 

3 ☐ 

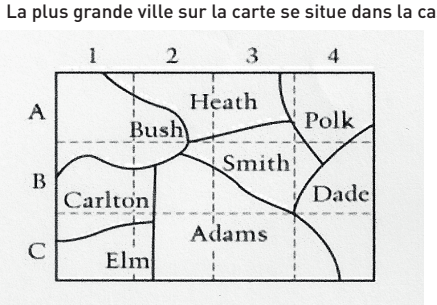
4 ☐ 

Certaines questions supposent une rotation de la figure (d1 et d5 par exemple) afin de trouver une réponse ▸ **Figure 15**. Ceci est impossible sur support numérique. Il est également difficile de poser une équerre sur l'écran alors que les outils de géométrie sont autorisés. Il n'est pas non plus possible de prolonger les droites dans le cadre de la recherche du parallélisme. Ce type d'items au contenu en phase avec les programmes officiels ne peut être testé sans avoir recours à un logiciel de géométrie dynamique. Pour cet item, le taux de réussite sur support papier est de 93 % et sur support numérique de 56 %. Sur support papier, les élèves prennent un brouillon pour effectuer les calculs intermédiaires. Sur support numérique, les élèves sont moins enclins à utiliser le brouillon de par l'environnement numérique et une certaine partie de l'activité de l'élève consiste à manipuler la souris. L'effort de mémorisation est donc plus important et source d'erreur. Pour l'item de la **figure 16**, le taux de réussite sur support papier est de 94 % et sur support numérique de 78 %. Le type de tâche de la **figure 17** correspondant à l'étude de la représentation graphique d'une fonction est régulièrement étudié en classe en utilisant le support numérique. Les élèves le retrouvent donc dans son environnement habituel. Les taux de réussite sont de 43 % sur support numérique et de 17 % sur support papier.

Bien que les items soient moins bien réussis sur support numérique, il ne faut pas en déduire la nécessité d'évaluer uniquement sur support papier. L'évaluation sur support numérique en mathématiques serait pertinente à condition de prendre en compte l'environnement numérique dans la construction des items, de mettre à disposition des élèves des outils (tableur, logiciel de géométrie dynamique, grapheur, etc.) pour résoudre des problèmes selon les méthodes préconisées par les programmes officiels et de les utiliser dans le contexte habituel de la classe.

▸ **Figure 13** **Second item dont le type de tâche induit un raisonnement à plus d'une étape**

Sur la carte ci-dessous sont indiquées 8 régions.
La plus grande ville sur la carte se situe dans la case B3.



D'après la carte, dans quelle région peut se trouver cette ville ?
Cocher la bonne réponse.

1 ☐ Adams ou Carlton
2 ☐ Adams ou Smith
3 ☐ Carlton ou Elm
4 ☐ Dade ou Polk
5 ☐ Polk ou Smith

Figure 14 **Item dont le type de tâche nécessite le recours à une schématisation de la situation**

On dispose de trois sacs de tailles différentes :

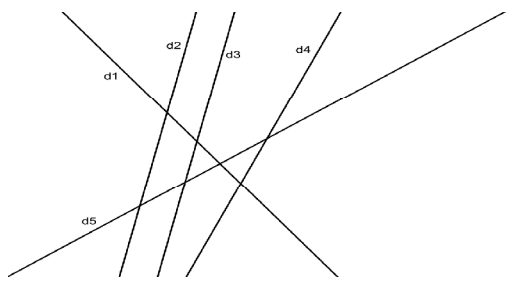
- Le plus petit sac contient 5 billes,
- Le sac de taille moyenne contient 50 billes,
- Le grand sac contient 500 billes.

Dans chaque sac, il n'y a qu'une seule bille noire. Sans regarder et au hasard, on prend une bille de chacun des sacs.
Quel sac doit-on choisir pour avoir le plus de chance de tirer une bille noire ?

1 ☐ **Le sac contenant 5 billes.**
2 ☐ **Le sac contenant 50 billes.**
3 ☐ **Le sac contenant 500 billes.**
4 ☐ **Il n'y a aucune différence.**

Figure 15 **Item dont le type de tâche nécessite le recours à des instruments de mesure**

On donne la figure suivante :



Pour chaque ligne du tableau, cocher la bonne réponse.

		Parallèles	Sécantes mais non perpendiculaires	Perpendiculaires
1	d1 et d2 semblent	☐ 1	☐ 2	☐ 3
2	d2 et d3 semblent	☐ 1	☐ 2	☐ 3
3	d4 et d5 semblent	☐ 1	☐ 2	☐ 3
4	d3 et d4 semblent	☐ 1	☐ 2	☐ 3

Conclusion de l'expérience 2

À l'issue de cette étude, les items semblent se répartir selon trois grandes catégories :
– les items « dématérialisables » ;
– les items spécifiques au support papier ;
– les items spécifiques au support numérique.

Lorsqu'un item permet une réponse directe, les taux de réussite constatés sont identiques sur les deux supports.

Les items spécifiques sur support papier sont ceux qui font appel à un raisonnement « papier-crayon » avec les outils associés (tracés géométriques, mesures, schématisation d'une situation, etc.).
Les items spécifiques sur support numérique sont ceux qui mettent en œuvre les fonctionnalités des outils numériques (géométrie dynamique, calculs avec tableur, utilisation d'un grapheur, etc.).

Les items spécifiques à chacun des deux supports peuvent être complémentaires pour un même objet d'étude. Par exemple, construire une médiatrice à la règle et au compas sur support papier ne mobilise pas les mêmes connaissances que construire la même médiatrice à l'aide d'un logiciel de géométrie dynamique.

Figure 16 **Item dont le type de tâche nécessite des calculs intermédiaires**

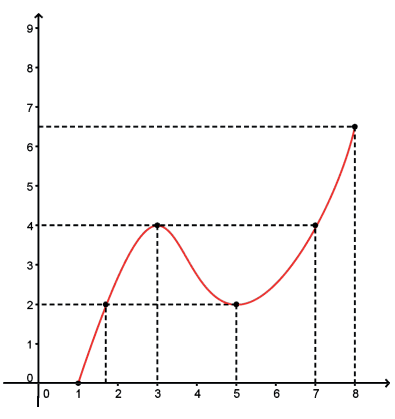
Manon pense à un nombre, elle le double, puis ajoute 10. Elle trouve 60.

Le nombre auquel Manon a pensé est...

1 ☐ 20
2 ☐ 25
3 ☐ 35
4 ☐ 140

Figure 17 **Item dont le type de tâche demandé relève d'une méthode d'apprentissage**

On a représenté ci-dessous la courbe représentative d'une fonction *f* définie pour tous les nombres compris entre 1 et 8.



	Vrai	Faux
1 a pour image 0 par la fonction *f*.	☐ 1	☐ 2
7 est un antécédent de 4 par la fonction *f*.	☐ 1	☐ 2
3 est un antécédent de 4 par la fonction *f*.	☐ 1	☐ 2
$f(3) = 4$	☐ 1	☐ 2
$f(2) = 5$	☐ 1	☐ 2

CONCLUSION

Cette expérience permet de mettre en évidence que la transition entre support « papier » et support « numérique » n'est pas sans conséquence.
Trois variables influent particulièrement sur la réussite aux items :
– la structure de l'item (la longueur des textes proposés, le nombre de documents, le type de documents, la mise en page et l'ergonomie intrinsèque) ;
– le type de tâches mises en jeu (raisonnement nécessitant des étapes intermédiaires et capacité à « naviguer » dans le support numérique) ;
– les contraintes liées à la spécificité du support (utilisation d'outils différents : le brouillon, le tableur, le grapheur, etc.).
Dans l'optique d'une évaluation des acquis des élèves, il est nécessaire de prendre en compte les critères mis en évidence dans cette étude.

BIBLIOGRAPHIE

Bessonneau P., 2012, *Évaluation de la compréhension de l'écrit sur support informatique et comparaison avec des épreuves de type papier-crayon*, 24e colloque de l'Admée Europe, Luxembourg.

Bunch M. B., Cizek G. J., 2007, *Standard Settings,* Londres, Sage Publications, 352 p.

Colmant M., Daussin J.-M., Bessonneau P., 2011, « Compréhension de l'écrit en fin d'école, Évolution de 2003 à 2009 », *Note d'information,* n° 11.16, MENJVA-DEPP.

Commission européenne, 2012, *First European Survey on Language Competences – Technical Report.*

Dierendonck C., Loarer E., Rey B., 2014, *L'évaluation des compétences en milieu scolaire et en milieu professionnel*, Bruxelles, De Boeck, 359 p.

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

Laveault D., Grégoire J., 2002, *Introduction aux théories des tests en psychologie et en sciences de l'éducation*, Bruxelles, De Boeck, 336 p.

OCDE, 2011, *Résultats du PISA 2009 : Élèves en ligne – Technologies numériques et performance*, vol. 6, PISA, Éditions OCDE.

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

Sautory O., 1993, *La macro CALMAR – Redressement d'un échantillon par calage sur marges*, Paris, Insee.

Wang H., Shin C. D., 2009, "Computer-Based and Paper-Pencil Test Comparability Studies", *Test, Measurement and Research Services Bulletin*, No. 9, Pearson Education.