From paper-pencil to tablet-based assessment: a comparative study at the end of primary school

Nadine Grapin, Nathalie Sayac

From paper-pencil to tablet-based assessment: a comparative study at the end of primary school

Nadine Grapin[1] and Nathalie Sayac[2]

[1] Université Paris-Est-Créteil, Laboratoire de didactique André Revuz, Paris, France;
nadine.grapin@u-pec.fr

[2]Université de Rouen Normandie, Laboratoire de didactique André Revuz, Paris, France;
nathalie.sayac@univ-rouen.fr

Assessments are increasingly being designed on a digital artefact (computer or tablet), while in France the equipment rate in primary schools is still low. Some computer-based assessment tasks use specific software functionalities (the dynamic aspect of a geometric figure for example), while others "migrate" from paper and pencil (PP) to digital artefact without using specific functionalities of dedicated software. In this research, we are interested in the validity of assessment tasks designed on tablet (especially when students do not usually use a tablet in the classroom and/or in assessment situations) and in the effects of migration (from PP to tablet) on student performance and procedures from one medium to another.

Keywords: Assessment, validity, comparative analysis, tablet-based assessment.

Introduction

Most standardized large-scale assessments like PISA or TIMSS have ever migrated (or are going to migrate) from paper-pencil (PP) to digital artefact (DA), like computer or tablet. Several reasons could be quoted for justifying such evolution and designing digital tests like increasing efficiency and consequently reducing costs or giving instant feedbacks, especially in the case of formative assessments (Threlfall, 2007). In France, since 2019, at the end of elementary school (Grade 5), a representative sample of students takes with tablets (and not with paper-pencil, as they did before), a national test, which aims to assess mathematics skills and knowledge and to observe their evolution at six-year intervals.

All French elementary schools have been equipped with computers and Internet access, but students haven't regular access to these digital environments: on average, there are 7.8 students per computer at elementary school (no data is produced about tablets) (Ministère de l'Education Nationale, 2018). These observations made us wonder about the validity of computer or tablet-based tests: if students are not used to doing mathematics with such artefacts, we can ask if the test assess really what it has to assess (mathematic skills and knowledge) or does it assess other competencies, like digital skills. More specially, is the test itself valid (especially when students do not usually use a tablet in the classroom and/or in assessment situations)? is longitudinal comparison performance (in the case of large-scale assessments) relevant? are mathematical processes to solve tasks the same when a similar problem is on PP or DA?

Previous research and theoretical background

In their review of assessment in mathematics education, Nortvedt and Buchholtz (2018) point to the development of digital assessments with parallel advances in psychometric models and the

development of adaptive assessments. At the same time, they also highlight the limitations of such assessments, particularly because they often only assess knowledge on simple tasks (and do not allow for the assessment of students' problem-solving skills). The authors conclude that "technology can also limit what is assessed" (p. 560). In general, the design of digital assessments raises validity issues, and we will explain how we address them. We do not review in this text all the research about digital assessments, but we limit our subject to the comparison between paper-pencil and digital-based assessments.

Comparative research between paper-pencil-based assessments and digital ones

Numerous studies, especially the Anglo-Saxon ones, aim to compare success scores between PP and DA were carried out at different school levels and on various disciplines. First, we can observe that most of these studies are based on data analysis without considering students' procedures (see, for example, Hamhuis & al., 2020). Second, their results are divergent, and they do not allow us to conclude, on the fact that a medium (DA or PP) would promote or not student success (Lemmo & Mariotti, 2017). For example, in France, from the results of two large-scale assessments, Bessoneau, Arzoumanian, and Pastor (2015) identified two variables that particularly influence the success in a mathematical item depending on the medium used: the structure of the item (length of texts, number of documents, etc.) and the type of tasks proposed. In particular, the items presenting a syntactic and linguistic complexity relative to the statement are better succeeded on PP whereas, in the case of an item requesting a direct taking of information (in a table or on a graph), success is better on a DA. In addition, problems requiring several steps of resolution are more successful on PP.

About such performance comparisons, we share Lemmo and Mariotti's (2017) point of view: "task comparability cannot be measured only in terms of students' outcomes, but it is also established by the comparison between the solving strategies that they use" (p. 3541).

About validity and legitimacy, Threlfall et al. (2007) explore in their research several aspects of "what may be lost and gained by undertaking mathematics assessment on the computer" (p. 336) and their conclusion, based on student's performance is :

> It is not only that translating paper and pencil items into the computer format sometimes undermines their validity as assessments, it is also that some paper and pencil items are less valid as assessments than their computer equivalents would be. (p. 335)

Let us now explain how we study the validity of a test from a didactic point of view while considering the modality of assessing (PP or DA).

Validity and instrumental genesis

For studying the validity of assessment tasks, we have developed in previous research a methodology with two complementary approaches (Grapin, 2016): one epistemological and didactical and another psycho-didactical. The first approach provides us with evidence of validity based on the a priori analysis: for each task, we list, among other things, the different solving procedures, the possible errors, and their origins, but also the complexity of the tasks (Sayac & Grapin, 2015). This analysis allows us to study whether the solving of the task mobilizes the mathematical knowledge that we want it to assess. The psycho-didactical approach is focused on student activity, i.e., that they develop

when carrying out the task; in our case, this includes their mathematical activity (their solving strategies), but it takes also into account the process of instrumental genesis (Rabardel, 1995). Moreover, "instrumental genesis is not the same for all students; it depends on their relationships with both mathematics and computer technologies" (Defouad 2000, as cited in Trouche (2005)). We hypothesize that students who use these artefacts routinely in the private sphere will have easier use of them in the school sphere; we also assume that students who have been able to use these artefacts in classroom situations will have a different instrumental genesis than others. So, to study the psycho-didactic validity of assessment tasks with a DA, it is therefore essential to determine students' use of the artefact (whether at school or home) and to observe how they solve the task with this artefact.

For this comparative research, we also have to determine how and when two tasks could be considered equivalent. Riplay (2009) distinguishes migratory and transformative approaches to switch from PP assessment media to a digital one. The migratory approach consists of the transition of PP task to DA without any modification; in the transformative approach, the original PP-based tests are transformed with the integration of specific functionalities of the artefact. The migration of tasks also requires considering the functionalities of the software or the application used for the test.

Research goals

We have chosen to focus our research on the comparison between PP and tablet-based assessment because one of the national large-scale assessments in France at the end of primary school migrated from PP to tablet in 2019. This raised questions about the comparability of results from one year to the next, but above all for us, several questions about validity. In this paper, we focus on two aspects of our research: the way we designed the test and analyzed its validity (1) and the analysis of the first results considered instrumental genesis (2).

(1) We have designed the test to compare the students' strategies in the case of migratory and transformative approaches: even if only some knowledge related to general tablet functionalities (virtual keyboard, drag and drop, virtual eraser in the draft) are needed to write or to provide the answer, we have considered that there was a change to the task (we will give examples in the following sessions). Under these conditions, is, a priori, the student's mathematical activity the same when solving a problem on a tablet or with PP? what are the differences in terms of mathematical strategies? And finally, are the same knowledge assessed?

(2) Since French schools are still poorly equipped digitally, we assume that students who regularly use tablets at home may have more advanced instrumental knowledge than others; since the time required must be short in an assessment's context, it is therefore possible that tests on a DA generate academic inequalities linked to the artefact itself. Is that so? How can the use (regular or not) of tablets impact students' scores and procedures? What are the implications for task validity?

The first question will be principally dealt with in the following part (methodology and design of the tests), and the second, in the part dedicated to results.

Methodology and design of the tests

To study more specifically the students' activity in a tablet assessment situation and to be able to compare it with that on PP, we conducted in June 2021 a study at the end of Grade 5 with 80 French

pupils from priority education schools and from “ordinary” schools (choosing two types of schools will enable us to observe whether inequalities are generated by the artefact itself).

Administration of tests and survey

All the students took the same two tests (PP and tablet). Each test consists of solving 23 tasks whose required knowledge of whole numbers and decimals, arithmetic, and problem-solving. The paper-pencil-based test took place during a regular classroom math session. The tablet-based test took place after PP one with the two researchers in the classroom. We observed how students were using the tablet and the difficulties they might encounter, depending on the tablet’s functionalities involved in certain items. During this observation phase, we focus on the student's instrumental genesis and study whether it interferes with the student's mathematical activity. We don't ask the students individually about their procedures because the a priori analysis of the tasks enables us to infer their strategies from the proposed answers.

Moreover, we ask students if they have a tablet at home and how often they use it (at home and in the classroom). The answers to this short questionnaire will also make it possible to judge the validity of the tasks and to ensure that they do not generate inequalities between students (especially those who have a tablet at home and those who do not).

Design of both tests

The choice of the didactic variables’ values made it possible to design mathematical equivalent tasks; we describe, in this paragraph, with examples, how we have designed such tasks in migratory and transformative approaches.

First, we have chosen, as equivalent tasks in a migratory approach, multiple-choice questions, as the example below (Table 1).

PP task	**Tablet task**
What is the result of 45,83 × 10 ? *Tick the correct answer.* □ 450,830 □ 450,83 □ 45,830 □ 458,3	Quel est le résultat de 23,85 x 10? *Coche la bonne réponse.* ☐ 230,850 ☐ 230,85 ☐ 23,850 ☐ 238,5

Table 1: Example of a QCM in PP and tablet environments

We can observe, in this first example, that the two tasks (PP and tablet) involve the same mathematical knowledge with the same level of complexity; the wrong answers correspond to the same type of errors both on the PP task (PPT) and on the tablet task (TT).

When the transition of an item from the PP to tablet involved the use of the virtual keyboard or drag and drop, for example, we considered these modifications to be in a transformative approach. Nevertheless, we were careful to design mathematically similar tasks by choosing appropriate values

for didactic variables, but let us explain, using the following example, how this type of transition can impact student answers. For assessing knowledge about writing numbers and units, we design the two following tasks (Table 2):

<table>
<tr>
<th>PP task</th>
<th>Tablet task</th>
</tr>
<tr>
<td>Connect each number on the left with its equal number on the right.

372 centièmes — 0,372
372 dixièmes — 3 720
372 unités — 37,2
372 dizaines — 3,72
372 millièmes — 372</td>
<td>Move the left or right label so that the number on the left is equal to the number on the right

Fais correspondre chacun des nombres de gauche avec celui qui lui est égal à droite.

483 centièmes — 0,483
483 dixièmes — 4830
483 unités — 48,3
483 dizaines — 4,83
483 millièmes — 483</td>
</tr>
</table>

Table 2: Example of a task using tablet functionalities

In this example, in the case of the TT, students can forget numbers or make mistakes but he or they cannot rely on one label on the right with two others on the left (unlike with the PPT). The treatment of the answers, especially in this case, is also simplified with the tablet.

We also want to study how students solve arithmetic problems, especially to observe how they use a draft on a tablet: on PP, they can easily make a diagram, write the operation, and use paper as a draft. With the tablet, we had provided a draft zone, but students need to understand pictograms (Fig. 1) for being able to draw, erase, organize their calculations.

Figure 1: Pictograms for using draft zone

Two types of problems were designed: a division problem (text of problem 1 - PPT: *"9 students of 6 years old must share 1,557 masks. How many masks will each student have?"*) and a number problem (text of problem 2 - PPT: *"Six 4-year-old students must share 6,000 sheets of paper. How many sheets of paper will each student have?"*). Students cannot use a calculator either in PP or with the tablet. For solving problem 1, they have to use a draft for calculating 1557:9 but for problem 2 they can mentally answer. With problem 1, we'd like to study how students use the draft of the tablet, and with problem 2, we'd like particularly to observe whether the tablet promotes mental calculation procedures.

Results

Results for all students

For all 80 students, the average success score in PP is 16 correct answers (out of 23 tasks) and 14.5 on the tablet. Now let's look at the difference in success question by question: we studied the number of correct answers per question, and we calculated the difference between the number of correct answers of PPT and the number of correct answers of TT (Figure 2). We observe for example that 4 students out of 80 did better on question 21 (QCM – the task is given in table 1) on PP than on a tablet.

On the whole test, only 2 tasks (Q1 and Q19) of 23 are better performed on a tablet than in PP, but only by 1 or 2 more students.

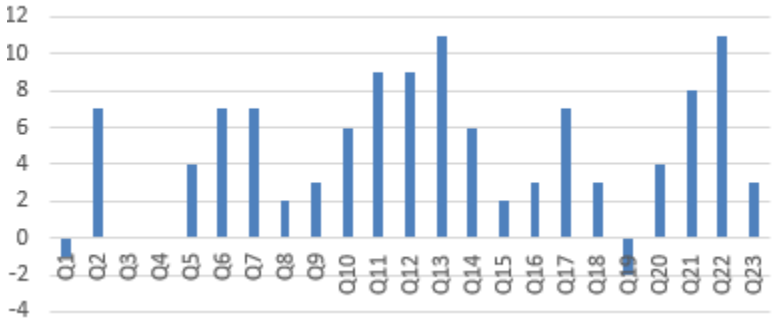

Figure 2: Difference between the success score in PP and on the tablet per question

It’s not surprising that the division problem (problem 1 – Q13, quoted before) is more successful on PP than on a tablet, but we must study exactly what are the errors and procedures in PP and on a tablet for better understanding this result. We have the same analysis to do with other questions, especially when the difference between performance on PP and with the tablet is important. We’ll then be able to determine the validity of such assessment tasks, whether PPT or TT.

Per student, we observe also that the difference in success score on the 23 questions between PP and tablet can vary between -8 and +14. 16 students have the same score in the two tests but 7 students have a score difference of 4 points (Figure 3).

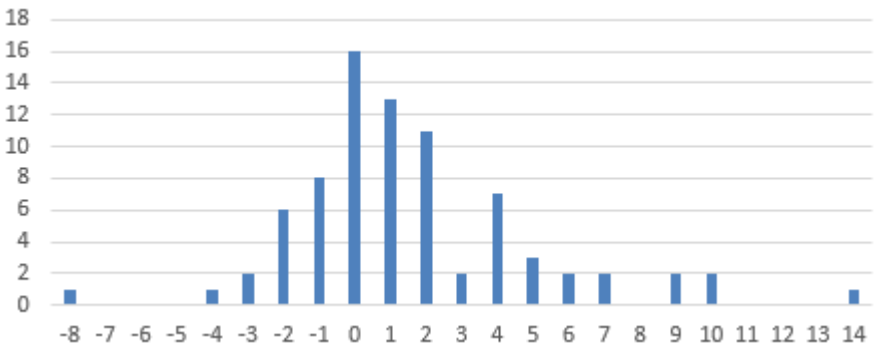

Figure 3: Number of students by the difference in success scores between PP and tablet

Results for students according to their use of the tablet

During the observation phase, we noticed that most students use the tablet by themselves, without help; only instructions on how to write the comma, use the "drag and drop" and erase in the draft have been given by the researchers. The answers of students confirm that 80% of the students regularly use a tablet at home (more than once a week) and only 17 students use a tablet less than once a month.

For these 17 students, we observe that the average success score both on PP (16,5) and tablet (14,5) is better than the average score of all students. We are currently further studying the responses and procedures of these students. We can observe for example, that, for problem 1 (division problem), which requires a little more advanced use of the tablet and the draft area, 4 students who had correctly solved the problem in PP were mistaken on the tablet (either because they did not answer anything, or because they made a mistake in the division or did not finish it). We cannot give general conclusions from this example, but during the presentation, we'll present the detailed results, and we'll try to show the relationship between the regularity of tablet use, the students' mathematical activity, and their performance.

Conclusion

To complete these first results, we will reproduce this experimentation on a larger sample of students, taking care to change the order of the two modalities (PP then tablet *vs* tablet then PP). We also wish to integrate tasks that use other functionalities of the tablet (such as the zoom to place a number on a graduated line or the integrated calculator to perform operations in problem-solving).

The question of the validity of the assessment tasks is raised on the DA, as on PP. DA offers several possibilities for designing new types of tests especially diagnostic and formative ones with automatic feedbacks (for example, Sirejacob & al., 2019), but if the designers of these tests do not consider the specificities of the DA, the validity of the test is not guaranteed: students' mathematical knowledge is not correctly assessed and all the feedbacks are not suitable. Our research aims to identify points of vigilance about the validity of tablet-based assessment and the comparison between PP and tablet performance and strategy.

This research also allows us to better study *a priori* the complexity of a task on a tablet (Sayac, 2018; Sayac and Grapin, 2015) by adding a specific dimension related to the instrumental genesis and the functionalities of the support involved in the resolution of the task.

References

Bessonneau, P., Arzoumanian, P., Pastor, J. M. (2015). Une évaluation sous forme numérique est-elle comparable à une évaluation de type « papier-crayon » ? *Éducation et formations*, *86-87*, 159-182. https://dx.doi.org/10.48464/ef-86-87-08

Grapin, N. (2016). *Validity of large-scale mathematics assessment: a didactical analysis.* 13th International Congress on Mathematical Education, Hamburg.

Hamhuis, E., Glas, C., Meelissen, M. (2020). Tablet assessment in primary education: Are there performance differences between TIMSS' paper-and-pencil test and tablet test

among Dutch grade-four students? *British Journal of Educational Technology, 51*(6), 2340-2358. https://doi.org/10.1111/bjet.12914

Lemmo, A., Mariotti, M-A. (2017, 1-5 February). *From paper and pencil to computer-based assessment: some issues raised in the comparison,* In. Dooley, T. & Gueudet, G. (Eds.), Proceedings of the Tenth Congress of the European Society for Research in Mathematics Education, Dublin, *Ireland: DCU Institute of Education & ERME,* 3540 – 3547.

Ministère de l'Education Nationale (2018). Repères et références statistiques. Enseignement formation recherche. [Benchmarks and Statistical References. Teaching Training Research.] https://cache.media.education.gouv.fr/file/RERS_2018/28/7/depp-2018-RERS-web_1075287.pdf

Nortvedt, G-A., Buchholtz, N-F. (2018) Assessment in mathematics education: issues regarding methodology, policy and equity. *Zentralblatt für Didaktik des Mathematik,* 50-4. http://dx.doi.org/10.1007/s11858-018-0963-z

Rabardel, P. (1995). *Les hommes et les technologies ; approche cognitive des instruments contemporains*. Armand Colin: Paris.

Ripley, M. (2009). Transformational computer-based testing. In F. Scheuermann & F. Björnsson (Eds.), *The transition to computer-based assessment,* p. 92–98. Luxembourg: Office for Official Publications of the European Communities. https://doi.org/10.2788/60083

Sayac, N. (2018). Assessment in mathematics: A French study based on a didactic approach, In D.R. Thompson, M. Burton, A. Cusi, D. Wright editors, *Classroom Assessment in Mathematics – Perspectives from Around the Globe (ICME 13 – Monograph).* Springer. https://doi.org/10.1007/978-3-319-73748-5

Sayac, N., Grapin, N. (2015). Analyse didactique d'une évaluation externe en mathématiques : quels outils pour quels enjeux ? *Recherches en didactique des mathématiques, 35(1),*101-126.

Sirejacob, S., Chenevotot-Quentin, F., Grugeon-Allys, B. (2019) PÉPITE Online automated assessment and student learning: The domain of equations in the 8th grade, In *Jankvist, U. T., Van den Heuvel-Panhuizen, M., & Veldhuis, M. (Eds.). Proceedings of the Eleventh Congress of the European Society for Research in Mathematics Education (CERME11, February 6 – 10, 2019). Utrecht, the Netherlands: Freudenthal Group & Freudenthal Institute, Utrecht University and ERME.*

Threlfall, J., Pool, P., Homer, M., Swinnerton, B. (2007). Implicit aspects of paper and pencil mathematics assessment come to light through the use of the computer. *Educational Studies in Mathematics,66-3,* 335 -348. http://www.jstor.org/stable/27822709

Trouche, L. (2005). Instrumental genesis, individual and social aspects, *In Dominique Guin, Kenneth Ruthven, Luc Trouche, The didactical challenge of symbolic calculators: turning a computational device into a mathematical instrument,* 197 – 230. Springer, Mathematics Education Library. https://doi.org/10.1007/0-387-23435-7_9