Hybrid teaching: A tale of two populations

Gerd Kortemeyer [*]
Educational Development and Technology, ETH Zurich, 8092 Zurich, Switzerland

Wolfgang Bauer [†] and Wade Fisher [‡]
Department of Physics and Astronomy, Michigan State University, East Lansing MI 48824, USA

(Received 6 July 2022; accepted 17 October 2022; published 3 November 2022)

In a partially flipped, hybrid introductory physics course where students had a free choice between attending any lecture session in person or via video conferencing, and where recordings of the lecture sessions were made available for asynchronous viewing, a total of 16 learner attributes and their relationships were investigated. Five of these attributes reflect participation choices, while eleven attributes reflect assessment outcomes on different course components. In line with the "no significant difference phenomenon," correlations between exam scores and participation choices were weaker than correlations with, for example, prior knowledge as evidenced by pretest scores. Overall, in terms of correlations, participation, and assessment attributes clustered together, respectively, with clicker questions being a connecting attribute between the clusters. Performance aside, we found two populations in the course, which, divided along the line of above and below average in-class attendance, exhibited other distinct behavior attributes mostly related to investment of time and effort in the course.

DOI: 10.1103/PhysRevPhysEducRes.18.020130

I. INTRODUCTION

For over two decades, the "no significant difference phenomenon" [1] has been an adage in what used to be called "tele-education:" learning outcomes hardly depend on the course-delivery method. The COVID-19 pandemic revived this discussion, as many courses had to switch to "tele-education" from one week to the other, and in the aftermath, universities have been considering what the "new normal" should look like [2–4].

COVID-19 has been characterized as a large-scale experiment in education [5,6], but one might argue that the term "experiment" implies controlled conditions, protocols, and environments, while the COVID-19 response can hardly be characterized as such. The real work might start in the aftermath: new avenues for teaching have gained acceptance, which can now be systematically investigated. The course under investigation in this study happened during the endemic phase of COVID-19, where maximum flexibility was offered to the students: they were expected to read materials online prior to class, given free choice whether to attend lecture sessions in class or via Zoom, and recordings of the lecture sessions were made available.

[*]Also at Michigan State University, East Lansing, USA.
kgerd@ethz.ch
[†]bauerw@msu.edu
[‡]fisherw@msu.edu

II. SETTING

The study was carried out in a first-semester physics course for scientists and engineers, taught by W. B. and W. F. The course is calculus-based and deals with introductory mechanics. There were four lectures per week; due to the COVID-19 pandemic, the course was offered in a hybrid fashion: students were given a free choice whether to attend lecture inside the classroom or via Zoom. Transmission was in real time, and in both settings, students could participate in answering clicker questions via a smartphone app, for which they received differential credit based on correctness (base points for participation and extra points for correct answers). Lectures were also recorded and made available for learners to watch anytime.

The course was taught in partial flipped-classroom mode [7], where the students were expected to read the materials [8] before class and answer some basic prereading questions, which were scored; however, part of the classroom time was still used for content transmission, so the questions and materials were used for preparation [9] rather than complete substitution of all traditional lecturing (as, for example, in Just-in-Time-Teaching models [10]). There was weekly, graded online homework with multiple allowed tries at the end of a topic week [11], and there were six quizzes ("midterm exams") distributed across the semester weeks [12] in addition to the final exam. All quizzes and exams were carried on site under supervision. Not part of the grade was a conceptual pre- and post-test, namely, the Force and Motion Conceptual Evaluation (FMCE) [13].

There was no lab directly associated with the course. Instead, lab is offered as a separate course with a different course number, which can be taken concurrently or in a subsequent semester.

III. METHODOLOGY

Table I lists the sixteen learner attributes that we considered in our study, as well as their labels (shown in italics) and ranges. Five of these attributes are self-reported by the learners on the following survey, which was administered online via the course management system between the last lecture and the final exam:

1. In a typical week, how many lectures did you attend on site in the lecture hall? (*inClass*)
2. In a typical week, how many lectures did you attend online via Zoom? (*zoom*)
3. In a typical nonexam week, how many minutes of recorded lectures did you watch? (*tRecNoEx*)
4. In a typical exam week, how many minutes of recorded lectures did you watch? (*tRecEx*)
5. After the pandemic is over, would you prefer to continue with your current mixture of zoom and classroom sessions? (*cont*)

Students were aware that per research protocol their survey answers were embargoed until after grades were submitted to the Office of the Registrar. 285 students (64% of the enrolled students) consented to participate in the study and filled out this survey.

The remaining ten attributes were gathered from the course management system:

- The number of solved (*nSol*) or unsolved (*nNoSol*) online homework problems; previous research shows that these are only weakly correlated to exam scores [14], since they are partly tainted by unproductive behavior (copying and guessing) [15].
- The numbers of tries for solved (*aSolTry*) and unsolved (*aNoSolTry*) online homework problems might indicate some level of persistence and commitment, even though they are are not always used productively [16]. Homework copying is a problem in physics courses [17], and a small number of attempts on solved problems has been used as an indicator for copying [18]. Initial concerns that copying of homework might increase during COVID-19 appear to be unfounded [18], suggesting that the propensity to copy might be an attribute of the learner rather than the course setting.
- Performance on clicker questions (*clicker*) are meaningful indicators of learning [19]. Here, students attending in class have the added benefit of being able to discuss the questions with their neighbors; these discussions were shown to increase rather than decrease the efficacy of these questions in terms of discrimination [20]. Apparently, there may be a strong beneficial aspect to the peer instruction conducted during these discussion with neighbors; taking advantage of this benefit would have been more cumbersome for students participating via Zoom, as they would have needed to text each other or use online group chats in addition the built-in plenary chat window.
- The reading problems (*preRead*) were part of the flipped-classroom design of this course. As opposed to the end-of-chapter type online homework problems, these questions were on the *knowledge* and possibly *comprehension* levels of Bloom's taxonomy [21];

TABLE I. Learner attributes used for this study, including their range, averages, and standard deviation.

Name	Description	Values	Average	StdDev
aNoSolTry	Average number of tries for homework problems that are not solved	0…99	4.7	4.3
aSolTry	Average number of tries for homework problems that are solved	0…99	2.3	0.6
click	Clicker participation and correctness	0…142.4	102.6	24.7
cont	Should class continue to be hybrid?	yes = 1, no = 0	0.83	0.38
final	Final exam score	0…16	11.8	3.3
gainTest	Normalized change of FMCE scores	−1…1	0.57	0.46
inClass	Times per week in classroom	0…4	1.9	1.5
nNoSol	Number homework problems not solved	0…258	11.8	11.1
nSol	Number homework problems solved	0…258	231.1	35.7
postTest	FMCE post-test score	0…47	41.5	9.0
preRead	Preclass reading score	0…4300	3567	954
preTest	FMCE pretest score	0…47	36.0	9.6
quizzes	Sum of the six quiz scores	0…50	33.9	8.8
tRecEx	Minutes watched recordings in exam weeks	0…999	52.3	103.0
tRecNoEx	Minutes watched recordings in nonexam weeks	0…999	29.7	63.6
zoom	Times per week via Zoom	0…4	2.1	1.5

besides getting students to read the materials in the first place, it was reported that generally students are making productive use of this kind of questions [9].

- The scores from six quizzes (*quizzes*) and the final exam (*final*) are the result of randomized multiple-choice examinations with a mixture of numerical and conceptual questions.
- The FMCE [13] pretest and post-test scores (*preTest*, *postTest*) scores are the result of nongraded, voluntary participation in working through this conceptual inventory of mechanics concepts. The attribute *gainTest* is the normalized change of the FMCE scores [22], which is equivalent to the Hake normalized gain [23] for students who score better on the post-test than on the pretest,

$$gain\,Test = \begin{cases} \frac{post\,Test - pre\,Test}{100\% - pre\,Test} & \text{for } post\,Test > pre\,Test \\ \frac{post\,Test - pre\,Test}{pre\,Test} & \text{else} \end{cases}. \tag{1}$$

The Hake gain of the averages $\langle g \rangle = (\langle post\,Test \rangle - \langle pre\,Test \rangle)/(100\% - \langle pre\,Test \rangle) = 0.5$ positions the course in the realm of interactive-engagement courses in terms of overall learning gains.

245 of the 285 students who completed the survey also completed the pre- and post-FMCE, and these form the sample for this study (55% of the enrolled students).

IV. RESULTS

A. Attendance

Figure 1 shows the combinations of in-class and Zoom lecture sessions that the students stated to have attended per week on the average. While 58 students attended all lectures online (combination “$0+4$”), other popular choices were half and half (combination “$2+2$”) and all lectures face to face (combination “$4+0$”). Only 18 students stated to have regularly skipped lecture sessions (combined less than four sessions per week), while 17 students stated to have attended more than four of the four sessions—they may have misunderstood the questions or filled them out carelessly, or they might have tried to express combinations like “basically all in the lecture hall ($in\,Class = 4$), but occasionally via Zoom ($zoom = 1$), which would explain why 11 out of these 17 students stated five sessions total. Finally, students concurrently taking the lab course may have added up the hours, even though that course has a different course number. In any case, students clearly took advantage of the flexibility that free choice of in-class and Zoom attendance provided.

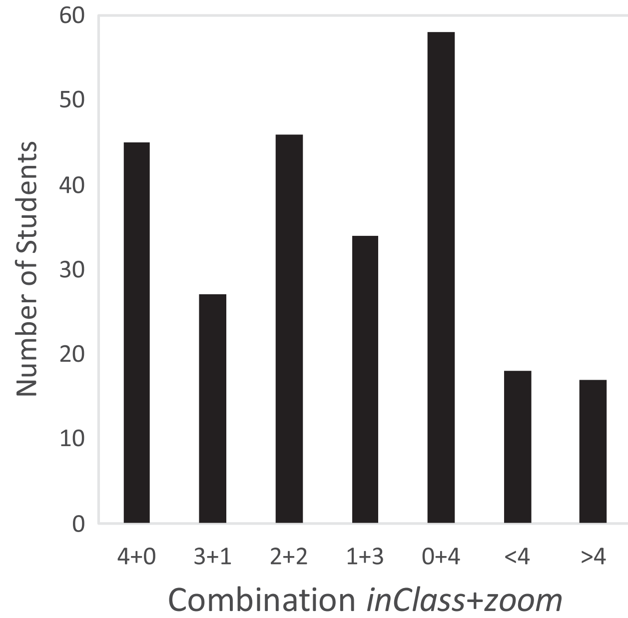

FIG. 1. Histograms of reported times per week attending lecture in the classroom and via Zoom. Combinations are labeled “*inClass* + *zoom*.” The bin “< 4” counts students who stated to have attended less than the four sessions per week, i.e., skipped class. The bin “> 4” counts students who stated to have attended more than four of the four sessions, i.e., filled out the survey incorrectly.

B. Averages

Table I also shows the averages and standard deviations of the attributes. As already seen in Fig. 1, students on the average visited the classroom less than twice a week, but with a wide distribution (1.9 ± 1.5 times). Usage of the recordings also varied widely; each lecture session has a length of 50 min, but as many students tend to watch at $1.5\times$ or $2\times$ speeds [24], one lecture could also equate to 33 or 25 min of linear video consumption—in reality, of course, students also skip around, pause the video, and repeat difficult segments [25]. While averages in exam weeks (before quizzes or the final) are higher than in nonexam weeks (one or two lectures depending on playback speed versus about half of that), statistically this difference vanishes in the large standard deviations.

The averages show that the video conference and recording scenario added much flexibility for the students, which might explain that over 80% of the students would like to continue this hybrid scenario.

C. Correlations between attributes

Table II shows the correlations between the learner attributes. Several of these are almost trivial; for example, the positive correlations between clicker, quiz, and final exam scores, and the correlations of those with online homework performance. Some correlations worth exploring might be as follows:

TABLE II. Correlations between the attributes listed in Table I. Moderate and higher correlations are indicated in boldface.

	aNoSolTry	*aSolTry*	*click*	*cont*	*final*	*gainTest*	*inClass*	*nNoSol*	*nSol*	*postTest*	*preRead*	*preTest*	*quiz*	*tRecEx*	*tRecNoEx*	*zoom*
aNoSolTry	**1**															
aSolTry	**0.33**	**1**														
click	−0.01	0.18	**1**													
cont	0	−0.09	0.02	**1**												
final	−0.04	0.16	0.29	−0.13	**1**											
gainTest	0.06	0.04	0.07	−0.08	0.06	**1**										
inClass	−0.05	0.16	**0.37**	−0.15	0.08	0.02	**1**									
nNoSol	−0.02	−0.03	−0.17	0.05	**−0.47**	−0.10	−0.01	**1**								
nSol	0.12	0.17	**0.35**	−0.02	**0.47**	0.12	0.06	**−0.74**	**1**							
postTest	0.01	0	0.05	−0.10	0.21	**0.73**	0.03	−0.11	0.08	**1**						
preRead	0.03	0.22	0.24	0.05	0.21	0.06	0.10	**−0.39**	**0.53**	−0.01	**1**					
preTest	−0.05	−0.03	0.14	−0.05	**0.31**	−0.03	0.12	−0.10	0.04	**0.41**	0	**1**				
quiz	−0.12	0.11	**0.33**	−0.01	**0.76**	0.04	0.14	**−0.50**	**0.48**	0.19	**0.30**	**0.40**	**1**			
tRecEx	0.02	−0.09	**−0.30**	0.08	−0.09	0.02	−0.17	0.01	−0.06	0	−0.11	−0.12	−0.22	**1**		
tRecNoEx	0.10	−0.02	−0.19	0.05	−0.05	0.01	−0.16	0.06	−0.10	−0.01	−0.12	−0.04	−0.19	**0.55**	**1**	
zoom	0.02	−0.12	−0.23	0.19	−0.11	0.01	**−0.84**	0	−0.03	−0.03	−0.06	−0.13	−0.16	0.19	0.21	**1**

- Usage of video recordings is moderately negatively correlated with clicker scores. There may be many explanation: the asynchronous nature of recordings versus the synchronous "think-on-your-feet" nature of clicker questions might appeal to different learners, or learners not doing well on clicker questions might watch the videos to review.
- Usage of video in and outside of exam weeks is strongly positively correlated, suggesting that some learners generally have a preference for watching them regardless of the point in the semester, while others do not.
- Usage of video recordings is negatively correlated with attendance of in-class lectures, but positively correlated with Zoom attendance; a preference for the online medium carries over to the recordings. It is also consistent with earlier findings that students not attending face-to-face lectures tend to make more use of online materials [26].
- Clicker scores are positively correlated with in-class attendance and negatively correlated with Zoom attendance, suggesting that either learners pay closer attention in-class or profit from discussing questions with their neighbors. However, it is also possible that better-performing students are also the ones who bother walking to the classroom.
- In-class attendance is weakly positively correlated with quiz and exam scores; this could be interpreted as students learning better in person; however, it is also positively correlated with the pretest, which may once again suggest that simply the better-performing or more interested students tend to go to class.
- The desire to continue the hybrid scenario is weakly negatively correlated with quizzes and the final. One possible explanation might be that online attendance, which is positively correlated with the desire to continue hybrid mode, might be perceived as allowing students to invest less time and effort into the course, and less effort also goes along with lower exam scores.

Figure 2 graphically shows the correlations in Table II using a force-directed Fruchterman-Reingold graph [27,28]; the attributes (vertices) are modeled to have a repulsive force by default (repelling each other), but are pulled together by the absolute value of the correlation, modeled as springs (edges). After simulating these forces for several iterations, the graph settles into a (local) energy minimum [29], and clusters of strongly correlated vertices become visible.

The graph suggests two separated clusters: what may be called the performance cluster, consisting of the quiz, exam, pre-reading and homework scores, and the learning behavior cluster, consisting of in-class and Zoom attendance, as well as the usage of recordings. In other words, the attributes reflecting learning behavior with respect to the hybrid nature of the course form an intercorrelated cluster that has only weak correlations with a cluster of intercorrelated attributes reflecting performance. The clicker

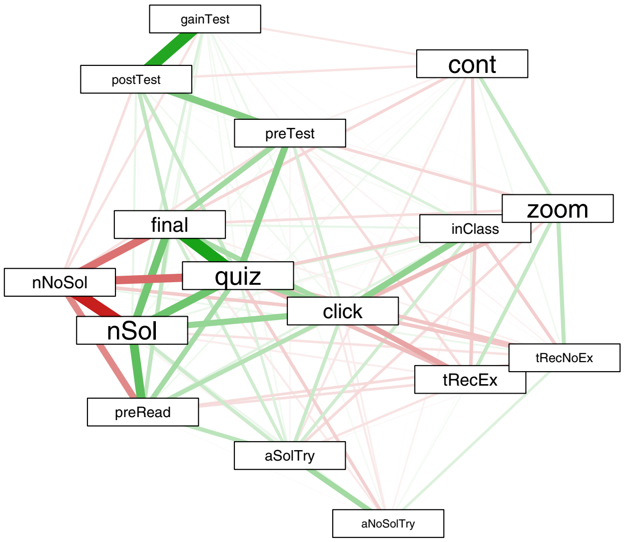

FIG. 2. Force-directed graph of the correlations Table II. Green edges represent positive correlations, red edges negative correlations, and the saturation and width denote strength.

performance attribute provides the strongest bridges between these clusters.

The two attributes describing the number of tries on homework appear disconnected, and so does the attribute reflecting the desire to continue the hybrid mode.

D. Predictors of exam score

Using the final exam as a measure of learning, 80% of the variation of the final exam score could be explained by the remaining 15 attributes ($R^2 = 0.797$). However, by far the strongest predictor is the quiz scores ($p < 0.0001$). While that is hardly surprising, the next most reliable predictor is the desire to continue hybrid mode: students who want to continue hybrid mode are predicted to lose one out of sixteen points on the final exam (regression coefficient −1.04)—not a strong effect, but reliable ($p < 0.01$).

When combining the closely correlated quiz and final exam scores, 70% of the variation of these total exam scores could be explained by the remaining 14 attributes ($R^2 = 0.697$). The most reliable predictors now become the pre-test FCME score and the number of unsolved homework problems ($p < 0.0001$). Table III shows the multiple linear regression results for all attributes. Apparently, being good at physics to begin with and solving all of the homework is a better predictor for exam success than attending lecture (regardless of in-class or via Zoom).

E. Learner similarity and clustering

The correlations suggest there are different classes of learners, having distinctly different preferences and expectations for the course. To explore possible groupings across attribute combinations, the learner attributes were linearly scaled between 0 and 1, with 0 being mapped to the smallest occurring attribute value and 1 to its highest occurring value; the result was considered a vector. The pairwise similarity between the attribute vectors can be calculated to construct a similarity matrix using cosine similarity [30,31], which can be visualized as the force-directed graph [27,28] in Fig. 3. Learners are represented as the nodes in this graph, connecting edges are only drawn if their cosine-similarity is higher than 0.95.

TABLE III. Multiple linear regression coefficients and p values for the prediction of the combined quiz and final exam scores.

	Coefficient	p value
Intercept	12.954	0.084
aNoSolTry	−0.383	0.006
aSolTry	2.270	0.030
click	0.065	0.018
cont	0.266	0.859
gainTest	−1.350	0.512
inClass	−1.307	0.067
nNoSol	−0.296	<0.0001
nSol	0.070	0.008
postTest	0.096	0.400
preRead	<0.0001	0.995
preTest	0.340	<0.0001
tRecEx	−0.008	0.219
tRecNoEx	−0.002	0.859
zoom	−1.478	0.039

In the graph, the *inClass* attribute is indicated by the colors of the nodes, where below-average weekly classroom attendance (zero or 1 time per week) is indicated by blue colors, while above-average classroom attendance is indicated by beige colors (2, 3, or 4 times per week).

There are only very few distinct clusters discernible:

- The cluster at the top of the graph (nodes 4, 51, 244, 183, etc.) consists exclusively of learners who indicated that they attended all lectures inside the classroom ($inClass = 4$). The average of their final exam scores is significantly beyond course-wide average ($\overline{final} = 14.2 \pm 1.2$; see Table I for course-wide averages), and so are their post-test scores ($\overline{postTest} = 46.7 \pm 0.5$). This seems to be a cluster of high-achieving students.
- The attribute averages of the learners in the cluster on the left side of the graph (nodes 136, 187, 17, 114, etc.) are generally consistent with the course-wide averages, with two exceptions: none of the students in this cluster indicated that they would want to continue hybrid mode ($cont = 0$), i.e., all of these students want to discontinue the online offerings, but at the same time, they had below-average class attendance ($\overline{inClass} = 0.7 \pm 0.8$). This seems to be a cluster of students whose performance was average, but who in retrospect might regret not having gone to class.
- The outliers at the bottom of the graph (nodes 139, 164, 2, 73, etc.) have little in common, except that they

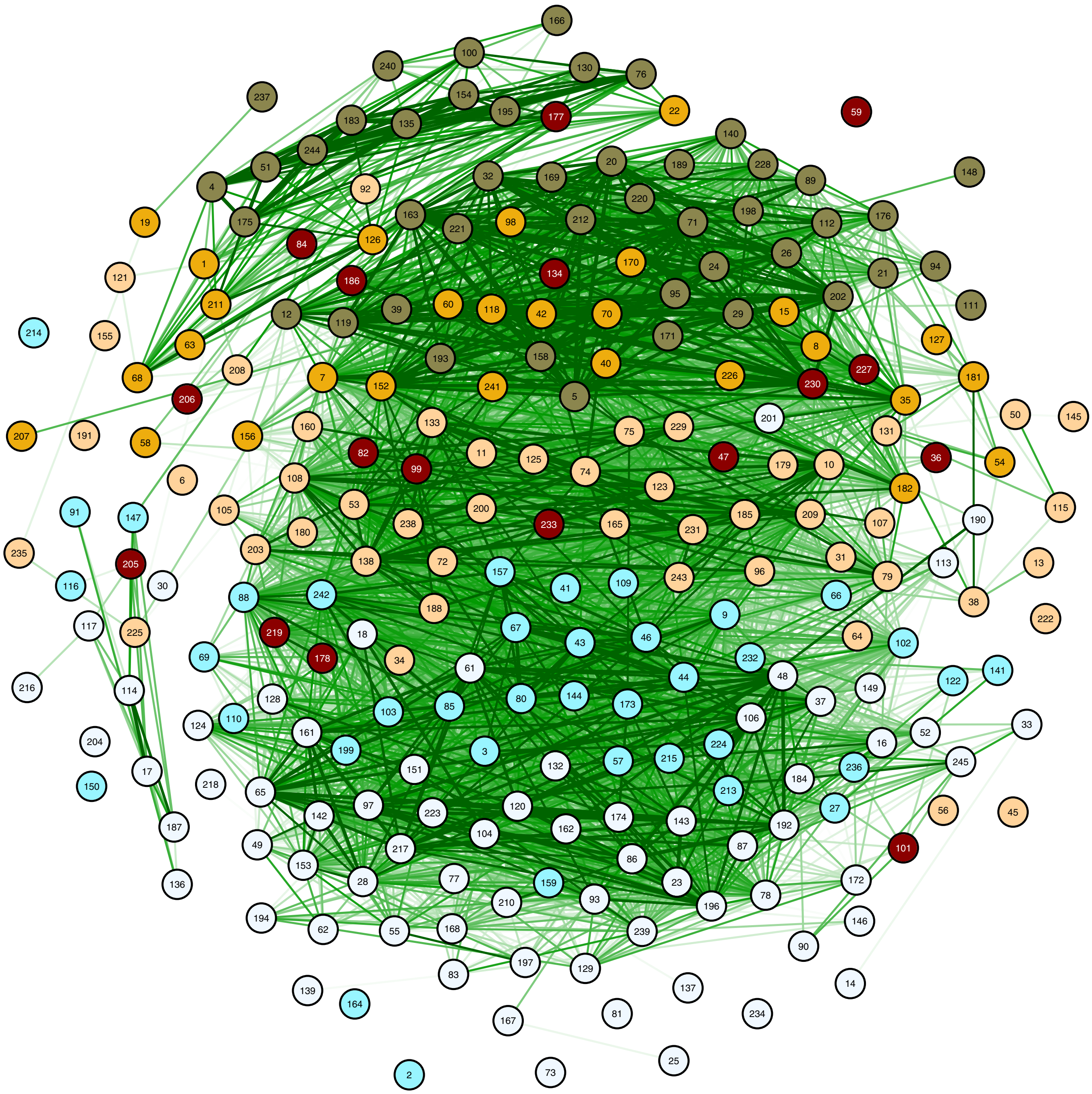

FIG. 3. Force-directed graph of learner similarity based on cosine similarity between the scaled attributes. The light blue color indicates learners who report not visiting the classroom at all, the slightly darker blue learners who report on the average visiting the classroom once per week. The beige colors, with increasing darkness, indicate learners who report on the average visiting the classroom two, three, or four times per week. The dark red indicates learners who had reported more than five sessions per week (rightmost category in Fig. 1). The edges, with increasing thickness and saturation, indicate similarities beyond 0.95.

rarely went to class and on the average did worse on the quizzes ($\overline{quizzes} = 21.9 \pm 7.9$). Their performance on the final exam, however, was widely distributed ($\overline{final} = 8.0 \pm 5.1$), and while below course-wide average, not significantly so.

Any proposals as to the underlying, latent attribute of cluster formations (e.g., regret) are speculative and would need to be confirmed through interviews of cluster members.

There is, however, a large-scale pattern of diminishing classroom attendance from the top to the bottom of the

TABLE IV. Fisher's exact test comparison of populations with above- and below-average classroom attendance with respect to above- and below-average values on the remaining attributes.

	Odds ratio	p value
aNoSolTry	0.8	0.5005
aSolTry	2.1	0.0062
click	3.7	<0.0001
cont	0.6	0.1264
final	1.1	0.7931
gainTest	0.9	0.5986
nNoSol	0.8	0.5094
nSol	1.5	0.1659
postTest	1.1	0.887
preRead	1.8	0.0443
preTest	1.6	0.0884
quiz	1.8	0.0272
tRecEx	0.5	0.0082
tRecNoEx	0.7	0.148
zoom	<0.01	<0.0001

graph, which is striking, since this difference in one of two dimensions in the graph is caused by one in sixteen dimensions of the attribute vector. In other words, learner similarity across all attributes aligns with classroom attendance. How do students with above-average classroom attendance differ from those with below-average classroom attendance?

F. Population differences

To find out how these populations differ, Fisher's exact test can be carried out on contingency tables for being above or below the average of attributes (see Table I). As Table IV shows, students who visit the classroom more than twice a week (above-average classroom attendance) are

- twice as likely to invest above-average numbers of tries on solved homework,
- almost four times as likely to have above-average clicker points,
- almost twice as likely to have above-average reading scores,
- 1.6 times as likely to have above-average pretest scores,
- almost twice as likely to have above-average quiz scores, and
- twice as likely to only make below-average use of recordings in exam weeks.

Many of these likelihoods reflect investment of time and effort, rather than necessarily measures of learning. For example, formative assessment activities such as preclass reading, clicker points (where credit is given also for wrong answers), number of homework tries and even homework performance reward effort, diligence, and persistence as much or even more than correctness of answers. Above-average face-to-face attendance is associated with above-average investment in these time-and-effort components of the course, more so than with above-average performance.

As an aside, it is worth considering the implications of the finding regarding the number of tries needed to solve homework problems. While an argument was made that this suggests persistence and investment of time and effort, this could also suggest that the students visiting the classroom might be having a harder time with the homework than the online students. There might be at least two explanations for this:

- The online students might not find it necessary to come into the classroom because physics comes to them more easily.
- The online students are also better connected online in general, where they can find the answers to most if not all of the homework; a low number of tries has been associated with copying [15].

Since the course-wide average number of tries to solve a homework problem is 2.3, three tries—which is a reasonable number for somebody just learning the content—is already above average.

It cannot be overemphasized that these likelihoods do not imply causation: some students may not have invested time into face-to-face attendance and these activities simply because they were able to follow the course and do well on quizzes without this investment. By the reverse token, students who came to class might simply have a higher interest in physics.

V. DISCUSSION

No significant difference could be found in exam performance between students who primarily attended lecture sessions face to face versus those who attended online. However, this does not necessarily mean that the two attendance modes are equivalent, since students were free to choose: it is an unanswered question how students who chose to come to class would have performed if forced to attend online and vice versa. Efforts to answer this question by considering pre-COVID-19 data, when all lectures were face to face, and mid-COVID-19 data, when all lectures were forced online, were inconclusive due to the sheer amount of other variables that changed—COVID-19 was not a controlled experiment, and conducting a controlled experiment in future semesters that forces students into particular attendance modes does not seem ethical.

While no significant difference could be found in exam performance, face-to-face and online audiences appear to be two different populations. Students with above-average face-to-face attendance appear to generally invest more time and effort into the course, but we are not able to make any statements regarding causality; latent attributes such as "aptitude," "expert likeness," "expectations," or "interest" were not investigated; this would have required epistemological surveys such as MPEX [32] or CLASS [33].

Face-to-face instruction has intangible benefits, such as socialization, peer support, motivation, self-regulation, self-efficacy [34] and less risk of mental health problems [35]. Slightly higher clicker scores associated with in-class instruction might be symptoms of the high-investment behavior, but they could also be the result of fruitful discussions with peer learners, which are less likely in online scenarios. The development of expertlike epistemologies was not assessed, and one might speculate that face-to-face instruction could be less detrimental than online instruction [36,37]; this would be a worthwhile future study.

Overall, the "no significant difference phenomenon" with respect to course-delivery medium was once again confirmed. Rather than focusing on the medium, considerations regarding the mode of instruction should continue to move into the foreground: there is a "very significant difference phenomenon" when it comes to using activating, research-based instructional strategies [38]—in this study, FCME gains were consistent with those of other courses using activating techniques, independent of how the students attended the course. Motivating students to spend more time with the physics content leads to better performance, independent of the delivery method.

VI. CONCLUSION

The real question is whether to continue hybrid mode beyond COVID-19 or not. As our study indicates, this mode enabled two populations to emerge: a low-investment population, which generally has a preference for online sessions and recordings, and a high-investment population, which tends to spend more time on course activities and prefers traditional face-to-face lecture sessions. Should instructors enable this low-investment behavior? If they trust exams as indicators of learning, the answer is a clear "why not?," as apparently no harm came to this population.

External factors, such as prior knowledge of physics, are much stronger predictors of exam success, and there seems to be little justification for making students invest more time and effort than they apparently need by making them come to class and stopping to make available recordings. By the reverse token, the study provides no basis for discontinuing face-to-face instruction: there is a population who wants or needs to invest this extra time and effort; we do not know what would have happened to this population had it been forced online.

It appears that rather than focusing on the medium, instructors should continue to focus on the mode of instruction, i.e., the techniques and strategies employed to foster learning. Barring possible future results regarding the development of learner epistemologies, and as long as the effort on the part of the instructors and support personnel is not prohibitive, it seems that the hybrid modes of instruction should be continued.

ACKNOWLEDGMENTS

We thank the students in our course for their willingness to participate in this study. We also thank Anna Kortemeyer for helpful discussions.

[1] T. L. Russell, The No Significant Difference Phenomenon: As Reported in 355 Research Reports, *Summaries and Papers* (North Carolina State University, Raleigh, NC, 1999).

[2] G. Kortemeyer, That's one giant step for a university, one small leap for digitization, Bull. VSH-AEU **46**, 33 (2020).

[3] M. O. Schapiro, Let's not return to normal when the "new normal" finally arrives, Chronicle Higher Educ. **68**, The Review (2021), https://www.chronicle.com/article/lets-not-return-to-normal-when-the-new-normal-finally-arrives.

[4] S. I. Hofer, N. Nistor, and C. Scheibenzuber, Online teaching and learning in higher education: Lessons learned in crisis situations, Comput. Hum. Behav. **121**, 106789 (2021).

[5] M. S. Thomas and C. Rogers, Education, the science of learning, and the COVID-19 crisis, Prospects **49**, 87 (2020).

[6] B. Dunrong and L. Jin, Temporary action or new model experiment? Teaching at Chinese universities in the time of COVID-19, Int. Higher Educ. **102**, 18 (2020), https://www.internationalhighereducation.net/api-v1/article/!/action/getPdfOfArticle/articleID/2908/productID/29/filename/article-id-2908.pdf.

[7] S. Amanah, F. Wibowo, and I. Astra, Trends of flipped classroom studies for physics learning: A systematic review, J. Phys. Conf. Ser. **2019**, 012044 (2021).

[8] W. Bauer and G. Westfall, *University Physics with Modern Physics* (McGraw-Hill Higher Education, New York, 2013).

[9] C. E. Heiner, A. I. Banet, and C. Wieman, Preparing students for class: How to get 80% of students reading the textbook before class, Am. J. Phys. **82**, 989 (2014).

[10] G. M. Novak, E. T. Patterson, A. D. Gavrin, and W. Christian, *Just in Time Teaching* (Prentice Hall, Upper Saddle River, NJ, 1999).

[11] G. Kortemeyer, E. Kashy, W. Benenson, and W. Bauer, Experiences using the open-source learning content management and assessment system LON-CAPA in introductory physics courses, Am. J. Phys. **76**, 438 (2008).

[12] J. T. Laverty, W. Bauer, G. Kortemeyer, and G. Westfall, Want to reduce guessing and cheating while making students happier? Give more exams!, Phys Teach **50**, 540 (2012).

[13] R. K. Thornton and D. R. Sokoloff, Assessing student learning of Newton's laws: The Force and Motion Conceptual Evaluation and the evaluation of active learning laboratory and lecture curricula, Am. J. Phys. **66**, 338 (1998).

[14] B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer, and W. F. Punch, Predicting student performance: An application of data mining methods with an educational web-based system, in *Proceedings of the 33rd Annual Frontiers in Education, FIE 2003* (IEEE, Bellingham, WA, 2003), Vol. 1, pp. T2A–13.

[15] G. Kortemeyer, Extending item response theory to online homework, Phys. Rev. ST Phys. Educ. Res. **10**, 010118 (2014).

[16] G. Kortemeyer, An empirical study of the effect of granting multiple tries for online homework, Am. J. Phys. **83**, 646 (2015).

[17] 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).

[18] Z. Chen, Measuring the level of homework answer copying during COVID-19 induced remote instruction, Phys. Rev. Phys. Educ. Res. **18**, 010126 (2022).

[19] G. Kortemeyer, The psychometric properties of classroom response system data: A case study, J. Sci. Educ. Technol. **25**, 561 (2016).

[20] M. Aka, M. Akveld, A. Caspar, G. Kortemeyer, and M. Valkering-Sijsling, In-class formative assessment in an introductory calculus class, eleed **13** (2020), https://eleed.campussource.de/archive/13/5122.

[21] B. S. Bloom, Taxonomy of educational objectives: The classification of educational goals, *Handbook I Cognitive Domain* (Longmans, London, 1956).

[22] J. D. Marx and K. Cummings, Normalized change, Am. J. Phys. **75**, 87 (2007).

[23] R. R. Hake, Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, Am. J. Phys. **66**, 64 (1998).

[24] D. H. Murphy, K. M. Hoover, K. Agadzhanyan, J. C. Kuehn, and A. D. Castel, Learning in double time: The effect of lecture video speed on immediate and delayed comprehension, Appl. Cogn. Psychol. **36**, 69 (2022).

[25] L. O. Campbell, T. Planinz, K. Morris, and J. Truitt, Investigating undergraduate students' viewing behaviors of academic video in formal and informal settings, Coll. Teach. **67**, 211 (2019).

[26] G. Kortemeyer, Work habits of students in traditional and online sections of an introductory physics course: A case study, J. Sci. Educ. Technol. **25**, 697 (2016).

[27] T. M. Fruchterman and E. M. Reingold, Graph drawing by force-directed placement, Software: Practice and Experience **21**, 1129 (1991).

[28] S. Epskamp, R package: qgraph, R Foundation for Statistical Computing, https://cran.r-project.org/web/packages/qgraph/qgraph.pdf (2022).

[29] G. Kortemeyer, Virtual-reality graph visualization based on Fruchterman-Reingold using Unity and SteamVR, Inf. Visualization **21**, 143 (2022).

[30] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria (2018).

[31] K. Watanabe and R. Cannoodt, R package: proxyC, R Foundation for Statistical Computing, https://cran.r-project.org/web/packages/proxyC/proxyC.pdf (2022).

[32] E. F. Redish, R. N. Steinberg, and J. M. Saul, Student expectations in introductory physics, Am. J. Phys. **66**, 212 (1998).

[33] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).

[34] I. Marzoli, A. Colantonio, C. Fazio, M. Giliberti, U. S. di Uccio, and I. Testa, Effects of emergency remote instruction during the COVID-19 pandemic on university physics students in Italy, Phys. Rev. Phys. Educ. Res. **17**, 020130 (2021).

[35] M. Dew, L. Ford, D. T. Nodurft, T. Erukhimova, and J. Perry, Student responses to changes in introductory physics learning due to the COVID-19 pandemic, Phys. Teach. **59**, 162 (2021).

[36] K. E. Gray, W. K. Adams, C. E. Wieman, and K. K. Perkins, Students know what physicists believe, but they don't agree: A study using the CLASS survey, Phys. Rev. ST Phys. Educ. Res. **4**, 020106 (2008).

[37] K. A. Slaughter, S. P. Bates, and R. K. Galloway, The changes in attitudes and beliefs of first year physics undergraduates: A study using the class survey, Int. J. Innovation Sci. Math. Educ. **19**, 29 (2011).

[38] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, Phys. Rev. ST Phys. Educ. Res. **10**, 020119 (2014).