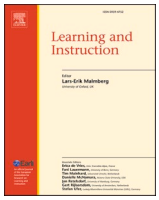
Learning and Instruction 97 (2025) 102100

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc



How do teachers process technology-based formative assessment results in their daily practice? Results from process mining of think-aloud data

Sarah Bez [a,b,*], Fabian Burkart [b], Martin J. Tomasik [c], Samuel Merk [b]

[a] University of Tübingen, Department of Education, Münzgasse 22-30, D-72070, Tübingen, Germany
[b] Karlsruhe University of Education, Department of school and teaching development, Bismarckstraße 10, D-76133, Karlsruhe, Germany
[c] University of Zurich, Institute of Education, Freiestrasse 36, CH-8032, Zurich, Switzerland

ARTICLE INFO

Keywords:
Think-aloud
Technology-based formative assessment
Processing
Teachers
Data-based decision making
Process mining

STRUCTURED ABSTRACT

Background: Technology-based formative assessments are considered promising in terms of reducing teachers' workload and providing validity advantages but little is known how teachers use the assessment results to inform their instruction in their daily practice.
Aims: We explored how teachers process technology-based formative assessment results using think-aloud methodology in an ecologically valid setting.
Sample: Forty-eight experienced in-service teachers participated in the study.
Methods: We asked the teachers to verbalize their thoughts while they processed students' formative assessment results as they usually do. Screencasts of the verbalizations and assessment results were recorded. Based on these, trained raters coded the main steps of processing and which specific aspects of the results were noticed based on a deductive-inductive coding scheme. Cluster analyses were applied to explore differences among teachers, and process mining was conducted to explore the main processes.
Results: We found four main steps: *noticing results*, *comparing with personal perspective*, *analyzing errors* and *constructing instructional implications*. Relative durations of these steps vary substantially among teachers. Cluster analyses indicate that processes were differentiated according to the complexity of summarizing and building relationships between single data points. The fitted process model revealed low dependency values in general and indicates that noticing results on its own seemed to be insufficient for constructing instructional implications.
Conclusions: This study generates the hypothesis that analyzing errors and comparing results with the personal perspective are important for teachers for next instructional decisions.

Formative assessment is considered to support the learning of students (e.g., Kingston & Nash, 2011; Lee et al., 2020; Xuan et al., 2022). In recent decades, it has received increased attention in the context of both policy and practice around the world (Birenbaum et al., 2015) and various technology-based systems have been developed and implemented in different countries (e.g., The Netherlands, Faber et al., 2017; Faber & Visscher, 2018; Switzerland, Tomasik et al., 2018). In the context of adaptive teaching, formative assessment is highlighted as essential because (ongoing) assessing students' individual characteristics and their current understanding seem to be necessary prerequisites for setting adequate individual learning goals as well as providing tailored instruction and giving individual support based on the assessment results (Corno, 2008; Hardy et al., 2019). Given the rise of digitization in education, technology-based formative assessments, in particular, are attributed some promising opportunities, especially related to validity, including task pools with calibrated items and sophisticated psychometric models (e.g., McLaughlin & Yan, 2017; Spector et al., 2016). At the same time, several current reviews point out that teachers play a key role in the successful use of formative assessments in the classroom (Heitink et al., 2016; Schildkamp et al., 2020; Yan et al., 2021), especially with regard to their adequate noticing of and interpretation of the assessment results: If teachers do not notice important

This article is part of a special issue entitled: Chronicles of cognition published in Learning and Instruction.

* Corresponding author. University of Tübingen, Department of Education, Münzgasse 22-30, D-72070, Tübingen, Germany.
E-mail addresses: sarah.bez@uni-tuebingen.de (S. Bez), fabian.burkart@ph-karlsruhe.de (F. Burkart), martin.tomasik@ife.uzh.ch (M.J. Tomasik), samuel.merk@ph-karlsruhe.de (S. Merk).

https://doi.org/10.1016/j.learninstruc.2025.102100
Received 14 October 2023; Received in revised form 20 October 2024; Accepted 7 February 2025
Available online 27 February 2025

aspects of the assessment results or misinterpret the information encoded in the data, one can conclude that setting adequate learning goals as well as providing tailored instruction and giving individual support will be inappropriate. From the perspective of consequential validity, providing valid formative assessments based on technology is not sufficient on its own, and adequate interpretations and inferences are essential (Kane, 2013). However, little is known about teachers' actual activities in terms of using (formative) assessment data in practice (Hebbecker et al., 2022; Mandinach & Schildkamp, 2021), especially the ways teachers notice and interpret technology-based formative assessments and how they construct implications for adaptive teaching in their daily practice. Against this background, we investigated how teachers process technology-based formative assessment results with a strong focus on ecological validity and a process perspective using think-aloud methodology.

1. Teachers' processing of technology-based formative assessment

1.1. Formative assessment

Formative assessment can be conceptualized as a process, in which students' learning is diagnosed with the goal of adapting future instruction and supporting students in achieving their learning goals, especially by providing feedback (Black & Wiliam, 2009). For this end, assessments are used *for* learning, that means that teachers use classroom assessments to inform and adapt instruction to foster student learning (Black & Wiliam, 2018). Although there are many different approaches and methods, according to Black and Wiliam's definition, the central characteristic of formative assessment is "that evidence about student achievement is elicited, interpreted and used (…) to make decisions about the next steps in instruction that are likely to be better, or better founded, than the decisions (…) in the absence of the evidence" (Black & Wiliam, 2009, p. 9). In general, formative assessment can be seen as part of data-based decision making in education (Schildkamp, 2019). Concerning the empirical evidence that formative assessment can foster learning, one can find a range of effect sizes: For example, Black and Wiliam (1998) name effect sizes $0.40 < d < 0.70$ as typical for formative assessment, Kluger and DeNisi (1996) calculated $d = 0.41$ in their meta-analysis and Kingston and Nash (2011) report a weighted mean effect size of $d = 0.20$ with a 95% CI[0.19, 0.21] in their meta-analysis. Recently published meta-analyses (Lee et al., 2020; Xuan et al., 2022) report substantial positive effect sizes on learning (Lee et al., 2020: $d = 0.29$; Xuan et al., 2022 [focus on reading]: $d = 0.19$, 95% CI[0.14, 0.22]) after adjusting for publication bias. There is an ongoing debate about the effectiveness of formative assessment (Bennett, 2011; Dunn & Mulvenon, 2009; Kingston & Nash, 2011) and several methodological issues are discussed, especially the quality of the studies concerning internal validity, the selection criteria in meta-analysis and the nature of formative assessment (McMillan et al., 2013). Although the magnitudes of the effect sizes differ and, to date, it seems to be difficult to average the size of the effect, all mentioned reviews and meta-analyses report positive effects.

1.2. Benefits of technology in the context of formative assessment

As outlined in the previous section, substantive evidence indicates that formative assessment can foster instruction and learning. With the rise of technology, it is argued that technology can provide several benefits for formative assessment. First, it is emphasized that technology-based formative assessment can reduce the time and effort required to conduct formative assessments and therefore relieve teachers, although using technology can be challenging for teachers (Adachi et al., 2018). For example, item banks that include various task types (Conole & Warburton, 2005; McLaughlin & Yan, 2017) can reduce the burden on teachers by saving them the time and effort needed to create assessments on their own (e.g., Maier et al., 2016; Russell, 2010). Moreover, such platforms can automatically correct students' answers (Conole & Warburton, 2005). Based on that, they can provide (more or less elaborated) feedback for all students and teachers (Spector et al., 2016) by visually displaying the assessment results of single students or the learning group as a whole. Second, it is emphasized that technology-based formative assessment systems can achieve a high level of validity as they can use large item banks with tasks of calibrated difficulty (Spector et al., 2016) and can integrate objective scoring rules during the analysis of students' answers (Tomasik et al., 2018). They can also rely on sophisticated psychometric models that enable adaptive testing (Conole & Warburton, 2005) which seems to be positive for students' motivation and emotion in tests without a time limit but not in general (Frey et al., 2024). Moreover, technology-based formative assessment systems can provide comparisons with other learners and the measurement of the learning progress over time at the individual level (Conole & Warburton, 2005). In particular, the valid, longitudinal measurement of the learning progress of all students is a promising opportunity because it hits the central pedagogical concept of formative assessment: The valid measurement of learning progress in comparison to learning goals enables teachers to adapt instruction continuously as they can evaluate feedback, instruction and learning goals based on the learning progress of individual students.

1.3. (Capturing) teachers' processing of assessment results

Given the advantages of technology-based formative assessment in terms of reducing teachers' workloads and increasing validity, along with the increasing availability of such systems (Faber et al., 2017; Faber & Visscher, 2018; Hall et al., 2015; Hebbecker et al., 2022; Tomasik et al., 2018), nevertheless, teachers play an important role when it comes to the successful use of these systems in classrooms. In particular, teachers' processing of assessment results is crucial from both a conceptual and an empirical point of view, as is outlined in the next section.

In general, formative assessment is considered a part of data-based decision making (Schildkamp, 2019). Current conceptual models of data-based decision making (e.g., Coburn & Turner, 2011; Mandinach & Gummer, 2016; Marsh, 2012; Schildkamp, 2019) differ in their details but they all conceptualize a sequential process with several steps that follow each other (Bez et al., 2023; Hebbecker et al., 2022): First, data must be collected. Then, teachers have to process the data, which means that they have to notice relevant aspects of the data and interpret them in order to formulate corresponding instructional implications. These implications have to be put into practice and evaluated in their outcomes. Since processing is the first step after data collection that can be simplified and improved through technology-based systems, one can conclude that appropriate processing is crucial for the following steps. Similarly, from the perspective of consequential validity (Kane, 2013), one can argue that noticing and interpreting technology-based formative assessment results is a necessary prerequisite for making adequate inferences in terms of constructing pedagogical implications and instructional decisions. If teachers do not notice important aspects in assessment results that are provided with high construct validity by a technology-based assessment system or if they misinterpret information, the derived instructional conclusions and decisions as well as the following specific instructional actions will be necessarily inadequate and therefore consequential validity is diminished.

From an empirical perspective, several systematic reviews have emphasized the importance of teachers, especially their data literacy, for successful formative assessment in practice (Heitink et al., 2016; Schildkamp et al., 2020; Yan et al., 2021). Research focusing on teachers' use of data indicates that teachers often find this challenging (Mandinach & Schildkamp, 2021) and generally tend to have difficulties reading and interpreting data and deriving conclusions for instructional decisions (e.g., Kippers et al., 2018; van den Bosch et al., 2017; Zeuch et al., 2017). However, little previous research has provided insights

into teachers' actual activities regarding (formative) assessment data in practice (Hebbecker et al., 2022; Mandinach & Schildkamp, 2021), especially how teachers process (assessment) data in their daily practice (Bez et al., 2021; Goffin et al., 2022). To understand teachers' actual processing, retrospective self-reports are limited because they can be biased and are restricted in their opportunity to provide insights into processing *in vivo* (Parry et al., 2021; Paulhus, 1991). Test scores (e.g. from data literacy tests) do not necessarily have to correspond with the performance of teachers in practice (Blömeke et al., 2015), and inferences based on intervention studies with a focus on internal validity (e.g., on data literacy) have (necessarily) limitations related to ecological validity (Shadish et al., 2002). Log-file analyses might be promising in investigations of technology-based formative assessments as they describe actual behavior objectively (Parry et al., 2021). However, the construct validity of measures such as dwell time or click sums remains difficult (Hebbecker et al., 2022) as their interpretations are often ambiguous. That means that inferences based on log data about the processing of teachers, in terms of what they notice in assessment results and how they interpret the information and construct instructional implications, are limited.

In this context, think-aloud is considered an appropriate methodological approach (Espin et al., 2017), because it can provide non-reactive insights into everyday life cognitive processes (Ericsson & Simon, 1998; Fox et al., 2011). In particular, concurrent think-aloud seems adequate because, in contrast to retrospective think-aloud, it addresses processes in the working memory (Leighton, 2017). According to the framework of Ericsson and Simon (1998), it is important to differentiate different types of concurrent verbalization procedures. Asking participants to just verbalize their thoughts spontaneously does not affect cognitions, whereas requesting additional explanations or descriptions elicits additional cognitive processes, changes sequences and therefore might affect performance. Fox et al. (2011) investigated this in their meta-analysis and report, in line with the framework, that think-aloud does not affect performance and generally increases time. However, prompting participants to describe or explain their thoughts in concurrent think-aloud settings significantly affects performance in terms of increasing performance. There are some think-aloud studies addressing teachers' reading and interpreting of (formative) assessment data (Espin et al., 2017; Goffin et al., 2023; van den Bosch et al., 2017; Wagner et al., 2017). Regarding the think-aloud approach, in all of the mentioned studies the participants were asked to think aloud and explain their thoughts, too. Moreover, only Goffin et al. (2023) and van den Bosch et al. (2017) used the assessment results of the students of the participating teachers. To sum up, concurrent think-aloud without eliciting descriptions or explanations can provide non-reactive (i.e. not affecting performance) insights into processing related to daily life. Previous studies are limited to provide insights concerning teachers' processing of assessment results in an ecologically valid setting due to the type of concurrent think-aloud and the assessment results used in the studies.

1.4. The present study

As outlined in the previous sections, technology-based formative assessments are promising in terms of their ability to relieve teachers and their advantages in terms of validity. A crucial prerequisite for the success in practice is teachers' processing of assessment results to inform their instruction for adaptive teaching. However, little is known about how teachers process technology-based formative assessments in their daily practice. Against this background, our research goal is to gain insights into teachers' processing in an ecologically valid setting using think-aloud methodology. Thus, our research questions are as follows.

- Research Question 1 (RQ 1): Which steps and aspects show teachers as they process technology-based formative assessment results?
- Research Question 2 (RQ 2): Can different groups of teachers be identified according to which aspects of formative assessment results are addressed and which steps of processing are shown?
- Research Question 3 (RQ 3): Which typical processes in teachers' think-aloud data can be identified?

To investigate these questions, we conducted an exploratory concurrent think-aloud study with in-service teachers and asked them to verbalize their thoughts while they processed the latest assessment results of their own students as they usually do. We used video calls to get in touch with the teachers in their daily environment. To prevent biases and reactivity, we avoided making directed prompts (e.g., for explanations, descriptions) as they are considered to affect cognitive sequences and performance (Fox et al., 2011; Leighton, 2017).

2. Method

2.1. Participants

We conducted think-aloud sessions with $N = 48$ in-service teachers ($M_{age} = 44.1$, $SD_{age} = 11.6$, $M_{teaching\ expertise} = 16.5$ years, $SD_{teaching\ expertise} = 10.5$ years, 56% identified themselves as male). All participating teachers are located in Switzerland and are voluntary users of the technology-based formative assessment platform Mindsteps (www.mindsteps.ch). Of the teachers, 31% work as primary school teachers, 69% work as secondary school teachers, and 90% teach at least one STEM subject (94% teach at least one language, 80% teach a social science subject, and 50% teach art). Participants were invited to participate via email.

2.2. The technology-based formative assessment platform

Mindsteps (www.mindsteps.ch) provides curriculum- and competence-based assessments for different subjects (Maths, German, French, English, and Science) from grade three to nine. Based on Item Response Theory and an extensive bank of calibrated items, the system provides criterion-based test scores including trajectories of individual students and classes (Berger et al., 2019; Tomasik et al., 2018). The items are mainly closed questions or short open-ended questions which are both automatically corrected by the system. The results are displayed in dashboards for students and teachers. Students receive feedback on the task level (knowledge of correct results) and on the summarized level of their overall test results. Teachers can view the assessment results of their students in different graph types on the class level (summarized overview of single assessments or progress) and the individual level (single items and answers of single assessments as well as competence levels and progress of students; please see Appendix A for exemplary screenshots).

2.3. Design

Executing concurrent think-aloud sessions, the participants were asked to verbalize their thoughts while they processed students' assessment results as they usually do (for details, please see Section 2.4). During the sessions, the participants logged into their formative assessment platform account and shared their screens. The participants' screens as well as their verbalizations were recorded. Trained raters coded these audiovisual data (verbalizations and screencasts) using an inductive-deductive developed coding scheme (see Section 2.4). The resulting data formed the basis for further statistical analyses (see Section 2.5). The ethics committee of the respective university approved the study.

2.4. Data collection and coding procedure

The one-to-one think-aloud sessions were conducted by two trained

interviewers and were structured based on the guidelines given in a detailed manual (a translated version can be found on the Open Science Framework [OSF], https://osf.io/r24jd/). The sessions were divided into several parts: After a short check of the technical setup, participants were introduced to the think-aloud procedure, and a short warm-up was conducted, as recommended in the literature (e.g., Leighton, 2017). Participants were asked to directly express everything that was going through their mind while looking at the assessment results as they usually do. They were informed that the interviewer would not interrupt them until they have finished and the interviewer is interested in their personal thoughts that can not be right or wrong. Then, as a short (warm-up) practice in concurrent think-aloud, participants were asked to verbalize all their thoughts while looking at a simple bar chart displaying weather data from Switzerland and pointing on the graph and tell the interviewer when they have finished. To prevent teachers from struggling in the warm-up practice, we decided to use a very simple bar chart and not a graph displaying assessment data because previous studies show that teachers tend to struggle in processing assessment data (see Section 1.3.). After that, participants logged in the platform and the concurrent think-aloud phase of the session concerning the assessment results was prompted as follows: "Please express everything that you are thinking while you are looking at the current assessment results of your class the way you usually do and point with your mouse at the corresponding areas. When you have finished, please say 'I have finished.'" In this phase, cameras of both the teacher and the interviewer were turned off in order to avoid unintended affections or distractions through facial expressions. The interviewer strictly stayed in the background and only said "please keep talking" or "please use your mouse", if necessary (Padilla & Leighton, 2017). When the teacher was finished, the interviewer asked the participant what the key information was in the results, whether they derived any conclusions from the assessment data and whether they would take specific instructional or general actions based on the assessments. The current paper focuses on the data obtained from the concurrent think-aloud phase with the participants. During the think-aloud phase, the participants clicked through the platform and looked at different graphs encoding the assessment results of their students. Appendix A provides exemplary screenshots of the different graph types.

To provide insights into the processing on the micro level, in the first run of coding, we maintained the sequential structure of the think-aloud data; therefore, the audiovisual data (recorded screencasts and verbalizations) were coded as timed-event codings (Bakeman & Quera, 2011), accurate to the second using MAXQDA 2022 (VERBI software, 2021) for data analysis. The coding scheme was inductive-deductive. The deductive part based on the framework provided by Coburn and Turner (2011) who differentiate noticing, interpreting and constructing implications. It was captured which graph type and, independently of this, which processing step was addressed. Two trained raters coded independently from each other and then discussed any disagreements until they reached a consensus. Due to a coding agreement of $0.58 \leq \alpha \leq 0.95$ (Hayes & Krippendorff, 2007) and the need for economic efficiency, 60% of the think-aloud sessions were coded using this procedure, and 40% were coded by one rater. During the coding process, we faced the known issue of co-occurrence and exclusivity of codes (Bakeman & Quera, 2011): In some cases, two steps were closely entangled, so we followed Bakeman and Quera (2011) and determined these cases later in the analysis.

To capture which aspects in the rich and extensive assessment results provided by the assessment platform were noticed, we conducted a second run of coding. It was captured based on an inductive-deductive developed coding scheme, whether specific aspects given in the displayed assessment results were noticed at least once. We derived key aspects of the displayed assessment results as deductive codes (e.g., mean of class, dispersion) and revised the coding scheme according to the aspects that were additionally noticed by the teachers (e.g., match individual assessment result to name of the student). This run of coding covered only think-aloud data related to graph 2 (see Appendix A) because this graph type was the one that was addressed mostly by the teachers (see Appendix A). We used Cohen's κ to calculate coding agreement and percentage agreement if κ could not be calculated, occurring when one rater never assigned a code. Due to a coding agreement ($0.49 \leq \kappa \leq 1.0$; *outlier* $0.25 \leq \kappa \leq 0.42$; $83\% \leq$ *percentage agreement* $\leq 100\%$), considering that Cohen's κ is reduced in the case of low prevalence of codes (Bakeman & Quera, 2011), and the need for economic efficiency, again 60% of the think-aloud sessions were coded independently by two trained raters, who discussed any disagreements until consensus was reached, and 40% were coded by one rater. The coding scheme (translated and shortened) is provided on OSF (https://osf.io/r24jd/).

2.5. Data analysis

Our research questions focus on teachers' processing of formative assessment data. Data analysis was conducted using R (v.4.3.1; R Core Team, 2023). To investigate which steps of processing can be found in the think-aloud data and how prevalent they are (RQ 1), we first preprocessed the timed-event codings to time samplings (Bakeman & Quera, 2011) with the interval of 1 s, calculated descriptive statistics and visualized the codings in their sequences for each teacher.

To explore the patterns of different groups of teachers in terms of the relative durations of the main steps and the aspects they addressed (RQ 2), we applied cluster analyses. As the data basis contained both continuous variables (e.g. proportions of main steps) and binary variables (e.g. noticed aspects), we decided to use Gower's coefficient (Gower, 1971) as similarity measurement (Kaufman & Rousseeuw, 2005), and we chose the average and complete linkage algorithms for clustering. Additionally to a visual inspection of the results in heatmaps with dendrograms, partitions of clustering solutions were compared using the Adjusted Rand Index (Hubert & Arabie, 1985) to aim for an appropriate and robust result.

To explore the think-aloud data (based on the codings) from a process perspective (RQ 3), we applied process mining because it can capture a typical process in event data and describe it in a sequential model (Reimann, 2009). Due to the exploratory design of the study and the absence of a preexisting process model, we decided to use discovery and not conformance checking (van der Aalst, 2016) and applied the Flexible Heuristics Miner Algorithm (Weijters & Ribeiro, 2011). This algorithm is appropriate for think-aloud data in educational research (e.g., Hartmann et al., 2022; Sonnenberg & Bannert, 2019) since it discovers dependencies in the ordering of events, takes loops into account, and, in contrast to other process mining algorithms, is able to deal with low-structured or noisy data (van der Aalst, 2016; Weijters & Ribeiro, 2011). The basic concept behind this algorithm is to retrieve a causal process model based on dependencies between events using frequency-based metrics (Weijters & Ribeiro, 2011). That is, if event A is very often directly followed by event B in the data, but event B is very rarely directly followed by event A, the value of the dependency measure for 'event A followed by event B' is high. This indicates that a causal dependency between event A and event B (in this order) is likely. The values of the dependency measures are always between −1 (low dependency) and 1 (high dependency). To discover the main process in teachers' processing, we explored a dependency graph based on a dependency matrix using the heuristicsmineR package (Mannhardt & Janssenswillen, 2023). Due to the very limited knowledge about how to set various parameters (a priori), especially in low-structured domains, we followed Weijters and Ribeiro (2011) and specified a complete model with low frequency and dependence thresholds and a simplified model with high(er) thresholds. The reproducible documentation of the complete data analysis is provided on OSF (https://osf.io/r24jd/).

3. Results

3.1. Research question 1

We found four main steps in the teachers' processing, as follows: *Noticing results* refers to the teachers' verbalizations when they notice the formative assessment results provided by the system, for example, "Tom got 546 points on the scale" or "mean of 477 points is quite good and concerning dispersion I see they are roughly all between 250 and 850". *Comparing system results with the teacher's personal perspective* covers the verbalizations of teachers when they compare the system's results or specific aspects with their own point of view, for example, "I am surprised because Hannah is normally one of the best in class", or "this is exactly what I expected". *Analyzing errors* captures expressions when teachers analyze errors or try to elaborate on the mistakes students make or their misconceptions, for example, "funny, he has an error at bar charts, probably a comprehension problem" or "the other addition tasks were correct, so the problem here was this specific task type". *Constructing instructional implications* covers teachers' verbalizations when they construct conclusions for general or instructional future actions, for example, "many students had problems with these tasks so I will repeat this topic in the next lesson'' or "I will talk to the student about the results". The relative durations of these steps of processing during think-aloud are provided in Fig. 1. All the teachers noticed the results but not all of the participants analyzed errors, made comparisons with their own perspectives or formulated instructional implications. The relative durations of the steps show obvious variance among teachers, especially regarding *noticing results* and *analyzing errors*.

Fig. 2 shows the percentages of teachers who noticed specific aspects of the results presented in graph 2 of the formative assessment system (see Appendix A, class overview) at least once. Results show that the highest percentages belong to those aspects of the results which can be directly noticed without complex cognitive elaborations (e.g., matching results to students' names, contrary to elaborating dispersion in the results according to the average distance of data points from the mean). Thirty percent of the teachers grouped the best and about 25% of the teachers grouped the weakest students, but only 9% grouped the middle of the displayed results. Concerning verbalizations that address common concepts within descriptive statistics, the highest percentage addressed the mean (about 55%) whereas only about 10% and fewer of the teachers mentioned range, outliers or dispersion based on the average distance from the mean.

To get a first impression of processing during think-aloud, we visualized the sequence of the main steps (respectively, the codings) for each teacher (Fig. 3). This graphical overview shows substantial variance among teachers in terms of the sequences. For example, the visualized think-aloud data of several teachers show many recursions and iterations in their sequences (e.g., teacher 37 and teacher 42).

3.2. Research question 2

Beyond the descriptive statistics of processing (RQ 1), we investigated whether the teachers could be differentiated into groups according to their processing. To this end, we conducted cluster analyses based on the relative durations of the main steps and the aspects of the results that teachers noticed in their processing using average and complete linkage algorithms (based on Gower's distance matrix coefficient [Gower, 1971],) and visualized the results using a heatmap with dendrograms (see Fig. 4).

According to Fig. 4, it is plausible to assume that there are three main groups (group A, B, C) of teachers in the data. This partition, based on average linkage, was compared to the three-group solution of complete linkage to aim for a robust result. The Adjusted Rand Index of $r = 0.9$ (values can theoretically range from -1 to 1) indicates a very high similarity between these two partitions (Hubert & Arabie, 1985). Group A (43% of participants, depicted at the top of the heatmap) notices no or only a few aspects that are directly given in the results of individual students and showed relatively low relative durations for each of the main steps of processing. In contrast, group B noticed the best and the weakest student, and group C formed groups of students according to the best and weakest results. Differences in processing can be characterized mainly according to the complexity of summarizing and building relationships between single and directly given data points (e.g., noticing the best/weakest students vs. forming corresponding groups). This is plausible according to the concept of graph literacy (Friel et al., 2001; Galesic & Garcia-Retamero, 2011). It differentiates between levels of

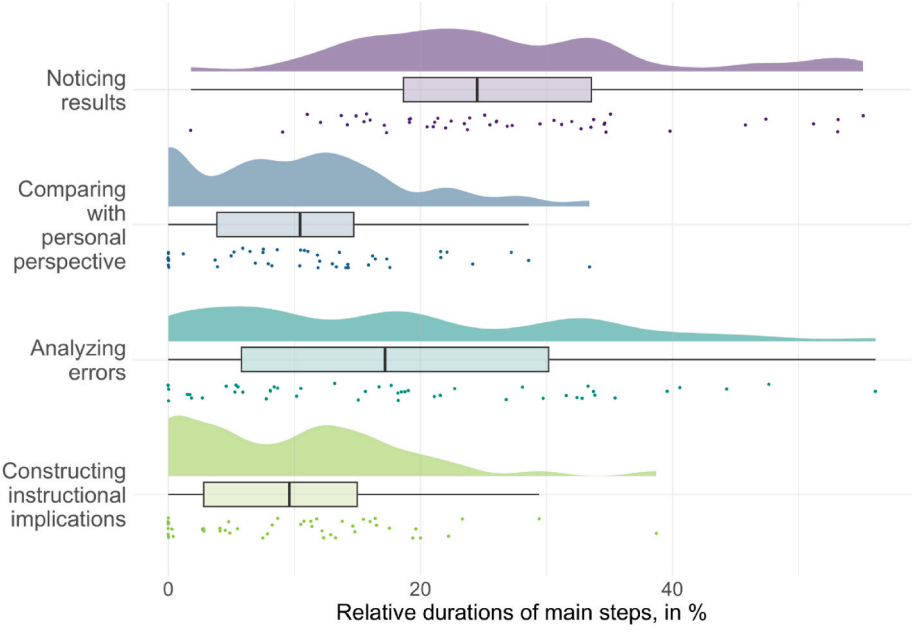


Fig. 1. Relative durations of main steps of processing during think-aloud
Note. Comparing with personal perspective is an abbreviation for *Comparing results with the teacher's personal perspective*. The Co-occurrences of main steps and the relative durations of addressing different graph types can be found in Appendix A.

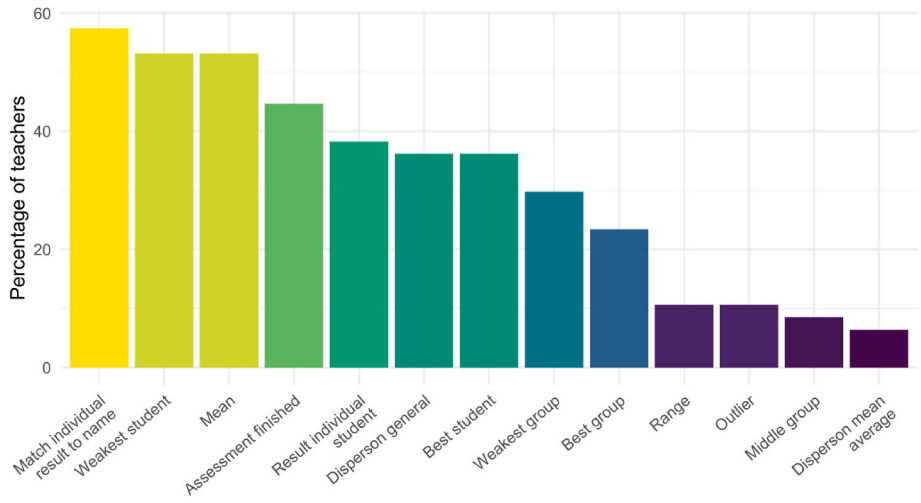


Fig. 2. Percentage of teachers who noticed specific aspects of the results
Note. The percentage of teachers who verbalized specific aspects of the results presented in graph 2 (class overview, see Appendix A) at least once in think-aloud.

Fig. 3. Sequences of main steps per teacher
Note. A larger version of Fig. 3 with better resolution can be found on the OSF, https://osf.io/r24jd/.

complexity in reading and comprehending graphs based on building relationships and summarizing data points as well as on making inferences on data: Reading data (the lowest level) covers the extraction of directly given entities in graphs, e.g., noticing means or finished assessments. Reading between data (middle level) captures building relationships or summarizing data points, e.g. forming groups. Reading beyond data (the highest level) means summarizing the graph as a whole and making inferences. Noticing of group A and B can be considered to reflect a rather low level of complexity, whereas the noticing cognitions of group C reflects mainly a middle level of graph literacy.

3.3. Research question 3

To explore the typical processes in teachers' processing of formative assessment results, we mined a dependency graph based on a dependency matrix using process mining (see Fig. 5). We followed Weijters and Ribero (2011) and specified a complete model (including low-reliable dependency relations) and a simplified model (including only high-reliable dependency relations). The complete model (see Fig. 5), including low thresholds on frequency and performance, is considered more appropriate to the data in this study, because higher

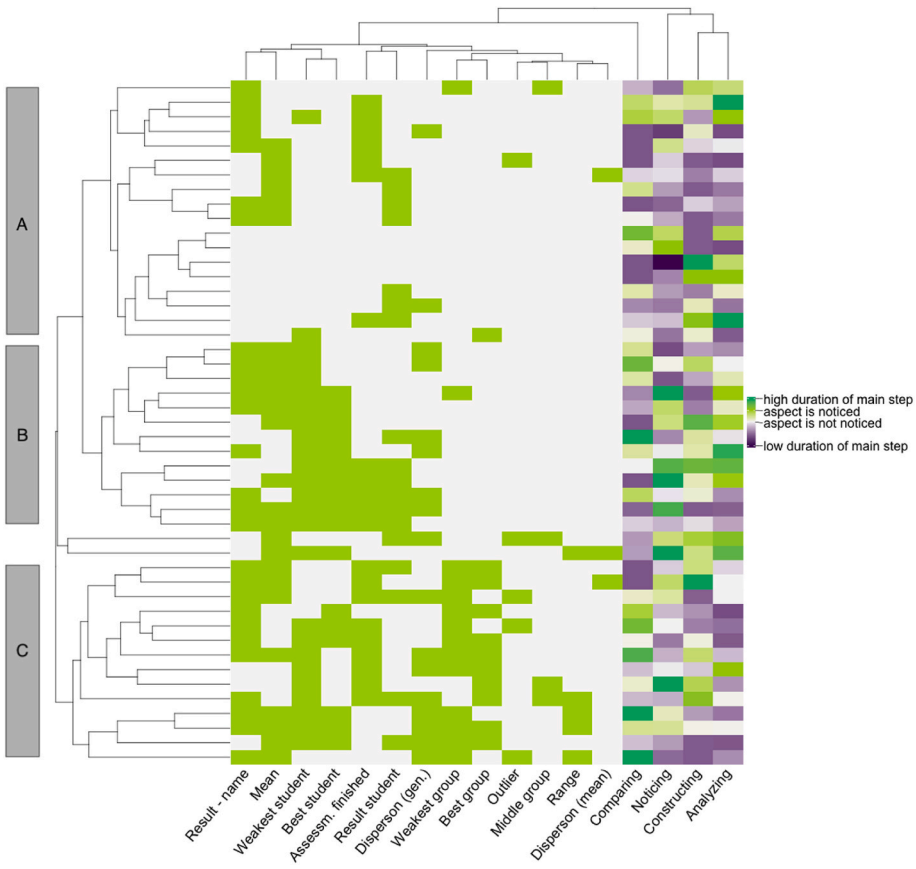


Fig. 4. Visualization of cluster analysis

Note. Heatmap and dendrograms using the average linkage algorithm based on Gower's similarity measure. Each row represents the data of one teacher. The gray boxes with letters on the left visualize groups (group A at the top, group B in the middle, and group C at the bottom).

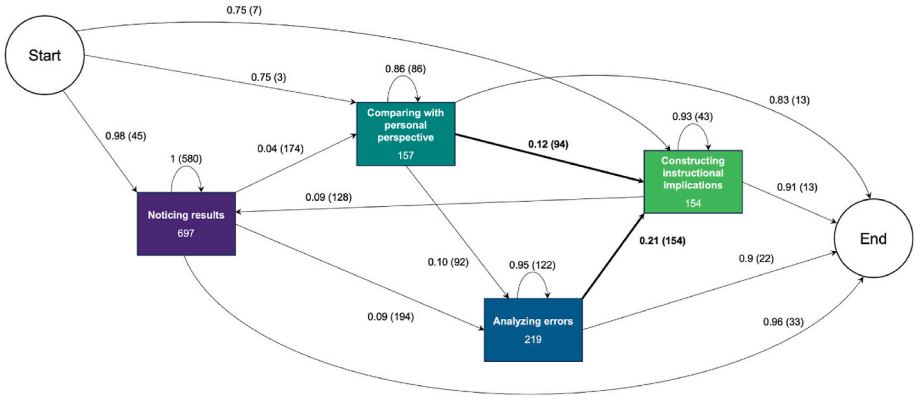


Fig. 5. Process model of processing

Note. The process model represented as a dependency graph including frequencies. The boxes represent event classes (steps of processing, such as *noticing results*), and the numbers in the boxes indicate the frequency of the step's occurrence. The arcs show the dependencies between the steps. The left numbers near the arcs are dependency measures (indicating the strength of the dependency, between −1, low, and 1, strong); the right numbers (in brackets) display the frequencies of the transitions.

default thresholds on dependency may lead to an oversimplified model (see reproducible documentation of data analysis for details, https://osf.io/r24jd/).

Generally, the model shows low dependency measures, except for length-one-loops. The latter indicates that codings of steps of processing tend to be followed by another coding event from the same step. Although dependency measures from the antecedents *analyzing errors* and *comparing the results with the teacher's personal perspective* to the consequent *constructing instructional implications* are low, the model shows high frequencies for the corresponding arcs and no arc from *noticing results* to *constructing instructional implications*. This can be interpreted as an indication that strictly noticing of assessment results is not sufficient for constructing instructional implications; rather, the analysis of errors in the results and the comparison of the assessment results with the personal perspective are important steps between *noticing results* and *constructing instructional implications*. The process model indicates a process in which information derived from formative assessment results is enriched by the teachers' own impressions as a

form of contextual information and knowledge as well as by elaborations of students' mistakes, e.g. misconceptions or motivational problems that may lead to a specific pattern of student responses, and, based on that, conclusions concerning instructional implications and actions are derived.

4. Discussion

4.1. Summary of findings, conclusions and practical implications

In this study, we explored how teachers process technology-based formative assessments for instruction in an ecologically valid setting with a strong focus on a process perspective using concurrent think-aloud. Concerning RQ 1, we found four main steps in the teachers' processing: *noticing results*, *comparing the results with the teacher's personal perspective*, *analyzing errors,* and *constructing instructional implications*. The relative durations of these steps in the think-aloud data varied substantially among the teachers. The step *comparing the results with the teacher's personal perspective* can be related to research focusing on beliefs and intuitions in data-based decision making of teachers, which has shown that data-based decision making can be affected by biases (e.g., Mandinach & Schildkamp, 2021; Vanlommel et al., 2017). This step also corresponds to the model developed by Mandinach and Gummer (2016), in which they highlight that data use for teaching means that data are combined, understood, and integrated with different domains of professional knowledge, including knowledge of educational contexts and the characteristics of students, as well as content, curriculum and pedagogical (content) knowledge. The latter corresponds to the step *analyzing errors*, which covers elaborations of students' mistakes involving corresponding domains of professional knowledge. To illustrate this, we give the following example on a conceptual level: A teacher notices that, on average, her class correctly solved 60% of the tasks provided by the technology-based formative assessment system (*noticing results*). She elaborates on what this result means in terms of average knowledge and competencies of the students, their misconceptions, patterns of errors, et cetera (*analyzing errors*). She considers this in relation to her own perspective (*comparing the results with the teacher's personal perspective*), for example, challenges the students experienced in past lessons, her students' characteristics, and the assessment situation, and formulates future instructional implications (*constructing instructional implications*), for example, revising prior knowledge in the class, addressing specific misconceptions, or adapting learning goals for specific students.

In this study, teachers addressed many different aspects in the exhaustive assessment results provided by the system. Specific aspects that were verbalized by most of the teachers at least once have in common that they can be directly noticed without complex cognitive elaborations. This corresponds with the results of other studies focusing on the graph literacy (Friel et al., 2001; Galesic & Garcia-Retamero, 2011), which showed low to middle levels of performance of teachers (e.g., Bez et al., 2021; Zeuch et al., 2017). Regarding RQ 2, cluster analyses using different algorithms indicate three groups of teachers who differed in terms of the complexity they displayed in summarizing and building relationships between single data points, which is consistent with the concept of graph literacy. Concerning practical implications of this finding, this could be taken as a starting point for developing adaptive support and training for teachers, for example, to help them to build groups of students for adaptive instruction based on the formative assessment results. Concerning RQ 3, we discovered a process model by applying the Flexible Heuristics Miner. The complete model (based on low thresholds for frequency and performance) reveals low dependency values in general, except for length-one-loops. Due to the antecedents *comparing the results with the teacher's personal perspective* and *analyzing errors* to the consequent *constructing instructional implications*, one can derive as a hypothesis that these steps play an important role for the step *constructing instructional implications*.

Conceptual models of data-based decision making consist of a sequential structure and differentiate noticing data, interpreting and transforming information to instructional decisions (see Section 1.3). This corresponds to the steps of processing and their sequences that were found in this study. In addition to the current conceptual models, the processing of many teachers show many recursions and iterations in their sequences and some teachers do not show complete sequences (see Fig. 3). If further research confirms this, then adding corresponding loops in the models may be reasonable to indicate that processing assessment results in reality might be more recursive, iterative and heterogeneous than 'linear-straightforward' sequential models imply.

For the practical development and improvement of technology-based formative assessment platforms, the findings of this study imply that dashboards displaying assessment results should be intuitive and not too complex, considering the complexity of noticing of most of the teachers. Moreover, since analyzing student errors and comparing results with the personal perspective seem to be important for constructing instructional implications, user interfaces might provide corresponding hints and support for teachers.

4.2. Limitations

Several limitations of this study arise due to the exploratory nature of the study and a potentially positive selected sample, which included only teachers who voluntarily use a specific technology-based formative assessment system. These limitations limit the generalizability of the findings, and all generated hypotheses need further confirmatory investigation in other contexts, for example, with teachers using different formative assessment platforms. Pivotal limitations of the study related to the think-aloud methodology lie in the prerequisite that it is not assumed that all thoughts of participants are captured (Fox et al., 2011). Furthermore, analyses based on a timed-event coding procedure in think-aloud must consider that think-aloud affects time (Fox et al., 2011), so time and durations have to be interpreted relatively. Due to this focus, we did not investigate the consistency or the plausibility of processing, in terms of the quality of derived instructional implications based on the previous noticing results, the error analysis or the teachers' comparisons with their personal perspectives. Concerning our cluster analyses, we aimed to achieve robust and validated results by using different algorithms and indices to compare different partitions, in addition to visual inspections and theoretical considerations. However, obtaining appropriate cluster analyses remains challenging in general (Kettenring, 2006), and although different algorithms indicate the groups and they are plausible from a theoretical point of view, these predictions might not correspond to stable and qualitative teacher-group differences. Regarding discovering a process model using process mining, one has to consider the low values in dependency measures in the resulting model. This could be due to noisy data or low-quality data coding or no causal dependencies of the corresponding steps of processing of the teachers. Further research could provide useful insights on the causes of the low values in the process model.

4.3. Further research and methodological reflections

As already implied in previous sections, further research should focus on processing relating to the consistency and plausibility as well as the interplay between domains of professional knowledge and processing. Furthermore, as already mentioned, insights into user cognitions in daily practice can provide valuable information and implications for further research in improving technology-based formative assessment platforms for teachers. For example, teachers may find having to click through several individual task results per student to find patterns of mistakes for individuals and groups of students, such as related to specific misconceptions, to be cognitively exhausting and time consuming, and technology could better facilitate this. We propose that assessment tasks be developed and implemented that are very good at diagnosing

students' misunderstandings or general difficulties in different areas. Based on that, technology could cluster groups of students according to their results and groups of tasks according to competence levels respectively knowledge domains and display this in the results dashboard. This may relieve teachers' processing for adaptive instruction because it would allow them to more easily capture which students are having difficulties in which domains. We propose clustering in this context because elaborating which students have (no) difficulties in which domains and which students in a class have similar results corresponds closely with the basic idea of classification based on multivariate data (Everitt et al., 2011; Kettenring, 2006), in terms of clustering objects (students) according to their similarities and dissimilarities in variable values (results in test items).

In this final section, methodological reflections and recommendations are outlined and discussed. Generally, we would argue that concurrent think-aloud is a valuable methodological approach for investigating cognitions in technology-rich settings with a focus on ecological validity. In our study, we conducted think-aloud sessions via video calls rather than in a laboratory setting with the aim of reducing the time and effort required from the participants and reaching them in their daily environment. However, further investigation is needed to understand the potential differences in remote think-aloud sessions versus face-to-face think-aloud sessions. Concerning the coding of think-aloud data, it is evident that the chosen coding approach determines further analyses: In some cases, 'plain coding' without taking time into consideration and analyzing the resulting frequencies can be sufficient. The timed-event coding of videotaped (or audiotaped) think-aloud sessions requires more resources and time and can be challenging in terms of achieving adequate reliability, in our experience. However, the timed-event coding approach in combination with new methodological approaches and developments in data science, for example, process mining as it is applied in this study, enables the maintenance of a sequential structure and to investigate sequential processes in cognition in educational contexts. Process mining can be used for discovery, as we did in this study, but also conformance checking and enhancement as other types of process mining (van der Aalst, 2016) are worth considering in future research in learning and instruction.

CRediT authorship contribution statement

Sarah Bez: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Fabian Burkart:** Data curation, Formal analysis, Validation, Visualization, Writing – review & editing. **Martin J. Tomasik:** Conceptualization, Funding acquisition, Investigation, Methodology, Resources, Supervision, Validation, Writing – review & editing. **Samuel Merk:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Resources, Software, Validation, Supervision, Writing – review & editing.

Research data availability statement

The data and the materials of this study as well as the reproducible documentation of data analysis (RDA) are provided on OSF, see https://osf.io/r24jd/.

Funding sources

This study was funded by the Suzanne and Hans Biäsch Foundation.

Declaration of competing interest

The authors declare no conflicts of interest.

Appendix A

Supplementary materials to this article can be found online at https://doi.org/10.1016/j.learninstruc.2025.102100.

References

Adachi, C., Tai, J. H.-M., & Dawson, P. (2018). Academics' perceptions of the benefits and challenges of self and peer assessment in higher education. *Assessment & Evaluation in Higher Education, 43*(2), 294–306. https://doi.org/10.1080/02602938.2017.1339775

Bakeman, R., & Quera, V. (2011). *Sequential analysis and observational methods for the Behavioral Sciences*. Cambridge University Press.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Berger, S., Verschoor, A. J., Eggen, T. J. H. M., & Moser, U. (2019). Development and validation of a vertical scale for formative assessment in mathematics. *Frontiers in Education, 4*. https://doi.org/10.3389/feduc.2019.00103

Bez, S., Poindl, S., Bohl, T., & Merk, S. (2021). Wie werden Rückmeldungen von Vergleichsarbeiten rezipiert? Ergebnisse zweier Think-Aloud-Studien [How are results of statewide assessments perceived? Findings from two think-aloud studies]. *Zeitschrift für Pädagogik, 67*(4), 551–572. https://doi.org/10.3262/ZP2104551

Bez, S., Tomasik, M. J., & Merk, S. (2023). Data-based decision making in einer digitalen Welt: Data Literacy von Lehrpersonen als notwendige Voraussetzung [Data-based decision making in a digital world: Data literacy of teachers as a necessary prerequisite]. In K. Scheiter, & I. Gogolin (Eds.), *Bildung für eine digitale Zukunft* (pp. 339–362). Springer VS. https://doi.org/10.1007/978-3-658-37895-0_14.

Birenbaum, M., DeLuca, C., Earl, L., Heritage, M., Klenowski, V., Looney, A., Smith, K., Timperley, H., Volante, L., & Wyatt-Smith, C. (2015). International trends in the implementation of assessment for learning: Implications for policy and practice. *Policy Futures in Education, 13*(1), 117–140. https://doi.org/10.1177/1478210314566733

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom sssessment. *Phi Delta Kappan, 80*(2), 139–144. https://doi.org/10.1177/003172171009200119

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice, 25*(6), 551–575. https://doi.org/10.1080/0969594X.2018.1441807

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies. *Zeitschrift für Psychologie, 223*(1), 3–13. https://doi.org/10.1027/2151-2604/a000194

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement: Interdisciplinary Research and Perspectives, 9*(4), 173–206. https://doi.org/10.1080/15366367.2011.626729

Conole, G., & Warburton, B. (2005). A review of computer-assisted assessment. *Research in Learning Technology, 13*(1), Article 1. https://doi.org/10.3402/rlt.v13i1.10970

Corno, L. (2008). On teaching adaptively. *Educational Psychologist, 43*(3), 161–173. https://doi.org/10.1080/00461520802178466

Dunn, K., & Mulvenon, S. (2009). A critical review of research on formative assessment: The limited scientific evidence of the impact of formative assessment in education. *Practical Assessment, Research and Evaluation, 14*(7). https://doi.org/10.7275/jg4h-rb87

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture and Activity, 5*(3), 178–186. https://doi.org/10.1207/s15327884mca0503_3

Espin, C. A., Wayman, M. M., Deno, S. L., McMaster, K. L., & Rooij, M. de (2017). Data-based decision-making: Developing a method for capturing teachers' understanding of CBM graphs. *Learning Disabilities Research & Practice, 32*(1), 8–21. https://doi.org/10.1111/ldrp.12123

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470977811

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

Faber, J. M., & Visscher, A. J. (2018). The effects of a digital formative assessment tool on spelling achievement: Results of a randomized experiment. *Computers & Education, 122*, 1–8. https://doi.org/10.1016/j.compedu.2018.03.008

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin, 137*(2), 316–344. https://doi.org/10.1037/a0021663

Frey, A., Liu, T., Fink, A., & König, C. (2024). Meta-analysis of the effects of computerized adaptive testing on the motivation and emotion of examinees. *European Journal of Psychological Assessment, 40*(5), 427–443. https://doi.org/10.1027/1015-5759/a000821

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education, 32*(2), 124–158. https://doi.org/10.2307/749671

Galesic, M., & Garcia-Retamero, R. (2011). Graph literacy: A cross-cultural comparison. *Medical Decision Making, 31*(3), 444–457. https://doi.org/10.1177/0272989X10373805

Goffin, E., Janssen, R., & Vanhoof, J. (2022). Teachers' and school leaders' sensemaking of formal achievement data: A conceptual review. *The Review of Education, 10*(1), e3334. https://doi.org/10.1002/rev3.3334

Goffin, E., Janssen, R., & Vanhoof, J. (2023). Principals' and teachers' comprehension of school performance feedback reports. Exploring misconceptions from a user validity perspective. *Pedagogische Studiën, 100*(1), Article 1. https://doi.org/10.59302/ps.v100i1.13991

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics, 27*(4), 857–871. https://doi.org/10.2307/2528823

Hall, T. E., Cohen, N., Vue, G., & Ganley, P. (2015). Addressing learning disabilities with UDL and technology: Strategic Reader. *Learning Disability Quarterly, 38*(2), 72–83. https://doi.org/10.1177/0731948714544375

Hardy, I., Decristan, J., & Klieme, E. (2019). Adaptive teaching in research on learning and instruction. *Journal for Educational Research Online, 11*(2). https://doi.org/10.25656/01:18004

Hartmann, C., Rummel, N., & Bannert, M. (2022). Using HeuristicsMiner to analyze problem-solving processes: Exemplary use case of a productive-failure study. *Journal of Learning Analytics, 9*(2), Article 2. https://doi.org/10.18608/jla.2022.7363

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77–89.

Hebbecker, K., Förster, N., Forthmann, B., & Souvignier, E. (2022). Data-based decision-making in schools: Examining the process and effects of teacher support. *Journal of Educational Psychology, 114*(7), 1695–1721. https://doi.org/10.1037/edu0000530

Heitink, M. C., van der Kleij, F. M., Veldkamp, B. P., Schildkamp, K., & Kippers, W. B. (2016). A systematic review of prerequisites for implementing assessment for learning in classroom practice. *Educational Research Review, 17*, 50–62. https://doi.org/10.1016/j.edurev.2015.12.002

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*(1), 193–218. https://doi.org/10.1007/BF01908075

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. https://doi.org/10.1111/jedm.12000

Kaufman, L., & Rousseeuw, P. J. (2005). *Finding groups in data. An introduction to cluster analysis*. John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470316801

Kettenring, J. R. (2006). The practice of cluster analysis. *Journal of Classification, 23*(1), 3–30. https://doi.org/10.1007/s00357-006-0002-6

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28–37. https://doi.org/10.1111/j.1745-3992.2011.00220.x

Kippers, W. B., Poortman, C. L., Schildkamp, K., & Visscher, A. J. (2018). Data literacy: What do educators learn and struggle with during a data use intervention? *Studies In Educational Evaluation, 56*, 21–31. https://doi.org/10.1016/j.stueduc.2017.11.001

Kluger, A., & 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*, 254–284. https://doi.org/10.1037/0033-2909.119.2.254

Lee, H., Chung, H. Q., Zhang, Y., Abedi, J., & Warschauer, M. (2020). The effectiveness and features of formative assessment in us K-12 education: A systematic review. *Applied Measurement in Education, 33*(2), 124–140. https://doi.org/10.1080/08957347.2020.1732383

Leighton, J. P. (2017). *Using think-aloud interviews and cognitive labs in educational research*. Oxford University Press.

Maier, U., Wolf, N., & Randler, C. (2016). Effects of a computer-assisted formative assessment intervention based on multiple-tier diagnostic items and different feedback types. *Computers & Education, 95*, 85–98. https://doi.org/10.1016/j.compedu.2015.12.002

Mandinach, E. B., & Gummer, E. S. (2016). What does it mean for teachers to be data literate: Laying out the skills, knowledge, and dispositions. *Teaching and Teacher Education, 60*, 366–376. https://doi.org/10.1016/j.tate.2016.07.011

Mandinach, E. B., & Schildkamp, K. (2021). Misconceptions about data-based decision making in education: An exploration of the literature. *Studies In Educational Evaluation, 69*, Article 100842. https://doi.org/10.1016/j.stueduc.2020.100842

Mannhardt, F., & Janssenswillen, G.. *heuristicsmineR* (Version 0.3.0) [computer software]. https://CRAN.R-project.org/package=heuristicsmineR.

Marsh, J. (2012). Interventions promoting educators' use of data: Research insights and gaps. *Teachers College Record, 114*, 1–48.

McLaughlin, T., & Yan, Z. (2017). Diverse delivery methods and strong psychological benefits: A review of online formative assessment. *Journal of Computer Assisted Learning, 33*(6), 562–574. https://doi.org/10.1111/jcal.12200

McMillan, J. H., Venable, J. C., & Varier, D. (2013). Studies of the effect of formative assessment on student achievement: So much more is needed. *Practical Assessment, Research and Evaluation, 18*(2). https://doi.org/10.7275/TMWM-7792

Padilla, J.-L., & Leighton, J. P. (2017). Cognitive interviewing and think aloud methods. In B. D. Zumbo, & A. M. Hubley (Eds.), *Understanding and investigating response processes in validation research* (pp. 211–228). Springer.

Parry, D. A., Davidson, B. I., Sewall, C. J. R., Fisher, J. T., Mieczkowski, H., & Quintana, D. S. (2021). A systematic review and meta-analysis of discrepancies between logged and self-reported digital media use. *Nature Human Behaviour, 5*(11), 1535–1547. https://doi.org/10.1038/s41562-021-01117-5

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measurement of personality and social psychological attitudes* (pp. 17–59). Academic Press. https://doi.org/10.1016/B978-0-12-590241-0.50006-X.

R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Reimann, P. (2009). Time is precious: Variable- and event-centred approaches to process analysis in CSCL research. *International Journal of Computer-Supported Collaborative Learning, 4*(3), 239–257. https://doi.org/10.1007/s11412-009-9070-z

Russell, M. K. (2010). Technology-aided formative assessment of learning: New developments and applications. In H. Andrade, & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 125–138). Routledge.

Schildkamp, K. (2019). Data-based decision-making for school improvement: Research insights and gaps. *Educational Research, 61*(3), 257–273. https://doi.org/10.1080/00131881.2019.1625716

Schildkamp, K., van der Kleij, F. M., Heitink, M. C., Kippers, W. B., & Veldkamp, B. P. (2020). Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. *International Journal of Educational Research, 103*, Article 101602. https://doi.org/10.1016/j.ijer.2020.101602

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

Sonnenberg, C., & Bannert, M. (2019). Using Process Mining to examine the sustainability of instructional support: How stable are the effects of metacognitive prompting on self-regulatory behavior? *Computers in Human Behavior, 96*, 259–272. https://doi.org/10.1016/j.chb.2018.06.003

Spector, J. M., Ifenthaler, D., Sampson, D., Yang, L. J., Mukama, E., Warusavitarana, A., Dona, K. L., Eichhorn, K., Fluck, A., Huang, R., Bridges, S., Lu, J., Ren, Y., Gui, X., Deneen, C. C., Diego, J. S., & Gibson, D. C. (2016). Technology enhanced formative assessment for 21st century learning. *Journal of Educational Technology & Society, 19*(3), 58–71.

Tomasik, M. J., Berger, S., & Moser, U. (2018). On the development of a computer-based tool for formative student assessment: Epistemological, methodological, and practical issues. *Frontiers in Psychology, 9*, 1–17. https://doi.org/10.3389/fpsyg.2018.02245

van den Bosch, R. M., Espin, C. A., Chung, S., & Saab, N. (2017). Data-based decision-making: Teachers' comprehension of curriculum-based measurement progress-monitoring graphs. *Learning Disabilities Research & Practice, 32*(1), 46–60. https://doi.org/10.1111/ldrp.12122

van der Aalst, W. M. P. (2016). Process mining. *Data science in action* (2nd ed.). Springer. https://doi.org/10.1007/978-3-662-49851-4

Vanlommel, K., Van Gasse, R., Vanhoof, J., & Van Petegem, P. (2017). Teachers' decision-making: Data based or intuition driven? *International Journal of Educational Research, 83*, 75–83. https://doi.org/10.1016/j.ijer.2017.02.013

VERBI software. (2021). *MAXQDA 2022 [computer software]*. Berlin, Germany: VERBI Software. www.maxqda.com.

Wagner, D. L., Hammerschmidt-Snidarich, S. M., Espin, C. A., Seifert, K., & McMaster, K. L. (2017). Pre-service teachers' interpretation of CBM progress monitoring data. *Learning Disabilities Research & Practice, 32*(1), 22–31. https://doi.org/10.1111/ldrp.12125

Weijters, A. J. M. M., & Ribeiro, J. T. S. (2011). Flexible Heuristics miner (FHM). *IEEE symposium on computational intelligence and data mining (CIDM)*. https://doi.org/10.1109/CIDM.2011.5949453

Xuan, Q., Cheung, A., & Sun, D. (2022). The effectiveness of formative assessment for enhancing reading achievement in K-12 classrooms: A meta-analysis. *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.990196

Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., & Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assessment in Education: Principles, Policy & Practice, 28*(3), 228–260. https://doi.org/10.1080/0969594X.2021.1884042

Zeuch, N., Förster, N., & Souvignier, E. (2017). Assessing teachers' competencies to read and interpret graphs from learning progress assessment: Results from tests and interviews. *Learning Disabilities Research & Practice, 32*(1), 61–70. https://doi.org/10.1111/ldrp.12126