Digital Dashboards for Summative Assessment and Indicators Misinterpretation: A Case Study

Matthieu Cisel

Digital Dashboards for Summative Assessment and Indicators Misinterpretation: A Case Study

Matthieu Cisel
Institut Des Humanités Numériques, CY Cergy Paris Université

Abstract

Over the last decade, teachers in France have been increasingly pressured to use digital learning environments, and to shift from grade-based to skill-based assessment. Educational dashboards, which measure student input electronically, could foster such a transition by providing insights into learners' performances. However, such dashboards could also foster data misinterpretation during the summative assessment process, should the indicators that they display be used without a proper understanding of what they reflect. This article presents a methodology to detect potential mistakes in the interpretation of the indicators in the context of inquiry-based learning. During the design of a learning environment, we analyzed, through analytics and classroom observations in primary and middle schools, the issues that could arise from the use of a dashboard. Our data suggest that the amount of information practitioners needed to collect to make indicators relevant was burdensome, making the dashboard unfit for assessment purposes at the scale of a classroom.

Keywords: dashboard, learning analytics, skill evaluation, case study

Résumé

Au cours de la décennie écoulée, les enseignants de France ont été de plus en plus poussés, d'une part, à adopter l'évaluation par compétences, et d'autre part, à utiliser des applications numériques. Les tableaux de bord qui mesurent les actions des utilisateurs de ces applications peuvent faciliter cette transition en apportant des indications sur les performances des apprenants. Néanmoins, dans une perspective d'évaluation sommative, une question se pose quant à la capacité des enseignants à interpréter correctement les indicateurs mis à leur disposition. Cet article présente une étude de cas à visée méthodologique qui a pour objectif d'identifier de potentiels problèmes d'interprétation d'indicateurs lors de l'évaluation d'une démarche d'investigation. Durant la conception d'une application numérique, le *CNEC*, nous avons analysé, par le moyen de traces d'interaction et d'une étude de terrain au collège et à l'école primaire, les éléments susceptibles d'affecter l'utilisation d'un tableau de bord. Nous montrons que la quantité de données à collecter pour rendre pertinent l'usage des indicateurs rend leur utilisation compliquée dans un contexte d'évaluation sommative à l'échelle d'une classe entière.

Mots-clés : tableau de bord, traces d'interaction, évaluation par compétences, étude de cas

Declaration of Interest Statement

The authors do not have any known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Introduction

Since the early 2000s, skill-based assessment has gained momentum in the educational system in France. For elementary schools and middle schools, lawmakers and authorities have written in official texts the necessity of such a transition and developed digital tools to foster it (Bulletin Officiel de l'Education Nationale [BOEN], 2007, 2016). This evolution reflects the growing importance of competence-based approaches to evaluation, both at the European level (Eurydice, 2012) and at an international level (UNESCO, 2007). One of the recurrent criticisms is that it lacks grounding in empirical data that would reflect learners' performances in an objective manner, despite the development of various tools, such as e-portfolios, that could be instrumental from that point of view (McMullan et al., 2003).

Learning environments that collect pupils' written productions can provide various indicators through educational dashboards. Such indicators could be used as an empirical grounding for skill assessment. Field observations in French middle schools (Cisel & Baron, 2019) suggested that some teachers, when provided with a dashboard, already followed such an approach for summative assessment in mathematics. However, no comprehensive study has been carried out so far to determine how widespread this practice was. While the improvement of the objectivity of summative assessment represents a strong rationale for the development of dashboards, there is, to our knowledge, scarce research on their relevance and their accuracy in real-life settings. In this article, we address this issue with a case study in the field of inquiry-based learning (Abd-El-Khalick et al., 2004), through the experimentation of a learning environment, the *Cahier Numérique de l'Eleve-Chercheur* (*CNEC*).[1]

There are at least two elements that should be taken into account with regard to the relevance of the dashboard for skill-based summative assessment. First, the chosen indicators must reflect the intended constructs adequately (e.g., pupils' skills). Secondly, the learning environment should be able to register most of learners' actions that are relevant to build those indicators. It can prove challenging, since, as Verbert et al. (2014) stated, "Comprehensive tracking is difficult in more closed learning management sys-

1 The learning environment is available at this URL https://www.*cnec*.fr/accueil. Its name can be translated to *Student-Researcher Digital Notebook* (Cisel & Barbier, 2021b).

tems, as they typically cover only the tip of the iceberg." Indeed, learning environments are blind to what happens in the classroom outside of the platform. A second issue must also be considered. For instance, if the mobile device used to write a hypothesis changes hands and if the learners do not log out and log in, the *CNEC* will attribute productions from diverse authors to only one person. Key events such as this can hinder any effort to perform an accurate assessment of learners' skills, which led us to pose the following research questions.

In the case of the *CNEC*, how can we identify the discrepancies between the indicators that were designed and the skills they mean to reflect? How can teachers reduce such discrepancies to analyze summative assessment accurately? To what extent do such discrepancies threaten the possibility of using educational dashboards to implement assessment of students over long periods?

We will see that, to avoid skewing the indicators displayed in the *CNEC*'s dashboard, teachers would have needed to restrain themselves from interacting spontaneously with pupils in a way that makes summative assessment over long periods practically unfeasible. While our results are likely to be valid for other learning environments, we limit our generalization to the *CNEC*. Indeed, the answer to our questions largely depends upon the features of the environment at hand. However, the methodology that we followed, using both field observations and learning analytics, could inspire other learning environment designers.

A Prototype Developed Within a Consortium

We addressed the question of indicator misinterpretation in a set of partnering elementary and middle schools, and more precisely in the context of Student-Question-Based Inquiry, a type of project where pupils design research questions, hypotheses, and protocols (Herranen & Aksela, 2019). This approach offers various opportunities to evaluate students' skills (Xhakaj et al., 2016) through the *CNEC*'s dashboard. Indeed, several indicators can be derived from learners' writing activity during inquiry-based learning, including number of words for a given idea, number of submitted ideas (for instance, hypotheses), and the amount of feedback necessary to reach an acceptable hypothesis. These reflect students' ability to propose relevant and rigorous scientific statements, a key skill to be assessed in the French educational system (BOEN, 2016).

The *CNEC* was developed between 2016 and 2019 within a consortium, *Les Savanturiers du Numérique,* composed of *Tralalère* (the company that codes and owns the software), a research laboratory, two school districts that allowed us to have access to partnering teachers, and the *Savanturiers* program (Cisel & Barbier, 2021a). It supports iterative writing of scientific claims, such as research questions, hypotheses, and protocols (Cisel & Barbier, 2021b). *CNEC* is largely inspired by the *Knowledge Forum*, a learning environment for collaborative writing whose first prototype was released in the late 1980s (Scardamalia & Bereiter, 2006). As such, it relies on a variety of tools to scaffold the writing process of students' inquiry. Students are asked to work in groups to produce research questions, hypotheses, and protocols. Figure 1 shows a sample interface from the Research Notebook*,* a module of the *CNEC* that allows instructors to see which written production ought to be reviewed by the teacher or the students.

Figure 1

A Screenshot of an Interface of the Research Notebook

The dashboard is meant to foster the acquisition of scientific reasoning through a set of scaffolds (Quintana et al., 2005) and through iterative writing (Vardi, 2012; Zhao & Chan, 2014). In iterative writing, the teacher provides feedback on student work. Learners are then required to submit a new version until the instructor is satisfied. In Figure 1, we can see the *Fiche-Recherche* of the *CNEC*, a module that provides an overview of a given student's different written productions (question, hypothesis, protocol). Symbols are meant to show if the written production is considered satisfactory (smiling face), in need of revision (unhappy face), or has not yet been assessed (hourglass).

Research publications have focused on the conditions under which feedback can foster the acquisition of writing skills (Hattie & Timperley, 2007; Kluger & DeNisi, 1996). Learning environments have been used to help instructors design such feedback (Bywater et al., 2019), notably in a context of iterative writing (Scardamalia & Bereiter, 2006). From that perspective, we focused on the amount of feedback necessary for learners to reach a satisfactory result, an indicator displayed in the *CNEC*'s dashboard.

Educational Dashboards in the Scientific Literature

There is abundant literature on the opportunities offered by dashboards in educational contexts. They have been used to identify disengaged learners (Hu et al., 2014), to support engagement (Silvola et al., 2021), to track interactions between pupils, or to track progress that cannot be reliably observed by other means (Scheffel et al., 2017). Supporting teachers' inferences about pupils' misconceptions (Xhakaj et al., 2016) or inquiry skills (Specht et al., 2013), learning gains (Mottus et al., 2015; Wang & Han, 2021), and students' self-regulation (Aguilar et al., 2021; Han et al., 2021) also represent common topics covered in the literature. Dashboards were designed to address some of these assessment gaps.

From the methodological point of view, two research axes on dashboards are worth mentioning. The first one is focused on the rationale of the design process, and how they grounded the design of the dashboard on learning theories (Schwendimann et al., 2016; Sedrakyan et al., 2018). The second axis corresponds to the evaluation of dashboards (Iandoli et al., 2014) in various settings, from the point of view of its acceptability, its usability, or its utility (Nielsen, 1993). Within this axis, we can distinguish research work where evaluation relies on mock-ups (Ali et al., 2012; Scheffel et al., 2017; Ste-

phens-Martinez et al., 2014), from research work that involves experiments in real-life settings (Faber et al., 2017). While a part of our work relied on mock-ups (Cisel & Baron, 2019), the present article focuses solely on classroom experiments.

The question of the misinterpretation of dashboards' indicators has been addressed in the published literature. For instance, Phillips et al. (2011) described in a pilot study on *Lectopia*, a learning environment archiving lectures, the discrepancy between their original interpretation of students' behaviour in a distance education setting, and the actual meaning of their actions obtained through qualitative interviews. This research highlights the need to deepen our understanding, in more varied educational settings, of dashboards' shortcomings when practitioners use them to assess skills. We believe that this topic requires further investigation, and research work that features field observations. To our knowledge, this is one of the first articles that addressed the gap in our understanding in a classroom setting.

Methods

In the following paragraphs, we first present the *CNEC* and its indicators, the characteristics of partnering teachers, and the protocol that we followed to identify potential sources of misinterpretation during the assessment process.

The CNEC and Its Indicators

Activities featuring the CNEC. The indicators of this study were designed based on the learning analytics produced by various modules of the *CNEC,* which were intended for subsequent use. We will focus on only one of them, the Fiche-Recherche (*Research Notebook*) (Figure 1). It enables iterations between a student and the instructor for a given idea. A teacher can either validate this idea, or ask the pupil, or the group of pupils, for a new version until they are satisfied with the written production. The typology of ideas that a learner can write in this module corresponds to the classical steps of the inquiry circle (Pedaste et al., 2015), namely research question, hypothesis, protocol, data, and data interpretation. In the *CNEC*, ideas can be submitted either by a single learner or by a group of learners.

Rationale behind the choice of the indicators. We wanted to analyze a limited set of skills: "being able to design a sound scientific hypothesis," and "being able to design a research protocol," without thoroughly analyzing the written text. Indeed, automated content assessment would require complex machine learning techniques. Three indicators were chosen to be part of the dashboard after the focus groups. The first one is the number of ideas submitted by a student or a group of students. This indicator reflects learner engagement. The second one is the number of iterations required before the first validation of an idea by the instructor. This practice incentivizes students to improve the quality of their writing and it reflects their ability to understand instructions when designing a hypothesis or a protocol (i.e., the smaller the number of iterations, the higher their ability to understand instructions and to produce satisfying written productions). This practice is designed to discourage students from submitting a high number of low-quality ideas to appear more engaged. The third indicator is the number of words per idea, which is an indicator of depth and nuance in understanding.

The first logs associated to user tests were archived in June 2017, and the company kept tracking logs for two years. However, no dynamic dashboards were implemented in the *CNEC* at the time of the user tests since they were not used to monitoring the situation in real time. The visualizations (e.g., Figure 2) were produced afterwards and used in focus groups to discuss their pros and cons in a context of summative assessment. We display in Figure 2 a mock dashboard based on learning analytics, featuring the indicators that were selected as a focus for this study.

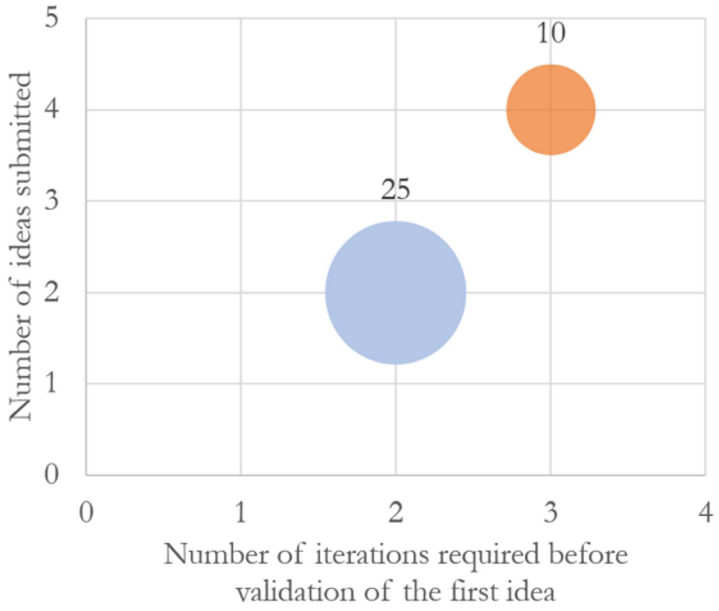
This dashboard allows users to visualize metrics for different students along three dimensions. In Figure 2, we can see, for instance, two students who submitted ideas in the hypothesis section of the *Research Notebook*. The first student (in blue) submitted two hypotheses, and it took two iterations for the first idea to be accepted. There are 25 words on average per idea. These indicators can help us determine to some extent how the teacher uses the *CNEC*. For instance, in the previous example, we know that the practitioner asked pupils to write and send a new iteration of their hypothesis and did not merely collect first drafts of the ideas through the learning environment.

We presented this mock dashboard to clarify how data interpretation can be impacted when these indicators poorly reflect what really happens in the classroom. If the teacher decided that the best students were the ones who reached an acceptable hypothesis with fewer iterations, then he or she would give a better grade to the blue

student. However, it is possible that the practitioner gave more oral feedback to the blue student and forgot about it. Since this feedback was not taken into account by the learning environment, the indicators wrongfully suggest that the orange student required more help to reach an acceptable hypothesis.

Figure 2

Mock Version of the Dashboard Designed for the Research Notebook (Hypothesis Section)



Observation Protocol

To assess how useful our dashboard was for assessment purpose, we carried out a series of observations in classrooms and organized a set of focus groups (Krueger, 2014). A group of teachers from three elementary schools (students aged six to 11) and four middle schools (students aged 11 to 15) partnered with the *Savanturiers du Numérique* consortium. They allowed researchers to observe their projects and carried out user tests of the first prototypes of the *CNEC,* after being extensively trained with regard to its features.

We contrasted observation protocols for two situations: sessions during which the *CNEC* was used, and sessions without the *CNEC*. The former will be labelled as *user tests*, while the latter will be labelled as *regular class sessions*.

Longitudinal follow-up of projects. We did a longitudinal follow-up of seven science projects during which the *CNEC* was used, in the 2017–2019 period (Table 1), in both Paris and Créteil (Paris suburbs) school districts. Observations included both user tests and regular class sessions, since teachers were free to use the learning environment as they saw fit. Some teachers engaged extensively in user tests of the *CNEC*, while others were followed mostly to capture the instructional design chosen by the teacher, and therefore to provide better understanding of how the *CNEC* could be used.

Savanturiers projects, during which instructors would make students propose research questions, hypotheses, and protocols, typically lasted from January to June and included 10 to 15 sessions (Cisel & Barbier, 2021a). We report in Table 1 the characteristics of the projects, like the level of teaching, the number of teachers per classroom and the average number of students present in the class. The exact names of the institutions and of the teachers were modified to ensure their anonymity.

At the time of this study, the *CNEC* was too prototypic to be used during an entire *Savanturiers* project. Technical issues would have disrupted classroom activity for an unreasonable amount of class sessions, and user tests involving the learning environment were therefore scattered across institutions. A given teacher would host up to five user tests, but most of them used the *Research Notebook* module only once or twice during the project (Table 1). While user tests enabled us to capture possible sources of indicator misinterpretation, regular class sessions allowed us to understand in which context the *CNEC*, and, therefore, the dashboard, could be used in a context of summative assessment.

Observation protocol for regular class sessions. To understand to what extent the *Research Notebook* module could be used in the classroom, notably in a context of assessment, we needed to determine what portion of a *Savanturiers* project was dedicated to iterative writing, both at the scale of the entire project, and at the scale of a classroom session. This approach enables us to assess the potential use of the technology, regardless of the existence of actual user tests.

Video recordings were not allowed by the schools that we partnered with; we therefore took notes in real time on the lesson progress, following the teacher, and keeping track of the different activities that would be organized during a class session. These sessions typically lasted between 50 minutes and 2 hours. In the reports, sessions were divided into sequences that were homogeneous from the point of view of the activity that was carried out. The timing of each sequence was tracked, and there were between five and 10 of them within a given session.

Table 1

A synthesis of partnering elementary and middle schools, with periods and number of observations per Savanturiers project

Name of the institution	Level	School District	# Teachers (# students)	# Observations (period)
Bouliers	E.S. (3rd grade)	Paris	2 (14 x 2)	8 (2017–2019)
Clignancourt	E.S. (4th grade)	Paris	1 (23)	8 (2018–2019)
Victor Dupont	M.S. (6th grade)	Créteil	2 (27)	4 (2018–2019)
André Girault	M.S. (7th grade)	Paris	3 (15 x 2)	9 (2016–2017)
Jean Sébastien	M.S. (7th grade)	Paris	2 (15 x 2)	3 (2018–2019)
Saint-Victor	M.S. (6th grade)	Paris	1 (28)	6 (2017–2018)
Peupliers	E.S. (5th grade)	Créteil	1 (23)	8 (2016–2017)

Note: E.S.: Elementary School. M.S.: Middle School

Two levels of coding were applied. First, each session was coded following a labelling system based on *Savanturiers* decomposition of an inquiry project. It corresponds to the different steps of an inquiry project as described by Pedaste et al. (2015): research question, hypothesis, experimental design, data collection, data interpretation. Additionally, it includes steps that are more specific to a class project. During the introduction phase, pupils are explained the context surrounding a *Savanturiers* project. During the communication phase, they present their group work in front of an audience.

A second level of coding was applied to classify each sequence within a given session. We used Nvivo 11 and based this step on a taxonomy of activities that had been validated by a group of researchers working on *Savanturiers* project. For the present contribution, we do not provide the full taxonomy, as we only need to distinguish between iterative writing-related activities, and other types of activities (reminding students what had been done in a previous class, summarizing what was done during the session, etc.). For each session, we tracked the amount of time that was used for iterative writing, compared to the total duration of the session.

Additional data collection during user tests. For the 19 user tests featuring the *CNEC*, we followed the same protocol as for regular class sessions, but we added an inci-

dent detection protocol that would allow us to detect potential sources of indicator misinterpretation in a context of summative assessment. An incident is defined as any event occurring within the classroom that may affect an indicator of the dashboard. Events like technical bugs or logistical issues (loss of wi-fi connection, change of the electronic device due to low battery, etc.) were all grouped into one category: "Technical issues." Even with a stable version of the *CNEC*, they are likely to occur in a classroom setting and to affect summative assessment.

Additionally, any event that would affect differentially students or groups of students in terms of writing performance was considered as a potential incident to register. For instance, teachers sometimes ignored the *CNEC* and iterated orally with students, and provided feedback on students' paper notebooks, when they could, and (given the fact that they were supposed to use indicators for summative assessment) should have done so in the learning environment. Such iterations could not be registered in the dashboard and were therefore considered as a source of data misinterpretation for the following indicator: "Number of iterations required before the first validation of an idea by the instructor." Indeed, most of the time, such events did not affect all students equally, which means that for each session, the teacher would have to take note of all such events, to correct the interpretation that they would make of the indicators.

All incidents involving the behaviour of the teacher were labelled as "Teacher-interaction-related incident." For each event, we noted how many learners had been affected by the incident. Learners worked in groups of three. An incident affecting the whole group would therefore affect the indicators of three students. Based on a chronological analysis of these events, we decided to select two user tests with a stark contrast in terms of number of issues likely to skew summative assessment. The first one, in Bouliers Elementary School, had the lowest number of incidents observed in a user test, while the second one, in Jean Sébastien Middle School, had the highest. For each incident, we counted how many pupils were impacted.

Learning analytics. During each user test, learning analytics were collected, but indicators were computed and displayed afterwards, in the laboratory. The goal of the study was to assess the relevance and robustness of the indicators that we had designed. User tests were used to support the reflections on the design of the dashboard, notably to identify events and teachers' decisions that could skew the indicators in the context of competency assessment. To illustrate the potential issues that derive from the use of

dashboards for summative assessment of students' skills, we used analytics from a user test in Bouliers Elementary School to show how teachers' actions would impact the indicators displayed in the dashboard. Once the coding phase was over, the quantitative analysis of learning analytics and user tests were carried out either with Excel or with R 4.0 (R Core Team, 2020).

Positionality of the Researchers

We were involved in the *CNEC* project from the design of its first specifications to its evaluation in the classrooms. Consistently with Baker's (2011) recommendations, we preferred to let teachers organize classroom activities in the way they saw fit. They were just required to use the modules of interest and to keep in mind the issue of skill-based assessment. Finally, it is important to note that the idea to use the dashboard in a context of summative assessment did not originate from the research team. We favoured the use of indicators for real-time classroom monitoring. Our study was originally designed as a demonstration of the flaws associated with indicators' use in a context of summative assessment and was later reported in a research article.

Results

In this section, we first described two user tests characterized by a stark contrast in terms of the number of incidents over the course of the session. We then focused on a specific incident that happened during a user test in which a teacher's behaviour skewed an indicator associated with the *Research Notebook*. We used learning analytics to show that without refraining themselves from certain interactions with learners, instructors hinder the accuracy of the depiction of learners' skills through the dashboard indicators. We finally proceeded to describe longitudinally a series of projects, encompassing both user tests and regular class sessions; our goal was to illustrate how often teachers would have to be careful when interacting with students, should they choose to use the *CNEC*'s dashboard as a tool for summative assessment.

Contrasting Two User Tests

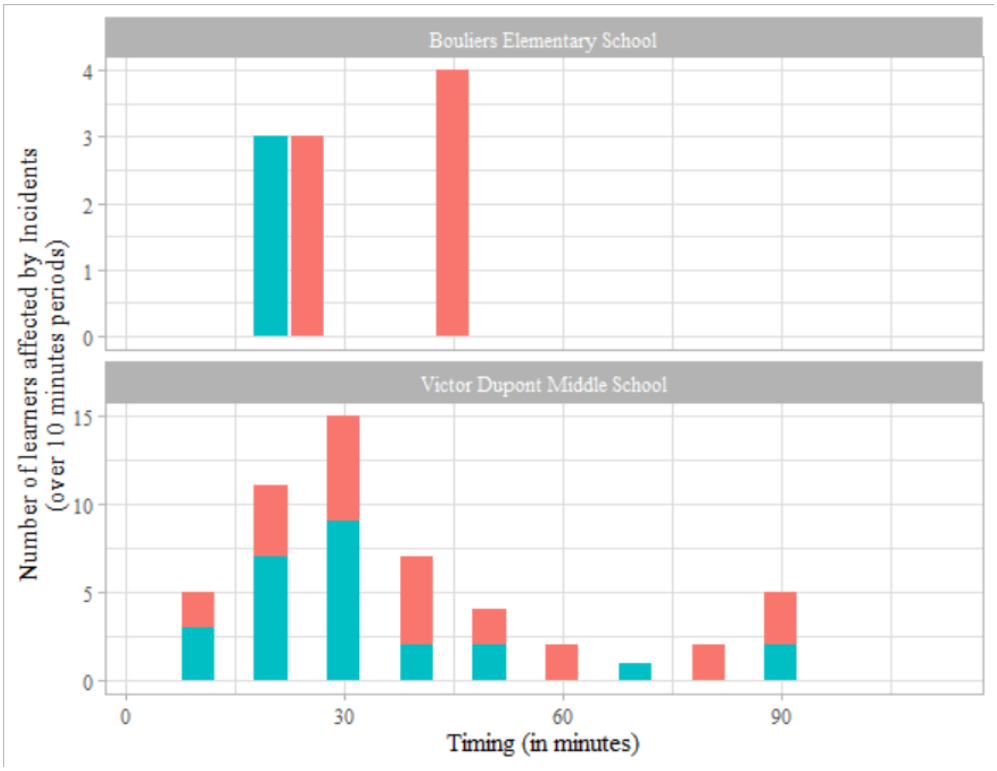
When evaluating inquiry skills through the InqITS learning environments, Gobert et al. (2013) were in perfect conditions to assess the ability of learners to design sound hypotheses. The only technical incidents that could have happened were linked to a hardware issue, since the software was working well. Teachers only had to ensure that learners understood the interfaces and refrain from interacting with them during the whole assessment process, which appeared manageable given that InqITS was used over short time periods. It is a stark contrast with what happened during our research project.

None of the user tests that featured the *Research Notebook* were devoid of incidents, despite the fact that teachers were trained to use the *CNEC* and to avoid skewing indicators. There were, however, substantial variations in terms of the number of incidents per session, as we can see in Figure 3. The first user test occurred in Bouliers Elementary School during an hour-long session; we witnessed only three incidents: a technical incident that impacted a group of three students, and oral interactions with two groups of three and four students, respectively.

By contrast, we observed 18 incidents (10 technical, eight linked to teacher–pupil interaction) in the case of the user tests that took place in Victor Dupont Middle School. Incidents mostly occurred at the beginning of the test, when instructors faced technical issues linked to pupils' computers (issues with their firewalls slowed down classroom activity). Such frequent incidents inevitably interfere with the depiction of learners' skills. The only solution that could be done to decrease this interference would be for such issues to decrease over sessions, and to find a way to make an entire session irrelevant from the assessment point of view. In the next section, we focus on one particular incident involving an interaction between a teacher and a group of pupils. We will use it to illustrate how an indicator can be skewed by a teacher–pupil interaction, therefore hindering the relevance of the use of the dashboard for summative assessment.

Figure 3

Number of Students Whose Indicators were Affected by Incidents, Over Time, in Two User Tests



Note: We made the distinction between technical incidents (blue) and teacher interaction-related incidents (orange), like oral interactions with students on written productions.

Impact of An "Incident" on Dashboard Indicators

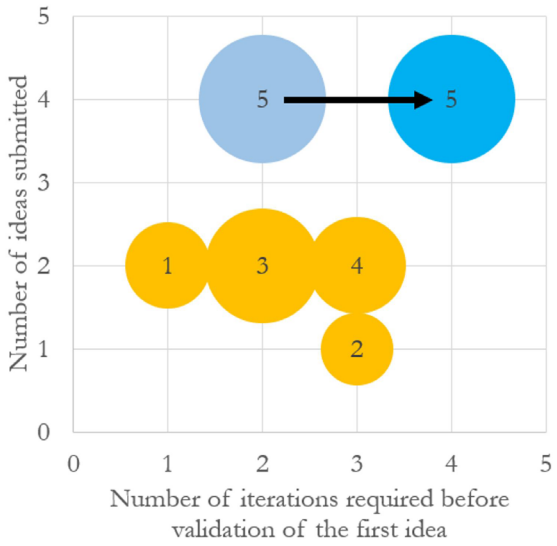
In Figure 4, we present a view of the dashboard for a user test that occurred in the Bouliers context, while learners were supposed to produce hypotheses in a collective manner, each one on their own research question. There were five groups of three to four students during this session. Instructors had decided that skill assessment would be collective, with the same grade being given to the whole group, notably based on the indicators provided by the *CNEC*.

For the sake of our demonstration, we focused on a specific interaction between an instructor and group five. Through learning analytics we observed that only two iterations were required before the first hypothesis was validated. However, we noticed based on our classroom observations that two additional cycles of iterations had occurred on

a pupil's notebook for this group, with oral help from the teacher (Figure 4, light blue). During a short intervention, the teacher helped students rephrase the hypothesis, and did twice with a pen what she had done virtually with other groups, using the interface of the *Research Notebook*. When we discussed this issue after the class, she said that it seemed easier for the students to grasp her feedback in that way, even if it meant affecting the outcome of this project for summative assessment. If these interactions had been registered (Figure 4, dark blue), four iterations would be displayed on the dashboard instead of two. It would possibly mean a lower grade, since it would suggest that students took longer to grasp instructions and feedback.

Figure 4

Incidents Affecting Research Notebook Indicators for a Selection of Students from Bouliers Elementary School



Note: The size of the bubbles represents the number of words for the idea that was accepted; we show an issue for group 5, due to an oral interaction with the teacher.

All interventions that can change the recorded number of iterations for a given student may affect the values of the indicators, and therefore the way that they will be assessed. For instance, when instructors allowed learners to show their production on the electronic device and to correct it before it was sent for evaluation, which happened in six of the user tests we monitored, it would similarly decrease the number of iterations registered by the *CNEC*. Practitioners willing to refrain themselves from skewing indica-

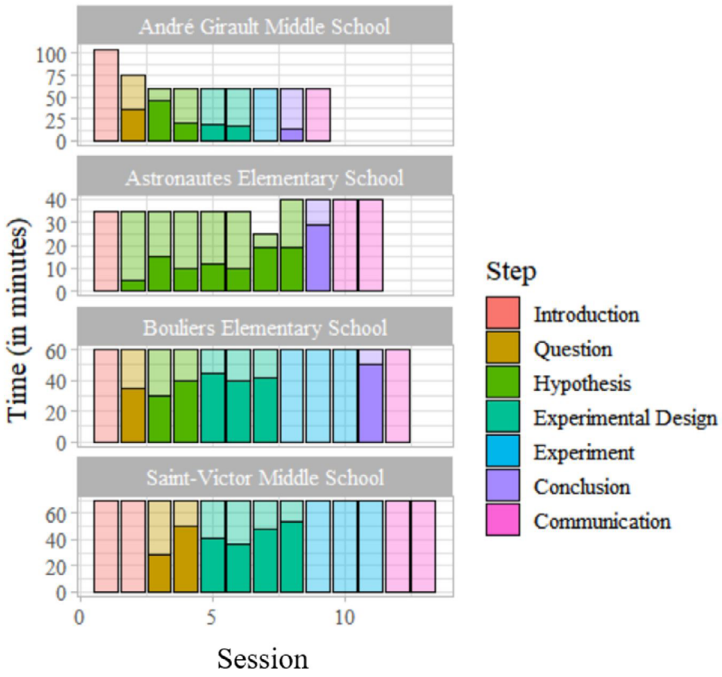
tors would have to control precisely the timing and the nature of the interactions that they have with learners. They would need to either avoid interacting with learners or interact in the same way with each group of learners. In the next section, we illustrate the difficulties that teacher interactions during summative assessment could pose in a *Savanturiers* project, given how common iterative writing phases are. Data collection periods would potentially span most of a given session, and most of the sessions of a given project.

Learner–Teacher Interactions and Their Impact on Indicators

A chronological description of four of the projects that we followed thoroughly (Figure 5) appeared to represent one of the most convincing demonstrations of the challenges posed by the use of a dashboard to assess skills in long-term projects. The goal underlying this analysis was to determine the proportion of the project during which practitioners and learners engaged in iterative writing over the various steps of the inquiry process (research question, hypothesis, experimental design, etc.).

Figure 5

Chronological Analysis of Four Savanturiers Projects, from the First to the Last Session



Note: We focus on the proportion of a given session that was dedicated to iterative writing (plain/solid colour), by contrast with other tasks (transparent).

The rationale is the following: Had practitioners actually used the *CNEC* for all sessions involving iterative learning, they would have had to refrain from spontaneously interacting with students in more than half of the sessions dedicated to the project (Figure 5). In Bouliers and Saint-Victor institutions, when learners engaged in iterative writing with the teacher, a step that typically began at the phrasing of the research question, iteration would represent most of the time of the session. For instance, in Bouliers, for a one-hour session, teachers would typically go from one group to the other during 30 to 40 minutes; the rest of the time (represented as transparent shading in Figure 5) would be dedicated to other tasks (maintaining order in the classroom, reminding learners about the schedule of the project, etc.).

In Peupliers Elementary School, the teacher focused solely on proposing hypotheses to explain a phenomenon—the loss of diversity at the global scale, since she had decided which question the whole classroom would pursue. Students iterated over the phrasing of various hypotheses, while browsing the web for plausible explanations. We can see that iterative writing, when implemented, took most of the time of a given session in both an elementary school (Bouliers), and a middle school (Saint-Victor). It suggests that dashboard-based assessment was practically unfeasible in this real-life inquiry, since teachers typically interacted in a spontaneous manner with students for such long periods. These interactions could not be overlooked in the focus of this study.

Discussion

Our results led us to conclude that using the dashboard in a context of summative assessment required instructors to be able, on the one hand, to collect numerous cues regarding unexpected events in the classroom, and, on the other hand, to refrain from automatically providing oral feedback to students. Dashboard-based assessment did not appear practical over long periods. The context of this research differs significantly from that of the works of Phillips et al. (2011). However, we reached similar conclusions on the issue of indicator misinterpretation. The characteristics of the classroom situation need to be collected to limit the risk of a superficial interpretation of quantitative indicators. Dashboard-based assessment works better in a situation with these additional cues that do not matter for the evaluation of the students. Field observations, mixing close monitoring of user tests, and an analysis of learning analytics used during them, enabled detection of some of the

discrepancies between what indicators intend to reflect, and pupils' actual skills. We went beyond the identification of sources of data misinterpretation and examined how they could be mitigated.

Since no data is impervious *per se* to misinterpretation, the challenge posed by their use in a context of summative assessment lies in the ability of the teachers to refrain from skewing the indicators. From that point of view, at least two approaches can be identified. Teachers can take notes of all events that could impact the interpretation of the indicators, and, when assessing skills, base their diagnosis upon a mix of notes and indicators. Our observations suggest, nevertheless, that this task could interfere with classroom management, given how frequently teachers would need to take notes during class sessions (Figure 3). It would, moreover, make the assessment process too complicated, and would defeat the purpose of providing a dashboard in the first place; one of the goals underlying its design is to save time by recording pupils' actions automatically.

The second approach consists in modifying directly, in real time, the logs of the learning environment, whether to delete events that should not have been recorded by the learning environment, or to add meaningful events that happened outside of it but were not recorded automatically, like iterations that were made orally. Some learning environments whose purpose is to closely track students' actions, like *Classcraft* (Sanchez et al., 2017), allow users to remove an action from the log of students' actions. One of the applications of such a feature is to correct a potentially skewed perception of a student's skills. If it is done immediately upon detection of the incident, it saves time by not compelling the practitioner to take notes, but this solution also interferes with the teaching activity, since it implies being able to modify logs while monitoring students.

When dashboards are used to regulate classroom activity (Verbert et al., 2014), notably to identify struggling or disengaged students (Hu et al., 2014), practitioners can collect cues on the fly to complement indicators without having to take notes on the specific details of the situations. Remaining doubts about how learners engaged in the activity can be dispelled through interactions with the pupils. This is not the case with skill-based assessment, since it usually occurs, at least in France, at the end of a semester or a year (i.e., days, weeks, or months after the situation during which the pupils were evaluated). The collection of contextual cues necessary to a relevant assessment is not possible anymore, nor is it possible to interact with the students to achieve a better understanding

of their activity. Indicators only cover a fraction of the activity. A superficial reading of the learning analytics can be insufficient to assess the situation, as Phillips et al. (2011) pointed out in their study on *Lectopia*. To mitigate the issues we highlighted earlier, among them technical issues, teachers need to collect data on the fly during interactions with students to supplement the dashboard.

Even if teachers were aware of the issues that we discussed in this article, they may rely too much on the indicators at their disposal when establishing a diagnosis, without critically interpreting how the characteristics of didactic situations, unexpected events, and unnoticeable learners' actions could affect such indicators. Moreover, in the absence of mitigation mechanisms, relying on indicators to ground their diagnosis would potentially constrain the interactions they have with learners. These solutions are neither optimal nor realistic in the context of K–12 education.

Finally, the dashboard is likely to incentivize teachers to behave like students were in a perpetual examination over the course of the project. It could lead to the standardization of project-based learning. That practice would be, to a large extent, contradictory with the principles of inquiry-based learning, which is to offer pupils more autonomy, and practitioners more freedom in their teaching approaches. Moreover, the oral or written feedback that teachers provide in real time represents a form of formative assessment that is valuable for the students. Hindering such pedagogical practices and establishing strict teacher–pupil interaction protocols for the sake of summative assessment's accuracy would probably be detrimental for the overall learning process. Since such complex protocols are unlikely to be followed over a long period, they would create a false sense of objectivity in the summative process. Therefore, we conclude that the relevance and the robustness of the indicators that had been designed for the *CNEC* are not adequate for summative assessment in our context. It is also probably the case in any type of project that spans several sessions, since it would require teachers to sustain, over a long period, a high level of vigilance with regard to how they interact with learners.

Should learning environments become increasingly used for the sake of summative assessment, dashboards that will allow monitoring of students' actions might become a valuable tool. It is possible that the rise of such an approach could contribute to the standardization of competency assessments, which could in turn lead to constraining pedagogical practices. Some authors have pointed out that teaching to the test can come

at the cost of practitioner creativity (Longo, 2010). This issue gains importance if the assessment encompasses year-long projects as is the case for *Savanturiers*. In that case, dashboards, if they were used for summative assessment, could hinder pedagogical innovation in day-to-day activities.

Conclusion

Limitations of the Present Study

As in any longitudinal study, we faced the trade-off between the number of observations that we could carry out per project, and the number of projects that could be followed. The small number of practitioners remains one of the strong limitations of our study. Moreover, even if our methods allowed us to identify potential sources of misinterpretation, we could not assess to what extent practitioners were able to detect and mitigate them. Finally, the type of incidents that we could detect strongly depended upon the nature of the indicators and upon the learning environment that teachers were using. While our case study can serve as an illustration of issues associated with dashboard-based summative assessment, our results, as is common in case studies, lack external validity.

Perspectives

In interviews that we had carried out in parallel to classroom observations (Cisel & Baron, 2019), a biology teacher from Cecile Middle School pointed out a potential risk associated with our indicators: "Submitting ideas, if the students understood that they are going to be evaluated by the number of ideas that [they submit], then they are going to submit 28 rubbish ideas." In other words, students could "game the system" (Baker et al., 2004, 2008). The authors "established links between gaming and learning" and developed models of gaming behaviour, in an article that triggered the launch of countless research works on the topic. Baker (2011) defines "gaming the system" as an attempt "to succeed in an educational task by systematically taking advantage of properties and regularities in the system used to complete that task, rather than by thinking through the material." For future research works, we could study how students, when they become aware that they are being assessed through indicators, could try to "game the system" based on their understanding of how

their actions are recorded by the system. Some students could, for instance, try to inflate a given indicator to increase a grade. A future area of research could be a reflection on how to detect and counter such strategies in the specific context of summative assessment.

References

Abd-El-Khalick, F., BouJaoude, S., Duschl, R., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., & Tuan, H. (2004). Inquiry in science education: International perspectives. *Science Education, 88*(3), 397–419. https://doi.org/10.1002/sce.10118

Aguilar, S. J., Karabenick, S. A., Teasley, S. D., & Baek, C. (2021). Associations between learning analytics dashboard exposure and motivation and self-regulated learning. *Computers & Education*, *162*, 104085. https://doi.org/10.1016/j.compedu.2020.104085

Ali, L., Hatala, M., Gašević, D., & Jovanović, J. (2012). A qualitative evaluation of evolution of a learning analytics tool. *Computers & Education, 58*(1), 470–489. https://doi.org/10.1016/j.compedu.2011.08.030

Baker, R. S. (2011). Gaming the system: A retrospective look. *Philippine Computing Journal*, *6*(2), 9–13.

Baker, R. S., Corbett, A. T., & Koedinger, K. R. (2004) Detecting student misuse of intelligent tutoring systems. In C. J. Lester, R. M. Vicari, & F. Paraguaçu (Eds.), *Proceedings of the 7th International Conference on Intelligent Tutoring Systems* (pp. 531–540). Springer.

Baker, R. S., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research, 19*(2), 185–224. https://www.learntechlib.org/primary/p/24328/

Bulletin Officiel de l'Education Nationale. (2007). Livret Personnel de Compétences [Personal Skills Gradebook]. *Bulletin officiel n° 22*. https://www.education.gouv.fr/bo/2007/22/MENE0754101D.htm

Bulletin Officiel de l'Education Nationale. (2016). Évaluation des acquis scolaires des élèves et livret scolaire, à l'école et au collège [Assessment of learners' skills and gradebooks for elementary and middle schools]. *Bulletin officiel n° 3.*

Bywater, J. P., Chiu, J. L., Hong, J., & Sankaranarayanan, V. (2019). The teacher responding tool: Scaffolding the teacher practice of responding to student ideas in mathematics classrooms. *Computers & Education, 139*(1), 16–30. https://doi.org/10.1016/j.compedu.2019.05.004

Cisel, M., & Barbier, C. (2021a). Mentoring teachers in the context of student-question-based inquiry: The challenges of the Savanturiers programme. *International Journal of Science Education*, *43*(17), 2729–2745. https://doi.org/10.1080/09500693.2021.1986240

Cisel, M., & Barbier, C. (2021b). Instrumentation numérique de la rédaction incrémentale : leçons tirées de la mise à l'épreuve du carnet numérique de l'élève chercheur [Iterative writing and digital technologies: Lessons drawn from an experimentation with the CNEC]. *Canadian Journal of Education/Revue canadienne de l'éducation*, *44*(2), 277–307. https://doi.org/10.53967/cje-rce.v44i2.4445

Cisel, M., & Baron, G. L. (2019). Utilisation de tableaux de bord numériques pour l'évaluation des compétences scolaires : une étude de cas [Using Digital Dashboards in the context of skill-based assessments in primary and secondary education: A case study]. *Questions Vives. Recherches en éducation*, (31). https://journals.openedition.org/questionsvives/3883

Eurydice. (2012). *Developing key competences at school in Europe: Challenges and opportunities for policy.* Publications Office of the European Union. https://eacea.ec.europa.eu/national-policies/eurydice/content/developing-key-competences-school-europe-challenges-and-opportunities-policy_en

Faber, J. M., Luyten, H., & Visscher, A. J. (2017). The effects of a digital formative assessment tool on mathematics achievement and student motivation: Results of a randomized experiment. *Computers & Education, 106*, 83–96. https:/doi.org/10.1016/j.compedu.2016.12.001

Gobert, J. D., Pedro, M. S., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences, 22*(4), 521–563.

Han, J., Kim, K. H., Rhee, W., & Cho, Y. H. (2021). Learning analytics dashboards for adaptive support in face-to-face collaborative argumentation. *Computers & Education*, *163*, 104041. https://doi.org/10.1016/j.compedu.2020.104041

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Herranen, J., & Aksela, M. (2019). Student-question-based inquiry in science education. *Studies in Science Education, 55*(1), 1–36. https://doi.org/10.1080/03057267.2019.1658059

Hu, Y.-H., Lo, C.-L., & Shih, S.-P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior, 36*, 469–478. https://doi.org/10.1016/j.chb.2014.04.002

Iandoli, L., Quinto, I., De Liddo, A., & Buckingham Shum, S. (2014). Socially augmented argumentation tools: Rationale, design and evaluation of a debate dashboard. *International Journal of Human-Computer Studies, 72*(3), 298–319. https://doi.org/10.1016/j.ijhcs.2013.08.006

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Krueger, R. A. (2014). *Focus groups: A practical guide for applied research.* SAGE.

Longo, C. (2010). Fostering creativity or teaching to the test? Implications of state testing on the delivery of science instruction. *The Clearing House, 83*(2), 54–57.

McMullan, M., Endacott, R., Gray, M. A., Jasper, M., Miller, C. M. L., Scholes, J., & Webb, C. (2003). Portfolios and assessment of competence: A review of the literature. *Journal of Advanced Nursing, 41*(3), 283–294. https://doi.org/10.1046/j.1365-2648.2003.02528.x

Mottus, A., Graf, S., & Chen, N.-S. (2015). Use of dashboards and visualization techniques to support teacher decision making. In T. Kinshuk & R. Huang (Eds.), *Ubiquitous learning environments and technologies* (pp. 181–199). Springer Berlin Heidelberg.

Nielsen, J. (1993). *Usability engineering.* Morgan Kaufmann Publishers.

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., & Tsourlidaki, E. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review, 14*, 47–61. https://doi.org/10.1016/j.edurev.2015.02.003

Phillips, R., Maor, D., Cumming-Potvin, W., Roberts, P., Herrington, J., Preston, G., & Perry, L. (2011, December 4–7). *Learning analytics and study behavior: A pilot study* [Conference presentation]. ASCILITE 2011 Conference, Tasmania. https://researchrepository.murdoch.edu.au/id/eprint/6751/

Quintana, C., Zhang, M., & Krajcik, J. (2005). A framework for supporting metacognitive aspects of online inquiry through software-based scaffolding. *Educational Psychologist, 40*(4), 235–244. https://doi.org/10.1207/s15326985ep4004_5

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Sanchez, E., Young, S., & Jouneau-Sion, C. (2017). Classcraft: From gamification to ludicization of classroom management. *Education and Information Technologies, 22*(2), 497–513. https://doi.org/10.1007/s10639-016-9489-6

Scardamalia, M., & Bereiter, C. (2006). Knowledge building: Theory, pedagogy, and technology. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 97–118). Cambridge University Press.

Scheffel, M., Drachsler, H., Toisoul, C., Ternier, S., & Specht, M. (2017). The proof of the pudding: Examining validity and reliability of the evaluation framework for learning analytics. In É. Lavoué, H. Drachsler, K. Verbert, J. Broisin, & M. Pérez-Sanagustín (Eds.), *Data driven approaches in digital education* (pp. 194–208). Springer.

Schwendimann, B. A., Rodriguez Triana, M. J., Prieto Santos, L. P., Shirvani Boroujeni, M., Holzer, A. C., & Gillet, D. (2016). Understanding learning at a glance: An overview of learning dashboard studies. In *LAK '16: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 532–533). Association for Computing Machinery.

Sedrakyan, G., Malmberg, J., Verbert, K., Järvelä, S., & Kirschner, P. A. (2018). Linking learning behavior analytics and learning science concepts: Designing a learning analytics dashboard for feedback to support learning regulation. *Computers in Human Behavior, 107,* 105512. https://doi.org/10.1016/j.chb.2018.05.004

Silvola, A., Näykki, P., Kaveri, A., & Muukkonen, H. (2021). Expectations for supporting student engagement with learning analytics: An academic path perspective. *Computers & Education*, *168*, 104192. https://doi.org/10.1016/j.compedu.2021.104192

Specht, M., Bedek, M., Duval, E., Held, P., Okada, A., & Stefanov, K. (2013). WESPOT: Inquiry based learning meets learning analytics. In *Proceeding: The Third International Conference on e-Learning* (pp. 15–20). Belgrade Metropolitan University. http://oro.open.ac.uk/42569/

Stephens-Martinez, K., Hearst, M. A., & Fox, A. (2014). Monitoring MOOCs: Which information sources do instructors value? In M. Sahamani (Ed.), *Proceedings of the First ACM Conference on Learning @ Scale Conference* (pp. 79–88). ACM. https://doi.org/10.1145/2556325.2566246

UNESCO. (2007). *Curriculum change and competency-based approaches: A worldwide perspective*. UNESCO Publications. http://www.ibe.unesco.org/en/services/online-materials/publications/recent-publications/single-view/news/curriculum-change-and-competency-based-approaches-a-worldwide-perspective-prospects-n-142/2842/next/4.html

Vardi, I. (2012). The impact of iterative writing and feedback on the characteristics of tertiary students' written texts. *Teaching in Higher Education, 17*(2), 167–179. https://doi.org/10.1080/13562517.2011.611865

Verbert, K., Govaerts, S., Duval, E., Santos, J., Van Assche, F., & Parra, G. (2014). Learning dashboards: An overview and future research opportunities. *Personal and Ubiquitous Computing, 18*(6), 1499–1514. https://doi.org/10.1007/s00779-013-0751-22

Wang, D., & Han, H. (2021). Applying learning analytics dashboards based on process-oriented feedback to improve students' learning effectiveness. *Journal of Computer Assisted Learning*, *37*(2), 487–499.

Xhakaj, F., Aleven, V., & McLaren, B. M. (2016). How teachers use data to help students learn: Contextual inquiry for the design of a dashboard. In K. Verbert, M. Sharples, & T. Klobučar (Eds.), *Adaptive and adaptable learning* (pp. 340–354). Springer. https://doi.org/10.1007/978-3-319-45153-4_26

Zhao, K., & Chan, C. K. K. (2014). Fostering collective and individual learning through knowledge building. *International Journal of Computer-Supported Collaborative Learning*, *9*(1), 63–95. https://doi.org/10.1007/s11412-013-9188-x