Measuring the level of homework answer copying during COVID-19 induced remote instruction

Zhongzhou Chen

Department of Physics, University of Central Florida, 4111 Libra Drive, Orlando, Florida 32816, USA

(Received 27 October 2021; accepted 16 February 2022; published 5 April 2022)

This paper examines the prevalence of rapid answer copying among university students completing online homework for an introductory level calculus-based physics course taught remotely during the COVID pandemic. We first compared the attempt duration distribution of 26 problems, between 42 students who self-reported as having completed the homework by themselves against the rest of the class. Significant differences were detected for 3 out of 26 problems. We then identified abnormally short problem attempts indicative of potential rapid answer copying, by fitting the attempt duration distribution of each problem with finite-mixture models, using mixtures of either normal or skewed distributions. We detected a significantly smaller fraction of short attempts from self-reporting students on only 3 out of 26 problems and found no statistically significant difference in percentage correct of short attempts between the populations. In conclusion, our analysis did not find evidence indicating widespread rapid answer copying among students. We also explored differences in learning behavior between the two populations by applying process mining to the event logs of one of the homework learning modules, which reveals that some students may have copied answers after spending a longer time or using multiple attempts on a given problem. However, this form of answer copying is also unlikely to be prevalent since the percentage correct on normal attempts is also similar between the two populations on most problems.

DOI: 10.1103/PhysRevPhysEducRes.18.010126

I. INTRODUCTION

One of the major concerns instructors have about online homework and online learning systems is that students may copy problem answers from other sources without actually trying to solve the problems [1–5]. The switch to remote learning during the COVID-19 pandemic further elevated the worry that answer copying could become more prevalent [2,6].

A number of earlier studies have identified rapid answer copying in online homework, especially physics online homework, by detecting abnormalities in students' log data from online learning systems. Since rapid answer copying takes less time than authentic problem solving, the distribution of problem-solving duration could be observed to have multiple peaks (see, for example, Fig. 4) with the shorter duration peak likely produced by students either guessing or answer copying [4,5,7–11].

However, those earlier studies have two shortcomings. First, the "true" attempt duration distribution from students who attempted the problem without answer copying was unknown. Not being able to contrast the attempt duration distribution between answer-copying and non-answer-copying students made it difficult to distinguish answer copying from other problem-solving behavior that could also generate short attempt duration, such as guessing. Second, a single cutoff time such as 30 s was used to distinguish between "short" and "normal" attempts for all problems, determined based on either the author's best estimate [3,5,11], or by analyzing cumulative duration data from all problems [7,8]. This "one size fits all" approach is clearly not ideal since single step conceptual problems, for example, can be solved much faster than multistep numeric problems. A uniform cutoff will overestimate the frequency of answer copying in the former case, and underestimate the frequency in the latter.

The current study examines the extent to which rapid answer copying is widespread among students taking an introductory level calculus-based physics course taught online during the pandemic, by analyzing students' attempt duration on 26 problems, administered in the form of online learning modules (OLMs) assigned as online homework. To address the first shortcoming of existing methods, we establish the important "ground fact" of the true attempt duration distribution for each problem without answer copying, by identifying a subgroup of students who self-reported as having completed all homework problems independently, using a survey administered at the end of the semester. To address the second shortcoming, we fitted the duration distribution of each individual problem using finite mixture models (FMM) to estimate the cutoff between short and normal attempts for each problem individually.

For students who self-reported having completed the problems independently, any short attempts identified by FMM should have resulted from either guessing or incorrect ways of solving the problem. In the current study, we assume that the frequency of those guessing attempts among self-reporting students is similar to the frequency among students who did not self-report but also solved the problems independently. In that case, if there were no or only a few cases of rapid answer copying among non-self-reporting students, the fraction of short attempts will be similar between self-reporting and non-self-reporting students. On the other hand, if the non-self-reporting population produced a significantly higher fraction of short attempts, or have a significantly higher correct rate on short attempts, then the difference is likely due to a fraction of the non-self-reporting students engaging in rapid answer copying.

More specifically, we hypothesize that if a substantial fraction of non-self-reporting students are engaged in rapid answer copying on a given problem, then we should be able to verify one or more of the following hypothesis:

H1: *The distribution of attempt time from non-self-reporting students will be significantly different from that of the self-reporting students, with self-reporting students spending longer on average answering the problems.*
H2: *Non-self-reporting students will be significantly more likely to submit a short attempt compared to self-reporting students. Short attempt is defined as attempts with duration shorter than the short-normal cutoff determined by FMM fitting for each module.*
H3: *Non-self-reporting students will have a significantly higher correct rate on short attempts compared to self-reporting students.*

II. METHODS

A. Instructional context

Data on students' problem solving behavior were obtained from a calculus-based Physics I course during Fall 2020 semester, taught entirely online using Microsoft teams [12,13]. Course contents were delivered through a combination of prerecorded instructional videos, OpenStax textbooks, and OLMs. Students had the option to work in groups on problem-solving worksheets during synchronous online class meetings, but attendance of class meetings was not required.

Two midterm exams were administered during weeks 6 and 11 of the 16-week semester, and a final exam was administered at the end of the semester. All exams were administered remotely, and students were required to turn on their webcams during the exam. In addition, biweekly 20-min quizzes were administered during the accompanying lab sessions, proctored by a TA over webcam.

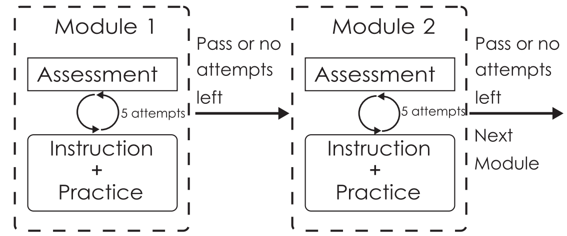

FIG. 1. Schematic illustration of the design of two online learning modules.

B. Design of online learning modules

Homework problems in this study are administered through OLMs [8,14,15]. Each OLM consists of an instructional component (IC) containing instructional text and practice problems, and an assessment component (AC) containing 1–2 problems (Fig. 1). Upon accessing a new module, students are required to make one attempt at the AC before being able to access the IC. This design could improve students' learning from the IC through the "preparation for future learning" effect [16,17], and also improve the interpretability of clickstream data [8,14]. Students are allowed 5 attempts on the AC and on each of the first 3 attempts an isomorphic problem is shown to them. Students cannot access the IC during an attempt on the AC. On average an OLM module is designed to be completed in 20–30 min. Eight to twelve OLM modules are assigned as a sequence covering a common topic such as mechanical energy, to be completed in 1–2 weeks. A student can proceed onto the next module in the sequence after either passing the AC or using up all the attempts on the current module.

A total of 70 OLMs belonging to 10 sequences were assigned as online homework in the Fall 2020 semester.

For the current study we selected three sequences, assigned at the beginning, middle, and end of the 15-week semester, with a total of 26 modules. The modules and naming conventions are listed in Table I. The first assigned module, 1D01, recorded activity from 250 students, whereas the last module, AM08, recorded activity from 209 students.

TABLE I. Sequences and modules selected for analysis in the current study.

Names	Topic	Assigned	Modules
1D01-08	1D motion	Week 2	8
E01-E10	Mechanical energy	Week 7	10
AM01-AM08	Angular momentum	Week 13	8

None of the three sequences were due right before a midterm exam.

Problem-solving duration in the current study is defined as the time spent on making an attempt on the AC of a given module. We do not distinguish between ACs with 1 or 2 problems, since the 2 problems in the same AC are closely related and can be seen as one bigger problem. In each sequence, the first 2–4 modules contain conceptual questions or one step numeric calculation in their AC, and the rest contain more elaborate numeric or symbolic calculation questions in the AC. All questions are in multiple-choice format.

C. Identifying short attempts using FMM

FMM is a model-based clustering algorithm that divides a population into subgroups according to one or more observable characteristics, by fitting the distribution of characteristic(s) with a finite mixture of normal or skewed probability distributions [18]. FMMs have been frequently used to detect abnormally short question attempts since the distribution of attempt durations are approximately log-normal. When two or more distinct problem-solving behaviors are present, the log of the attempt duration distribution can be fitted with the sum of two or more normal distributions (for example, in Fig. 4, E03), with the leftmost distribution corresponding to abnormally short attempts.

Many previous applications of FMM in detecting answer copying are based on normal distributions [10,14]. However, in some cases when the actual distribution of log duration is skewed, such as shown in Fig. 4: E02, using normal distribution may cause the algorithm to artificially add more components and identify clusters that may not exist. To overcome this shortcoming, we consider both normal and skewed distribution models using the R package mixsmsn [18]. For each duration distribution, the fitting algorithm first searches for the optimal number of components up to 4, using one of three families of distributions: normal, skewed normal, and skewed t. Then the best fit FMM of each family is compared based on four selection criteria: the Akaike information criterion (AIC) [19], the Bayesian information criterion (BIC) [20], the efficient determination criterion (EDC) [21], and the integrated complete-data likelihood (ICL) [22], and the distribution favored by 3 out of 4 criteria are selected. In the rare case that two models are each favored by two different criterion, the one favored by EDC is selected [8,23].

To better identify short attempts in the distribution, we used data from all attempts submitted by every student, since some students are observed to submit multiple guessing attempts on the same module. Students are also more likely to submit a short attempt on the mandatory first attempt prior to accessing the instructional component. We also included submission from both self-reporting and non-self-reporting students because the self-reporting population is relatively small, and that the difference in duration distribution is only significant on 3 out of 26 modules. Even on those 3 modules, the distributions had the same number of peaks at similar locations, as seen in the example shown in Fig. 2. The differences mostly lie in the magnitude of each peak.

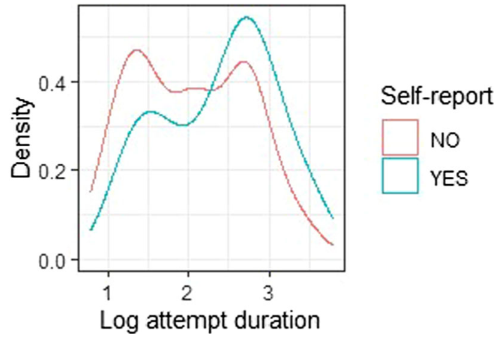

FIG. 2. Density distribution of log (base 10) attempt duration of the assessment component of module E08.

If the attempt distribution is best fitted with a 2 or more component FMM, then the intersection between the shortest and second shortest component is used as the cutoff between short and normal attempts. If a single component fit is favored, then the cutoff is set as either 2 standard deviations below the mean duration, or 15 s, whichever is longer. This is because a previous clinical study indicated that attempts under 15 s are likely to arise from complete random guessing [24]. Figure 3 shows examples of duration distributions fitted with either a one-component or a two-component model, with the cutoff indicated by a red vertical line.

D. Student self-report on homework completion

An end of semester survey was administered to all students after the last homework assignment was due. Two questions on the survey asked students to voluntarily disclose whether they completed all or most of the assignments by themselves and indicate the sequence on which

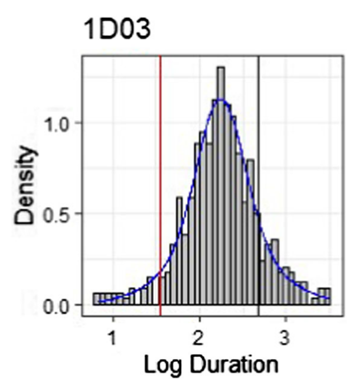

FIG. 3. Histogram and FMM fit profile for log attempt duration of module 1D03. The black line represents the cutoff originally determined by the algorithm based on a two-component fit, and the red line represents the adjusted cutoff (35 s).

they had external help on any module. These two questions were not mandatory for the survey. 42 students responded to the questions and consented to their response being used for research purposes. Of those, 33 indicated completing all homework modules independently, and 9 indicated having sought external help on 1 to 3 sequences. Those 9 students were excluded from the self-reporting population on all modules belonging to the indicated sequences. The normalized quiz, exam, and final course scores of the self-reporting students are not statistically different from the rest of the class at $\alpha = 0.05$ level according to Mann-Whitney U tests.

E. Hypothesis testing

To test hypotheses H1 and H2, we conducted statistical testing using the duration of either a student's first attempt, or of their correct attempt on the assessment of each OLM. Answer copying was more likely to take place on those two attempts, because students who decided to copy their answers without engaging with the problem were most likely to do so on the first attempt, and that a copied answer is significantly more likely to be correct. Using the correct attempt also minimizes the fraction of guessing attempts in the dataset. Difference in fraction of correct attempts among short attempts (H3) was tested using short first attempts. Since there were only a small number of short first attempts on many modules, we also conducted the test using all short attempts, which may artificially reduce the correct percentage of non-copying students, since they were more likely to submit multiple incorrect attempts.

Mann-Whitney U tests were used to compare the duration distribution between populations (H1), and Fisher's exact tests were used to compare the fraction of short attempts (H2) and the fraction of correct short attempts (H3).

III. RESULTS

A. FMM fitting of attempt duration

The attempt duration distribution of 11 problems were best fitted with skewed normal or skewed-t FMMs, and 13 were fitted with normal distribution FMMs. Eight problems were fitted with 1 component FMMs, and the rest are all fitted with 2 or more components FMMs.

For 4 problems, the short versus normal cutoffs as determined by FMM modeling were less than 15 s and were thus adjusted to 15 s. Twenty problems had cutoffs between 15 and 120 s, 2 problems had cutoffs beyond 120 s. We visually examined those two cases and found one of them, 1D03, to be an artifact resulting from the algorithm selecting a two-component normal distribution as the best fit for an obviously one component distribution, as shown in Fig. 4. Therefore, we adjusted the cutoff from 480 to 35 s, based on best estimates from a previous study [8].

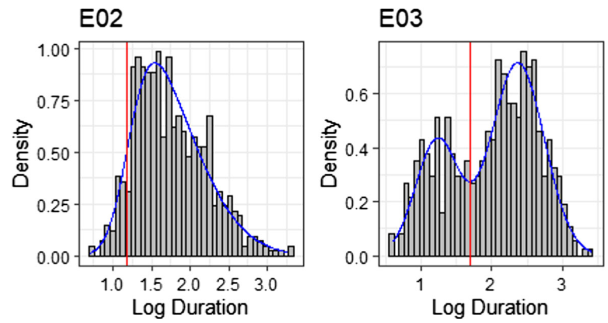

FIG. 4. Histogram and best FMM fit profile for log attempt duration (base 10) of modules E02 (1-component skewed normal) and E03 (two-component normal). Red line indicates the cutoff for short attempts estimated from the FMM fit.

B. Hypothesis testing

H1: Differences in attempt duration: As listed in Table II, in only 3 out of 26 modules did we detect a statistically significant difference in the distribution of first attempt durations between self-reporting and non-self-reporting students ($\alpha = 0.05$). No significant differences were detected for correct attempt duration on any problem. On all three modules, self-reporting students spent longer on average on their first attempt, as shown in the example in Fig. 2.

H2: Differences in fraction of short attempts: In Table III, we list the modules for which a significant difference was found comparing the fraction of short attempts submitted by self-reporting and non-self-reporting students. When comparing first attempts, only 2 out of 26 modules had a significant difference. In both cases non-self-reporting

TABLE II. Modules for which a significant difference in the distribution of attempt duration was detected.

Type	Module	p value
First attempt	1D07	0.02*
First attempt	E08	0.01*
First attempt	AM08	0.03*

* indicates $p < 0.05$.

TABLE III. Modules for which a significant difference between the fraction of short attempts among self-reporting (SR) and non-self-reporting (other) students were found. Columns SR and other list the fraction of short attempts detected for each population.

Module	Attempt	Cutoff	p value	SR	Other
E08	First	94.5	0.03*	0.32	0.47
AM08	First	15.6	0.00**	0.27	0.51
AM05	Correct	23.4	0.02*	0.27	0.46
AM08	Correct	15.6	0.04*	0.38	0.53

* indicates $p < 0.05$.
** indicates $p < 0.01$.

TABLE IV. Modules with marginally significant differences in fraction of correct short first attempts between self-reporting and non-self-reporting students.

Module	Attempt	p value	SR	Other
AM04	First	0.09	0.17	0.57
E09	First	0.07	0.24	0.48
AM05	All	0.06	0.30	0.51

students had higher fractions of short attempts (show in columns SR and other of Table III). When comparing all correct attempts, AM08 remains significant, but E08 is not. Instead, AM05 is significantly different.

H3: Differences in fraction of correct short attempts: When comparing the fraction of correct answers among short first attempts, we found no statistically significant differences between the self-reporting and non-self-reporting students on any problem, with two problems having p values less than 0.1. As listed in Table IV, the differences in correct fractions were greater than 0.2, but not significant likely because there were too few short attempts for sufficient statistical power on those problems. Even when we included all short attempts, there were still no statistically significant differences on any module, with one module being marginally significant, as listed in Table IV.

IV. DISCUSSION AND FUTURE DIRECTION

We analyzed attempt duration data from the assessment components of 26 modules, and compared between self-reporting and non-self-reporting students according to three hypotheses about rapid answer copying.

If we assume that students' self-report is trustworthy, and that the self-reporting population's problem-solving behavior is representative of students who completed homework independently, then our results suggest that rapid answer copying is uncommon and isolated on only a few modules. For H1, on only 3 out of 26 modules did we detect a difference in the distribution of attempt duration. For H2, on just two of the modules (E08 and AM08) did we find that self-reporting students are about 15%–20% less likely to submit a short attempt (H2). Both modules where either the last or close to the last module in the sequence, with AM08 being the last assigned module in the semester. This finding agrees with the findings of Warnakulasooriya *et al.* [9] that showed answer copying was more likely to occur on the last few problems in a long assignment.

Regarding H3, we did not detect a statistically significant difference in the fraction of correct short attempts on any of the modules. However, this could have been due to lack of statistical power on some modules, since the differences in correct percentage can be as large as 40% in the case of AM04.

Overall, the current analysis found little evidence of widespread rapid homework answer copying in our Physics I course taught during the pandemic, as significant differences were detected on just 2 out of 26 modules.

It must be emphasized that the current attempt-duration-based analysis only measures rapid answer copying, where students submit their answer without trying to solve the problem properly or even read the problem body. Alternatively, students may also copy answer after spending adequate time trying (and failing) to solve the problem. However, this form of answer copying is also unlikely to be overly prevalent, since we also checked the differences in correct percentage between the self-reporting and non-self-reporting students on normal duration first attempts and found only 2 modules to be statistically significant. In addition, it can be argued that answer copying after spending adequate time on solving the problem may reflect less of students' lack of motivation, and more of ineffective instructional material.

Another potential caveat of the current analysis is that it assumes that the fraction of guessing attempts among self-reporting and non-self-reporting students who completed the problems independently are similar. However, if significantly more students who guessed on problems chose to self-report, then the currently analysis will underestimate the fraction of answer copying in the class. Future studies could examine the validity of this assumption by comparing other aspects of problem-solving behavior indicative of answer copying, such as the frequency of making consecutive correct short first attempts on multiple problems. This behavior could be extracted and visualized using techniques such as process mining [25] and sequence mining [26].

Finally, regarding the true duration of problem solving, a previous study [24] found that students who complete the modules while being proctored in a classroom are significantly less likely to make an attempt under 15 s compared to the rest of the student population. An interesting future direction would be to compare and consolidate data from the current study and the previous study, to gain further insight into what actually causes short attempts of different durations.

ACKNOWLEDGMENTS

The authors would like to thank the Center for Distributed Learning at the University of Central Florida for developing and hosting the Obojobo platform for the implementation of OLMs. This research is supported by NSF Grant No. DUE-1845436.

[1] H. Busch, One method for inhibiting the copying of online homework, Phys. Teach. **55**, 422 (2017).
[2] T. Lancaster and C. Cotarlan, Contract cheating by stem students through a file sharing website: A covid-19 pandemic perspective, Int. J. Educ. Integr. **17**, 3 (2021).
[3] A. Edgcomb, F. Vahid, R. Lysecky, and S. Lysecky, Getting students to earnestly do reading, studying, and homework in an introductory programming class, Proc. Conf. Integr. Technol. into Comput. Sci. Educ. ITiCSE 171 (2017).
[4] G. Alexandron, J. A. Ruiperez-Valiente, Z. Chen, Pedro J. Muñoz-Merino, and D. E. Pritchard, Copying@scale: Using harvesting accounts for collecting correct answers in a MOOC, Comput. Educ. **108**, 96 (2017).
[5] D. J. Palazzo, Y.-J. Lee, R. Warnakulasooriya, and D. E. Pritchard, Patterns, correlates, and reduction of homework copying, Phys. Rev. ST Phys. Educ. Res. **6**, 010104 (2010).
[6] M. Amzalag, N. Shapira, and N. Dolev, Two sides of the coin: Lack of academic integrity in exams during the corona pandemic, students' and lecturers' perceptions, J. Acad. Ethics (2021).
[7] Z. Chen, G. Garrido, Z. Berry, I. Turgeon, and F. Yonekura, Designing online learning modules to conduct pre- and post-testing at high frequency, in *Proceedings of the 2017 Physics Education Research Conference, Cincinnati, OH* (AIP, New York, 2018), pp. 84–87.
[8] Z. Chen, M. Xu, G. Garrido, and M. W. Guthrie, Relationship between students' online learning behavior and course performance: What contextual information matters?, Phys. Rev. Phys. Educ. Res. **16**, 010138 (2020).
[9] R. Warnakulasooriya, D. J. Palazzo, and D. E. Pritchard, Time to completion of web-based physics problems with tutoring, J. Exp. Anal. Behav. **88**, 103 (2007).
[10] D. L. Schnipke and D. J. Scrams, Modeling item response times with a two-state mixture model: A new approach to measuring speededness, J. Educ. Measure. **34**, 213 (1997).
[11] S. L. Wise and X. Kong, Response time effort: A new measure of examinee motivation in computer-based tests, Appl. Meas. Educ. **18**, 163 (2005).
[12] Z. Chen, How the abrupt shift online led to a permanent shift in course design, https://www.aaas-iuse.org/resource/course-design/.
[13] Z. Chen, Obojobo sample canvas course, https://canvas.instructure.com/courses/1726856.
[14] Z. Chen, S. Lee, and G. Garrido, Re-designing the structure of online courses to empower educational data mining, in *Proceedings of the 11th International Educational Data Mining Conference, Buffalo, NY*, edited by K. Elizabeth Boyer and M. Yudelson (2018), pp. 390–396.
[15] K. M. Whitcomb, M. W. Guthrie, C. Singh, and Z. Chen, Improving accuracy in measuring the impact of online instruction on students' ability to transfer physics problem-solving skills, Phys. Rev. Phys. Educ. Res. **17**, 010112 (2021).
[16] D. L. Schwartz and J. D. Bransford, Efficiency and innovation in transfer, in *Transfer of Learning from a Modern Multidisciplinary Perspective (Current Perspectives on Cognition, Learning and Instruction)*, edited by J. P. Mestre (Information Age Publishing Inc., Charlotte, NC, 2005), pp. 1–51.
[17] J. W. Morphew, G. E. Gladding, and J. P. Mestre, Effect of presentation style and problem-solving attempts on metacognition and learning from solution videos, Phys. Rev. Phys. Educ. Res. **16**, 010104 (2020).
[18] M. O. Prates and C. R. B. Cabral, Mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions marcos, J. Stat. Softw. **30**, 1 (2009).
[19] Y. Sakamoto, M. Ishiguro, and G. Kitagawa, in *Akaike Information Criterion Statistics*, edited by D. Reidel (Taylor & Francis, Dordrecht, Netherlands, 1986), Vol. 81, p. 26853.
[20] A. A. Neath and J. E. Cavanaugh, The Bayesian information criterion: Background, derivation, and applications, Wiley Interdiscip. Rev. Comput. Stat. **4**, 199 (2012).
[21] P. A. A. Resende and C. C. Y. Dorea, Model identification using the efficient determination criterion, J. Multivariate Anal. **150**, 229 (2016).
[22] M. Marbac and M. Sedki, Variable selection for model-based clustering using the integrated complete-data likelihood, Stat. Comput. **27**, 1049 (2017).
[23] C. C. Y. Dorea, P. A. A. Resende, and C. R. Gonçalves, Comparing the Markov order estimators AIC, BIC and EDC, in *Transactions on Engineering Technologies: World Congress on Engineering and Computer Science 2014* (Springer Netherlands, 2015), pp. 41–54.
[24] M. W. Guthrie, T. Zhang, and Z. Chen, A tale of two guessing strategies: Interpreting the time students spend solving problems through online log data, *Proceedings of the 2020 Physics Education Research Conference*, virtual conference (AIP, New York, 2020), pp. 185–190.
[25] J. Saint, Y. Fan, S. Singh, D. Gasevic, and A. Pardo, Using process mining to analyse self-regulated learning: A systematic analysis of four algorithms, *LAK21 11th International Learning Analytics and Knowledge Conference* (ACM, New York, 2021), pp. 333–343.
[26] J. Jovanović, D. Gašević, S. Dawson, A. Pardo, and N. Mirriahi, Learning analytics to unveil learning strategies in a flipped classroom, Internet High. Educ. **33**, 74 (2017).