Borter, Natalie

Differential effects of additional formative assessments on student learning behaviors and outcomes

Studia paedagogica. 2023, vol. 28, iss. 3, pp. [9]-38

ISSN 1803-7437 (print); ISSN 2336-4521 (online)

Stable URL (DOI): https://doi.org/10.5817/SP2023-3-1
Stable URL (handle): https://hdl.handle.net/11222.digilib/digilib.79677
Access Date: 29. 03. 2024
Version: 20240320

STUDIA PAEDAGOGICA

vol. 28, n. 3, 2023
https://doi.org/10.5817/SP2023-3-1

DIFFERENTIAL EFFECTS OF ADDITIONAL FORMATIVE ASSESSMENTS ON STUDENT LEARNING BEHAVIORS AND OUTCOMES

Natalie Borter[a,b]

[a] Institute for Psychology, University of Bern, Switzerland
[b] School of Business, HSLU Luzern, Switzerland

ABSTRACT

It is well-established that formative assessments with accompanying feedback can enhance learning. However, the degree to which additional formative assessments on the same material further improve learning outcomes remains an open research question. Moreover, it is unclear whether providing additional formative assessments impacts self-regulated learning behavior, and if the benefits of such assessments depend on students' self-regulated learning behavior. The current study, conducted in a real-world blended learning setting and using a Learning Analytics approach, compares 154 students who completed additional formative assessments with 154 students who did not. The results indicate that the additional formative assessments led to an improvement in learning outcomes, but also had both positive and negative effects on students' self-regulated learning behavior. Students who completed additional formative assessments performed better on the assessments but reported lower levels of subjective comprehension and devoted more time to completing exercises. Simultaneously, they devoted less effort to additional learning activities (additional investment), such as class preparation and post-processing. Furthermore, the impact of additional formative assessments on learning success depended on students' self-regulated learning behavior. It was primarily the students who invested above-average time during formative assessments (time investment) who benefited from the additional exercises. Cluster analysis revealed that high-effort students (those with above-average time investment and above-average additional investment) gained the most from the extra exercises. In contrast, low-effort students and those who achieved high performance with relatively low effort (efficient students) did not benefit from additional formative assessments. In conclusion, providing students with additional formative assessments can enhance learning, but it should be done with caution as it can alter self-regulated learning behavior in both positive and negative ways, and not all students may benefit from it equally.

KEYWORDS

(indirect) testing effect; formative assessments; feedback; self-regulated learning behavior; individual differences

CORRESPONDING AUTHOR
Natalie Borter, Institute for Psychology, University of Bern, Fabrikstrasse 8, 3012 Bern, Switzerland
e-mail: natalie.borter@unibe.ch

Introduction

The testing effect, a well-established learning technique (Jensen et al., 2020), denotes the enhanced learning success observed when students actively engage with learned material, rather than relying on passive repetition or memorization (Schwieren et al., 2017). The effect is typically quantified by comparing the post-learning performance of learners who participated in active retrieval during the learning phase to those who did not engage in such practices. An instance of active retrieval is solving formative assessments, wherein learners apply their knowledge to solve problems (Boston, 2002).

A significant testing effect was demonstrated in both experimental and applied settings (Lamotte et al., 2021; Schwieren et al., 2017). It is evident in tasks ranging from relatively simple ones, like vocabulary memorization, to more complex tasks that involve applying theoretical knowledge to novel situations (Schwieren et al., 2017). The effect occurs when exercises during the learning phase are identical to those measuring learning success (Carpenter, 2011; Eriksson, et al., 2011; Karpicke & Roediger, 2007) and when non-identical exercises covering the same material are employed (Batsell et al., 2017; Foss & Pirozzolo, 2017; Francis et al., 2020; Jensen et al., 2020; McDaniel et al., 2013).

One open research question regarding the testing effect concerns whether multiple testing instances result in greater learning success than administering a single test. Multiple testing can include either the repeated use of the same test or the utilization of different tests covering the same content (Yang et al., 2021). A meta-analytic review conducted by Adesope et al. (2017) did not identify a significant difference between the testing effects of multiple tests versus a single test on the same material. However, a more recent review by Yang et al. (2021) discovered that, in applied contexts such as classrooms, the testing effect was more pronounced when the same test (or a similar test with the same content) was administered repeatedly. Additional empirical research is required to determine whether multiple tests on the same content yield a more pronounced testing effect (Yang et al., 2021).

Assessing the effectiveness of additional exercises on identical content is crucial for practical implementation, particularly given the laborious task of generating such exercises. Moreover, the provision of additional learning materials may not always lead to increased learning, and can even have a negative impact on learning in some cases by inducing cognitive overload or stress (Kossen & Ooi, 2021).

As additional formative assessments on the same content have the potential to influence self-regulated learning both positively and negatively, a learning analytics approach was employed to investigate the impact of additional formative assessments on self-regulated learning behavior. This approach involves the collection and analysis of data on students' learning behavior and progress, to enhance learning and teaching (Chatti et al., 2013; Ifenthaler, 2015; Leitner et al., 2017). The utilization of such data has attracted attention in the field of self-regulated learning, as it enables the monitoring of the learner's holistic action without interference in the process (Winne & Baker, 2013).

Prior research suggests that formative assessments have a positive effect on self-regulated study time allocation and monitoring (Clariana & Park, 2021; Fernandez & Jamet, 2017; Perry & Winne, 2006; Soderstrom & Bjork, 2014; Yang et al., 2017). Engagement in solving exercises enhanced students' monitoring of their knowledge, leading to a reduction in the overestimation of their knowledge and an increase in time allocated for studying (Soderstrom & Bjork, 2014). The positive impact of exercise solving on learning success was partially mediated by improved monitoring and learning behavior (Fernandez & Jamet, 2017). This is because additional formative assessments can provide feedback on learning status, aiding learners in metacognitive control, and adapting their self-regulated learning behavior (Clariana & Park, 2021; Perry & Winne, 2006).

Furthermore, the efficacy of additional formative assessments likely depends on individual differences in students' characteristics (Bertilsson et al., 2021). Prior knowledge and experience are also important factors, as students with more prior knowledge tend to benefit more from the testing effect than those with less prior knowledge (Cogliano, et al., 2019; Francis et al., 2020). According to the elaborative retrieval hypothesis, mental effort during recall predicts the magnitude of the testing effect, and practicing with exercises that are challenging but within the learner's abilities can enhance the effect (Carpenter, et al., 2009; Greving et al., 2020; Minear et al., 2018). In addition, students who already possess effective learning strategies may not benefit as much from formative assessments as those who lack such strategies (Robey, 2019). This is because they have already achieved high learning success even without formative assessments, and the effect of testing may not significantly enhance their learning outcomes.

Not only the number but also the timing of formative assessments can influence their effectiveness (Karpicke & Bauernschmidt, 2011). In line with current research on spaced learning (Greene, 2008; Jost et al., 2021), evenly distributed learning throughout the semester is more effective in promoting learning success than cramming right before a test (Adesope et al., 2017; Karpicke & Bauernschmidt, 2011).

In summary, the positive impact of testing on learning can be influenced by multiple factors, including individual student characteristics (e.g., hope of success, prior knowledge, investment, timing). It is plausible to assume that students with limited prior knowledge, low motivation, low investment, frequent incorrect responses, last-minute study habits, and shallow feedback processing may not benefit as much from extra exercises with feedback as other students. Therefore, various types of students may exist, and some may not experience the same benefits from additional formative assessments.

In this study, a clustering based on the data collected through the learning analytics approach was employed to examine whether the impact of solving additional formative assessments on end-of-semester knowledge test performance is influenced by student self-regulated learning behavior. Cluster analysis is a method that divides students into groups based on similarities within a cluster and dissimilarities between clusters (Dalmaijer et al., 2021; Shin & Shim, 2021).

To sum up, although there is significant evidence supporting the idea that formative assessments improve learning outcomes, it is still uncertain whether administering additional assessments on the same content leads to a more substantial effect and how this practice influences self-regulated learning behavior. Additionally, studies indicate that the benefits of additional formative assessments may differ based on students' self-regulated learning characteristics. As a result, this study aims to examine the following three hypotheses:

1. Administering additional formative assessments of the same content leads to higher performance on an end-of-course knowledge test.
2. Administering additional formative assessments of the same content influences self-regulated learning behavior.
3. The relationship between administering additional formative assessments and learning success depends on students' self-regulated learning behavior.

1 Method

1.1 Participants

To be included in the study, students had to take both the prior knowledge test at the beginning of the semester and the end-of-semester knowledge test. They also had to complete at least four of the five formative assessments (not including the additional formative assessments). In addition, students with very low performance in the formative assessments or in the end-of--semester knowledge test (clear outliners) were excluded (three students). Of the 276 students enrolled in the course in 2020, a total of 194 met the inclusion criteria, while in 2021 a total of 166 of the 234 students enrolled met the inclusion criteria. This represents approximately 70% of all enrolled students in both courses. Out of the total 360 students evaluated (194 + 166), 324 had no missing values and 36 students were missing one of the five formative assessments. To address this, missing values for the 36 students (21 from 2020's 194 and 15 from 2021's 166) were imputed using the mice function (van Burren & Groothuis-Oudshoorn, 2011) employing a predictive mean matching approach.

The aim of the current study was to compare students solving most additional exercises with students not solving any additional exercises. Thus, students without access to any extra exercises were labeled "non-solvers" while those who finished at least four out of the five additional assessments were referred to as "solvers". From the 2021 cohort, 12 students who completed fewer than four additional formative assessments were excluded. Accordingly, 194 students were identified as "non-solvers" and 154 as "solvers". The study was approved by the local ethics committee.

1.2 Procedure

The mandatory "Psychological Diagnostics" course for master's students in psychology focused on complex methodological content such as equivalence analysis and item response theory. Students were permitted to choose the learning approach they found most suitable for the specific learning situation, as their learning behavior was not explicitly manipulated. Consequently, this study examined the impact of extra exercises compared to any other learning approach, including no learning at all.

Due to ethical concerns, students were not randomly assigned to groups. Instead, this study adopted a quasi-experimental approach, evaluating the same course across two successive years, 2020 and 2021. The only variation introduced between the two years was the inclusion of optional additional formative assessments in 2021. The additional formative assessment consisted of new exercises covering the same content as the initial assessment. All other elements of the course, including initial exercises, podcasts, literature, and instructions, were kept consistent across both years. Participation in the study was voluntary. All students had access to the standard learning materials. However, those who volunteered to participate in the study received additional feedback post-exam: z-standardized values of all variables specified in the method section, enabling them to compare their performance with that of the entire group. Non-participants did not receive this supplementary information. All data were pseudo-anonymized using unique pseudonyms. Based on the pseudonyms, there were no students who attended the course in both years, 2020 and 2021.

A blended learning approach was employed in both years, allowing students to engage with course material at their own pace and participate in timely online discussions. The curriculum included 12 weekly podcast lectures, each 90 minutes long, a prior knowledge test, and five biweekly formative assessments covering two lectures each. Data collection occurred during the formative assessments.

Upon completing each exercise, students received immediate feedback and suggestions for supplementary resources, including relevant literature, lecture slides, podcast excerpts, and additional links or references. Students could ask further questions on the feedback page, which were addressed in a forum or, if necessary, through scheduled online discussions.

Students were advised to adhere to a one-week submission window for formative assessments, facilitating an even distribution of their learning throughout the semester. This structure was consistent over both years; however, the frequency of formative assessments varied. In 2020, formative assessments were assigned every two weeks, whereas in 2021, with the addition of supplementary formative assessments, they occurred weekly. Despite the change in frequency, the exercises, including the additional assessments, could be repeated and remained accessible to students until the final exam.

Two weeks before the final exam, a comprehensive end-of-semester knowledge test with new exercises covering the entire course content was administered. All exercises and self-reports were designed and hosted using Qualtrics (Qualtrics, Provo, UT). Response latency, accuracy, and time spent on feedback pages were assessed. Links to the Qualtrics questionnaires were embedded into the ILIAS learning management system, where all essential learning resources, such as podcasts and literature, were made available to students.

To identify clusters of response behavior, data from the initial formative assessments were utilized. Additional formative assessments were not included because they were only accessible to the solvers.

In the next section, the behavioral and self-reported learning analytics data gathered are presented. Self-reports were used to capture perceptions such as subjective knowledge, subjective investment, and subjective importance.

Furthermore, for variables over which I did not have full control, as I intentionally permitted students to learn in ways they found most suitable for the specific learning situation, such as downloading or printing materials; self-reports were employed. The drawbacks of self-reports were minimized by employing pseudonymization to reduce the impact of social desirability and by using precise questions to decrease the likelihood of recall errors. For our main emphasis, the formative assessments, offline solutions were not allowed, so behavioral data were utilized.

1.3 Behavioral data

Prior knowledge

Prior knowledge was assessed with 19 multiple-choice exercises. The test contained mainly theoretical exercises and calculations concerning real-world applications of the knowledge acquired during the bachelor's program (e.g., reliability, validity). One sum score was built for the 13 exercises covering theoretical exercises and one for the six exercises covering calculations.

Performance in the formative assessments

For each of the five formative assessments, covering different content such as item response theory, confirmatory factor analysis, equivalence analysis, and criterion-referenced testing, a sum score was built. The number of exercises per formative assessment ranged from 10 to 24. The exercises consisted primarily of multiple-choice exercises, in which the theoretical knowledge acquired from the podcast was applied to concrete situations. The sum scores of the five assessments were highly related, with a Cronbach's alpha of 0.76. The sum of all five formative assessments was used for further analyses.

End-of-semester knowledge test

The dependent variable of the current study was the performance in the end-of-semester knowledge test. It consisted of 22 exercises. In contrast to the exam, the knowledge test covered only the content of the formative assessments and was identical in 2020 and 2021. The correlation between the end-of-semester knowledge test and the final grades was comparable in both cohorts – 2020 ($r = 0.52, p < 0.05$) and 2021 ($r = 0.57, p < 0.03$).

Time investment

Response latency was recorded for each task and the feedback page. To reduce the effect of strong outliers, for each time measure, all values greater than the 95th percentile were trimmed to the 95th percentile. As the response latencies were still highly right-skewed, each time measure was logarithmized. Thereafter, all response latencies were z-standardized and the first strong principal component of the response latencies on the exercise page (explaining 43% of the variance), and the first strong principal component of the response latencies on the feedback page (explaining 46% of the variance) were extracted and used for further analyses.

Number of completions

The "completions initial exercises" variable was computed for the initial formative assessments, considering the count of completions for both identical and distinct formative assessments. For the variable "completions overall" the total number of completions (also including the additional formative assessments) was calculated. The number of completions overall was categorized into six groups: 1 = up to ten completions; 2 = 11–15 completions; 3 = 16–20 completions; 4 = 21–25 completions; 5 = 26–30 completions; 6 = more than 30 completions.

Questions for the forum

Across all exercises of the formative assessments, the frequency with which questions were posed by students on the feedback page was recorded. This variable was highly right-skewed, and therefore the values were logarithmized.

On time / regularity

As a measure of regularity, it was counted how often students finished the formative assessments during the recommended one-week submission window.

1.4 Self-reported data

Subjective knowledge

At the beginning of each formative assessment, students rated their subjective understanding of the content covered in the respective exercise session on a five-point scale (1 = I don't know this concept, 2 = I don't understand this concept well, 3 = I understand this concept less well, 4 = I understand this concept well, 5 = I understand this concept very well). First, the average of these ratings was taken for each formative assessment, and then the first strong principal component (explaining 65% of the variance) was extracted from the averaged ratings across all five formative assessments (excluding the additional formative assessment) and used for further analyses.

Subjective investment

After each formative assessment, students rated on a four-point scale their effort level in attempting to complete the exercises to the best of their ability (1 = I didn't try hard, 2 = I tried a little, 3 = I tried a lot, 4 = I tried hard). The average of these ratings was calculated and used for further analyses.

Lectures

At the beginning of each formative assessment, students indicated whether they had listened to the podcasts of the two lectures covered in the formative assessment (1 = I listened to neither of the two podcasts, 2 = I listened to parts of both podcasts, 3 = I listened to at least one of the podcasts completely, 4 = Yes, I listened to both podcasts completely). The mean value of this variable, computed across the five formative assessments, was utilized for subsequent analyses.

Reading forum

At the beginning of the end-of-semester knowledge test, students indicated on a three-point scale whether they had read the forum posts before (0 = I never read the forum, 1 = I read the forum only when I had questions, 2 = I read all forum posts at least once).

Compulsory literature

At the outset of each formative assessment, students specified their engagement with the mandatory literature, which, in combination with lectures, served as the foundational preparation for the assessment: (1) indicated they read at least some part of the mandatory literature, while (0) denoted they did not engage with it.

The mean value of this variable, computed across the five formative assessments, was utilized for subsequent analyses.

Relevance of content

On a four-point scale (false, somewhat false, somewhat true, true) students responded to the following questions about the content of the course:

- I find "Psychological Diagnostics" interesting.
- I think my knowledge of "Psychological Diagnostics" will be useful to me in the future.
- I think it is important to learn "Psychological Diagnostics" in psychology education.

The average of the three items was used for further analyses.

Learning hours during semester holidays

The students reported the number of hours they dedicated to studying for the exam following the final lecture of the semester. As data were highly right-skewed, they were logarithmized.

2 Results

All analyses were conducted in R version 3.6.1 (R Core Team, 2021).

2.1 Descriptive statistics

Given the quasi-experimental design of the study, it was crucial to establish that there were no initial differences between the students from 2020 and 2021 in terms of "prior knowledge" and "subjective relevance of the content" at the beginning of the course. To compare the means for these measures, an equivalence analysis was conducted (Bentler & Satorra, 2010). For "prior knowledge," a two-factor solution (theory and calculations) was compared to a one-factor solution. The significant Chi-square difference ($\Delta\chi^2(1) = 67.70$, $p < 0.001$) indicated that the two-factor model ($\chi^2(151) = 178.43$, $p = 0.063$, CFI = 0.934, RMSEA = 0.022, SRMR = 0.047) provided a better fit to the data than the one-factor model ($\chi^2(152) = 237.07$, $p < 0.001$, CFI = 0.796, RMSEA = 0.039, SRMR = 0.055). Consequently, prior knowledge is more accurately represented by a two-factor solution. The two factors, theory and calculations, were correlated ($r = 0.53$, $p < 0.01$).

A measurement invariance analysis using lavaan (Rosseel, 2012) confirmed scalar equivalence (configural vs. metric fit: $\Delta\chi^2(17) = 17.34$, $p = 0.43$; scalar vs. metric fit: $\Delta\chi^2(17) = 16.72$, $p = 0.47$; scalar model fit $\chi^2(336) = 367.15$, $p = 0.12$, CFI = 0.926, RMSEA = 0.023, SRMR = 0.066), allowing for comparison of the means between the two groups (2020 vs. 2021 course). Accordingly, prior knowledge in calculations was measured using the sum score of all items loading on the calculations factor, while the sum score of all items loading on the theory factor was employed as a measure of theoretical prior knowledge.

Non-solvers differed from solvers in both scales of prior knowledge, calculations ($t(358) = -2.14$, $p < 0.05$; non-solvers: $M = 4.62$, $SD = 1.45$; solvers: $M = 4.98$, $SD = 1.20$,), and theory ($t(286.03) = -2.15$, $p < 0.05$; non-solvers: $M = 7.77$, $SD = 1.39$; solvers: $M = 8.14$, $SD = 1.76$) as well as in the subjective relevance of the content ($t(325.37) = -2.30$, $p < -0.05$; non-solvers: $M = 2.96$, $SD = 0.64$; solvers: $M = 3.12$, $SD = 0.66$). To ensure comparability of prior knowledge and subjective relevance of the content between solvers and non-solvers, a matching approach was employed. The matching was conducted using the function matchit from the MatchIt package, with a nearest neighbor method, distance logit, and an "ATT" estimate (Pishgar et al., 2021). The 194 "non-solvers" were matched to the 154 "solvers". The matched samples, each consisting of 154 students, did not differ in the prior knowledge scale calculations (non-solvers: $M = 4.86$, $SD = 1.35$ versus solvers: $M = 4.98$, $SD = 1.20$, $p = 0.40$) and theory (non-solvers: $M = 7.97$, $SD = 1.39$ versus solvers: $M = 8.14$, $SD = 1.76$, $p = 0.34$) nor in subjective relevance

($t(325.37) = -2.30, p = .67$; non-solvers: $M = 3.09, SD = 0.59$; solvers: $M = 3.12$, $SD = 0.66$). Subsequent analyses were carried out exclusively on the matched samples.

In Table 1, mean (standard deviation), skewness and kurtosis of the variables considered in the study are provided for the entire sample ($N = 308$), for the solvers ($N = 154$) and for the non-solvers ($N = 154$). The skewness of all variables was between −3 and 3 and the kurtosis between 10 and −10. According to Kline (2011), this indicates approximately normally distributed variables. Parametric methods were applied in this study as they are generally robust to scale assumption violations, especially when likert scales have seven or more categories (Norman, 2010; Dolan, 1994; Robitzsch, 2020). The majority of our ordinal variables had seven or more categories due to aggregation. The sole exception, "reading forum" with three categories, showed negligible differences between Pearson and Spearman correlations (maximum difference: 0.0165; average difference: < 0.0016). Hence, parametric methods were used.

2.2 Solving additional formative assessments, self-regulated learning behavior and learning success

With a *t*-test I investigated whether the solvers performed better in the end-of-semester knowledge test than the non-solvers. As shown in Table 1, solvers reached a higher performance in the knowledge test than non-solvers ($t(305.06) = -2.92, p < 0.01, d = 0.33$), confirming the first hypothesis.

Consistent with the hypothesis, the findings indicate that engagement with additional formative assessments significantly influences self-regulated learning behavior (see Table 1). Specifically, it was observed that those who solved these assessments demonstrated enhanced performance, invested more time in the completion of exercises, and posed fewer questions about those exercises.

Albeit not statistically significant, in tendency, solvers demonstrated a lower level of subjective understanding and less dedication to reading the mandatory literature than the non-solvers.

When analyzing the "total completions", which is the total number of completed exercises from both the initial and the additional formative assessments (where multiple attempts were possible), solvers completed significantly more exercises. This was expected since they had access to both initial and additional assessments.

However, when considering the "initial completions" (which both groups could attempt multiple times), solvers completed fewer exercises than non-solvers. This suggests that while having access to additional assessments led to more completions overall, it resulted in fewer completions of the initial assessments that were available to everyone.

Table 1
Descriptive statistics for entire sample, non-solvers and solvers as well as correlations with end-of-semester knowledge test (r).

	Mean (*SD*)	Skew	Kurtosis	Non-solvers	Solvers	*p*	*r*
Knowledge test	17.83 (2.82)	−0.95	1.00	17.36	18.29	<0.01	–
Prior knowledge	12.98 (2.25)	−0.28	−0.14	12.83	13.12	0.25	0.33***
Formative assessments	48.21 (5.96)	−0.86	0.91	47.28	49.13	<0.01	0.51***
Completions initial exercises	12.84 (6.19)	1.65	3.69	13.64	12.04	<0.05	0.08
Completions overall	2.93 (1.54)	0.59	−0.71	2.23	3.63	<0.001	0.18**
Subjective understanding	0.03 (1.02)	−1.01	3.47	0.14	−0.08	0.06	0.28***
Time investment on exercises	0.07 (2.45)	−1.74	−6.57	−0.35	0.59	<0.01	0.10
Time investment on feedback	−0.02 (2.59)	−0.54	0.40	−0.04	−0.01	0.93	0.04
Subjective investment	3.16 (0.49)	−0.16	−0.31	3.20	3.13	0.17	0.15*
Completing on time	0.56 (0.37)	−0.17	−1.53	0.58	0.54	0.45	0.15*
Lectures	3.94 (0.15)	−2.30	3.80	3.94	3.94	0.64	0.24***
Read forum	0.84 (0.69)	0.21	−0.92	0.88	0.81	0.32	0.10
Compulsory literature	0.51 (0.40)	−0.06	−1.58	0.55	0.47	0.09	−0.06
Questions	0.18 (0.27)	2.12	4.86	0.23	0.13	<0.01	0.01
Relevance of content	3.11 (0.62)	−0.57	−0.02	3.09	3.12	0.67	0.22***
Learning hours after course	3.08 (0.99)	−1.21	2.24	3.15	3.01	0.22	0.06

Note. r – correlation between the corresponding variable and performance in the end-of-semester knowledge test; $*p < 0.05$, $**p < 0.01$, $***p < 0.001$. The remaining correlations were not significant ($p > 0.10$). For subjective understanding and the two-time investment measures, scores on the first principal component are reported.

2.3 Students' characteristics, solving additional formative assessments and learning success

To identify meaningful clusters of self-regulated learning behavior, understanding the interrelations of the learning variables detailed in the Method section was crucial. An exploratory factor analysis was conducted to reduce the variables to a few interpretable factors. By decreasing the number of variables in the model, the cluster analysis can more effectively detect clusters within the dataset (Dalmaijer et al., 2021). The *z*-standardized variables were inputted into the fa.parallel function from the psych package (Revelle, 2022), resulting in a three-factor solution that best described the correlations between the thirteen manifest variables. The factor solution, following an oblimin rotation, is presented in Table 2.

Table 2

Standardized loadings of the measures on the three factors extracted by exploratory factor analysis with oblimin rotation

Variable	Performance	Time investment	Additional investment	h^2
Formative assessments	**0.88**	0.06	−0.06	0.80
Subjective understanding	**0.52**	−0.10	0.20	0.31
Time investment exercises	0.26	**0.65**	−0.04	0.57
Time investment feedback	−0.08	**0.85**	0.03	0.71
Subjective investment	0.25	**0.38**	**0.33**	0.44
Completing on time	0.28	−0.23	**0.41**	0.25
Lectures	**0.37**	0.19	0.12	0.25
Read forum	−0.04	0.02	**0.42**	0.18
Compulsory literature	−0.13	0.09	**0.53**	0.30
Questions	−0.07	0.04	**0.43**	0.19
Prior knowledge	**0.49**	−0.06	−0.05	0.23
Relevance of content	**0.35**	−0.10	0.10	0.13
Learning hours after course	−0.10	0.12	**0.39**	0.18
R^2	0.14	0.12	0.09	
Proportion R^2	0.41	0.33	0.26	

Note. R^2 – variance explained by the corresponding factor, h^2 – explained variance of the corresponding measurement, loadings of at least 0.30 are in bold.

To comprehend the three factors, they will be described based on the measures exhibiting the highest loadings (Table 2). The first factor is associated with performance, as evidenced by substantial loadings of performance in formative assessments, subjective understanding, and prior knowledge. The second factor is connected to time investment, which includes time spent on exercise pages and feedback pages. This factor is related to the time investment in content learning, a critical self-regulation skill identified by Kim et al. (2018) to effort regulation (Baker et al., 2020) or organization (Mega et al., 2014).

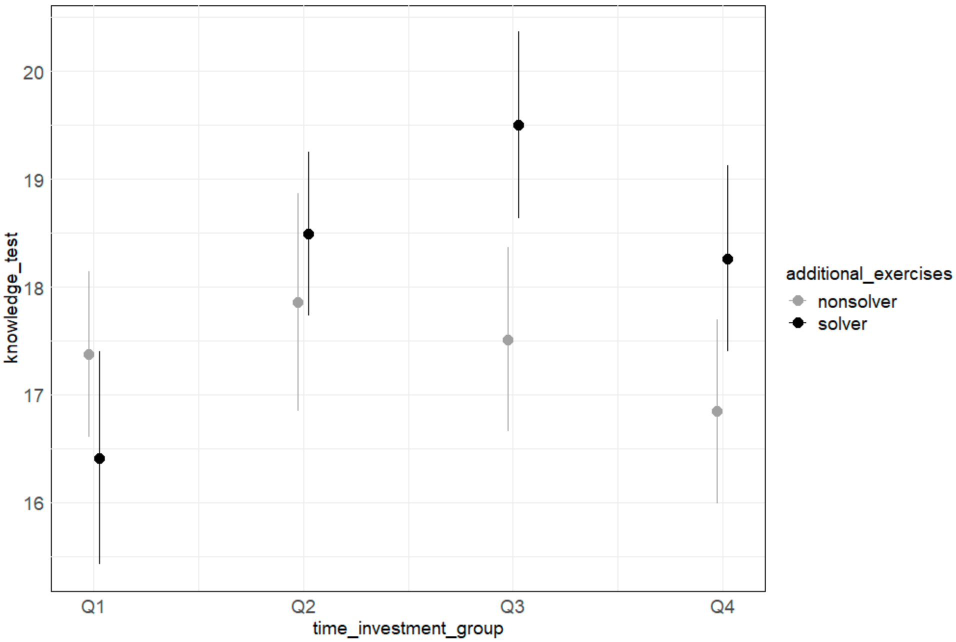
The third factor pertains to additional investment, as demonstrated by engagement in reading the literature, posing questions, reading the forum, dedicating learning hours during the semester break and timely completion of exercises. Accordingly, additional investment is a combination of help seeking (Kim et al., 2018), time management (Kim et al., 2018; Li et al., 2018) and investment in content learning (Kim et al., 2018).

The three factors were slightly correlated (performance and time investment $r = 0.27, p < 0.001$; performance and additional investment $r = 0.14, p < 0.05$; time investment and additional investment $r = 0.23, p < 0.001$) and together explained 35% of the variance. For further analyses, factor scores extracted using the regression method were used (DiStefano et al., 2009). Solvers scored higher on the performance factor ($t(305.35) = -2.03, p < 0.05, d = 0.013$) and lower on the additional investment factor ($t(305.37) = 3.55, p < 0.001, d = 0.04$) while there was no difference in scores on the time investment factor ($t(292-76) = -0.99, p = 0.32, d = 0.003$).

To examine whether the effect of additional formative assessments depends on students' self-regulated learning behavior, two approaches were employed. First, each of the three extracted factors was divided into four equal groups (quartiles) and the dependency of the effect of solving additional formative assessments on that split variable was investigated for each factor. Second, a cluster analysis was conducted across all three factors, and the dependency of solving additional formative assessments on cluster membership was examined.

In the first approach, a two-way ANOVA was conducted for each factor group (quartiles), with factor group membership and solving additional formative assessments as the between-subject factors and performance in the end-of-semester knowledge test as the dependent variable. A significant interaction would indicate that the effect of additional formative assessments on performance in the end-of-semester knowledge test depends on student characteristics. The interaction term was not significant for the performance factor ($F(3, 300) = 0.41, p = 0.75, \eta^2 = 0.003$) or the additional investment factor ($F(3, 300) = 0.23, p = 0.87, \eta^2 = 0.002$); however, it was significant for the time investment factor ($F(3, 300) = 4.14, p < 0.01, \eta^2 = 0.04$).

Figure 1 displays the interaction between the time investment group and solving additional formative assessments. In the lowest time investment quartile, solving additional formative assessments was associated with slightly lower performance in the end-of-semester knowledge test, whereas in all other quartiles, it was associated with higher performance. The performance difference in the end-of-semester knowledge test between non-solvers and solvers was −0.96 ($p = 0.13$) for the first quartile, 0.63 ($p = 0.32$) for the second quartile, 1.99 ($p < 0.01$) for the third quartile, and 1.42 ($p < 0.05$) for the fourth quartile. However, when applying a Bonferroni-corrected alpha level of 0.0125, the difference in the fourth quartile was no longer statistically significant. Overall, solving additional formative assessments appeared to be more beneficial for students who invested more time in solving the exercises.



Note. Q1 = first quartile, Q2 = second quartile, Q3 = third quartile, Q4 = fourth quartile; means and standard deviations are displayed.

Figure 1

Interaction between the completion of additional formative assessments and students' quartile ranking in time investment, in relation to performance on the end-of-semester knowledge test

It is important to note that solvers and non-solvers were not equally distributed across the four time investment groups ($\chi^2(3) = 10.44, p < 0.05$). Fewer solvers ($n = 29$) than non-solvers ($n = 48$) were in the first quartile, and more solvers ($n = 49$) than non-solvers ($n = 28$) were in the second quartile. In the other two groups, solvers and non-solvers were similarly distributed (either $n = 38$ or $n = 39$).

In the second approach, which is based on all three factors (performance, time investment, additional investment), a k-means cluster analysis was conducted to identify distinct student types. Initially, the number of clusters was determined using the NbClust function (Charrad et al., 2014), followed by the execution of the k-means cluster analysis using the stats package (R Core Team, 2021). The NbClust function helps determine the number of clusters in a dataset by evaluating 22 distinct fit indicators. Among these fit indicators, eight suggested a two-cluster solution and six recommended a three-cluster solution. Higher numbers of clusters were proposed by fewer than three fit indicators each. Consequently, both the two and three-cluster solutions were further examined. To circumvent local minima, 1,000 random starting positions were utilized.

For both the two and three-cluster solutions, an investigation was conducted to determine if the positive effect of additional formative assessments depended on cluster membership, or in other words, whether a significant interaction existed between cluster membership and the positive effect of solving additional formative assessments on performance on the end-of-semester knowledge test. To this end, a two-way ANOVA was performed for both the two and three-cluster solutions, with cluster membership and solving additional formative assessments as between-subject factors, and performance in the end-of-semester knowledge test as the dependent variable. The interaction was not significant for the two-cluster solution ($F(1, 304) = 1.09, p = 0.29, \eta^2 = 0.003$) but it was for the three-cluster solution ($F(2, 302) = 3.13, p < 0.05, \eta^2 = 0.02$, see Figure 2). Therefore, the three-cluster solution was further investigated. In addition to the significant interaction, there was a main effect of cluster membership ($F(2, 302) = 20.72, p < 0.001, \eta2 = 0.11$) and a significant main effect of completing additional formative assessments ($F(1, 302) = 9.77, p < 0.01, \eta2 = 0.03$).

In the three-cluster solution (see Table 3), one cluster ($n = 66$) exhibited low performance, low time investment, and relatively low additional investment.

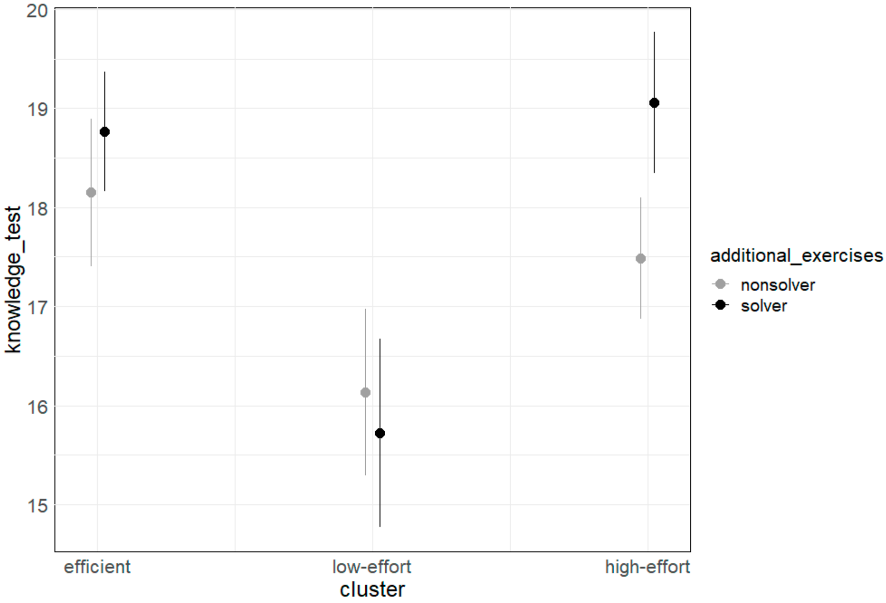


Figure 2
Interaction between the completion of additional formative assessments and the students' cluster membership in relation to their performance on the end-of-semester knowledge test

For subsequent analyses, this cluster will be denoted as the "low effort cluster." Another cluster (*n* = 120) was characterized by high performance, moderate time investment, and low additional investment. Accordingly, this cluster achieved high performance with comparatively low investment and is therefore referred to as the "efficient cluster." The last cluster (*n* = 122) exhibited above-average performance and considerable effort in both time investment and additional investment. This cluster will be referred to as the "high effort cluster" in subsequent analyses and discussions.

Table 3
Characterization of the three clusters identified as well as size of the entire sample (N) the sample of solvers, and the sample of non-solvers

	Cluster name	Performance	Time investment	Additional investment	*N* (non-solvers, solvers)
Low-performance, low-investment	low effort	−1.30	−1.02	−0.40	66 (37, 29)
High-performance, medium-investment	efficient	0.43	0.07	−0.70	120 (47, 73)
High-performance, high-investment	high effort	0.28	0.48	0.90	122 (70, 52)

Note. N = sample size of the entire sample and in parentheses sample size of non-solvers and solvers.

The performance difference in the end-of-semester knowledge test between non-solvers and solvers was −1.57 ($p < 0.01$) for the high effort cluster, −0.62 ($p = 0.21$) for the efficient cluster, and 0.41 ($p = 0.53$) for the low effort cluster. This pattern of results persists when alpha is adjusted for multiple testing. Accordingly, solving additional formative assessments appeared to be most beneficial for high effort students.

Again, solvers and non-solvers were not equally distributed across the three clusters ($\chi^2(2) = 9.26, p < 0.01$). A smaller proportion of solvers ($n = 52$) relative to non-solvers ($n = 70$) was observed in the high effort cluster, while a greater proportion of solvers ($n = 73$) compared to non-solvers ($n = 47$) was present in the efficient cluster. In contrast, the low effort cluster exhibited a more evenly distributed composition of solvers ($n = 29$) and non-solvers ($n = 37$).

Taken together, the effect of additional formative assessments depended on students' characteristics in both approaches. Both higher time investment alone and belonging to the high effort cluster resulted in a larger positive effect of additional formative assessment on the end-of-semester knowledge test. As shown in Table 4, the low effort cluster consisted mostly of students of the low time investment group (Q1), the efficient cluster consisted mostly of students with medium time investment (Q2, Q3) and the high effort cluster of high time investment students (Q3, Q4).

Table 4

Number of students in the three clusters depending on solving additional formative assessments and time investment group

Cluster	Time investment group				*Total*
	Q1	Q2	Q3	*Q4*	
Low effort	44 (26, 18)	13 (6, 7)	6 (4, 2)	3 (1, 2)	66 (37, 29)
Efficient	21 (14, 7)	40 (11, 29)	36 (15, 21)	23 (7, 16)	120 (47, 73)
High effort	12 (8, 4)	24 (11, 13)	35 (20, 15)	51 (31, 20)	122 (70, 52)

Note. Cells marked light gray contained at least thirty students. In parentheses (the number of non-solvers, the number of solvers).

3 Discussion

This study aimed to investigate the impact of additional formative assessments on students' self-regulated learning behavior and learning success, while also considering the varying impacts on different student groups. The completion of additional formative assessments covering identical content led to improved performance on the end-of-semester knowledge test. Moreover, these assessments had a differential impact on self-regulated learning behaviors across various variables. Notably, solvers exhibited enhanced performance in the formative assessments, yet reported lower levels of subjective comprehension (albeit not significantly so). They dedicated more time to completing exercises within the assessments, asked fewer questions about the exercises, and tended to engage less with the compulsory literature (albeit not significantly so). Furthermore, the influence of additional formative assessments on learning success depended on students' self-regulated learning behaviors. Both increased time investment individually and membership in the high-effort cluster contributed to a more substantial positive effect of additional formative assessments on the end-of- semester knowledge test outcomes.

3.1 Influence of additional formative assessments on self-regulated learning behavior and learning success

The positive effect of additional formative assessments on learning success is consistent with the findings of Yang et al. (2021), who conducted a meta-analytic overview. The current study extends the existing literature by demonstrating this beneficial effect in an applied setting, with complex exercises and even when the learning phase and assessment phase exercises were not identical but covered the same content.

This study found that solvers exhibited differences from non-solvers in certain aspects of self-regulated learning behavior. This was based on the initial assessments that both groups completed. Given that each formative assessment introduced new material, solvers' enhanced performance can only be attributed to an indirect testing effect, since apart from the additional formative assessments, all other conditions were identical for both groups. The indirect testing effect occurs when testing not only enhances performance on the tested material but also on new, related material (Fernandez & Jamet, 2017; Szpunar et al., 2008; Wissman et al., 2011). Consequently, the additional formative assessments impacted the solvers' self-regulated learning behavior with this new content.

In this context, the differential impact of additional formative assessments on self-regulated learning behavior offers interesting insights. Even though solvers performed better than non-solvers in formative assessments, they reported lower estimates of their understanding (albeit not significantly so) compared to non-solvers. This pattern of results indicates that solvers exhibit less overestimation of their own performance, a phenomenon known as the "illusion of knowing" (Avhustiuk et al., 2018), where learners tend to overestimate their understanding relative to their actual performance.

This pattern of results, combined with the solvers' higher time investment in solving formative assessments, indicates that the provision of additional formative assessments promotes better monitoring of one's knowledge, which is consistent with the observation made by Fernandez and Jamet (2017), and more time allocation for studying, which is in line with Soderstrom and Bjork, (2014). This can be attributed to the fact that the provision of additional formative assessments enables students to receive supplementary feedback on their learning status (Clariana & Park, 2021; Perry & Winne, 2006). This feedback assisted them in monitoring which behaviors in the initial assessments were most beneficial for their learning success in the additional formative assessments, prompting them to adjust their strategies and behaviors accordingly.

However, it is crucial to acknowledge the potential less beneficial effects of the additional learning material. The increased cognitive demands associated with the additional formative assessments, in terms of both the material's complexity and the volume of information, could lead to cognitive overload (Kossen & Ooi, 2021), and the extra exercises probably reduced the time available for students to fully engage with the material, causing them to adopt less elaborate learning strategies (e.g., less additional investment).

3.2 Student characteristics and the benefit of additional formative assessments

The impact of additional formative assessments on learning success depended on students' self-regulated learning behavior. It was primarily the students who invested above-average time during formative assessments that benefited from the additional exercises. Cluster analysis revealed that high-effort students (those with above-average time investment and above-average preparation/post-processing) gained the most from the extra exercises.

This outcome aligns with previous research by Greving et al. (2020), which demonstrated that the beneficial effect of solving exercises was most pronounced when retrieving information from memory was difficult but successful. In the high effort cluster, the retrieval of information from memory was generally successful, as overall performance in the investigated formative assessments was high. Furthermore, the retrieval of information from memory was difficult, as indicated by the above-average time investment (Dodonov & Dodonova, 2012; Dunst et al., 2014; Goldhammer, 2015) and the above-average additional effort (e.g., asking numerous questions in the forum).

The observed results align with the retrieval elaboration hypothesis (Carpenter et al., 2009). The high effort cluster demonstrated high time investment, additional investment, and above-average performance in formative assessments. Increased investment is typically linked with enhanced elaboration (Goldhammer et al., 2021). Likely due to their substantial investment, further elaboration or learning occurred during the initial formative assessments. The retrieval of this newly learned or elaborated content through additional formative assessments led to a more pronounced testing effect.

The high effort of this cluster may be correlated with high expectations of success, which is associated with a stronger positive impact of testing (Heitmann et al., 2022). Additionally, their regular learning behavior might also contribute to a more pronounced testing effect (Adesope et al., 2017; Karpicke & Bauernschmidt, 2011).

On the other hand, students in the low effort and efficient clusters did not show significant positive effects from additional formative assessments on their learning success. Low performers probably do not utilize the extra assessments effectively, while efficient learners do not require them, having already comprehended the material (Bjork et al., 2013).

For the low-effort cluster, this lack of effect might be attributed to the difficulty of the assessments, low motivation, or low elaboration of learning content (Carpenter et al., 2009; Heitmann et al., 2022; Minear et al., 2018). Exercises were probably too difficult for those students and retrieval of information was often unsuccessful, as indicated by the low performance in the formative assessments (Minear et al., 2018). According to the Yerkes-Dodson law (Yerkes & Dodson, 1908), when exercises become too difficult, motivation, response latencies and performance decrease (Borter et al., 2016; Dunst et al., 2014; Goldhammer, 2015). Their lack of prior knowledge may have posed challenges in integrating and elaborating on new but related content (Cogliano et al., 2019; Francis et al., 2020). In addition, especially for this group, the increased cognitive demands associated with the additional formative assessments, in terms of both the material's complexity and the volume of information, might have led to cognitive overload (Kossen & Ooi, 2021) or to hasty and unelaborated learning behavior due to the higher investment requirements imposed by the additional formative assessments.

In the efficient cluster, the absence of a significant positive effect could be due to either high ability and abstraction or the assessments being too easy for these students (Goldhammer, 2015) and accordingly no elaboration was needed. Even though retrieval from memory was quite successful in this cluster as indicated by the high performance in the formative assessments, it was not difficult (average time investment, very low additional investment e.g., asking questions). The exercises were probably not difficult enough for those students and after the first formative assessments no additional exercises were needed, as the students already grasped the content. Beside the possibility that formative assessments were too easy for students in this cluster, the high performance associated with rather low investment might be a sign of high ability or abstraction (Goldhammer, 2015). In this case, additional exercises are probably not necessary, as students understand the content on an abstract level and do not need different exercises from different contexts covering the same content. When low exercise difficulty is the reason for the missing effect of testing in this cluster, more difficult exercises would lead to a testing effect, whereas when high abstraction is the reason, more difficult exercises would probably not lead to a stronger testing effect. To differentiate between the two possibilities, further research is needed.

In addition, it was shown that students with poorer learning strategies show a larger testing effect than students with good strategies (Minear et al., 2018, Robey, 2019). The efficient cluster might have particularly good learning strategies as indicated by the high performance reached with rather low investment.

3.3 Solvers and non-solvers not equally distributed across time investment groups or clusters

The impact of solving additional formative assessments on self-regulated learning behavior led to an uneven distribution of students across time investment groups or clusters. Fewer solvers than non-solvers were found in the very low time investment group (Q1), while more solvers than non-solvers were present in the second time investment group (Q2). Furthermore, solvers more frequently belonged to the efficient cluster and less frequently to the high effort cluster.

On one hand, the additional formative assessments might have resulted in high effort students sacrificing additional investment (e.g., asking questions, reading literature) to invest more time in solving formative assessments (indirect testing effect, better monitoring, prioritizing different learning materials). Due to the positive effect of additional formative assessments, this resulted in higher performance. Higher performance in combination with lower additional investment is the behavioral pattern associated with the efficient cluster and led to a shift from the high effort to the efficient cluster (e.g., in Table 4, more solvers in the efficient cluster and higher time investment groups).

On the other hand, solving additional formative assessments prompted low investment students to invest more time in solving exercises and to achieve higher performance in the formative assessments (indirect testing effect). This combination of medium time investment, higher performance, and low additional investment is associated with the efficient cluster (e.g., in Table 4, there are more solvers in high time investment groups of the efficient cluster but fewer in the low effort low time investment group).

In conclusion, due to an indirect testing effect, solvers demonstrated improved monitoring associated with more efficient learning, and as a result, many solvers were part of the efficient cluster, which is linked to high performance on the end-of-semester knowledge test. Additionally, the availability of numerous formative assessments for solvers may have forced them to make decisions on where to allocate their time (Yang et al, 2017). As they spent more time on the exercises and solved a greater number of them, they reduced other activities (additional investment, fewer repetitions of the first formative assessments, but more repetitions when including additional formative assessments).

3.4 Practical relevance of the findings

As a lot of time is invested in solving additional formative assessments and not all students profit from them, it seems unethical to suggest additional assessments to all students. In the future, approaches from adaptive learning analytics (Mavroudi et al., 2018) should be implemented into the course. As indicated by the results of this study, for students with above average time investment, additional formative assessments should be suggested as adding formative assessments probably improves their learning success. For students with below-average time investment, it is important to know whether below-average time investment is associated with low or high performance in the formative assessments. If it is associated with high performance, there is no need to suggest the additional formative assessments as they probably would not lead to greater learning success. However, more difficult exercises might lead to even greater learning success in this cluster, but future research is needed to test those predictions. When low time investment is linked to low performance in formative assessments, interventions to increase content understanding, content elaboration, improve learning strategies, enhance monitoring, or adjust time allocation should be suggested. Only after successfully making these improvements should additional assessments be recommended.

When deciding whether to create additional formative assessments for a course, it is essential to consider that although many students benefited from the extra assessments and nearly all students solved them when available, the effect sizes were relatively small, and providing additional formative assessments influenced students' behavior in both beneficial and less beneficial ways. The present study highlights the importance of considering individual differences in students' self-regulated learning behavior when implementing additional formative assessments.

3.5 Measurement considerations

To investigate learning as comprehensively as possible, a variety of variables were measured, some of which were highly related. Therefore, variables of the same type (e.g., response latencies for exercises) were reduced to a single score. Observations of the same type can be interpreted as a sampling of observations, and combining them leads to a more reliable measure (Goldhammer et al., 2021). For example, when combining 100 response latencies, the influence of measurement error (e.g., taking a coffee break while solving an exercise, leading to longer response latency) is reduced. Moreover, high correlations between similar measures, as indicated by a strong first principal component, suggest that the different variables measured the same construct. The summarized measures of the same type were combined in a factor analysis. First, this resulted in well-interpretable factors (performance, time investment, additional investment), and second, fewer but more reliable measures lead to a better performance in cluster analysis (Dalmaijer et al., 2021). Based on these three factors, three clusters were built. The clusters found were similar to previous studies, in which clusters based on effort and/ or processing depth (Jovanović et al., 2017; Kovanovic et al., 2015; Li et al, 2020; Ning & Downing, 2015; Parpala et al., 2021; Sun & Xie, 2020; van Alten et al., 2021; Vanslambrouck et al, 2019; Zheng et al., 2020) based on regularity of learning (Kim et al., 2018; Parpala, 2021), on prior knowledge (Khayi & Rus, 2019), on the pace of learning (Munje et al., 2020), and on performance and learning behavior were found (Waspada et al., 2019). Accordingly, the three clusters of this study fit well into previous research.

3.6 Future work

Future research could investigate how cluster membership and learning behavior evolves throughout the semester and whether adaptive hints or instructions can help students find the learning behavior or strategy that maximizes their learning success. The consistency of these clusters across various courses needs to be investigated. Furthermore, the psychological traits associated with cluster membership should be understood. It has been suggested by a recent study (Heitmann et al., 2022) that quizzing might not be beneficial for learners exhibiting a low hope of success, an attribute that might be prevalent in some of the clusters identified.

Additionally, the behavior data of the extra formative assessments should be examined, and exercise difficulty should be considered. Future research could benefit from a deeper exploration of the potential impact of assessment length on learner engagement, to discern if longer formative assessments might introduce variability in self-regulated learning. Furthermore, integrating various theories of self-regulation into our understanding of self-regulated learning behavior warrants further investigation. In addition, determining whether the positive effect of additional formative assessments can be attributed to an indirect testing effect, a direct testing effect, or a combination of both would be of significant interest in future research.

3.7 Limitations

The study's limitations primarily stem from its quasi-experimental approach in a real-world setting. Consequently, it is challenging to determine the generalizability of the findings to other courses. Furthermore, not all students in the course participated or met the inclusion criteria, which may have affected the results. Additionally, principal component analysis, exploratory factor analysis, and cluster analysis are exploratory instruments bearing the risk of false discoveries (Moosbrugger & Kelava, 2012). As a result, it is necessary to confirm or disprove these exploratory and course-specific findings in future research.

Conclusion

In conclusion, additional formative assessments led to an overall better performance in the end-of-semester knowledge test. However, this effect depended on students' characteristics. Above-average time investment was associated with a more beneficial effect of solving additional formative assessments. As indicated by the results of the cluster analysis, solvers characterized by above-average time investment and additional investment (high effort cluster) benefited from additional formative assessments, while below-average time investment was associated either with low investment/understanding (low effort cluster) or high understanding with relatively low investment (efficient cluster). In both these clusters, no positive effect of additional formative assessments was identified. Furthermore, engaging in additional formative assessments led to changes in self-regulated learning behavior, both positive and negative, resulting in a higher proportion of solvers in the efficient cluster, which is associated with high performance on the end-of-semester knowledge test. Taken together, solving additional formative assessments is beneficial for some but not all students and is associated with both beneficial and less beneficial changes in self-regulated learning behavior.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Avhustiuk, M. M., Pasichnyk, I. D., & Kalamazh, R. V. (2018). The illusion of knowing in metacognitive monitoring: Effects of the type of information and of personal, cognitive, metacognitive, and individual psychological characteristics. *Europe's Journal of Psychology, 14*(2), 317–341. https://doi.org/10.5964/ejop.v14i2.1418

Baker, R., Xu, D., Park, J., Yu, R., Li, Q., Cung, B., Fischer, Ch., Rodriguez, F., Warschauer. M., & Smyth, P. (2020). The benefits and caveats of using clickstream data to understand student self-regulatory behaviors: Opening the black box of learning processes. *International Journal of Educational Technology in Higher Education*, *17*(3), 1–24. https://doi.org/10.1186/s41239-020-00187-1

Batsell Jr., W. R., Perry, J. L., Hanley, E., & Hostetter, A. B. (2017). Ecological validity of the testing effect: The use of daily quizzes in introductory psychology. *Teaching of Psychology*, *44*(1), 18–23. https://doi.org/10.1177/0098628316677492

Bentler, P. M., & Satorra, A. (2010). Testing model nesting and equivalence. *Psychological Methods*, *15*(2), 111–123. https://doi.org/10.1037/a0019625

Bertilsson, F., Stenlund, T., Wiklund-Hörnqvist, C., & Jonsson, B. (2021). Retrieval practice: Beneficial for all students or moderated by individual differences? *Psychology Learning & Teaching, 20*(1), 21–39. https://doi.org/10.1177/1475725720973494

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Borter, N. (2016). *Aufgabenkomplexität und Intelligenz* [Dissertation, Universität Bern]. https://boris.unibe.ch/101386/

Boston, C. (2002). The concept of formative assessment. *Practical Assessment, Research, and Evaluation*, *8*(9). https://doi.org/10.7275/kmcq-dj31

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(6), 1547–1552. https://doi.org/10.1037/a0024140

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology*, *23*(6), 760–771. https://doi.org/10.1002/acp.1507

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, *61*(6), 1–36. https://doi.org/10.18637/jss.v061.i06

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2013). A reference model for learning analytics. *International Journal of Technology Enhanced Learning*, *4*(5–6), 318–331. https://doi.org/10.1504/IJTEL.2012.051815

Clariana, R. B., & Park, E. (2021). Item-level monitoring, response style stability, and the hard-easy effect. *Educational Technology Research and Development*, *69*, 693–710. https://doi.org/10.1007/s11423-021-09981-8

Cogliano, M., Kardash, C. M., & Bernacki, M. L. (2019). The effects of retrieval practice and prior topic knowledge on test performance and confidence judgments. *Contemporary Educational Psychology*, *56*, 117–129. https://doi.org/10.1016/j.cedpsych.2018.12.001

Dalmaijer, E. S., Nord, C. L., & Astle, D. E. (2021). *Statistical power for cluster analysis.* arXiv. https://doi.org/10.48550/arXiv.2003.00381

DiStefano, C., Zhu, M., & Mîndrilã, D. (2009). *Understanding and using factor scores: Considerations for the applied researcher*, *14*(20). https://doi.org/10.7275/DA8T-4G52

Dodonov, Y. S., & Dodonova, Y. A. (2012). Response time analysis in cognitive tasks with increasing difficulty. *Intelligence*, *40(5)*, 379–394. https://doi.org/10.1016/j.intell.2012.07.002

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, *47*(2), 309–326. https://doi.org/10.1111/j.2044-8317.1994.tb01039.x

Dunst, B., Benedek, M., Jauk, E., Bergner, S., Koschutnig, K., Sommer, M., Ischebeck, A., Spinath, B., Arendasy, M., Bühner, M., Freudenthaler, H., & Neubauer, A. C. (2014). Neural efficiency as a function of task demands. *Intelligence*, *42*, 22–30. https://doi.org/10.1016/j.intell.2013.09.005

Eriksson, J., Kalpouzos, G., & Nyberg, L. (2011). Rewiring the brain with repeated retrieval: A parametric fMRI study of the testing effect. *Neuroscience Letters*, *505*(1), 36–40. https://doi.org/10.1016/j.neulet.2011.08.061

Fernandez, J., & Jamet, E. (2017). Extending the testing effect to self-regulated learning. *Metacognition and Learning*, *12*, 131–156. https://doi.org/10.1007/s11409-016-9163-9

Foss, D. J., & Pirozzolo, J. W. (2017). Four semesters investigating frequency of testing, the testing effect, and transfer of training. *Journal of Educational Psychology*, *109*(8), 1067–1083. https://doi.org/10.1037/edu0000197

Francis, A. P., Wieth, M. B., Zabel, K. L., & Carr, T. H. (2020). A classroom study on the role of prior knowledge and retrieval tool in the testing effect. *Psychology Learning & Teaching*, *19*(3), 258–274. https://doi.org/10.1177/1475725720924872

Goldhammer, F. (2015). Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement: Interdisciplinary Research and Perspectives*, *13*(3–4), 133–164. https://doi.org/10.1080/15366367.2015.1100020

Goldhammer, F., Hahnel, C., Kroehne, U., & Zehner, F. (2021). From byproduct to design factor. On validating the interpretation of process indicators based on log data. *Large-Scale Assessments in Education*, *9*, 1–25. http://nbn-resolving.de/urn:nbn:de:0111-pedocs-250050

Greene, R. L. (2008). Repetition and spacing effects. In H. L. Roediger (Ed.), *Learning and memory: A comprehensive reference. Vol. 2: Cognitive psychology of memory* (pp. 65–78). Elsevier.

Greving, S., Lenhard, W., & Richter, T. (2020). Adaptive retrieval practice with multiple-choice questions in the university classroom. *Journal of Computer Assisted Learning*, *36*(6), 799–809. https://doi.org/10.1111/jcal.12445

Heitmann, S., Grund, A., Fries, S., Berthold, K., & Roelle, J. (2022). The quizzing effect depends on hope of success and can be optimized by cognitive load-based adaptation. *Learning and Instruction*, *77*. https://doi.org/10.1016/j.learninstruc.2021.101526

Ifenthaler, D. (2015). Learning Analytics. In J. M. Spector (Ed.), *The SAGE encyclopedia of educational technology* (Vol. 2, pp. 447–451). SAGE publication. https://madoc.bib.uni-mannheim.de/38809/

Jensen, J. L., McDaniel, M. A., Kummer, T. A., Godoy, P. D. D. M., & St. Clair, B. (2020). Testing effect on high-level cognitive skills. *CBE—Life Sciences Education*, *19*(3). https://doi.org/10.1187/cbe.19-10-0193

Jost, N. S., Jossen, S. L., Rothen, N., & Martarelli, C. S. (2021). The advantage of distributed practice in a blended learning setting. *Education and Information Technologies*, *26*, 3097–3113. https://doi.org/10.1007/s10639-020-10424-9

Jovanović, J., Gašević, D., Dawson, S., Pardo, A., & Mirriahi, N. (2017). Learning analytics to unveil learning strategies in a flipped classroom. *The Internet and Higher Education*, *33*, 74–85. https://doi.org/10.1016/j.iheduc.2017.02.001

Karpicke, J., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *37*(5), 1250–1257. https://doi.org/10.1037/a0023436

Karpicke, J., & Roediger, H. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*(4), 704–719. https://doi.org/10.1037/0278-7393.33.4.704

Khayi, N. A., & Rus, V. (2019). Clustering students based on their prior knowledge. In F. Collin, A. Merceron, & M. Desmarais (Eds.), *Proceedings of the 12th International Conference on Educational Data Mining.* Université du Québec à Montréal, Polytechnique Montréal (pp. 246–251). https://educationaldatamining.org/edm2019/proceedings/

Kim, D., Yoon, M., Jo, I.-H., & Branch, R. M. (2018). Learning analytics to support self-regulated learning in asynchronous online courses: A case study at a women's university in South Korea. *Computers & Education*, *127*, 233–251. https://doi.org/10.1016/j.compedu.2018.08.023

Kline, R. B. (2011). Principles and practice of structural equation modeling (3rd ed.). Guilford Press.

Kossen, C., & Ooi, C.-Y. (2021). Trialling micro-learning design to increase engagement in online courses. *Asian Association of Open Universities Journal*, *16*(3), 299–310. https://doi.org/10.1108/AAOUJ-09-2021-0107

Kovanovic, V., Gasevic, D., Joksimovic, S., Hatala, M., & Adesope, O. (2015). Analytics of communities of inquiry: Effects of learning technology use on cognitive presence in asynchronous online discussions. *The Internet and Higher Education*, *27*, 74–89. https://doi.org/10.1016/j.iheduc.2015.06.002

Lamotte, M., Izaute, M., & Darnon, C. (2021). Can tests improve learning in real university classrooms? *Journal of Cognitive Psychology*, *33*(8), 974–992. https://doi.org/10.1080/20445911.2021.1956939

Leitner, P., Khalil, M., & Ebner, M. (2017). Learning analytics in higher education– A Literature Review. In A. Peña-Ayala (Ed.), *Learning analytics: Fundaments, applications, and trends. Studies in systems, decision and control* (pp. 1–23). Springer. https://doi.org/10.1007/978-3-319-52977-6_1

Li, S., Chen, G., Xing, W., Zheng, J., & Xie, C. (2020). Longitudinal clustering of students' self-regulated learning behaviors in engineering design. *Computers & Education*, *153*, Article 103899. https://doi.org/10.1016/j.compedu.2020.103899

Li, H., Flanagan, B., Konomi, S. & Ogata, H. (2018). Measuring behaviors and identifying indicators of self-regulation in computer-assisted language learning courses. *Research and Practice in Technology Enhanced Learning*, *13*(19), 1–12. https://doi.org/10.1186/s41039-018-0087-7

Mavroudi, A., Giannakos, M., & Krogstie, J. (2018). Supporting adaptive learning pathways through the use of learning analytics: Developments, challenges and future opportunities. *Interactive Learning Environments*, *26*(2), 206–220. https://doi.org/10.1080/10494820.2017.1292531

McDaniel, M. A, Thomas, R. C, Agarwal, P. K, McDermott, K. B, & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology*, *27*(3), 360–372. https://doi.org/10.1002/acp.2914

Mega, C., Ronconi, L., & De Beni, R. (2014). What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement. *Journal of Educational Psychology*, *106*(1), 121–131. https://doi.org/10.1037/a0033546

Minear, M., Coane, J. H., Boland, S. C., Cooney, L. H., & Albat, M. (2018). The benefits of retrieval practice depend on item difficulty and intelligence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(9), 1474–1486. https://doi.org/10.1037/xlm0000486

Moosbrugger, H., & Kelava, A. (2012). *Testtheorie und Fragebogenkonstruktion*. Springer. https://doi.org/10.1007/978-3-642-20072-4

Ning, H. K., & Downing, K. (2015). A latent profile analysis of university students' self-regulated learning strategies. *Studies in Higher Education*, *40*(7), 1328–1346. https://doi.org/10.1080/03075079.2014.880832

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education*, *15*(5), 625–632. https://doi.org/10.1007/s10459-010-9222-y

Parpala, A., Mattsson, M., Herrmann, K. J., Bager-Elsborg, A., & Hailikari, T. (2021). Detecting the variability in student learning in different disciplines—A person-oriented approach. *Scandinavian Journal of Educational Research*, *66*(6), 1020–1037. https://doi.org/10.1080/00313831.2021.1958256

Perry, N. E., & Winne, P. H. (2006). Learning from Learning Kits: gStudy traces of students' self-regulated engagements with computerized content. *Educational Psychology Review*, *18*, 211–228. https://doi.org/10.1007/s10648-006-9014-3

Pishgar, F., Greifer, N., Leyrat, C., & Stuart, E. (2021). MatchThem: Matching and weighting after multiple imputation. *The R Journal*, *13*(2), 294–305. https://doi.org/10.32614/RJ-2021-073

R Core Team (2021). *R: A language and environment for statistical computing*. R foundation for statistical computing. Retrieved from https://www.R-project.org

Revelle, W. R. (2022). *psych: Procedures for psychological, psychometric, and personality research. R package version 2.2.5*. Northwestern University. Retrieved from https://CRAN.R-project.org/package=psych

Robey, A. (2019). The benefits of testing: Individual differences based on student factors. *Journal of Memory and Language*, *108*. https://doi.org/10.1016/j.jml.2019.104029

Robitzsch, A. (2020). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. *Frontiers in Education*, *5*. https://doi.org/10.3389/feduc.2020.589965

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The testing effect in the psychology classroom: A meta-analytic perspective. *Psychology Learning & Teaching*, *16*(2), 179–196. https://doi.org/10.1177/1475725717695149

Shin, D., & Shim, J. (2021). A systematic review on data mining for mathematics and science education. *International Journal of Science and Mathematics Education*, *19*, 639–659. https://doi.org/10.1007/s10763-020-10085-7

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language*, *73*, 99–115. https://doi.org/10.1016/j.jml.2014.03.003

Sun, Z., & Xie, K. (2020). How do students prepare in the pre-class setting of a flipped undergraduate math course? A latent profile analysis of learning behavior and the impact of achievement goals. *The Internet and Higher Education*, *46*. https://doi.org/10.1016/j.iheduc.2020.100731

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. III. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1392–1399. https://doi.org/10.1037/a0013082

van Alten, D. C. D., Phielix, C., Janssen, J., & Kester, L. (2021). Secondary students' online self-regulated learning during flipped learning: A latent profile analysis. *Computers in Human Behavior*, *118*. https://doi.org/10.1016/j.chb.2020.106676

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

Vanslambrouck, S., Zhu, C., Pynoo, B., Lombaerts, K., Tondeur, J., & Scherer, R. (2019). A latent profile analysis of adult students' online self-regulation in blended learning environments. *Computers in Human Behavior*, *99*, 126–136. https://doi.org/10.1016/j.chb.2019.05.021

Waspada, I., Bahtiar, N., & Wibowo, A. (2019). Clustering student behavior based on quiz activities on moodle LMS to discover the relation with a final exam score. *Journal of Physics: Conference Series*, *1217*. https://doi.org/10.1088/1742-6596/1217/1/012118

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2011). The interim test effect: Testing prior material can facilitate the learning of new material. *Psychonomic Bulletin & Review*, *18*, 1140–1147. https://doi.org/10.3758/s13423-011-0140-7

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, *147*(4), 399–435. https://doi.org/10.1037/bul0000309

Yang, C., Potts, R., & Shanks, D. R. (2017). The forward testing effect on self-regulated study time allocation and metamemory monitoring. *Journal of Experimental Psychology: Applied*, *23*(3), 263–277. https://doi.org/10.1037/xap0000122

Yerkes, R. M., & Dodson, J. D. (1908). The Relation of Strength of Stimulus to Rapidity of Habit Formation. *Journal of Comparative Neurology & Psychology, 18,* 459–482. https://doi.org/10.1002/cne.920180503

Zheng, J., Xing, W., Zhu, G., Chen, G., Zhao, H., & Xie, C. (2020). Profiling self-regulation behaviors in STEM learning of engineering design. *Computers & Education*, *143*. https://doi.org/10.1016/j.compedu.2019.103669