Editors' Suggestion

Cheat sites and artificial intelligence usage in online introductory physics courses: What is the extent and what effect does it have on assessments?

Gerd Kortemeyer [*]
Rectorate and AI Center, ETH Zurich, 8092 Zurich, Switzerland

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

(Received 27 December 2023; accepted 7 May 2024; published 23 May 2024)

As a result of the pandemic, many physics courses moved online. Alongside, the popularity of Internet-based problem-solving sites and forums rose. With the emergence of large language models, another shift occurred. One year into the public availability of these models, how has online help-seeking behavior among introductory physics students changed, and what is the effect of different patterns of online resource usage? In a mixed-method approach, we investigate student choices and their impact on assessment components of an online introductory physics course for scientists and engineers. We find that students still mostly rely on traditional Internet resources and that their usage strongly influences the outcome of low-stake unsupervised quizzes. We empirically found distinct clusters of help-seeking and resource-usage patterns among the students; the impact of students' cluster membership on the supervised assessment components of the course, however, is nonsignificant.

DOI: 10.1103/PhysRevPhysEducRes.20.010145

I. INTRODUCTION

The general assumption in teaching introductory physics courses is that we need to teach a few essential concepts, for example, conservation laws, and, based on those, to derive basic equations that govern the motion of all objects in our universe, from atomic nuclei to galaxies. Almost all of us teaching professionals subscribe to the notion that it aids the students' learning processes to flesh out these basic concepts with exercises, in class and as homework. While simply solving a large number of traditional homework problems is not sufficient to learn physics [1], mastering our subject requires time spent on wrestling with conceptual questions, and familiarity with basic equations is acquired by rearranging and combining them and solving for unknown quantities.

"I think, however, that there isn't any solution to this problem of education other than to realize that the best teaching can be done only when there is a direct individual relationship between a student and a good teacher–a situation in which the student discusses the ideas, thinks about the things, and talks about the things. It's impossible to learn very much by simply sitting in a lecture, or even by simply doing problems that are assigned. But in our modern times we have so many students to teach that we have to try to find some substitute for the ideal." These sentences were written by Richard Feynman in 1963 [2]!

Some of us, the authors included [3,4], have spent enormous amounts of time constructing learning management systems, audience feedback systems, simulation software, and other apps to help with giving the students an opportunity to spend meaningful time on task. The ultimate goal may have been to construct a software system that acts as a personal tutor for each individual student. The rapid rise of large language models (LLMs) and artificial intelligence (AI) systems gives hope that the substitute for the ideal that Feynman described might become reality. Alternatively, students might just use their new AI tools to sidestep their learning process and just let the AI complete the assignments for them.

For the last 25 years, Michigan State University has been running online sections of the introductory physics courses [5]. For decades, even preceding online courses, the prevalent notion regarding course delivery modes (distance or in-person) had been "no significant difference" [6–8]. More recently, we confirmed this for the courses under investigation in this study when we found no significant difference

[*]kgerd@ethz.ch
Also at Michigan State University, East Lansing, MI 48823, USA.
[†]bauerw@msu.edu

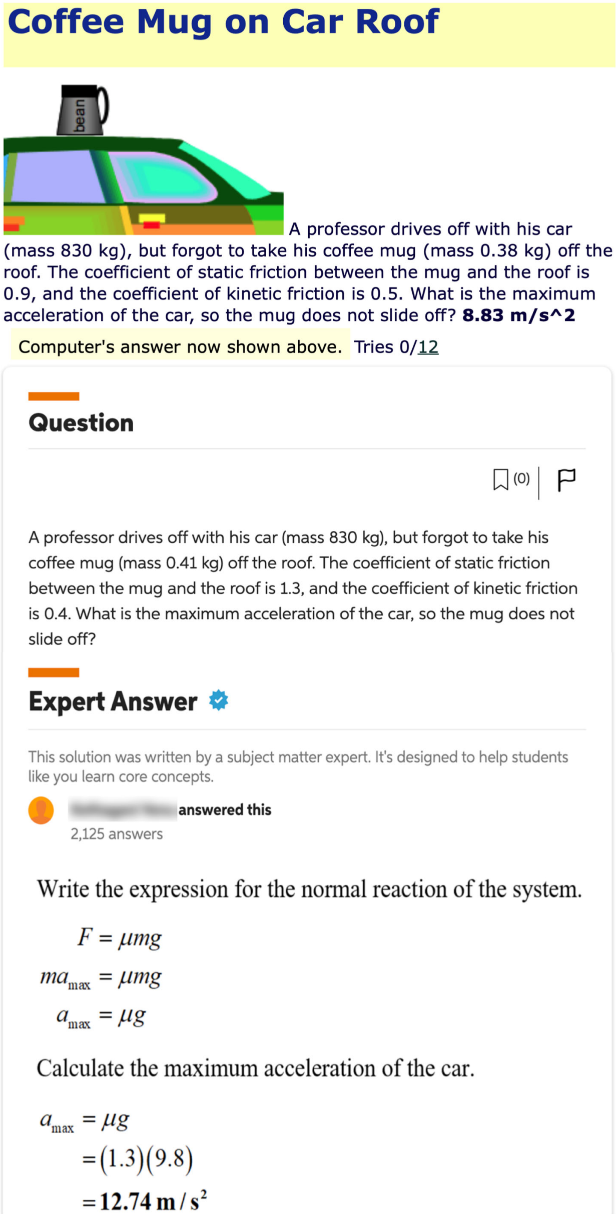

FIG. 1. A problem written and copyrighted by one of the authors (G. K.) in LON-CAPA [18] (top panel) and a typical solution found on a problem-solving site (bottom panel). The problem numbers are randomized, so the student would need to identify their own numerical values and insert them. Depicted is 1 of 50 answers found on the site.

between attendance choices with regard to learning success [9] and preparation for subsequent courses [10].

Since the courses first came online, the amount of additional resources available online has greatly increased [11], particularly during COVID-19. This brings about a whole set of additional challenges when administering online assessments [12–14]. The solution to virtually any introductory-physics problem is available shortly after it is published [15], including on commercial sites like Chegg [16,17]; a typical example can be seen in Fig. 1. Students find these resources useful for homework or unsupervised online exam problems that have been "recycled" from earlier in the semester (e.g., homework problems making an encore appearance on exams).

Instructors have been fighting these Internet sites and forums by editing the problem content, removing any references to problem numbers, randomizing the solutions, selecting problems from older editions of textbooks, or writing new problems every time [19]. All of these countermeasures rely on there being some time delay between publishing a problem and it appearing online (even contracted problem solvers need a little time) and on the sites having one static version of the problem (e.g., not being able to adapt to randomizations).

These same countermeasures against cheat sites will not work against artificial intelligence. Tools like ChatGPT [20] and Bard [21] deliver solutions *ad hoc* and on-demand, they are immune against changing wording or numbers, and they solve problems independent of them having been published days ago or appearing for the first time on an online exam. This is illustrated in Fig. 2, where the original friction problem has been modified by introducing additional distractors and different randomized numbers. Not only does the availability of these solutions depend on some "expert" having solved the problem before, but as opposed to the forum answer in Fig. 1, the response includes the actual numbers encountered by the student and the physics explanation is arguably better and more helpful.

Chegg can serve as a proxy for the popularity of online problem-solving sites: while the share price of Chegg (NYSE: CHGG) greatly increased when courses went online due to the onset of the pandemic in 2020, with the appearance of ChatGPT in late 2022, they dropped below prepandemic levels [22,23], see Fig. 3.

Physics may still have been spared from this, since large language models as "calculators for words" are still notoriously bad at math, and other online resources may still be more reliable. However, it has been shown that even older versions of popular AI chatbots can (barely) pass the assessment components of introductory physics courses [24] (though still struggling with math [25]) and exhibit respectable performance on nongraphical elements of physics concept tests [26], and newer versions make less calculation errors [27]. Students start to trust physics answers provided by AI [28].

In our study, we investigate if one year into the public availability of powerful large language models, online help-seeking behavior of students in an introductory physics course has shifted from traditional resources to AI. We also investigate the possible impact on the assessment components of such a course for student subpopulations with different help-seeking patterns.

You

A professor drives off with his car (mass 830 kg), but forgot to take his coffee mug (mass 0.38 kg) off the roof. The coefficient of static friction between the mug and the roof is 0.9, and the coefficient of kinetic friction is 0.5. What is the maximum acceleration of the car, so the mug does not slide off?

ChatGPT

To find the maximum acceleration of the car without causing the coffee mug to slide off, we need to consider the forces acting on the mug. The maximum force of static friction (which prevents the mug from sliding) is given by:

$$F_{\text{friction, max}} = \mu_{\text{static}} \times F_{\text{normal}}$$

where:

- $F_{\text{friction, max}}$ is the maximum static frictional force.
- μ_{static} is the coefficient of static friction (0.9 in this case).
- F_{normal} is the normal force, which is equal to the weight of the mug in this scenario.

The weight of the mug (F_{normal}) is given by:

$$F_{\text{normal}} = m \times g$$

where m is the mass of the mug (0.38 kg) and g is the acceleration due to gravity (approximately $9.81\,\mathrm{m/s^2}$).

Once the mug starts to slide, it will experience kinetic friction, which is not relevant for calculating the maximum static friction.

The maximum force of static friction is what keeps the mug from sliding off. This force is equal to the horizontal force exerted by the car's acceleration. Therefore, the maximum acceleration $a_{\max}$ of the car without causing the mug to slide is given by setting the maximum static frictional force equal to the horizontal force $F_{\text{horizontal}}$:

$$F_{\text{friction, max}} = F_{\text{horizontal}}$$
$$\mu_{\text{static}} \times m \times g = m_{\text{mug}} \times a_{\max}$$

Solving for $a_{\max}$, we get:

$$a_{\max} = \mu_{\text{static}} \times g$$

Let's calculate $a_{\max}$.

The maximum acceleration of the car, so that the coffee mug does not slide off the roof, is approximately $8.83\,\mathrm{m/s^2}$. [>-]

You

As a summer job, you work as a waiter in a restaurant, pushing around carts (mass 40 kg) with ordered food items. You depend on tips for fast service. Right now, you need to deliver a huge plate of spare ribs with fries (mass 0.35 kg). The coefficient of static friction between the plate and the cart is 0.7, and the coefficient of kinetic friction is 0.3. During your speedy service, what is the maximum acceleration of the cart, so the plate does not slide off? Neglect that the ribs and fries could fly off the plate, that's a separate problem.

ChatGPT

To solve this problem, we need to calculate the maximum acceleration of the cart that will prevent the plate of spare ribs and fries from sliding off. The key concept here is friction. Frictional force is what keeps the plate in place, and it's determined by the coefficient of static friction and the normal force acting on the plate.

Here are the steps to calculate the maximum acceleration:

1. **Calculate the Frictional Force:** The maximum frictional force F_{friction} that prevents sliding is given by the formula:
   $F_{\text{friction}} = \mu_s \times F_{\text{normal}}$
   Where μ_s is the coefficient of static friction, and F_{normal} is the normal force. In this case, the normal force is simply the weight of the plate of spare ribs and fries, which is the mass multiplied by gravity (9.8 m/s²).
2. **Relate Frictional Force to Acceleration:** According to Newton's second law, force equals mass times acceleration. The frictional force here is what provides the maximum possible acceleration without sliding. So, we can write:
   $F_{\text{friction}} = m \times a_{\max}$
   Where m is the mass of the plate of spare ribs and fries, and $a_{\max}$ is the maximum acceleration.
3. **Solve for Maximum Acceleration:** By combining these equations, we can solve for the maximum acceleration.

Let's perform the calculations.

The maximum acceleration of the cart, to ensure that the plate of spare ribs and fries does not slide off, is approximately $6.86\,\mathrm{m/s^2}$. This means you can accelerate the cart up to this limit without causing the plate to slide. [>-]

FIG. 2. The problem from Fig. 1 in its original form (left panel) and after on-the-fly modification (right panel), correctly solved by GPT-4 on the first attempt.

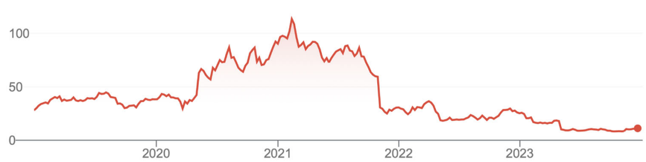

FIG. 3. Historic share prices in USD of Chegg (NYSE CHGG) as a proxy for the popularity of online problem-solving sites. Prices greatly increased when courses went online as a result of the pandemic but fell again dramatically coinciding with the emergence of large language models.

II. SETTING

Michigan State University is a public, large-enrollment (>50,000 students) R-1 university. Almost 78% of the undergraduate population is from Michigan. The online courses in this study were taught asynchronously using a variety of multimedia components [5].

The study takes place in a calculus-based introductory physics course sequence for scientists and engineers, where both a first-semester mechanics and a second-semester E&M course were offered during the fall semester 2023.

Each course offered several asynchronous video lessons every week and online homework using LON-CAPA [18]. No explicit statements were made regarding the use of external resources on the homework. Discussing homework with peers or course faculty and staff was actively encouraged, and the department offers a learning center for just this purpose.

The courses had 11 low-stakes weekly exams ("quizzes") [29], of which 9 were conducted online and 2 of which had to be taken on-campus under supervision. Faculty sanctioned the use of the textbook and the LON-CAPA materials during these exams, no other resources were allowed.

The course also included a high-stakes on-campus final exam. The final exams included five questions that were randomized duplicates of problems that had been assigned earlier in the semester. As a resource for the students, the course also offered an on-campus and online help room staffed by course faculty and staff.

At the end of the semester, a survey was given asking students to report how frequently they consulted artificial intelligence tools and other online resources during homework and online quizzes, and how often they conversed with fellow students and course faculty and staff while working on homework.

III. METHODOLOGY

A. Survey administration

The survey contained the following items, which for the numerical answers had sliders ranging from 0% to 100%:

- Homework: Estimate the percentage of your homework and lecture problems, for which the following is true:
  - — You used AI tools like ChatGPT, Khanmigo, … to solve them (*HwkAI*).
  - — You used Internet resources like help sites or forums to solve them (*HwkInt*).
  - — You consulted other student to solve them (*HwkPeer*).
  - — You consulted the TAs/prof to solve them (*HwkFac*).
- Online exams: Estimate the percentage of your online exam problems, for which the following is true:
  - — You used AI tool to solve them (*OnlAI*).
  - — You used Internet resources like help sites or forums to solve them (*OnlInt*).
- Your opinion: Please tell us what you think about using AI in online classes; what should ideally be done; what should not be done?

The survey was administered online during the last week of the semester, but results were not viewed or analyzed until the grades for the course had been turned in. A nominal participation credit was given for submitting the survey, regardless of whether or not the students agreed to be a part of the study. The students were aware of this protocol as part of the informed consent, and data were only analyzed for students who agreed to participate. The study was approved under MSU-IRB-STUDY00009987.

B. Considered variables

We compiled a range of variables that capture various aspects of student performance and behavior, shown in Table I. Key performance metrics include *Hwk* (homework score), *OnlExams* (score from online exams), *CamExams* (score from on-campus exams with supervision), and *Final* (score from the final exam, also conducted on-campus with supervision). These scores are presented as percentages, reflecting the students' achievement in each respective assessment. Additionally, *Sem5* represents the scores for five problems initially available online, *Final5* for the same problems when included in the final exam, and *Diff5* indicates the score difference between these two settings. Thus, *Diff5* can be used as a proxy for the retention of

TABLE I. Summary of variables in the dataset.

Variable	Description	Minimum	Maximum	Mean	Std. Deviation
CamExams	On-campus exam score (%)	0.00	100.00	53.80	22.73
Diff5	Difference in online/final duplicate problem scores	−100.00	40.00	−39.00	33.80
DiffExams	Difference in online/on-campus exam scores	−81.55	23.57	−29.30	20.04
Final	Final exam score (%)	0.00	100.00	49.19	27.64
Final5	Final exam duplicate problem score (%)	0.00	100.00	51.22	33.37
Hwk	Homework score (%)	25.00	100.00	86.60	15.68
HwkAI	AI usage during homework (%)	0.00	100.00	17.08	23.94
HwkFac	Discussions with faculty during homework (%)	0.00	100.00	16.32	24.29
HwkInt	Internet usage during homework (%)	0.00	100.00	50.00	30.18
HwkPeer	Peer discussions during homework (%)	0.00	100.00	32.79	31.44
OnlAI	AI usage during online exams (%)	0.00	88.00	8.07	17.54
OnlExams	Online exam score (%)	43.69	98.06	83.10	11.77
OnlInt	Internet usage during online exams (%)	0.00	100.00	23.38	29.39
Sem5	Online problem score (%)	20.00	100.00	90.23	17.33

concepts between the semester and the final exam. Finally, *DiffExams* quantifies the score difference between online and on-campus quizzes, offering insight into performance variations across different assessment environments.

The dataset also encompasses variables related to the use of digital resources and student interactions. *HwkAI* and *OnlAI* denote the self-reported percentage of problems for which artificial intelligence (AI) tools were used during homework and online quizzes, respectively. Similarly, *HwkInt* and *OnlInt* represent the usage of other Internet resources in these settings. Finally, *HwkPeer* and *HwkFac* quantify the self-reported extent of peer discussions and interactions with faculty during homework.

C. Statistical methods

Data were downloaded from the course management system, and survey results were merged using Python scripts. Calculations for this project were carried out using ChatGPT-4 Advanced Data Analysis [20] and R [30] (in particular qgraph [31] and CTT [32]).

IV. RESULTS

A. Response rate

The first and second semester courses were completed by 156 and 183 students, respectively. Of these, 90 and 131 students agreed to participate in the study, bringing the total to 221 participants.

B. Online versus on-campus exams

Figure 4 shows the score distributions for the nine exams that were conducted online and the two exams that were conducted on campus under supervision. In a t test, these distributions are significantly different ($p \approx 4.7 \times 10^{-47}$). While the distribution of *CamExams* appears almost perfectly normal, the distribution of *OnlExams* exhibits a pronounced ceiling effect, as students cannot reach more than 100%.

The conditions under which these exams were conducted led to vastly different outcomes, and an immediate assumption would be that this is related to the use of external resources during unsupervised assessments. As an example, for nine of the ten questions on the last online exam, solutions could be found on Chegg within about 1 min each.

C. Usage of resources during unsupervised assessments

Figure 5 illustrates the average self-reported usage of AI and other Internet resources, as well as self-reported consultation with peers and course personnel.

Overall, students report less usage of resources during exams than during homework but not significantly. The only one-sigma significant differences are between usage of AI and talking to faculty on the one hand, and using other Internet resources while working on homework;

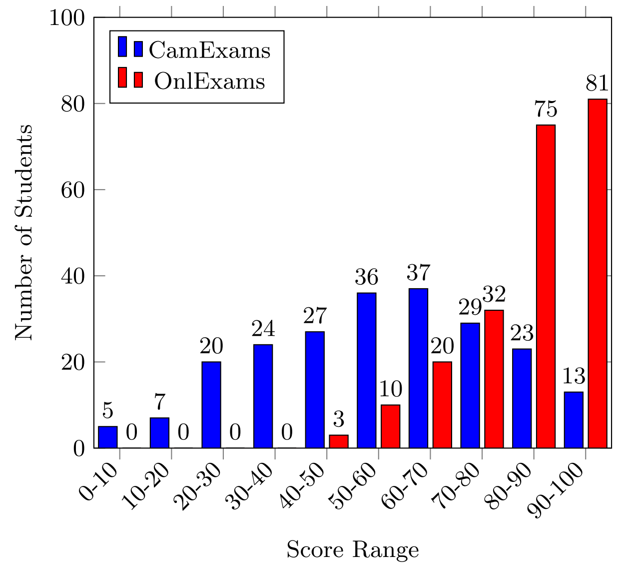

FIG. 4. Comparison between score distributions for the nine low-stakes exams that were conducted online (red) and the two low-stakes exams that were conducted on campus under supervision (blue).

students have not yet adopted AI and stick with "traditional" problem-solving sites. On average, students use other Internet resources for half of the homework problems. The distributions of these types of resource usages,

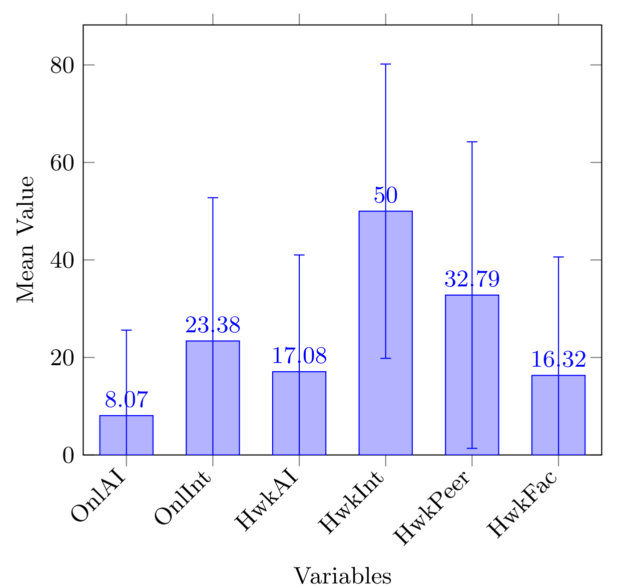

FIG. 5. Self-reported usage of resources while working on homework and online exams (in percent of problems). Most prevalent is the usage of the Internet during homework (*HwkInt*), followed by discussing homework with peers (*HwkPeer*). Students consulted faculty as frequently as AI while doing homework (*HwkFac* versus *HwkAI*). The Internet was also frequently used during online exams (*OnlInt*) while AI was used sporadically in this scenario (*OnlAI*).

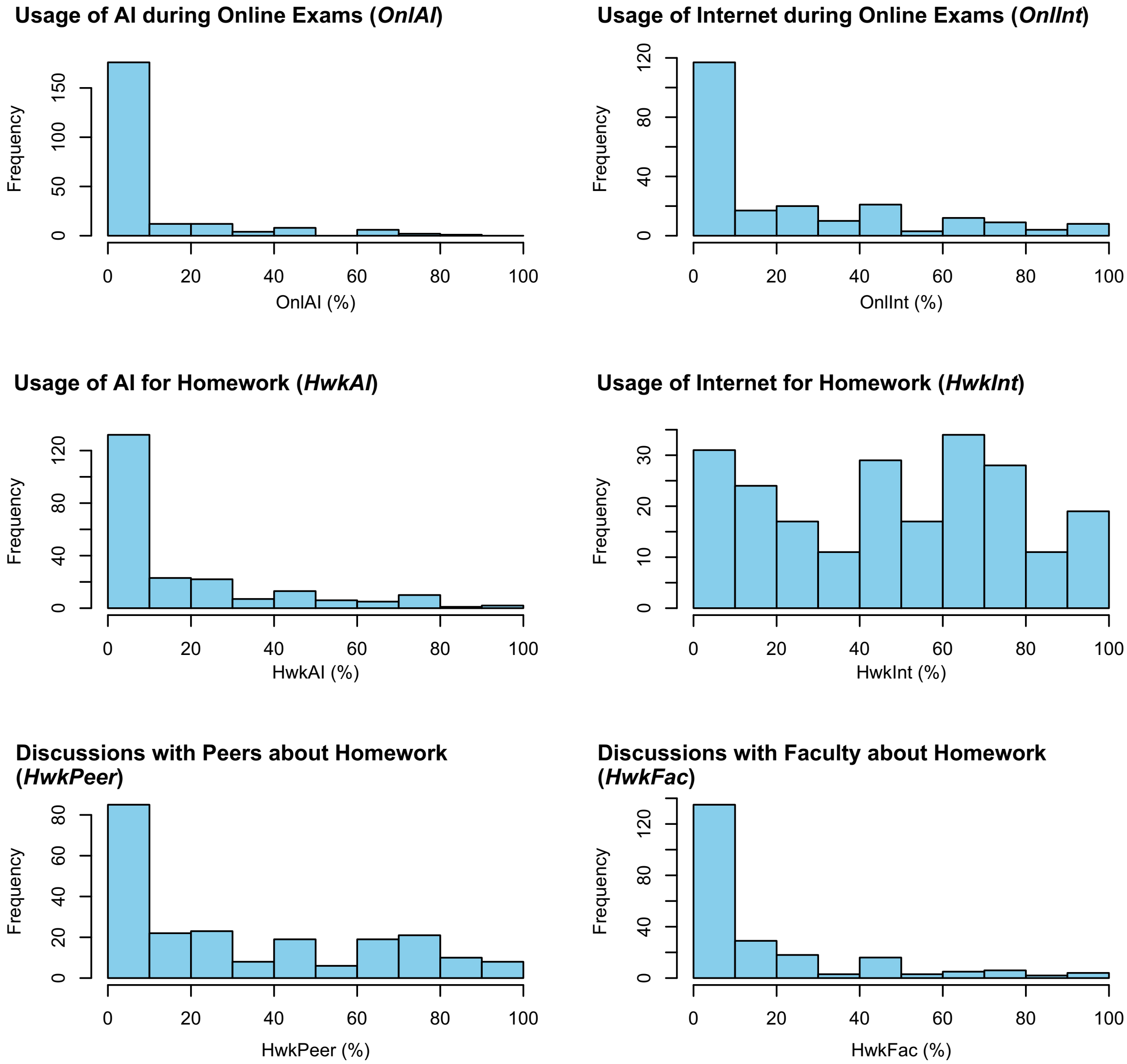

FIG. 6. Distribution of the survey responses behind the means reported in Fig. 5. With the exception of using the Internet for homework, across the types of resource usages, there is a strong faction of students who report completely refraining from them (156 students (71%) reported 0% usage for *OnlAI*, 95 students (43%) for *OnlInt*, 100 students (45%) for *HwkAI*, 20 students (9%) for *HwkInt*, 57 students (26%) for *HwkPeer*, and 105 students (48%) for *HwkFac*). By the reverse token, 29% of the students reported at least some usage of AI and 57% at least some usage of the Internet during online exams, which was prohibited.

however, are very different, see Fig. 6. For no variables are these response distributions normal; given the prevalence of "no usage" in several of the distributions, some of them could be considered binary: no usage versus any usage, which suggests that there are different clusters (subpopulations) of resource usage.

Of course, these data are self-reported, and underreporting of academic dishonesty by approximately one-third has been reported before [33]. As no explicit rules or regulations were violated by using external resources for solving homework, one can hope that their self-reported use reflects their actual use. A study conducted in an earlier, pre-AI semester of the same course estimated external resource usage ("copying") between 37% and 46% based on item response theory (IRT) models [34], which is lower than the usage self-reported here.

For the online exams, on the other hand, rules were violated by using the Internet or AI. Figure 4 raises doubts that all students trusted the research protocol, and thus the self-reported use might only be the minimum actual use (for the record, the protocol was strictly observed, and no student suffered any repercussions). Also, some students might genuinely underestimate their dependence on external resources or be in denial.

To get an estimate of this possible underreporting, the variable *CamExam* was taken as an indicator of the students' actual performance. Assuming the most efficient cheating scenario, where every incorrectly solved problem

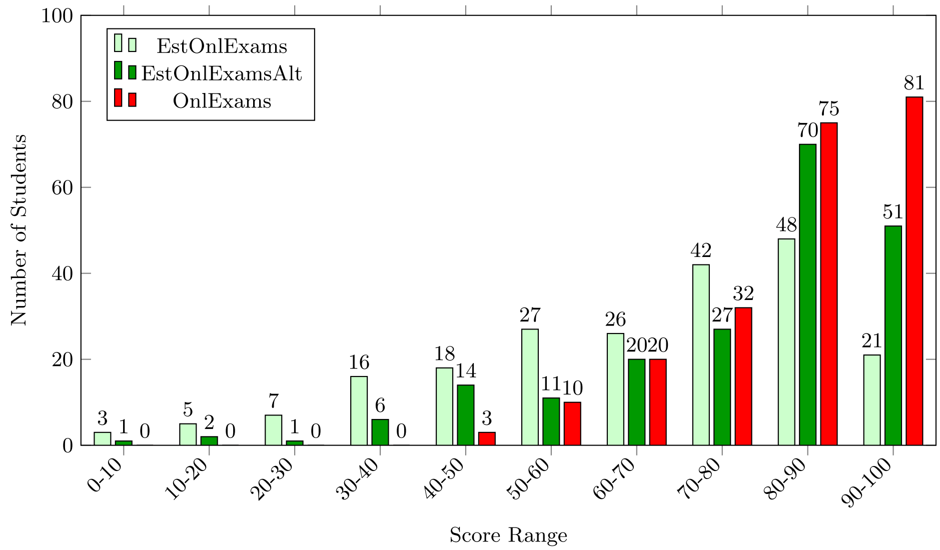

FIG. 7. Comparison between score distributions for the nine low-stakes exams that were conducted online (*OnlExams*, red) and an estimated score distribution based on the supervised on-campus exams (blue in Fig. 4) and the reported usage of external resources during online exams (Fig. 6) using Eq. (1) (*EstOnlExams*, light green), as well as the reported usage of external resources during homework (Fig. 6) using Eq. (2) (*EstOnlExamsAlt*, dark green).

would be solved after cheating, the maximum of the self-reported *OnlAI* and *OnlInt* was applied to simulate cheating at the reported percentage. This upper estimate is thus given by

$$EstOnlExams = CamExams + (100 - CamExams) \cdot \max(OnlAI, Onl\mathrm{Int})/100 \quad (1)$$

and results in the score distribution shown in light green in Fig. 7. It is evident that the self-reported usage of external resources does not explain the actual distribution. The discrepancy is especially blatant for the very high scores (>80%), hinting that the students who used external resources the most were also the ones who underreported it the most. An alternative model would be that students exhibit the same behavior during online exams as during online homework, which would lead to

$$EstOnlExamsAlt = CamExams + (100 - CamExams) \cdot \max(HwkAI, Hwk\mathrm{Int})/100 \quad (2)$$

and the distribution shown in dark green in Fig. 7. This suggests that actual usage of external resources during online exams may actually be closer to that during online homework but still does not match for the students with very high online exam scores. It is important, however, to keep in mind that Eqs. (1) and (2) are only unverified models, and we will proceed using the self-reported values.

D. Linear relationships

The simple hypothesis, "the more you cheat on formative assessments, the worse you do on the summative assessment" would to first order entail a linear relationship between *Final* and other attributes in the study, particularly those related to cheating. Disregarding the derived variable *DiffExams* and all variables related to the five duplicate problems, in a linear regression of all remaining assessment and survey variables, only *CamExam* emerges as a statistically significant predictor of the final exam score ($p \approx 7.3 \times 10^{-16}$); *OnlExams* comes close to statistical significance with $p = 0.05$. Table II shows the outcome of this linear regression, which has a weak correlation of $R^2 = 0.36$.

As seen in Figs. 4 and 6, *CamExam* is also the only variable that appears normally distributed. To ascertain that

TABLE II. Linear regression for *Final*.

	Estimate	Std. dev.	t	p
Hwk	−0.08	0.13	−0.6	0.55
OnlExams	0.33	0.17	1.96	0.05
CamExams	0.69	0.08	8.74	7.31×10^{-16}
HwkAI	−0.06	0.09	−0.68	0.50
HwkInt	−0.05	0.07	−0.70	0.49
HwkPeer	−0.04	0.05	−0.78	0.44
HwkFac	0.01	0.07	0.20	0.84
OnlAI	0.06	0.12	0.46	0.65
OnlInt	0.05	0.07	0.74	0.46

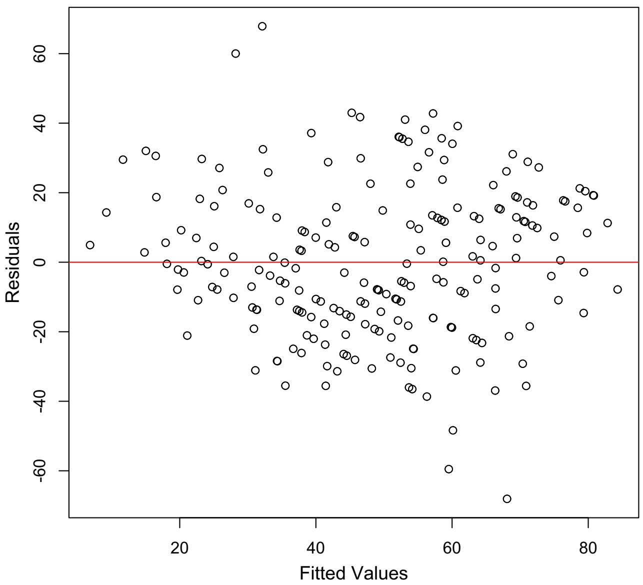

FIG. 8. Residual plot for the linear regression in Table II.

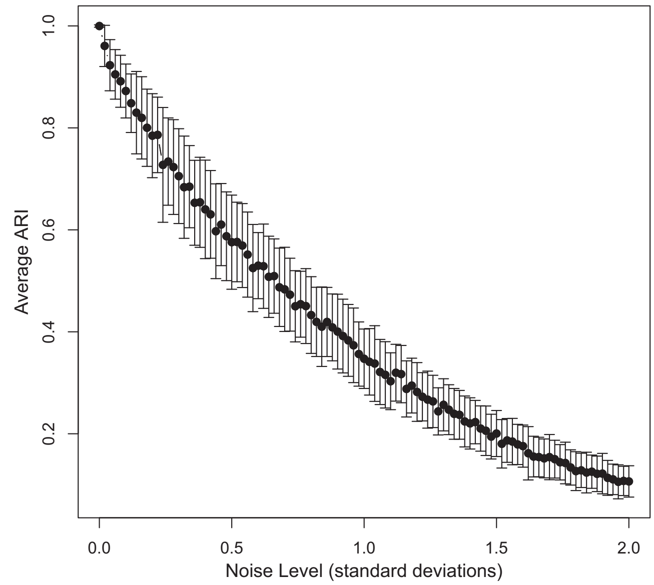

FIG. 9. Stability of the clusters in Table III. The plot shows the average adjusted Rand index of a comparison between the original cluster assignments to those resulting after applying up to two standard deviations in random noise to the data with 100 runs each.

the low predictive power of the other attributes is not simply an artifact of their non-normal distributions (including the evident ceiling and floor effects), a residual analysis was carried out. Figure 8 shows the result. The residuals appear to be centered on zero, which suggests that the model does not have a bias. Also, there does not appear to be a clear pattern or systematic structure to the residuals. This implies that the model captures the linear relationship well without missing a potential nonlinear relationship. The spread of the residuals is wide, corresponding to the low R^2-value, however, it seems relatively even across the range of fitted values. There is not an obvious sign of increasing or decreasing variance (heteroscedasticity) in the residuals as the fitted values increase.

What this means, though, is that the simplistic "higher justice" hypothesis of cheating during formative assessment leading to worse summative assessments (and thus worse grades) could not be confirmed across all course participants. Any possible relationships must be nonlinear in nature or depend on subpopulations.

E. Subpopulations of resource usage patterns

Using k-means clustering and elbow method on the scaled attributes, we identified four different clusters as indicated in Table III. Cluster 1, the smallest group, is comprised of students who appear to prefer human interaction to any online resources, and these students mostly adhere to rules for the online exams. Students in cluster 2, the largest cluster, state that they make little use of resources overall and that they most closely adhere to rules for the online exams. Students in cluster 3 make heavy use of Internet resources other than AI in both homework and exams, thus not following rules. Finally, students in cluster 4 use all available resources and disregard exam rules.

While the result might seem intuitively correct, the stability of these clusters was confirmed by applying random perturbations ("noise") to the data and calculating the adjusted Rand index (ARI) [35] for the identified cluster memberships in comparison to the original clustering. Figure 9 shows the result for 100 runs per noise level with perturbation strengths up to 50% of the standard deviations (0.5σ) applied to the survey attributes. While an initial plateau in the graph would have been desirable, the near-linear falloff of the ARI still indicates excellent agreement up to 0.1σ and still moderate to good agreement at 0.5σ (an ARI of 0 would indicate largely random

TABLE III. Number of members and mean values of variables in identified clusters (subpopulations) of resource usage.

Cluster	No. of members	*HwkAI*	*HwkInt*	*HwkPeer*	*HwkFac*	*OnlAI*	*OnlInt*	Interpretation
1	27	9.6	39.9	53.1	66.4	3.6	14.3	Using mostly peer discussions and help rooms
2	96	8.3	29.3	20.6	5.7	1.2	3.1	Using resources on homework but not exams
3	56	5.1	80.1	35.8	7.4	1.9	51.3	Heavily using Internet for homework and exams
4	42	58.0	63.7	43.6	20.3	34.8	38.2	Using all resources including AI everywhere

clustering while a negative ARI would indicate that the clustering method or parameters are not suitable for the attribute dataset).

F. Correlations of attributes

Figure 10 shows a Fruchterman-Reingold [31,36] representation of the correlation matrix between the variables. Indicated in light bluish-gray are the online, unsupervised assessments, in green the on-campus, supervised assessments, and in dark gray the differences in scores between selected subsets of assessments. The percentages of AI usage are indicated in beige, usage of Internet resources in yellow, and discussions with humans in orange. Green edges denote positive correlations, red edges negative correlations, and the thickness their absolute strength. Due to the force-directed nature of Fruchterman-Reingold graphs, closely correlated vertices tend to cluster while unrelated vertices tend to be further apart from each other. As a consequence, weakly correlated vertices tend to be further apart from each other.

It is apparent that the scores achieved online and those achieved under supervision each cluster together, but both are disconnected from the survey answers. We would have expected a very different picture. For example, we would have expected the vertices representing measures of cheating during homework (*HwkInt* and *HwkAI*) to be clustering with final exam performance, and we would have expected measures of cheating during midterm quizzes to be clustered with quiz performance. Instead, there is literally a disconnect between assessment performance and self-reported cheating, which appear to simply cluster together (despite resulting from distinct subpopulations, see Sec. IV E).

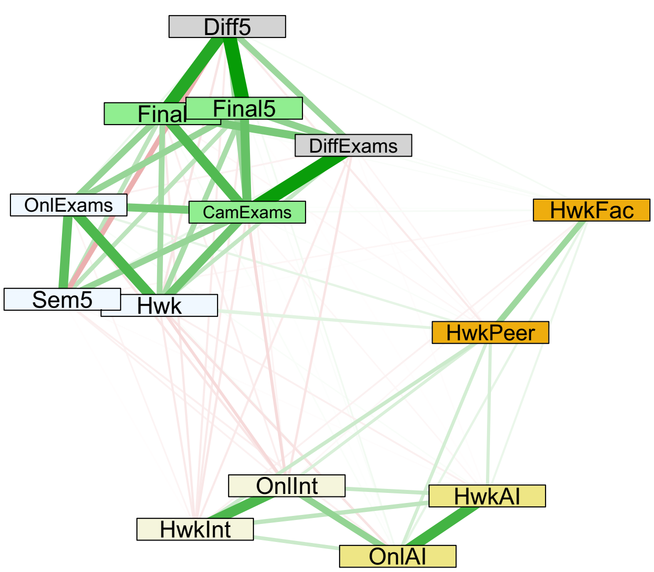

FIG. 10. Fruchterman-Reingold [31,36] representation of all correlations between the variables in Table I for all students. Scores in supervised settings are indicated in green (final exam score (*Final*), score on the five duplicate final exam questions (*Final5*), and low-stake exams on campus (*CamExams*)). Scores in unsupervised settings are indicated in light bluish-gray (low-stake exams online (*OnlExams*), homework (*Hwk*), and score of the five duplicate final questions when they first appeared during the semester (*Sem5*)). The differences $DiffExams = CamExams - OnlExams$ and $Diff5 = Final5 - Sem5$ are indicated in dark gray.

Also within the usage clusters (Table III), there are very few significant correlations between resource usage and assessment performance. Figure 11 shows the statistically significant correlations between the variables ($p < 0.05$). For the heavy Internet users (cluster 3), the usage of Internet resources other than AI during online exams (*OnlInt*) is significantly negatively correlated with the scores on the exams (*OnlExams*) ($r = -0.28$; $p = 0.04$), which may indicate that relying on the Internet, the students were not able to quickly enough find what they needed to correctly solve the problems, including replacement of the numbers by their values. For the users who made use of all resources everywhere (cluster 4), the use of AI during online exams (*OnlAI*) is significantly positively related to *Diff5* ($r = 0.31$; $p = 0.04$); this means that AI usage during online exams is positively correlated with doing better on the final exam instance of duplicate problems than on their first occurrence in the course.

Notably, within all usage clusters, significant correlations between the supervised final exam and the unsupervised assessment components of the course remain. This means that despite even the heaviest use of external resources, if not as a significant predictor, unsupervised assessment still retains formative relevance.

Using *Final* as a proxy for learning success, we found no significant performance differences between the clusters. We also investigated if higher or lower performing quartiles of the students may have benefitted or been harmed by the use of online resources, but we found no difference in that regard between these populations, either.

G. Item analysis

The use of external resources is detrimental to the validity of assessments, as problems (in this context, often referred to as "items") discriminate less between students who generally have a good grasp of a concept and those who have not; even students who have not understood the concepts get the problem correct. Table IV shows the average item parameters for assessment items on homework and unsupervised online exams. The mean reflects the average percentage of correctly solved items and the point biserial ("pBis") discrimination of these items. The point biserial ranges from -1 to 1, where negative values usually denote invalid assessment items; in reality, the range is smaller and it is further limited for nonuniform and nonnormal distributions [37] (pBis does not become invalid

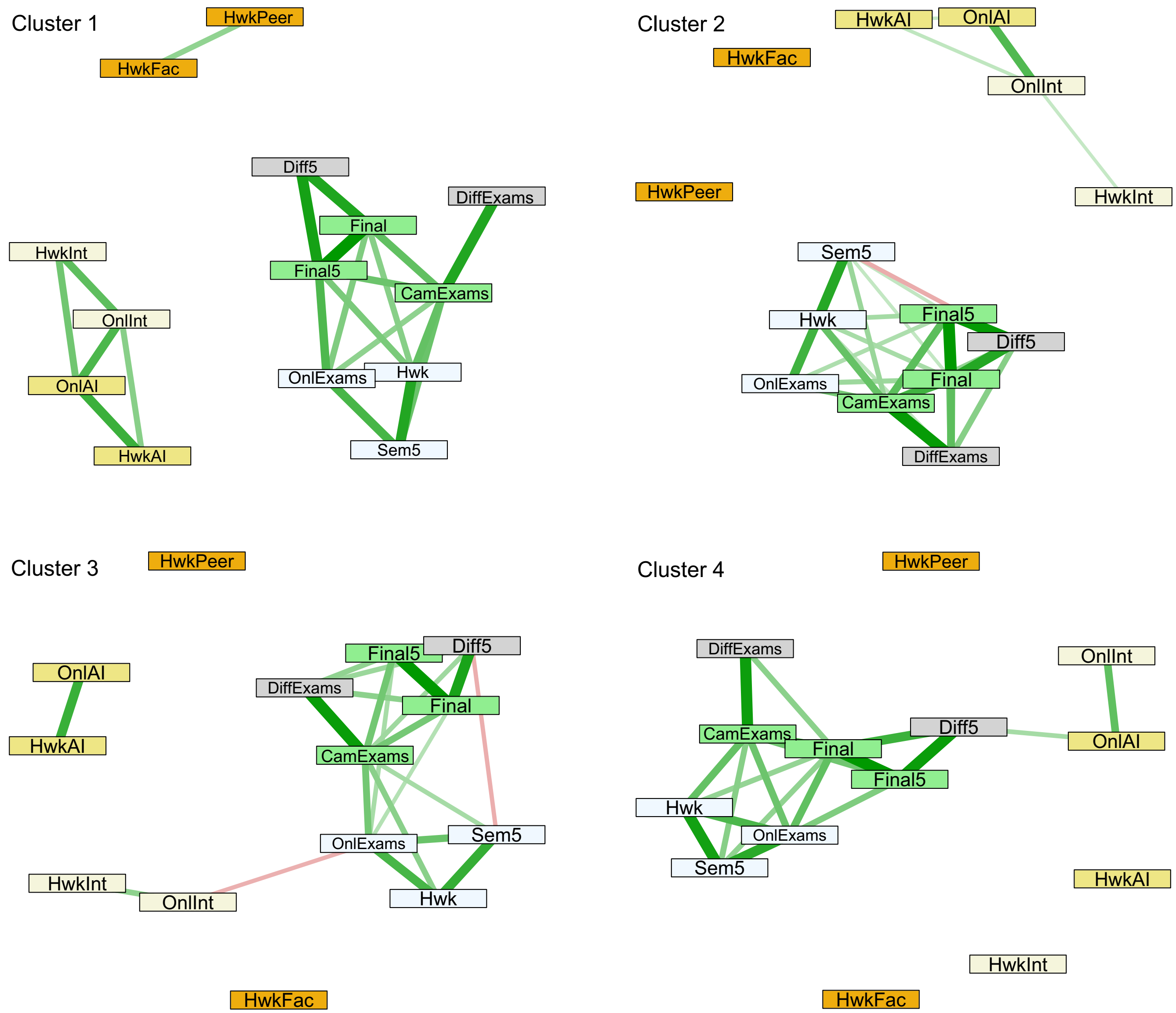

FIG. 11. Fruchterman-Reingold [31,36] representation of the statistically significant correlations ($p < 0.05$) between the variables in Table I for the clusters in Table III (note the rotation and handedness of these representations are random). For the students in cluster 1 who mostly limit help seeking to homework discussions with peers (*HwkPerr*) and faculty (*HwkFac*), all performance measures are significantly positively correlated. For the students in cluster 2 who used resources only for homework, this is also true (the negative correlation between the performance on the duplicate problems during the semester (*Sem5*) and the difference in scores between final exam and semester on those problems (*Diff5*) is merely autocorrelation). For the heavy Internet users in cluster 3, there is a significant negative correlation between usage and performance, while for the students in cluster 4 who used all resources everywhere, the correlations become more fragmented.

TABLE IV. Average problem-solving mean and average point biserial correlations for homework and online exams by clusters.

	Homework		Online exams	
Cluster	Mean	pBis	Mean	pBis
1	0.87	0.46	0.82	0.33
2	0.86	0.45	0.82	0.30
3	0.84	0.40	0.81	0.28
4	0.85	0.49	0.83	0.26

under non-normality; its limited range simply reflects the maximum “left-over” discrimination possible).

The values only insignificantly differ between different clusters of online resource usage patterns. Overall, they are consistent with the low predictive power of these unsupervised assessments, but they still indicate that items can provide feedback to students and instructors (a point biserial in the range of 0.3–0.5 would usually be considered “moderate”); this result agrees with findings in pre-AI times [38].

H. Student comments about AI

Based on the replies to the open-ended question on the survey, many students recognize AI as a valuable tool for assisting in learning, particularly for understanding complex topics and guiding problem solving. They particularly value its ability to quickly provide information without having to flip through textbook materials or scroll through videos. They appreciate AI's ability to provide alternate explanations and solutions, which can be especially helpful when traditional teaching methods fall short. However, there's a consensus that AI should not replace genuine learning and effort. Students suggest that AI's role should be that of an assistant rather than a solution provider, and its usage should be context dependent. For instance, in major-related courses, students advocate for minimal AI use to ensure a solid understanding of essential concepts. Conversely, in subjects that are less critical for them, they see AI as a more acceptable aid.

Concerns about academic integrity and the potential for AI to promote laziness and dependency are prominent. Students worry that reliance on AI for problem solving or essay writing could lead to a superficial understanding of course material and hinder the development of critical thinking skills. They propose a balanced approach, where AI is used judiciously to enhance learning without becoming a crutch. This balance involves using AI for initial guidance or concept clarification while avoiding its use for directly solving assignments or exams. Some students commented on having some of the exams during the semester being in person as beneficial. Ironically, but of course also very temptingly, based on style considerations, it appears that several students filled out the free-response question using ChatGPT.

Finally, the practicality of regulating AI use in online education is a significant concern. Some students acknowledge that while AI tools like ChatGPT may not be sophisticated enough currently to solve complex academic problems accurately, they could still be misused. Several students find ChatGPT is not yet trustworthy but expect this to improve in the future. There's an acknowledgment that AI is a part of the evolving educational landscape, and rather than outright banning it, educators should find ways to integrate it responsibly into the curriculum. This integration could involve designing assessments that still require a deep understanding of the material, even with AI assistance, and teaching students how to use AI ethically and effectively as part of their learning toolkit.

V. DISCUSSION

A. Seeking help from traditional cheat sites or AI

Students are making extensive use of external resources when working on unsupervised assessments. One year after becoming available, students have not made the jump from "traditional" problem-solving sites to AI. The reasons might be manifold: the free version of GPT (at the time of writing version 3.5) is much less powerful than version 4, which is only available to subscribers. A GPT subscription costs \$20/month, while for example, Chegg costs \$14.95/month with promotion sales for half that price. Students may also be used to the traditional resources from high school and carry over their habits to college. For questions that contain figures or graphs, on forums, it is sufficient to submit only the text in order to locate the question, while with AI tools, these illustrations need to be described in words [24,26] (this will change as the multimodal capabilities of these systems develop further [39]). Finally, with traditional sites, students would find the exact problem with the expected answer, while all large language models still hallucinate.

B. The surprising nonimpact of cheating

A surprising result of this study is the lack of correlations between self-reported resource usage and assessment outcomes. While the discrepancy between the score distributions of online and on-campus assessments (Fig. 4) could be explained by the usage of AI and other online resources (Fig. 5), one would have expected a correlation [33,40]: the more resources are used, the higher the discrepancy; this, however, is not the case. This null result may be due to students being worried about punitive measures in spite of the strict research protocol or students underestimating their reliance on external resources. We also would have expected a significant outcome analyzing the recurring questions (variables *Final5* and *Diff5*), since presumably, students who cheated would not remember the earlier solutions when encountering these problems on the final exam; also here, we found nothing but null results.

Finally, it is also surprising but encouraging that despite all of the external resource usage, a significant correlation remains between unsupervised and supervised assessments; while the best and only statistically significant predictor of the score on the final exam are the scores on the supervised in-semester exams, the other assessments have not lost their formative relevance.

C. Supervised versus nonsupervised exams

From the results, it is clear that high-stake exams like the final exam cannot be conducted in unsupervised settings. This finding is in line with earlier studies [41,42] (with some notable exceptions [43]). At the moment, the time it takes to look up solutions on the Internet may still be a hindrance to overly relying on those resources (as was also found by the negative correlation between the extent of using the Internet and performance on online exams among students who heavily rely on external resources), on the long run, AI will likely be reliable enough that answers to any introductory physics problem, including newly created ones, can be obtained instantaneously.

Simple usage of lockdown browsers such as Respondus [44] or Safe Exam Browser [45] is no remedy in an online setting, as students can simply use another machine or their phone to access sites such as ChatGPT [20], Gemini [46], or Chegg [16]. Instead, these lockdown browsers, which limit access to local disks and particular Internet sites, are useful in supervised on-campus settings where students use their own devices (on-campus BYOD exams). AI-detection tools are not only unreliable [47,48], they are also nonapplicable for physics exams, as those tools attempt to hone in on language features and narrative structure, which would not be copied by the students when working on solving physics exams.

During online quizzes, which are low-stakes but nevertheless have exam character, the usage of external resources was prohibited; unfortunately, the grade distributions in Fig. 4 make it blatantly obvious that these restrictions were not observed, and as a matter of fairness to all honest students, it cannot naïvely be assumed that the outcome would be much different for high-stakes online exams. The situation is aggravated by AI, which can solve unpublished problems. Thus, if instructors prohibit use of external resources in online assessments, there is no alternative to intrusive proctoring systems that use cameras and microphones.

D. Student awareness of pitfalls

Students are well aware of the possible pitfalls associated with AI usage. While they argue that it will be part of their professional lives, they support a balanced approach to its use, in particular overdependence and overreliance. They are also aware of reliability and trustworthiness issues, which agrees with earlier findings regarding students' ability to judge the quality of answers [28,49]. Overall, though, some statements about using these tools for learning purposes may have to be taken with the same grain of salt as the statement about the expert solution in Fig. 1 being "designed to help students […] learn core concepts;" students may believe these statements, but still not act accordingly [50]. Remarks about courses that are "not critical for them" suggest that nonsanctioned external resource usage may be particularly strong when the goal is simply to pass the course [51].

VI. LIMITATIONS

Students self-selected into this study, and students who are self-aware that they are using external resources in nonconstructive ways may have refrained from participating. As all help seeking involves third parties, we were unable to obtain data beyond self-reported usage; among the participants, we suspect that resource usage was underreported. As the survey and consent form were administered at the end of the semester, students who dropped out of the course earlier were not considered.

VII. OUTLOOK

The null results in this study leave many open research questions. While performance on formative assessments remains correlated with summative assessment, how can it be that undermining the formative assessment through cheating does not seem to have a significant impact? How exactly are students interacting with external resources? Which hidden variables exist, and which unobserved actions happen in preparation for the summative exam? This study, which joins a long list of "no-significant-difference" studies regarding instructional methods [6,9,10,52], might call for a more holistic view than individual surveys, log analyses, and performance scores provided. A future study might have to observe students while doing homework and preparing for exams and literally look over their shoulders while they make use of the resources at their disposal; the ethical, logistical, and research-design aspects of such an intrusive study are of course tremendous.

VIII. CONCLUSION

Our findings reveal that despite the emergence of LLMs, students predominantly rely on traditional Internet resources for unsupervised quizzes. However, this reliance does not significantly affect supervised assessments.

Our data indicate that the unsupervised online exams and supervised on-campus exams yield markedly different outcomes, presumably due to the use of external resources in unsupervised settings. Interestingly, there is no strong correlation between self-reported resource usage and exam performance, suggesting other factors at play or potential underreporting of resource usage. Despite heavy resource use, significant correlations between supervised and unsupervised assessments persist, indicating that unsupervised assessments, while having no significant predictive properties, retain formative value.

While students advocate for a balanced approach to AI use, emphasizing its role as an assistant rather than a solution provider, the study underscores the necessity of carrying out high-stakes exams in supervised settings to ensure academic integrity if external resources are not allowed.

Only one quarter of our survey respondents claims to have utilized at least some minimal form of AI tools in the solution of their homework problems. AI systems, though, are becoming exponentially more competent. They will rapidly penetrate all education settings. Our present results do not show significant causal effects on student learning success from using AI tools. In this sense, our study was conducted perhaps a bit early. But it is clear, nevertheless, that all of us need to rethink our course offerings, and particularly our assessment tools, due to the rise of AI tools. Now is the time to help shape AI environments into the perfect one-on-one tutor for students, similar to what Feynman envisioned, instead of a means to avoid learning physics.

ACKNOWLEDGMENTS

We would like to thank the students who participated in this study. We would also like to thank Christine Kortemeyer for her help with the manuscript. We would also like to thank the reviewers of this journal for their valuable insights and suggestions.

[1] E. Kim and S.-J. Pak, Students do not overcome conceptual difficulties after solving 1000 traditional problems, Am. J. Phys. **70**, 759 (2002).

[2] R. Feynman, R. Leighton, and M. Sands, *The Feynman Lectures on Physics* (Addison Wesley, Reading, MA, 1965).

[3] G. Kortemeyer and W. Bauer, Multimedia collaborative content creation (mc3): The MSU lecture online system, J. Eng. Educ. **88**, 421 (1999).

[4] W. Bauer and G. D. Westfall, *University Physics with Modern Physics (ebook)* (McGraw-Hill Higher Education, New York, NY, 2023).

[5] G. Kortemeyer, Over two decades of blended and online physics courses at Michigan State University, eleed **10** (2014), https://www.eleed.de/archive/10/4011.

[6] T. L. Russell, *The "No Significant Difference" Phenomenon as Reported in 248 Research Reports, Summaries, and Papers* (North Carolina State University, 1997).

[7] J. Cavanaugh and S. J. Jacquemin, A large sample comparison of grade based student learning outcomes in online vs. face-to-face courses, Online Learn. **19**, 1 (2015).

[8] E. Bergeler and M. F. Read, Comparing learning outcomes and satisfaction of an online algebra-based physics course with a face-to-face course, J. Sci. Educ. Technol. **30**, 97 (2021).

[9] G. Kortemeyer, W. Bauer, and W. Fisher, Hybrid teaching: A tale of two populations, Phys. Rev. Phys. Educ. Res. **18**, 020130 (2022).

[10] G. Kortemeyer, C. Kortemeyer, and W. Bauer, Taking introductory physics in studio, lecture, or online format: What difference does it make in subsequent courses, and for whom?, Phys. Rev. Phys. Educ. Res. **19**, 020148 (2023).

[11] H. Ibrahim, F. Liu, R. Asim, B. Battu, S. Benabderrahmane, B. Alhafni, W. Adnan, T. Alhanai, B. AlShebli, R. Baghdadi *et al.*, Perception, performance, and detectability of conversational artificial intelligence across 32 university courses, Sci. Rep. **13**, 12187 (2023).

[12] F. R. A. Ahmed, T. E. Ahmed, R. A. Saeed, H. Alhumyani, S. Abdel-Khalek, and H. Abu-Zinadah, Analysis and challenges of robust e-exams performance under COVID-19, Res. Phys. **23**, 103987 (2021).

[13] M. Farrokhnia, S. K. Banihashem, O. Noroozi, and A. Wals, A swot analysis of ChatGPT: Implications for educational practice and research, Innov. Educ. Teach. Int. **61**, 460 (2023).

[14] J. A. Oravec, Artificial intelligence implications for academic cheating: Expanding the dimensions of responsible human-ai collaboration with ChatGPT, J. Interact. Learn. Res. **34**, 213 (2023), https://www.learntechlib.org/primary/j/JILR/v/34/n/2/.

[15] C. Ruggieri, Students' use and perception of textbooks and online resources in introductory physics, Phys. Rev. Phys. Educ. Res. **16**, 020123 (2020).

[16] Chegg, Chegg, https://www.chegg.com [accessed December 2023].

[17] 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).

[18] 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).

[19] H. Busch, One method for inhibiting the copying of online homework, Phys. Teach. **55**, 422 (2017).

[20] OpenAI, ChatGPT, https://chat.openai.com/ [accessed December 2023].

[21] Google, Google Bard, https://bard.google.com/ [accessed December 2023].

[22] P. Prakash, Chegg's shares tumbled nearly 50% after the edtech company said its customers are using ChatGPT instead of paying for its study tools, Fortune (2023), https://fortune.com/2023/05/02/chegg-shares-tumble-students-fleeing-chatgpt-a-i/.

[23] D. Paresh, Chegg embraced ai. ChatGPT ate its lunch anyway, Wired (2023), https://www.wired.com/story/chegg-embraced-ai-chatgpt-ate-its-lunch-anyway/.

[24] G. Kortemeyer, Could an artificial-intelligence agent pass an introductory physics course?, Phys. Rev. Phys. Educ. Res. **19**, 010132 (2023).

[25] W. Yeadon and T. Hardy, The impact of AI in physics education: A comprehensive review from GCSE to university levels, Phys. Educ. **59**, 025010 (2024).

[26] G. Polverini and B. Gregorcic, Performance of ChatGPT on the test of understanding graphs in kinematics, Phys. Rev. Phys. Educ. Res. **20**, 010109 (2024).

[27] T. Kumar and M. A. Kats, ChatGPT-4 with code interpreter can be used to solve introductory college-level vector calculus and electromagnetism problems, Am. J. Phys. **91**, 955 (2023).

[28] M. N. Dahlkemper, S. Z. Lahme, and P. Klein, How do physics students evaluate artificial intelligence responses on comprehension questions? A study on the perceived scientific accuracy and linguistic quality of ChatGPT, Phys. Rev. Phys. Educ. Res. **19**, 010142 (2023).

[29] 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).

[30] R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria 2021).

[31] S. Epskamp, A. O. J. Cramer, L. J. Waldorp, V. D. Schmittmann, and D. Borsboom, qgraph: Network visualizations of relationships in psychometric data, J. Stat. Softw. **48**, 1 (2012).

[32] J. T. Willse, CTT: An R package for classical test theory functions (2014), https://CRAN.R-project.org/package=CTT.

[33] 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).

[34] E. Gönülateş and G. Kortemeyer, Modeling unproductive behavior in online homework in terms of latent student traits: An approach based on item response theory, J. Sci. Educ. Technol. **26**, 139 (2017).

[35] W. M. Rand, Objective criteria for the evaluation of clustering methods, J. Am. Stat. Assoc. **66**, 846 (1971).

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

[37] Y. Cheng and H. Liu, A short note on the maximal point-biserial correlation under non-normality, Br. J. Math. Stat. Psychol. **69**, 344 (2016).

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

[39] K. Wang, J. Pan, W. Shi, Z. Lu, M. Zhan, and H. Li, Measuring multimodal mathematical reasoning with math-vision dataset, arXiv:2402.14804.

[40] D. A. Kashy, G. Albertelli, W. Bauer, E. Kashy, and M. Thoennessen, Influence of nonmoderated and moderated discussion sites on student success, J. Asynchronous Learn. Netw. **7**, 31 (2003).

[41] M. Castro, S. Manoharan, U. Speidel, X. Ye, and J. Zu, Observations of cheating behaviours in online examinations and tools for mitigation, in *2023 IEEE Frontiers in Education Conference (FIE)* (IEEE, New York, 2023), pp. 1–7.

[42] M. Alguacil, N. Herranz-Zarzoso, J. C. Pernías, and G. Sabater-Grande, Academic dishonesty and monitoring in online exams: A randomized field experiment, J. Comput Higher Educ. (2023), 10.1007/s12528-023-09378-x.

[43] B. Sorenson and K. Hanson, Statistical comparison between in-person and online general chemistry exam outcomes: A COVID-induced case study, J. Chem. Educ. **100**, 3454 (2023).

[44] Respondus, Respondus Browser, https://web.respondus.com/he/lockdownbrowser/ [accessed December 2023].

[45] SEB Alliance, ETH Zurich, Save Exam Browser, https://www.safeexambrowser.org/ [accessed December 2023].

[46] Google, Gemini, https://deepmind.google/technologies/gemini/ [accessed December 2023].

[47] D. Dalalah and O. M. Dalalah, The false positives and false negatives of generative AI detection tools in education and academic research: The case of ChatGPT, Int. J. Manage. Educ. **21**, 100822 (2023).

[48] A. M. Elkhatat, K. Elsaid, and S. Almeer, Evaluating the efficacy of ai content detection tools in differentiating between human and AI-generated text, Int. J. Educ. Integrity **19**, 17 (2023).

[49] A. Shoufan, Exploring students'perceptions of CHATGPT: Thematic analysis and follow-up survey, IEEE Access (2023).

[50] 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).

[51] H. Lin, Learning physics vs. passing courses, Phys. Teach. **20**, 151 (1982).

[52] G. Kortemeyer, N. Dittmann-Domenichini, C. Schlienger, E. Spilling, A. Yaroshchuk, and G. Dissertori, Attending lectures in person, hybrid or online—how do students choose, and what about the outcome?, Int. J. Educ. Technol. Higher Educ. **20**, 19 (2023).