# Does presentation matter? Examining show-work presentation quality and student performance in introductory physics

Heather Mei,[1,*] Qiaoyi Liu,[1] Kentaro Fujita,[2] and Andrew F. Heckler[1]
[1]*Department of Physics, The Ohio State University, Columbus, Ohio 43210, USA*
[2]*Department of Psychology, The Ohio State University, Columbus, Ohio 43210, USA*

(Received 13 August 2025; accepted 5 October 2026; published 5 November 2026)

Physics problem-solving assessments often prioritize accuracy over presentation. However, the presentation quality of a student's written solution—such as its handwriting clarity and layout—may play an important role both as an indicator and facilitator of student problem-solving and achievement. In this study, we investigated the relationships between show-work presentation (SWP), conscientiousness, and student performance in a large-enrollment introductory calculus-based physics course. Using an emergent coding process, we first iteratively developed a reliable rubric for quantifying several presentation qualities of student solutions to two introductory mechanics final exam show-work questions. We found that the coding produced scores for 2 orthogonal factors, which we label as Form, measuring layout and legibility, and Annotation, measuring student notation of their work. Statistical analyses showed that Form was a moderate-to-weak positive predictor of various course performance measures ($r \sim 0.3$), including show-work score, exam score, nonexam scores, and course grade. Even when controlling for ACT Math and conscientiousness scores, Form still accounted for 4%–5% of the variance of these course performance scores. In contrast, Annotation was at best very weakly and inconsistently (in sign) correlated with these performance scores. Sex differences were also observed, with female students' mean scores for both Form and Annotation typically about one-half standard deviation higher than male students' scores. Simultaneous mediation models of the relation between sex and course outcomes included the mediators of ACT score, conscientiousness, and Form. Results indicate that Form partially mediated the relation between sex and course grade, show-work score, multiple-choice score, and nonexam score. Specifically, women tended to score higher on Form, and this effect in turn led to a positive effect on scores on all of these outcomes, typically compensating for other negative effects. The (simultaneous) mediations of ACT score and conscientiousness were much weaker, consistent with being vanishingly small. It is important to note that these are correlational results only, and control-treatment designs are called for to investigate potential causal links between Form and performance. Overall, by developing and implementing a validated rubric measuring show-work presentation quality, we find evidence that presentation quality may play a relatively small but reliable role in student performance on show-work problems and is a weak but reliable predictor of most components of course performance. These results suggest that show-work presentation may be an important factor to consider in physics instruction.

DOI: 10.1103/wp62-bwjs

## I. INTRODUCTION

Numerous studies in physics education over the last few decades have investigated students' problem-solving skills, particularly for open-response questions that require students to show their work (e.g., Refs. [1–10]). Such studies typically analyze the correctness of the solution, but they do not typically analyze the show-work presentation (SWP) quality as a distinct construct, even though prior research in other educational topics suggests that the presentation features of written solutions, such as handwriting [11,12], are associated with scores. In this study, we consider that presentation quality may play an important role in problem-solving, for example, because it may both facilitate and indicate how students reason about their problem-solving, and because it may influence how graders interpret this reasoning during scoring.

Although some education studies in physics have examined specific aspects of written solution presentation, such as the number of words or mathematical symbols [13], or

*Contact author: mei.228@osu.edu

*Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.*

the logical progression of a solution [3], we have found no studies that define and measure show-work presentation quality as a distinct construct separable from solution correctness. This presents an empirical and educational gap, since it remains unclear whether presentation quality in handwritten physics solutions can be reliably and meaningfully measured, whether it is associated with physics show-work problem performance in ways that are distinct from adjacent factors such as conscientiousness or prior preparation and, if that relationship exists, what mechanisms might account for that association. Here, we characterize SWP quality as specific, observable presentation features of students' written solutions and investigate whether it is related to conscientiousness, student performance, and previously documented sex-based differences in course outcomes.

In addition to the reasons stated above, there are several other theoretical and instructional perspectives that motivate this study of SWP. Let us begin with the theoretical perspectives. First, SWP may provide critical insights into important mechanisms behind variation in student show-work scores. Research in education and psychology suggests that clarity, organization, and structure of written work can significantly impact how students process information [11]. For example, in teaching writing skills, Graham *et al.* have shown that students who are taught specific writing strategies such as planning, drafting, and revising produce work that reflects a deeper understanding of the subject [14]. In another study on writing instruction, Connelly *et al.* showed that explicit training in writing presentation can improve clarity and content understanding, emphasizing the role of aesthetic presentation in academic success [12]. While student solutions to physics problems may not necessarily involve extensive essay writing, it is reasonable to consider the possibility that the same principles of clarity and structured organization may be helpful for processing and conveying reasoning in physics show-work problems. For example, Mayer has suggested that specific visual elements may also serve as cognitive cues that help students recall relevant concepts and problem-solving strategies under time pressure [15]. Conversely, Sweller suggested that an unstructured approach to solution writing for physics problems may add unnecessary cognitive load, drawing students' attention away from conceptual reasoning and potentially undermining student performance [16]. Further, Zimmerman suggested that noncognitive factors such as conscientiousness or self-regulation have been linked to more methodical solution writing [17].

Second, SWP may be an operational way of measuring the student-level psychological factor conscientiousness, though it is not yet clear whether it functions as an observable manifestation of conscientiousness, a distinct construct, or a potential pathway linking conscientiousness to student performance. It is well established that the noncognitive psychological factor of conscientiousness, as defined by the Big Five personality model [18], is positively correlated with academic achievement (e.g., Refs. [19–21]), even when controlling for prior academic achievement or measures of cognitive ability. This is important for our study because it is reasonable to consider that conscientiousness, which is often characterized by a "spectrum of constructs" such as "the propensity to be self-controlled, responsible to others, hardworking, orderly, and rule abiding" [22], may be related to SWP, which, we propose, is characterized by features such as visual clarity and spatial organization of the written solution. In this context, "responsible to others" may include keeping the needs of graders or other readers in mind, while "orderly" may relate to clear structuring or organization of the written solution [22]. For this reason, we conceptualize SWP as a possible behavioral manifestation of Big Five conscientiousness. We focus specifically on conscientiousness, rather than the other Big Five personality traits, because it is consistently associated with academic achievement and is the most conceptually relevant to observable features such as orderliness, carefulness, and structured presentation [18–22]. Therefore, one goal of this study is to determine the extent to which conscientiousness is correlated with a systematic measure of SWP because the result may indicate (or refute) a mechanism explaining how conscientiousness influences academic outcomes: does high conscientiousness lead to high presentation quality, and does this in turn lead to higher grades? Further, while there may be surface-level conceptual similarities between conscientiousness and presentation quality, we would also like to determine the extent to which SWP is a student-level measure distinct from conscientiousness.

Third, given the well-known and documented differences in conscientiousness [23] and handwriting quality [24,25] between male and female students, as well as documented differences in exam scores between male and female students [26,27], we are compelled to include considerations of differences in SWP between female and male students in our research study. Specifically, any observed male-female differences in SWP may provide critical insight into the nature of SWP and potential mechanisms explaining differences in scores both within and between male and female student populations. For example, given known differences in handwriting between female and male students [28,29], if no SWP differences were found in our study, this could provide important constraints on the nature of SWP. Further, to avoid potential bias in our results that would otherwise measure effects for only the majority population enrolled in calculus-based introductory physics (i.e., males), our experimental design ensured that equal numbers of male and female students were included in the study sample.

Fourth, moving to instructional perspectives, if it turns out that SWP is a causal factor in the physics solution

process, then improving SWP may improve performance on physics problems. Of course, both causality and malleability of SWP would have to be established in future studies, but first, in this study, we are interested in determining the extent to which there is any association between SWP and scores on physics show-work problems.

Finally, it could be that the process and result of grading may depend on the quality of the presentation. In other words, thinking beyond student-centered mechanisms, the presentation quality of the solution may matter to the grader. For example, assuming that two solutions have the same accuracy, the grader may award a lower score to the solution with lower presentation quality. Prior research outside of the physics context has shown that visual features such as handwriting legibility can significantly influence grader evaluations, even when the content remains unchanged [11,12,30,31].

Therefore, in this study, we will not focus on the content of student solutions to physics questions but rather on the presentation qualities of the content and the relation of those qualities with other student performance measures. The goal of this study is to examine SWP and its relationship with conscientiousness and student performance. Specifically, we first iteratively develop a reliable and productive coding scheme to characterize "SWP", namely, the presentation quality of show-work answers. We then code the SWP for two exam items and explore associations between SWP, conscientiousness, and student performance outcomes. Finally, we also examine differences in SWP between males and females and determine whether any differences in SWP mediate differences in exam grades between males and females in an introductory calculus-based physics class. We begin with further discussion of relevant literature and a specification of our research questions.

### A. Solution structure, handwriting, and graded performance

Scientific problem-solving has long been considered a crucial component of science, technology, engineering, and mathematics (STEM) education, but the evaluation of student responses often neglects the cognitive load associated with solution structuring. Stewart and Ballard found that features related to written presentation, such as the quantity of math, language, and graphical elements, accounted for a significant proportion of variance in both exam scores and conceptual understanding in an introductory undergraduate physics course [13]. Their findings suggest that even when content knowledge is present, the way in which a solution is communicated can influence performance outcomes. Cognitive load theory suggests that students who struggle with organizing their work may experience higher extraneous cognitive load, which diverts attention away from the actual problem-solving itself [16]. A study by Paas and van Merrienboer found that in mathematics and physics, students who clearly structure their solutions tend to perform better, not only because of better conceptual understanding but also because structured problem-solving reduces demands on working memory [32]. According to the visible learning theory, as researched by Hattie, making students' reasoning visible through organized solutions can significantly enhance both learning outcomes and grader accuracy by allowing graders to see the thought process behind student solutions [33]. A lack of formal training on problem presentation may contribute to inconsistencies in performance and assessment outcomes [34].

Research on handwriting legibility has also suggested that the legibility and organization of written responses can influence grading outcomes across different disciplines. When studying undergraduate handwriting essays for exams, slow or inefficient handwriting was found to constrain overall exam performance, limiting a student's ability to fully demonstrate their knowledge within high-pressure assessment environments [12]. The issue of grader bias may further complicate the evaluation of student work. Grading biases linked to handwriting legibility have been previously documented, with evidence supporting that neater, more structured responses are perceived as being higher quality even if the content remains the same [11]. These biases may disadvantage students who, for various reasons, struggle with handwriting or the physical organization of their solutions. These effects are particularly relevant in physics exam show-work questions, where students are often required to show detailed problem-solving steps in a time-pressured environment.

Moreover, research by Santangelo and Graham in K–12 education highlights the broader implications of handwriting instruction, suggesting that explicit teaching in handwriting enhanced legibility and fluency, which allowed students to communicate their reasoning more effectively [35]. Educators have expressed concerns about fairly evaluating student work with poor handwriting legibility, which can lead to misinterpretations of student ability, as found by Hammerschmidt and Sudsawad [36].

### B. Noncognitive factors and academic achievement

Beyond handwriting, show-work presentation may reflect underlying cognitive or behavioral traits, such as conscientiousness. The Big Five personality model, as discussed by John and Srivastava, defines conscientiousness as associated with diligence, organization, and responsibility [18]. A meta-analysis conducted on the relationship between personality and academic performance confirms that conscientiousness, while largely independent of intelligence, is a consistent predictor of academic achievement across diverse educational settings [20].

Multiple studies have supported this relationship in different educational contexts. Henderson *et al.* found that personality traits, including conscientiousness, mediated

the relationship between gender and self-efficacy in physics and mathematics [23]. Ivcevic and Brackett showed conscientiousness was a predictor of academic success in high school settings [37], while Dumfart and Neubauer found that conscientiousness has been identified as a key noncognitive predictor of academic achievement in middle school science through traits such as self-discipline and persistence [38]. Duckworth and Seligman found that self-discipline predicts academic performance [39], which further supports the role of conscientiousness-related traits in students' academic performance. Further evidence linking conscientiousness to academic performance is work by Zusho *et al.* [40], who found that students who score higher on conscientiousness may be more likely to engage in deliberate planning, thorough checking of work, and careful use of space while writing—practices that may collectively enhance clarity in written show-work solutions. Prior studies have also shown that self-regulated learners tend to engage in proactive planning and monitoring of their academic work, which may indirectly support more organized and methodical approaches to writing solutions during exams [17].

Rice *et al.* have found that perfectionism among STEM students can influence academic performance, with adaptive perfectionism promoting clarity and high standards while maladaptive perfectionism can increase anxiety [41]. Perfectionist tendencies may have mixed effects, either improving SWP through meticulous attention to detail while writing solutions with extra annotations to justify each step or possibly hindering SWP through lack of time to complete a solution during an exam. This tension between striving for completeness and managing time constraints highlights the importance of balancing efficient presentation with sufficient clarity to demonstrate understanding to the grader.

### C. Sex disparities in performance outcomes in STEM courses

As mentioned in the introduction, differences in performance between male and female students in physics education have been well documented, and there is evidence suggesting that female students display higher levels of conscientiousness in introductory mathematics and physics classes [23] but often underperform on exams in introductory STEM courses when compared to their male peers, potentially due to stereotype threat and testing anxiety [27,42]. A study by Balart and Oosterveen has shown that females demonstrate more sustained performance throughout test-taking than males, suggesting that female students may maintain higher levels of attention and effort across the duration of the exam, which could potentially contribute to more consistent and higher-quality show-work presentation, especially since show-work problems are often placed after multiple choice on exams [43]. These exam endurance-related differences may influence how students allocate cognitive resources during exams and potentially affect the clarity and organization of their written solutions. For instance, female students' sustained effort may allow them to devote more attention to presentation quality, which could in turn impact assessment outcomes. Additionally, a study by Ballen *et al.* suggests that assessment formats (i.e., low- versus high-stakes exams) may impact performance gaps between males and females [42]. There are relatively few studies examining the male-female performance differences based on item format, but a couple of studies point to females scoring ∼0.2–0.3 SD lower than males on multiple-choice (MC) items but about the same on show-work items [44,45]. This performance discrepancy raises the question of whether show-work presentation may play a role in differentially affecting female and male students' performance across different exam item types. Another study by Simmons and Heckler has shown significant sex disparities in introductory physics course grades, highlighting how different assessment components such as exams, quizzes, and homework can impact different demographic groups, including females and underrepresented minorities [26]. Examining specific graded components through the lens of SWP may offer insights into how assessment can be made fairer for both male and female students in physics courses.

Furthermore, systemic issues such as implicit bias and cultural stereotypes continue to create barriers for females in STEM [46,47]. Research by Greifeneder *et al.* suggests that handwriting legibility can bias graders' assessments through implicit gender-based inferences [31]. This study demonstrated that legible handwriting not only positively influences graders' evaluations but also activates gender stereotypes, further complicating grading fairness in STEM disciplines [31]. A study by Burr suggested that handwriting features such as neatness and roundness were stereotypically associated with femininity, and participants were able to identify a writer's gender at rates significantly higher than chance [48], which may prompt some graders to either be more lenient in an effort to support females in STEM or more critical based on the stereotype that females are less capable in STEM fields. Malouff and Thorsteinsson found graders may unintentionally place lower expectations for female students or interpret their solutions more critically due to gender stereotypes [30]. Compared to binary multiple-choice scoring, human graders can introduce greater potential for sex-based biases in scoring show-work exam problems.

By studying how SWP and sex impact different question types, we aim to understand whether show-work presentation alleviates or exacerbates existing differences in scores. A study by Hofer demonstrated that physics teachers may exhibit consistent biases against female students when grading identical physics solutions, though these biases tended to decrease with more teaching

experience [49]. While multiple-choice questions offer a structured format with no possible partial credit, show-work questions can reward clarity and incremental steps in problem-solving. These opportunities for students to communicate partial knowledge may reduce testing anxiety by decreasing the consequences of committing to one all-or-nothing answer choice. Show-work items allow students to showcase their intermediate reasoning steps and offer students partial credit, which could alleviate anxiety and improve performance [50,51]. Taken together, the findings discussed in this section highlight the need to examine how assessment structures impact academic performance across male and female groups.

### D. Research questions

To summarize, prior physics education research has extensively studied student solutions to problems in terms of the correctness of the solution or solution path, but there has not been much research attention paid to the presentation quality of physics solutions, even though in other areas, it has been determined that presentation aspects such as handwriting quality play a multifaceted role. The literature review above suggests that SWP may exist at the intersection of three processes: student affect and reasoning during problem-solving, the communication of that reasoning, and the interpretation of student work by the grader.

Therefore, in order to conduct a more systematic investigation into the multiple issues involving presentation and performance in introductory physics, this study aims to determine whether aspects related to the presentation quality of physics students' show-work solutions could be consistently and meaningfully coded and whether the coding results are associated with other student measures of physics course performance. In order to better understand the nature of SWP, we also aimed to determine the extent to which SWP was associated with the ostensibly similar, well-studied Big Five personality inventory trait of "conscientiousness," which is known to be associated with grades. Further, since there are known observed differences between males and females in exam performance and conscientiousness [20,23,26,38,42], we investigate the extent to which SWP may play a role in these differences. More specifically, our research questions are:

RQ1: What are useful indicators that characterize SWP, and do these indicators cluster into a small number of meaningful categories (i.e., factors)?

RQ2: To what extent do SWP indicators and factors predict student performance on the same SW question, other SW questions, multiple-choice questions, exam scores, and nonexam graded components, even controlling for ACT score and conscientiousness?

RQ3: To what extent is SWP correlated with conscientiousness? If they are correlated, to what extent does SWP mediate self-reported conscientiousness and student performance?

RQ4: To what extent is SWP different between male and female students? If there are differences, to what extent does SWP mediate the relation between sex and course performance outcomes?

### E. Transparency and openness

This study was preregistered on the Open Science Framework website as an observational (exploratory) study and includes the original hypotheses, sampling plan, coding rubric, and analysis plan [52]. To balance the number of male and female students, the preregistered plan for the analytic sample included 264 students (132 female and 132 male) selected through stratified random sampling from 1152 students enrolled in the course. In accordance with institutional review board policy, only students who provided informed consent for their data to be used in research were eligible for analysis. During manuscript preparation, we discovered that some nonconsenting students had inadvertently remained in the dataset due to a clerical oversight. These cases were removed prior to analysis, yielding a final analytic sample of $N = 225$. Additional exclusions were applied when students lacked data for a given analysis (e.g., missing ACT math scores). Analysis-specific sample sizes are reported in tables or diagrams. Because we found, contrary to our original hypothesis, that SWP was not substantially correlated with conscientiousness, we extended one part of the exploratory study beyond the preregistered analysis plan to determine whether SWP mediates gender differences in performance outcomes.

Consistent with Cohen's treatment of power and effect sizes, we report sensitivity analyses for the final samples for the primary inferential test (a multiple regression predicting course grade from SW2 Form while controlling for ACT math and conscientiousness) as a *post hoc* analysis check that estimates the smallest effect that could be detected with high probability given the final sample size and planned model [53]. With $N = 190$, $\alpha = 0.05$ (two tailed), and $1\text{-}\beta = 0.80$, the model was sensitive to $\Delta R^2 \geq 0.041$ ($f^2 = 0.043$, partial $r = 0.203$). Under these circumstances, there was a high probability of detecting effects where adding SW2 Form explained about 4% additional variance in course grade beyond the control variables. Effects smaller than this threshold may be present but would be less likely to reach statistical significance in this sample, so null results should be interpreted with this sensitivity analysis in mind.

## II. PARTICIPANTS AND MATERIALS

The data for this study were collected in Autumn 2022 from students enrolled in an introductory calculus-based physics course taught at the Ohio State University, a large public research university. The majority of students were

first-year engineering majors, and the remaining students were mainly other STEM majors. In total, 1156 students were enrolled in the course. Of these, demographic data were available from the university registrar for 1152 students. Based on the self-reported university registrar demographic data, the sex distribution included 271 females, 874 males, and 7 students who did not disclose their sex. Students who did not disclose sex were excluded from this study, as the analysis focused on having a relatively balanced male-to-female ratio.

To specifically explore sex-based differences in show-work presentation, particularly whether presentation influences performance disparities between male and female students, we intentionally balanced the male-to-female ratio in our sample. Because prior research indicates gender-based differences in conscientiousness [23], we aimed to obtain a balanced sample to reduce potential biases by selecting equal numbers of male and female students, despite the course's predominantly male enrollment. We also sought to investigate any sex-based differences in show-work presentation. To achieve this, stratified random selection was used with strata defined by recitation section and sex. More specifically, we randomly sampled an equal number of three male and three female students in each of the 44 recitation sections. Two recitation sections lacked at least three female students (one recitation only had one female student and another recitation section only had two female students), so three female students were randomly sampled from other recitation sections.

Students who did not consent to the use of their data for research ($N = 193$) were removed from the sample postselection. The final sample consisted of 225 students: 114 female and 111 male. The Wilcoxon rank-sum test was used to compare the selected sample of students with the remaining enrolled students on metrics such as the show-work question scores and overall course grade to determine the extent to which the selected sample was representative of the broader population.

As shown in Table I, male students from the study sample had lower mean course grades than uncoded male students (75.7 vs 78.9) with a Cohen's $d$ of $-0.28$ and a 95% confidence interval (CI) that did not overlap with zero. Female students' means in the sample were slightly higher than the uncoded female students in overall course grade (76.3 vs 74.9), but the difference was small and consistent with being negligible ($d = 0.12$, 95% CI $[-0.14, 0.38]$). For show-work question scores, there were no statistically significant differences between the study sample and the rest of the uncoded students.

Together, these results suggest that female students in this sample are representative of the broader female population in the physics course, while sampled males may slightly underperform relative to male students not in the study sample.

This study analyzed SWP for two of the four show-work questions on the final exam (see Appendix A). These two show-work questions were selected due to their shared focus on Newtonian mechanics, whereas the other questions covered other topics in thermodynamics and relativity. These questions were also chosen because Newtonian mechanics is a comparatively well-researched (thus familiar) domain in education research compared to thermodynamics or relativity. Newtonian mechanics also tends to be introduced earlier in introductory physics courses, making it more likely that students had sufficient practice with the content for how to present solutions to show-work problems. In contrast, topics such as thermodynamics and relativity were introduced later in the semester, and students received less time on the topics. The two show-work questions selected for analysis were worth equal point values on the final exam and appeared consecutively as Questions 1 and 2 (referred to as SW1 and SW2; see Appendix A).

Student performance on SW1 and SW2 was significantly different, with evidence indicating SW2 had substantially better psychometric characteristics. For example, for SW 1, only 174 students completed their solution and

TABLE I. Female and male student means and differences between the study sample and the comparison group, which included students who consented to their data use, had final exam scores and were not part of the study sample. Cohen's $d$ represents the standardized mean difference between the two groups, and confidence intervals were not corrected for multiple comparisons. Course grades were on a scale of [0,100], and the show-work questions, labeled SW1 and SW2, respectively, were graded on a scale of [0,1].

| | | Study sample avg. ($N = 225$) | Uncoded students avg. ($N = 724$) | Cohen's $d$, 95% CI | ACT reported avg. ($N = 190$) |
|---|---|---|---|---|---|
| Course grade | Male | 75.7 | 78.9 | −0.28, [−0.48, −0.07] | 75.6 |
| | Female | 76.3 | 74.9 | 0.12, [−0.14, 0.38] | 76.8 |
| SW1 | Male | 0.66 | 0.64 | 0.09, [−0.12, 0.29] | 0.69 |
| | Female | 0.55 | 0.49 | 0.15, [−0.11, 0.41] | 0.57 |
| SW2 | Male | 0.53 | 0.59 | −0.19, [−0.39, 0.02] | 0.53 |
| | Female | 0.58 | 0.56 | 0.06, [−0.20, 0.32] | 0.59 |

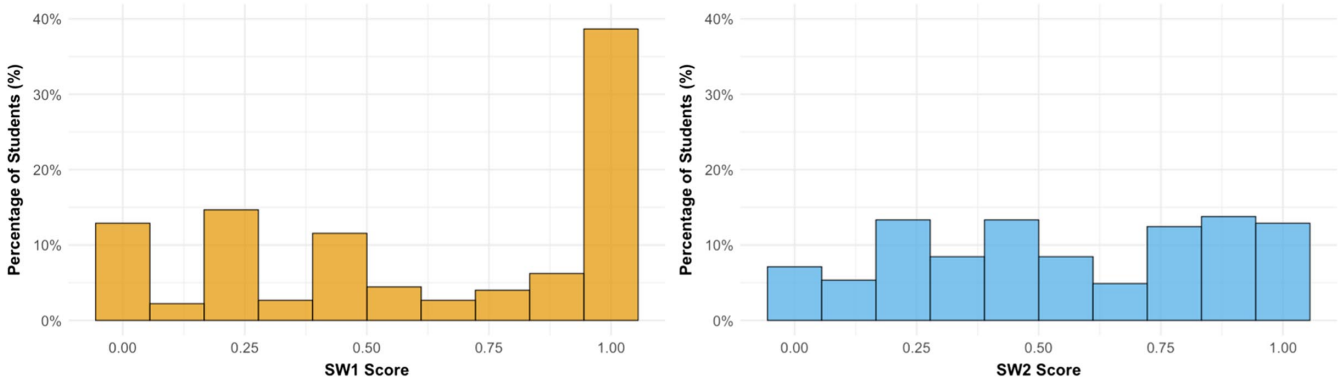



FIG. 1. The show-work score distributions for SW1 and SW2 for the 225 students chosen to be coded. The graphs are plotted on the same scale. SW1 has a ceiling effect compared to SW2. Note: SW1 has many students with the maximum score.

included a final answer from the final sample. The remaining 51 student solutions for SW1 generally included only one to two lines of work, such as copying information from the question itself or assigning variable names, without substantive progress toward a final answer. In contrast, SW2 had 207 completed student solutions since only 18 students did not complete the question. Both questions' score distributions are displayed in Fig. 1. SW1's score distribution was somewhat bimodal with a substantial ceiling effect, with 87 (39%) students receiving a full score, whereas SW2 was more evenly distributed and only had 29 (12.9%) students receiving a full score.

Further, as we shall see in Sec. IV A, SW2 was substantially more predictive of student performance than SW1. For example, SW2 was highly correlated ($r = 0.73$) with scores on all other SW exam problems (including midterms), but SW1 was more moderately correlated ($r = 0.56$) with the other SW scores. In fact, SW2 was also somewhat (and statistically significantly) more highly correlated with the course grade ($r = 0.64$, 95% CI [0.58, 0.70]) than SW1 was with the course grade ($r = 0.40$, 95% CI [0.31, 0.49]). Finally, when we also consider the high number of students at the ceiling for SW1, we expect that SW2 will be more useful (e.g., more discriminating) in our analysis.

A conscientiousness survey was administered online approximately one-third of the way through the semester at the start of an online homework assignment. The survey was graded solely for completion. The survey consisted of two items with one question being reverse-coded. Students responded using a five-point Likert scale for the following two items: "I see myself as dependable, self-disciplined" and the reverse-coded statement, "I see myself as disorganized, careless." These two items were taken from the Ten Item Personality Inventory (TIPI), a well-validated scale which measures Big Five personality dimensions [54]. These were the only two items available in the TIPI that assess conscientiousness. Prior research has shown that these two items are psychometrically sound and appropriate for use in educational studies, which is why they were included in the TIPI [54]. We examined reliability using the Spearman–Brown formula, as recommended for two-item scales [55]. The Spearman–Brown coefficient was $r_{SB} = 0.625$ for the sampled students, indicating low but acceptable reliability. Student scores from both items were then summed, and the total was normalized to a [0,1] scale for analysis.

## III. THE SWP MEASURE

In this section, we address RQ1 and describe our work in developing a measure of SWP. Broadly speaking, we envision the SWP construct to indicate the presence of visual features such as clarity, organization, and readability. SWP encompasses how students structure and communicate their problem-solving processes in their exam solution. This may involve how equations are laid out, whether reasoning is annotated or explained, and how neatly or legibly the work is written. SWP is distinct from correctness of the solution and instead reflects how the solution is presented.

To help further guide us, we also kept in mind potential reasons why a student might want to produce a high-quality presentation, including: to help organize their own thinking, to support a self-concept of doing quality work, or to facilitate easier interpretation and grading by the instructor, all of which could lead to a higher score. A coding rubric, outlined in Appendix D, was developed through an iterative process that integrated close inspection of students' written solutions with prior literature on observable presentation features such as organization, legibility, and communication of reasoning [3,10–13,16,30–36]. The literature directly motivated the use of the indicators related to handwriting legibility, layout and visual organization, text elaboration, defined variables, final answer markings, and use of diagrams [3,10–13,16,30–36]. Spacing and straight lines were treated as related aspects of visual organization that we chose to quantify more explicitly [11,13,14,31,35,36].

We then utilized an emergent coding technique to generate and refine potential indicators from the student

work for the SWP rubric rather than beginning from a predetermined rubric. Emergent coding is a coding technique that is used within classical grounded theory where codes are developed by the researchers examining the data [56]. This approach allows coding schemes to emerge from the data and ensures the coding categories reflect the content of these student solutions [56–58]. Utilizing this coding technique, two of the authors (H. M., Q. L.) independently reviewed a subset of student solutions and generated a list of potential indicators based on common patterns they observed. They then consulted with the other two authors of this study. First, they consulted with author A. F. H., who has over 25 years of physics teaching experience, to refine the indicators and the accompanying rubric. Next, they consulted with author K. F., a social psychologist, to discuss and finalize the rubric. Crossed-out work, smudging, photographic care, and holistic neatness were four indicators from emergent coding through directly inspecting the student solutions and submission format. These features were included because they could plausibly affect ease of grading and influence a grader's overall impression of the solution [11,30,31].

After combining the prior literature and repeated observable features, an initial set of 12 candidate indicators was assembled that captured the core presentation elements of students' written exam solutions (see Appendix D). The indicators in the rubric allowed us to more systematically and precisely quantify elements of show-work presentation quality.

The rubric was made up of scores of 0, 1, or 2 for each category, with 0 being the lowest score and 2 being the highest, and some categories incorporated examples of student work to help clarify differences. Prior to analysis, student responses were anonymized (the names were obscured) to prevent bias during the coding process. Coders began by independently solving SW1 and SW2 themselves and providing criteria they would consider important for coding. The rubric was then iteratively refined starting with "training sets" using sets of 10 solutions. During these sessions, coders identified discrepancies in their scoring and collaboratively discussed the differences. When substantial scoring inconsistencies were observed, additional details were added to the rubric to improve clarity and alignment. This process was repeated with subsequent sets of 10 practice solutions, until consistent scoring with a Cohen's kappa value greater than 0.8 was achieved for the practice set (see Table IX) . Notably, the rubric remained unchanged between coding SW1 and SW2.

To further refine our measure of SWP, we iteratively reduced the number of indicators to seven, each of which loaded onto 1 of 2 factors (see Appendix E). This process helped to maximize reliability and consistency in the ratings between SW1 and SW2 while maintaining conceptual consistency and relevance of the SWP construct. We briefly describe this process in the following paragraphs.

The 12 initial indicators were straight lines, spacing, crossed-out work, smudging, text elaboration, layout, handwriting legibility, photographic care, diagrams, final answer markings, defined variables, and holistic neatness. Holistic neatness was removed early on from the rubric, since it was initially intended to reflect a grader's first impression of the overall organization and clarity of a solution, rather than specific, observable features. Because of its subjective nature, it was difficult to define explicitly in the coding rubric, and the coders disagreed on how to apply it consistently.

To ensure that rating consistency was obtainable, we found that the interrater reliabilities for each of the remaining 11 items for both SW1 and SW2 were high (see Appendix B). There were some relatively small differences in interrater reliability between SW1 and SW2, but the values were typically above 0.9, with only three indicators less than 0.9, and all were above 0.82.

Next, we iteratively employed exploratory factor analysis (EFA). We used EFA as a diagnostic to check whether indicators clustered into conceptually meaningful dimensions of SWP, rather than defining the construct purely from a statistical factor solution. The EFA on the 11 items indicated weak, highly variable factor loadings (between SW1 and SW2) for four indicators—specifically diagrams, smudging, crossed-out work, and photographic care. There were additional reasons to remove these indicators. For example, diagrams and photographic care may depend on the question context and thus are not universal; namely, not all problems naturally need a diagram, and not all tests require uploading a picture of the work. Further, smudging and crossed-out work may be unavoidable when a student makes a mistake and wants to correct it. For example, students used pencils or pens as writing utensils, which contributes to different outcomes: either smudging or crossed-out work. Further, some students also wrote their answers digitally, such as through a tablet, and so they would not have issues such as smudging or photographic care as an indicator. As a result, these four indicators were removed from the model.

The remaining seven indicators were used for the final EFA analysis, as shown in Table II. Figure 2 shows a scree plot and parallel analysis indicating the seven indicators formed 2 factors. The indicators moderately loaded onto one of these 2 factors, similarly for both SW1 and SW2.

The first factor, which we label as "Form," consists of the indicators straight lines, spacing, layout, and handwriting legibility. We labeled this factor as Form because an inspection of these four indicators suggests a unifying construct of the clarity, structure, and organization of the content, namely, how words and equations are arranged two-dimensionally on the page and how legible the written content itself is. The second factor, which we label "Annotation," consists of final answer markings, defined variables, and text elaboration. We labeled this second

TABLE II. EFA tables for SW1 and SW2 with variance explained. Strong factor loadings have been bolded.

| | SW1 | | SW2 | |
|---|---|---|---|---|
| Variable | Form | Annotation | Form | Annotation |
| Straight lines | **0.51** | 0.10 | **0.48** | 0.26 |
| Spacing | **0.91** | −0.06 | **0.69** | −0.20 |
| Layout | **0.47** | 0.19 | **0.59** | 0.01 |
| Handwriting legibility | **0.40** | 0.19 | **0.50** | 0.26 |
| Final answer markings | −0.17 | **0.49** | 0.07 | **0.64** |
| Defined variables | 0.04 | **0.34** | −0.25 | **0.28** |
| Text elaboration | 0.02 | **0.86** | −0.09 | **0.70** |
| Variance explained | 0.21 | 0.17 | 0.20 | 0.17 |

factor as Annotation because its indicators reflect the extent to which the students used annotation in the form of circling items, defining terms, or explaining their steps to clarify the meaning of the problem or their solution process.

To provide a measure of reliability of the factors Form and Annotation, we present the Cronbach's alpha scores for SW1 and SW2 in Table III. Note that Form has a low but somewhat acceptable reliability around $\alpha = 0.6$. We also note that the within-student correlation between the Form ratings for SW1 and SW2 was $r = 0.62$, also indicating a moderate level of reliability of Form as a student-level attribute. It is important to note that Cronbach's alpha is sensitive to the number of items in a scale, and Form includes four indicators while Annotation only includes three. As such, the lower alpha for Annotation (around $\alpha = 0.4$) is not unexpected and does not necessarily imply poor internal consistency. The SW1-SW2 Annotation within-student correlation was $r = 0.39$, suggesting Annotation may be a less stable construct across items. Note that the indicator defined variables weakly loads onto Annotation, but we included it because it was conceptually consistent with the construct. Further, to provide more detail on the performance of individual indicators, we present a correlation table of the seven indicators and other student performance measures in Table X in Appendix C, and the defined variables indicator shows some interesting behavior, which we discuss later in Sec. IV A.

TABLE III. Cronbach's alpha for Form and Annotation.

| Factor | SW1 | SW2 |
|---|---|---|
| Form | 0.59 | 0.66 |
| Annotation | 0.45 | 0.34 |

In summary, we have constructed, thus operationally defined, a measure of SWP that is comprised of seven indicators that cluster into 2 orthogonal factors. The first factor is Form, which indicates a measure of two-dimensional display and legibility quality and is comprised of the moderately loading indicators: straight lines, spacing, layout, and handwriting. The second factor is Annotation, which indicates a measure of student notation of specific aspects of the solution and is comprised of three indicators: final answer marking, defined variables, and text elaboration. Because Form is shown to be a more reliable indicator, we will focus more on Form than Annotation for our subsequent analyses.

## IV. RELATIONS BETWEEN SWP AND SCORES

Here, we present key findings that address RQ2 and RQ3, highlighting the associations of the factors Form and Annotation with student performance outcomes. Most of our analyses focus on SW2 because SW1 is less reliable psychometrically since it shows weaker associations with outcomes (see Sec. II) and has a substantial number of students at ceiling.

We begin by addressing RQ2 and examine descriptive statistics, including correlations, among SWP factors and indicators and other student performance outcomes. We then investigate regression models to determine whether

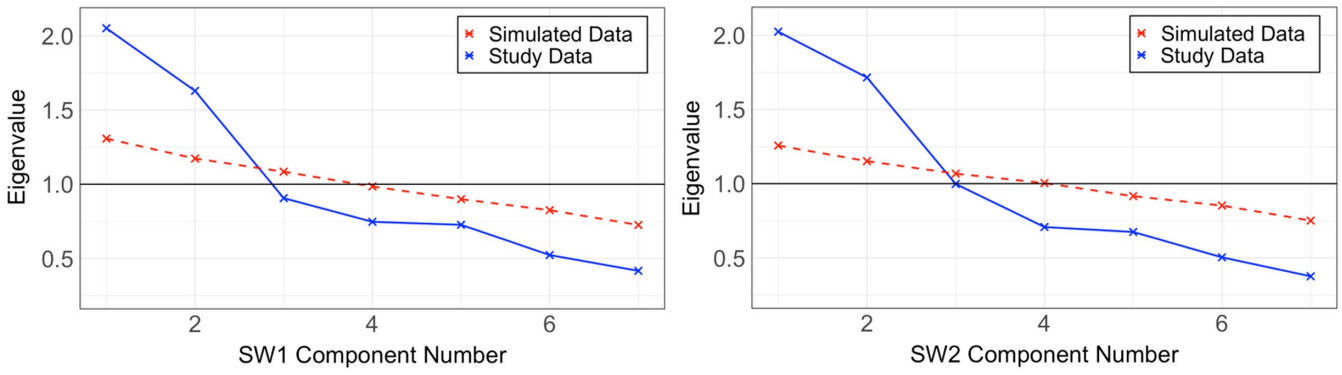



FIG. 2. Scree plots and parallel analysis simulations for SW1 and SW2. The blue line uses data from this study, while the red line represents eigenvalues obtained from a simulated randomized dataset with the same dimensions. The black line is the visual cutoff for eigenvalues equaling 1.

Form predicts student performance when both ACT math and conscientiousness are also included as predictors.

We note here that for the rest of the paper, we will compute the Form and Annotation scores as the unweighted mean score of their respective indicators. Essentially, we are using the unweighted "sum score" (as opposed to a weighted factor score) to determine the Form and Annotation scores. The simplicity of using the sum scores (as opposed to conducting latent-variable analyses) provides a level of clarity, transparency, and pragmatism to the analyses in the remainder of the paper while maintaining relevance in terms of empirical prediction [59]. For example, sum scores allow for simpler linear regression analysis (i.e., path analysis models compared to "full" structural equation models) and sum scores may provide better opportunities for cross-study comparison, since factor weights may be population specific [60]. Further, in our case, the weights may also be specific to SW questions, as implied by the variation in loadings of indicators onto factors for SW1 and SW2 in Table II, though the variations are typically relatively small and the patterns are qualitatively the same. Thus, in this paper, we forgo more elaborate latent-variable models that incorporate factor loadings for each indicator, leaving such investigations for future studies.

### A. Descriptive statistics

Table IV presents correlations among Form and Annotation for SW1 and SW2, conscientiousness, and several measures of student performance. There are several important correlations to point out. First, and perhaps most importantly, Form is reliably correlated ($r \sim 0.3$) with course grade, final exam score, nonexam grade components, SW2 score, and the mean of all SW scores (including on midterms and the final). Interestingly, SW2 Form is more strongly correlated with all SW scores ($r = 0.31$, 95% CI [0.19, 0.44]) than with all multiple-choice questions ($r = 0.15$, 95% CI [0.01, 0.29]; Steiger's $z = 2.40$, $p = 0.016$), increasing our confidence that Form is more predictive of SW performance than multiple-choice performance. Note the SW1 Form has a similar pattern of correlations, further indicating that the measure of Form is reliable across SW questions.

Second, Annotation is only weakly and unreliably correlated ($|r| \sim 0.15$) with student performance. In fact, SW2 Annotation is negatively correlated with final exam performance ($r = -0.15$, 95% CI $[-0.27, -0.01]$), but SW1 Annotation is positively correlated with exam performance ($r = 0.18$, 95% CI [0.03, 0.31]). This indicates that Annotation at best plays a weak and inconsistent role in prediction of student performance.

Third, somewhat surprisingly, both Form and Annotation are at best very weakly correlated with Conscientiousness. This is especially interesting in light of the fact that Conscientiousness is, as expected, correlated with course grade, nonexam components, and all SW scores. This finding that SWP and conscientiousness are both correlated with course performance but not with each other indicates that SWP and conscientiousness are distinct student-level characteristics. This lack of correlation between SWP and conscientiousness already provides a simple answer to RQ3: SWP does not mediate the relation between

TABLE IV. Pearson correlation table. Spearman correlations were used for sex. SW2 was included in both the final exam score and average show-work question score. All correlations utilized the study sample of 225 students except for the correlations involving ACT scores ($N = 190$), which had smaller sample sizes due to missing data. $P$-values were adjusted for multiple comparisons using Holm–Bonferroni's procedures within the Pearson correlation matrix and within the set of sex correlations. Bolded values denote Holm–Bonferroni adjusted $p < 0.05$.

| | Course grade | Final exam | Mean MC | Mean SW | Nonexam components | SW1 | SW2 | SW1 Form | SW1 Annotation | SW2 Form | SW2 Annotation | Conscientiousness | Sex |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Final exam | **0.84** | | | | | | | | | | | | |
| Mean MC | **0.57** | **0.77** | | | | | | | | | | | |
| Mean SW | **0.82** | **0.92** | **0.47** | | | | | | | | | | |
| Nonexam comp. | **0.65** | **0.31** | **0.13** | **0.35** | | | | | | | | | |
| SW1 | **0.40** | **0.53** | **0.28** | **0.56** | 0.02 | | | | | | | | |
| SW2 | **0.64** | **0.66** | **0.31** | **0.73** | **0.36** | **0.14** | | | | | | | |
| SW1 Form | **0.31** | **0.23** | **0.13** | **0.25** | **0.28** | 0.01 | **0.27** | | | | | | |
| SW1 Annotation | **0.22** | **0.18** | 0.11 | **0.18** | **0.13** | **0.18** | 0.11 | 0.09 | | | | | |
| SW2 Form | **0.36** | **0.29** | **0.15** | **0.31** | **0.29** | 0.04 | **0.30** | **0.65** | 0.05 | | | | |
| SW2 Annotation | −0.12 | **−0.14** | **−0.17** | −0.10 | 0.05 | −0.09 | 0.00 | 0.02 | **0.41** | −0.03 | | | |
| Conscientiousness | **0.27** | **0.21** | 0.08 | **0.24** | **0.29** | −0.01 | **0.29** | 0.09 | −0.07 | 0.09 | 0.12 | | |
| Sex (1 = Female) | 0.00 | −0.06 | −0.12 | −0.02 | **0.18** | **−0.17** | 0.07 | **0.26** | 0.05 | **0.27** | **0.20** | 0.15 | |
| ACT math[a] | **0.53** | **0.49** | **0.31** | **0.49** | **0.15** | **0.18** | **0.35** | 0.12 | 0.12 | **0.16** | −0.12 | 0.01 | −0.03 |
| Mean | 75.90 | 58.30 | 61.40 | 56.30 | 94.20 | 0.61 | 0.55 | 1.36 | 1.14 | 1.34 | 1.20 | 0.67 | ⋯ |
| Std. Dev. | 11.80 | 19.40 | 20.93 | 23.39 | 8.40 | 0.39 | 0.32 | 0.48 | 0.55 | 0.51 | 0.46 | 0.24 | ⋯ |

[a]Mean ACT math score for the 190 students was 28.4, and the standard deviation was 4.58.

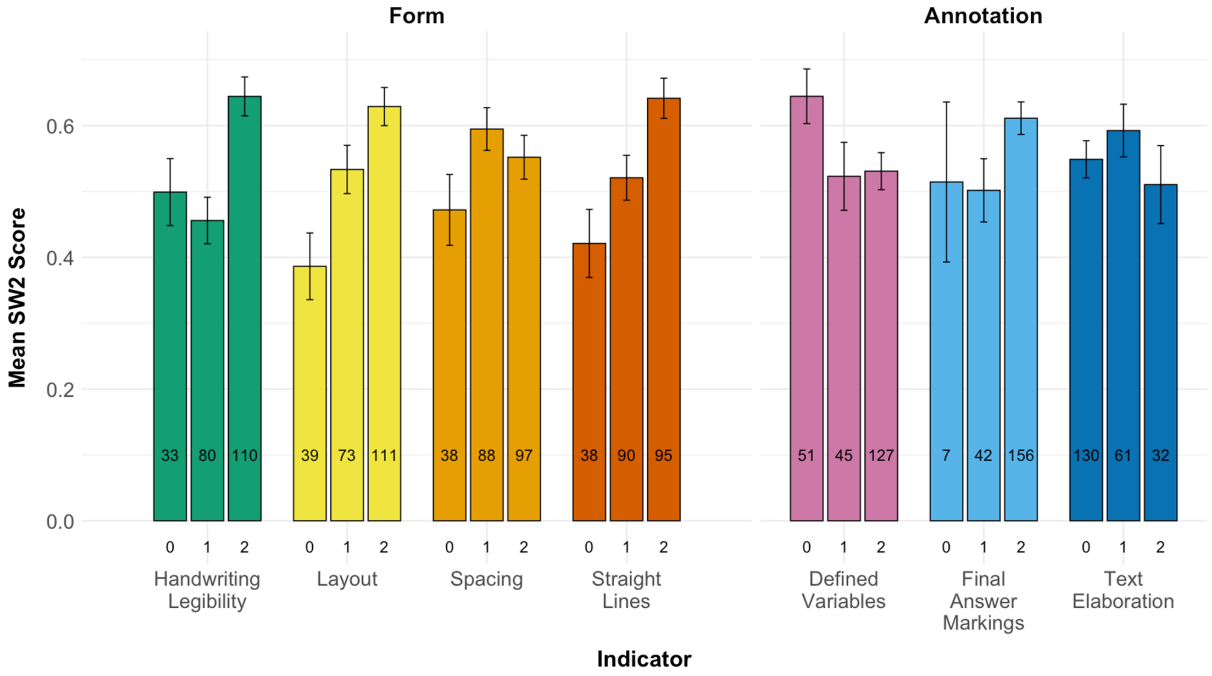



FIG. 3. Mean SW2 scores as a histogram by mean indicator rating bins with error bars representing standard error. Indicator ratings were averaged over the two raters. Mean indicator ratings labeled 0, 1, and 2 correspond to scores of 0 or 0.5, scores of 1 or 1.5, and scores equal to 2, respectively. Indicators comprising the Form factor are displayed on the left, and those comprising Annotation are on the right. The number of students per bin is indicated within each bin. Note for the indicator final answer marking, the 18 students who had no final answer (only one or two lines of work or only rewrote the question without substantive progress toward a final answer) were not included in this plot.

conscientiousness and exam performance because SWP and Conscientiousness are not even correlated.

Finally, we point out that sex is correlated with the SWP factor Form ($r \sim 0.27$), indicating that female students tend to have higher Form scores than males. We will address this difference between male and female students more in Sec. V.

To provide more detail and insight into the relation between SW score and SWP, we present in Fig. 3 the mean SW2 score for each of the three rating levels of each of the seven indicators. Inspection of this figure reveals that for each of the indicators comprising Form, the mean SW2 score typically increases with the indicator rating.

More specifically, a one-way analysis of variance (ANOVA) was conducted to assess differences in SW2 scores by indicator bin, and, to adjust for multiple comparisons, additional *post hoc* Dunnett's tests were then used to compare the 1 and 2 rating bins against the "comparison" bin 0. Significant main effects were found for layout ($p < 0.05$) for both bins. Handwriting legibility and straight lines each had the 2 bin significantly different ($p < 0.05$), but the 1 bin did not show any significant differences. The fourth Form indicator, spacing, had marginal differences. For layout, students in the 1 bin scored on average 0.15 higher (95% CI [0.01, 0.28]) on SW2 than the 0 bin, while students in the 2 bin scored on average 0.24 higher (95% CI [0.12, 0.37]). For handwriting legibility, only the 2 bin was significantly different, with students scoring 0.15 higher (95% CI [0.01, 0.28]) on average. Similarly, for straight lines, the 2 bin was also significantly different, with students scoring 0.22 higher (95% CI [0.09, 0.35]) on average. For spacing, the effects are smaller and consistent with zero. Students in the 1 bin scored on average 0.12 higher (95% CI [−0.01, 0.26]) than the 0 bin, while students in the 2 bin scored on average 0.08 (95% CI [−0.05, 0.21]) higher.

In contrast, for the indicators comprising Annotation, no consistent effect or pattern exists between the ratings of all the indicators and SW2 score. More specifically, no significant main effects were observed for final answer markings or text elaboration, while defined variables showed a marginally significant effect. For defined variables, students in the 1 bin scored on average 0.12 lower (95% CI [−0.27, 0.02]) than the 0 bin, while students in the 2 bin scored on average 0.11 lower (95% CI [−0.23, 0.00]). While the effect for text elaboration was not statistically significant, bin 2 students also trended lower, with an average difference of 0.11 (95% CI [−0.23, 0.00]) lower than the 0 bin.

### B. Regression results

In the last section, we saw that SWP Form was correlated with several student performance outcomes, including course grade, exam scores, and nonexam scores. In this section, we build several multiple regression models

to better understand the controlled variation among SW Form, conscientiousness, ACT score, and student course performance measures, as shown in Table V. This regression analysis will also help us to better understand our research questions involving mediation. We note several findings from the regression model results in Table V.

The first observation is that, consistent with prior research by Simmons and Heckler [26], ACT math score is a relatively strong predictor of course grade ($R^2 = 0.28$), is slightly less predictive of show-work scores ($R^2 = 0.22$), is more weakly predictive of multiple-choice scores, and is a very weak predictor of nonexam components.

The second set of observations involves the predictiveness of conscientiousness, controlling for ACT score. Consistent with prior research as mentioned in the introduction, conscientiousness (weakly) predicts course grade ($R^2 = 0.1$). However, it is interesting to note that conscientiousness is a stronger predictor of the nonexam component of the grade compared to ACT score ($R^2 = 0.1$ vs $R^2 = 0.02$). Further, conscientiousness does not reliably predict multiple-choice exam scores, though (accounting for multiple comparisons) it does reliably, though somewhat weakly, predict show-work scores ($R^2 = 0.04$), when controlling for ACT score.

Perhaps the most important results of the regression models involve the extent to which SW Form predicts student performance above and beyond what is predicted by ACT score and conscientiousness. Specifically, Form weakly but statistically significantly (accounting for multiple comparisons) predicts show-work score ($R^2 = 0.04$), but it does not statistically significantly predict multiple-choice score. This contrast in prediction provides a measure of discriminant validity for the SWP measure Form. Further, Form predicts nonexam scores more strongly than ACT score, even controlling for conscientiousness, and Form is also a weak but statistically significant predictor of course grade ($R^2 = 0.05$).

Altogether, these regression results demonstrate that ACT math, conscientiousness, and Form each predict distinct aspects of student performance. ACT math emerged as the strongest predictor of student performance for exam-related outcomes. In comparison, both conscientiousness and Form consistently and independently, though somewhat weakly, predicted course grade, show-work performance, and nonexam components, but they were not statistically significant predictors of multiple-choice performance.

## V. DIFFERENCES BETWEEN MALE AND FEMALE STUDENTS

### A. Descriptive statistics

To address RQ4, we begin by first comparing the mean scores of male and female students on each SWP indicator

TABLE V. Regression table for student outcomes scores. Mean multiple-choice and show-work scores are from the two midterm exams and the final exam.[a] All regression dependent and independent variables have been standardized and mean-centered, so all intercepts are equal to zero and have been removed from the table. The student sample in this regression analysis is comprised of the 190 students (of the 225 in the study) with ACT scores. Bolded values denotes $p < 0.05$, with no multiple comparisons assumed. Standard error values are in parentheses.

| Variables | Mean multiple choice | | | | Mean show work[a] | | | | SW2 score | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ACT math | ⋯ | **0.32 (0.07)** | **0.31 (0.07)** | **0.31 (0.07)** | ⋯ | **0.48 (0.07)** | **0.48 (0.06)** | **0.45 (0.06)** | ⋯ | **0.34 (0.07)** | **0.34 (0.07)** | **0.31 (0.07)** |
| Conscientiousness | ⋯ | ⋯ | 0.11 (0.07) | 0.10 (0.07) | ⋯ | ⋯ | **0.20 (0.06)** | **0.18 (0.06)** | ⋯ | ⋯ | **0.31 (0.07)** | **0.28 (0.07)** |
| SW2 Form | 0.12 (0.07) | ⋯ | ⋯ | 0.06 (0.07) | **0.25 (0.07)** | ⋯ | ⋯ | **0.16 (0.06)** | **0.30 (0.07)** | ⋯ | ⋯ | **0.22 (0.07)** |
| Adjusted $R$-squared | 0.01 | 0.09 | 0.10 | 0.10 | 0.06 | 0.22 | 0.26 | 0.28 | 0.09 | 0.11 | 0.20 | 0.25 |
| Residual Std. error | 1.00 | 0.95 | 0.95 | 0.95 | 0.97 | 0.88 | 0.86 | 0.85 | 0.96 | 0.94 | 0.89 | 0.87 |

| Variables | Nonexam components | | | | | Course grade | | |
|---|---|---|---|---|---|---|---|---|
| ACT math | ⋯ | **0.15 (0.07)** | **0.15 (0.07)** | 0.11 (0.07) | ⋯ | **0.53 (0.06)** | **0.54 (0.06)** | **0.49 (0.06)** |
| Conscientiousness | ⋯ | ⋯ | **0.32 (0.07)** | **0.29 (0.07)** | ⋯ | ⋯ | **0.31 (0.06)** | **0.28 (0.06)** |
| SW2 Form | **0.30 (0.07)** | ⋯ | ⋯ | **0.25 (0.07)** | **0.35 (0.07)** | ⋯ | ⋯ | **0.24 (0.06)** |
| Adjusted $R$-squared | 0.09 | 0.02 | 0.12 | 0.17 | 0.12 | 0.28 | 0.38 | 0.43 |
| Residual Std. error | 0.96 | 0.99 | 0.94 | 0.91 | 0.94 | 0.85 | 0.79 | 0.76 |

[a]Average show-work question scores exclude the score for show-work Question 2 from the final exam.

for both SW1 and SW2, as shown in Table VI. There are two main observations that we point out from this table. The first observation is the large and consistent difference in handwriting legibility for both SW questions. Specifically, female students tend to have substantially better handwriting scores than male students for both SW1 (Cohen's $d = 1.29$, 95% CI [1.00, 1.63]) and SW2 (Cohen's $d = 1.11$, 95% CI [0.84, 1.43]), even after correcting for multiple comparisons among all the indicators. The second observation is that while the 95% confidence intervals often overlap with zero, all but one of the 14 mean score differences are positive, indicating that female students tend to have higher SWP scores than male students.

We can get a better sense of the extent to which there is a collective pattern of differences in SWP indicators between male and female students by examining the Form and Annotation scores. These means are presented in Table VII. The results indicate that female students have consistently and substantially better Form scores than male students (even using Holm–Bonferroni correction for multiple comparisons among the four SWP factor scores in Table VII) for both SW1 (Cohen's $d = 0.59$, 95% CI [0.33, 0.86], $p = 1.6 \times 10^{-5}$) and SW2 (Cohen's $d = 0.60$, 95% CI [0.34, 0.88], $p = 1.1 \times 10^{-5}$). On the other hand, there is no difference between female and male students in Annotation scores for SW1 (Cohen's $d = 0.10$, 95% CI [−0.16, 0.36], $p = 0.456$), and, again including the Holm–Bonferroni correction for all 4 factor scores, there is a significant but smaller difference in Annotation scores for SW2 (Cohen's $d = 0.42$, 95% CI [0.16, 0.69], $p = 0.002$). Interestingly and perhaps unexpectedly, Table VII also shows that there was at best a small difference in conscientiousness scores between female and male students, and the confidence interval does slightly overlap with zero.

As far as student performance in the course, there may be some small but important differences between male and female students. For example, Table VII indicates that there is no significant difference in course grade between male and female students, though this may be due to the competing effects of female students having statistically significantly higher nonexam scores but slightly lower exam scores, consistent with findings from previous studies [26]. Examining the exam performance more closely, Table VII reveals that while there is no appreciable difference in the SW scores between male and female students, there may be weak evidence that female students score slightly lower (∼0.25 SD) on MC final exam items, although the 95% CI of the difference does slightly overlap with zero, so the effect is small or possibly negligible.

In summary, female students have higher Form scores on average and especially higher handwriting scores than male students. Further, female students score about the same on average as male students on SW problems, but female students may score slightly lower on multiple-choice items, though it is not clear whether this is a small effect (∼0.25 SD) with insufficient statistical power or if the difference is negligible. Note that our observed male-female effect size differences in MC and SW are consistent with the studies mentioned in Sec. I [44,45].

### B. Mediation models

In the previous section, we observed that female students have higher Form scores than male students. Because we have also observed that Form scores predict course performance, we are led to the second part of RQ4, namely, investigating the extent to which SWP may

TABLE VI. Comparison of male ($N = 111$) and female ($N = 114$) student mean scores on the seven SWP indicators for SW1 and SW2. 95% bootstrapped confidence interval values for female-male mean differences are reported and bolded if they do not overlap with zero. Note $p$-values and confidence intervals were not adjusted for multiple comparisons. The maximum score for each indicator is 2.

| | Indicator | F Mean (SD) | M Mean (SD) | F-M Mean | 95% CI | $p$-value |
|---|---|---|---|---|---|---|
| SW1 | Straight lines | 1.70 (0.54) | 1.56 (0.63) | 0.14 | [−0.02, 0.29] | 0.078 |
| | Spacing | 1.44 (0.65) | 1.33 (0.70) | 0.11 | [−0.06, 0.29] | 0.207 |
| | Layout | 1.14 (0.78) | 1.12 (0.82) | 0.02 | [−0.18, 0.23] | 0.829 |
| | Handwriting legibility | 1.67 (0.50) | 0.87 (0.71) | 0.79 | **[0.63, 0.95]** | <0.001 |
| | Final answer markings | 1.72 (0.58) | 1.73 (0.51) | −0.01 | [−0.18, 0.14] | 0.875 |
| | Defined variables | 1.00 (0.84) | 0.86 (0.76) | 0.14 | [−0.07, 0.35] | 0.193 |
| | Text elaboration | 0.98 (0.85) | 0.90 (0.73) | 0.08 | [−0.12, 0.29] | 0.441 |
| SW2 | Straight lines | 1.33 (0.72) | 1.22 (0.72) | 0.12 | [−0.07, 0.30] | 0.221 |
| | Spacing | 1.35 (0.71) | 1.26 (0.71) | 0.09 | [−0.09, 0.28] | 0.322 |
| | Layout | 1.49 (0.62) | 1.24 (0.80) | 0.25 | **[0.06, 0.44]** | 0.009 |
| | Handwriting legibility | 1.71 (0.49) | 1.01 (0.74) | 0.70 | **[0.53, 0.86]** | <0.001 |
| | Final answer markings | 1.75 (0.48) | 1.70 (0.54) | 0.05 | [−0.09, 0.19] | 0.465 |
| | Defined variables | 1.43 (0.81) | 1.27 (0.84) | 0.15 | [−0.06, 0.37] | 0.172 |
| | Text elaboration | 0.72 (0.77) | 0.39 (0.66) | 0.33 | **[0.14, 0.51]** | 0.001 |

TABLE VII. Descriptive statistics table comparing average sampled male ($N = 111$) and female scores ($N = 114$). The female-male mean difference is reported as a 95% confidence interval value and is bolded when it does not include 0 (not adjusted for multiple comparisons). There was one male and one female student, each who did not provide full conscientiousness survey data, so they were excluded from the conscientiousness survey row data. For the comparison for ACT math, only $N = 93$ female students and $N = 97$ male students provided their scores, so only those scores were utilized.

| | Range | F Mean (SD) | M Mean (SD) | F-M Diff. | F-M groups 95% CI | $p$-value |
|---|---|---|---|---|---|---|
| SW1 Form | [0,2] | 1.49 (0.41) | 1.22 (0.49) | 0.27 | **[0.15, 0.39]** | <0.001 |
| SW1 Annotation | [0,2] | 1.16 (0.59) | 1.11 (0.51) | 0.05 | [−0.09, 0.20] | 0.46 |
| SW2 Form | [0,2] | 1.47 (0.44) | 1.18 (0.52) | 0.29 | **[0.16, 0.42]** | <0.001 |
| SW2 Annotation | [0,2] | 1.29 (0.45) | 1.10 (0.46) | 0.19 | **[0.07, 0.31]** | 0.002 |
| Con. survey score | [0,1] | 0.76 (0.20) | 0.71 (0.18) | 0.05 | [0.00, 0.10] | 0.05 |
| SW1 | [0,1] | 0.54 (0.40) | 0.67 (0.36) | −0.13 | **[−0.23, −0.03]** | 0.01 |
| SW2 | [0,1] | 0.58 (0.31) | 0.52 (0.33) | 0.06 | [−0.03, 0.14] | 0.18 |
| Course grade | [0,100] | 76.17 (11.12) | 75.68 (12.45) | 0.50 | [−2.61, 3.60] | 0.75 |
| MC final exam score | [0,100] | 59.00 (20.47) | 63.85 (21.21) | −4.86 | [−10.34, 0.62] | 0.08 |
| SW final exam score | [0,100] | 57.96 (18.33) | 61.46 (19.40) | −3.50 | [−8.36, 1.46] | 0.17 |
| Avg. MC score | [0,100] | 55.34 (16.01) | 58.81 (15.57) | −3.47 | [−7.62, 0.68] | 0.10 |
| Avg. SW score | [0,100] | 60.31 (20.57) | 60.77 (21.36) | −0.45 | [−5.96, 5.06] | 0.87 |
| Final exam score | [0,100] | 57.14 (18.76) | 59.55 (20.08) | −2.41 | [−7.52, 2.70] | 0.35 |
| Exam scores | [0,100] | 63.43 (16.45) | 64.57 (16.29) | −1.13 | [−5.44, 3.17] | 0.60 |
| Nonexam scores | [0,100] | 95.56 (5.74) | 92.71 (10.28) | 2.85 | **[0.66, 5.05]** | 0.01 |
| ACT math score | [0,36] | 28.25 (4.75) | 28.53 (4.44) | −0.28 | [−1.60, 1.04] | 0.68 |

mediate male-female differences in course performance. Because it is established that ACT score and conscientiousness predict course outcomes, we would also like to include them in a mediation model, thus controlling for these potential additional mediators. Therefore, we would like to investigate the mediation model shown in Fig. 4 in order to investigate possible causal relationships and indirect effects [61]. Since these are correlational studies rather than control-treatment studies, we treat the mediation analyses as exploratory.

Since, as discussed at the beginning of Sec. IV, we are using unweighted means to determine Form and conscientiousness; the mediation analysis is essentially an extension of the linear regression analysis of Sec. IV. Specifically, we model all variables as observed (rather than latent), thus reducing the structural equation model to a path analysis model. Six outcome scores were investigated, as shown in Table VIII. The mean show-work question score was calculated from four show-work questions across two midterms (two per midterm) and three of the four show-work questions for the final exam. SW2 was excluded from the SW mean to further support the idea that SW Form is a more general student-level measure and any effects are not specific to a particular SW question. A parallel (rather than sequential) mediation model for Form and conscientiousness was chosen because conscientiousness and Form were not significantly correlated.

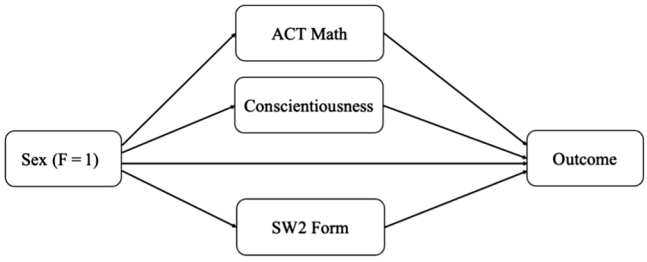



FIG. 4. Simultaneous mediation model diagram. The potential outcomes are SW2, final exam scores, average multiple choice, average show work, nonexam components, and course grade. The specific path coefficient values and indirect effects can be found in Table VIII.

Using the lavaan package in R [62], we created six distinct mediation models, with one for each outcome variable, as represented in Fig. 4. For each model, the predictor was sex (0 = male and 1 = female), and there were three parallel mediators: ACT math score, self-reported conscientiousness, and Form. These mediation analyses use observed-variable path models rather than latent-variable analysis. We did not fit a latent-variable structural equation modeling for Form in the mediation analyses. The quantity Form was calculated from the unweighted average of the indicators: straight lines, spacing, handwriting, and layout. Coefficients were standardized and standard errors were computed using bootstrapping (N = 1000), using the standardized solution function in the lavaan package. The mediation model results in Table VIII provide the estimates for the standardized path coefficients (direct effects) and indirect effects for all six outcome variables. All coefficients, standard errors, and confidence intervals have been standardized, except for sex, which was treated as a categorical variable (male = 0 and female = 1). We note that the fit statistics (including $R^2$) indicate that the models are a

TABLE VIII. Mediation model estimates. Both the final exam score and mean show-work question score were calculated without SW2, and all factor scores were from SW2. The student sample in this regression analysis is comprised of the 190 students (of the 225 in the study) with ACT scores. Bolded values represent $p < 0.05$, unadjusted for multiple comparisons. Coefficient estimates are reported with standard error in parentheses and 95% confidence intervals in brackets.

| Path | SW2 | Final exam | Mean MC | Mean SW | Nonexam | Course grade |
|---|---|---|---|---|---|---|
| *Direct effects* | | | | | | |
| Sex → ACT | −0.06 (0.15) [−0.32, 0.24] | −0.06 (0.15) [−0.32, 0.24] | −0.06 (0.15) [−0.32, 0.24] | −0.06 (0.15) [−0.32, 0.24] | −0.06 (0.15) [−0.32, 0.24] | −0.06 (0.15) [−0.32, 0.24] |
| Sex → Form | **0.60 (0.14) [0.33, 0.86]** | **0.60 (0.14) [0.33, 0.86]** | **0.60 (0.14) [0.33, 0.86]** | **0.60 (0.14) [0.33, 0.86]** | **0.60 (0.14) [0.33, 0.86]** | **0.60 (0.14) [0.33, 0.86]** |
| Sex → Con. | 0.23 (0.15) [−0.07, 0.50] | 0.23 (0.15) [−0.07, 0.50] | 0.23 (0.15) [−0.07, 0.50] | 0.23 (0.15) [−0.07, 0.50] | 0.23 (0.15) [−0.07, 0.50] | 0.23 (0.15) [−0.07, 0.50] |
| ACT → Outcome | **0.31 (0.07) [0.17, 0.43]** | **0.45 (0.06) [0.34, 0.57]** | **0.46 (0.06) [0.34, 0.59]** | **0.49 (0.06) [0.37, 0.62]** | 0.12 (0.068) [−0.05, 0.29] | **0.49 (0.06) [0.36, 0.63]** |
| Form → Outcome | **0.22 (0.07) [0.10, 0.35]** | **0.20 (0.07) [0.08, 0.32]** | **0.16 (0.06) [0.05, 0.27]** | **0.23 (0.06) [0.12, 0.34]** | **0.22 (0.07) [0.05, 0.39]** | **0.25 (0.06) [0.14, 0.37]** |
| Con → Outcome | **0.28 (0.07) [0.16, 0.40]** | **0.22 (0.06) [0.11, 0.34]** | **0.16 (0.06) [0.05, 0.27]** | **0.26 (0.06) [0.15, 0.37]** | **0.28 (0.07) [0.10, 0.48]** | **0.29 (0.06) [0.17, 0.41]** |
| Sex → Outcome | 0.00 (0.13) [−0.24, 0.26] | −0.21 (0.13) [−0.47, 0.02] | **−0.26 (0.13) [−0.50, −0.03]** | −0.15 (0.12) [−0.38, 0.11] | 0.16 (0.14) [−0.05, 0.38] | −0.08 (0.12) [−0.29, 0.14] |
| *Indirect effects* | | | | | | |
| Sex → ACT → Outcome | −0.02 (0.05) [−0.11, 0.08] | −0.03 (0.07) [−0.15, 0.12] | −0.03 (0.07) [−0.15, 0.11] | −0.03 (0.07) [−0.17, 0.12] | −0.01 (0.02) [−0.06, 0.04] | −0.03 (0.07) [−0.17, 0.12] |
| Sex → Con. → Outcome | 0.06 (0.04) [−0.02, 0.16] | 0.05 (0.04) [−0.01, 0.13] | 0.04 (0.03) [−0.01, 0.10] | 0.06 (0.04) [−0.02, 0.15] | 0.06 (0.06) [−0.01, 0.20] | 0.07 (0.05) [−0.02, 0.17] |
| Sex → Form → Outcome | **0.13 (0.05) [0.05, 0.24]** | **0.12 (0.05) [0.04, 0.22]** | **0.10 (0.05) [0.01, 0.19]** | **0.14 (0.05) [0.06, 0.23]** | **0.13 (0.06) [0.03, 0.27]** | **0.15 (0.05) [0.06, 0.26]** |
| Total effect | 0.18 (0.14) [−0.10, 0.46] | −0.06 (0.15) [−0.37, 0.22] | −0.16 (0.15) [−0.46, 0.13] | 0.02 (0.15) [−0.27, 0.32] | **0.35 (0.14) [0.09, 0.63]** | 0.104 (0.15) [−0.19, 0.39] |
| Adjusted $R^2$ | 0.24 | 0.32 | 0.28 | 0.40 | 0.17 | 0.43 |
| df | 3 | 3 | 3 | 3 | 3 | 3 |
| $\chi^2$ | 6.97 | 6.97 | 6.97 | 6.97 | 6.97 | 6.97 |
| RMSEA (90% CI) | 0.08 [0.00, 0.17] | 0.08 [0.00, 0.17] | 0.08 [0.00, 0.17] | 0.08 [0.00, 0.17] | 0.08 [0.00, 0.17] | 0.08 [0.00, 0.17] |
| SRMR | 0.05 | 0.06 | 0.05 | 0.06 | 0.05 | 0.06 |
| CFI/TLI | 0.95/0.82 | 0.96/0.86 | 0.95/0.84 | 0.97/0.89 | 0.93/0.76 | 0.97/0.89 |

reasonable fit to the data; especially, the standardized root mean squared residual (SRMR) values indicate a very good fit, which is the critical statistic for these path models. We also note that the Tucker Lewis index (TLI) statistic indicates a fit that is not good, but this is to be expected in a model in which several paths have very low strengths. We chose to keep all paths in the mediation model (Table VIII), since *a priori*, they were plausible mediation paths.

There were several notable results from the mediation analyses in Table VIII, above and beyond what was revealed in the analyses of Secs. IV A and IV B. First and most importantly, the results indicate that Form partially mediates the relation between sex and all of the performance outcomes. Specifically, the estimates for the indirect effect of sex → Form → outcomes are all positive, ranging in size from 0.10 to 0.15. In contrast, ACT score had a negligible indirect effect, and conscientiousness had at best a small and positive indirect effect, typically less than half the size of the mediation of Form, and with 95% confidence intervals overlapping with zero.

Second, the result that Form partially mediates the relation between sex and performance outcome (with a positive sign) is especially interesting when considering

several specific performance outcomes. For example, the total (combined) effect of sex on Final exam score is very small at $-0.06$ ($SE = 0.15$) (consistent with zero), but this is comprised of a positive indirect effect from Form at 0.12 (0.05), a negative (though possibly zero) direct effect of sex on final exam score at $-0.21(0.13)$, and a small positive (possibly zero) indirect effect of conscientiousness at 0.05 (0.04). In other words, controlling for Form, ACT, and conscientiousness, there is a negative association between sex and final exam score (female students score lower), and there is a compensating positive indirect effect of sex → Form → final exam.

There is a similar compensating effect for MC items. The total effect of sex on MC score is small and consistent with zero, though leaning negative $-0.16(0.15)$. This small total effect is comprised of a positive indirect effect from Form at 0.10 (0.05) and a negative direct effect of sex on MC score at $-0.26(0.13)$. In this case, it is clearer that, controlling for ACT score, Form, and conscientiousness, female students score lower than men on MC items, but this is somewhat compensated for by the positive indirect effect of Form. Since MC items are also scored automatically, this compensating effect cannot be attributed to grader bias. Instead, this suggests that Form may reflect a broader student-level attribute, such as structured or organized thinking, that positively influences performance across different assessment formats.

Finally, the nonexam scores indicate a different pattern: not one of competing effects, but one of aligned effects. Here, there is a positive total effect of sex on nonexam scores, and this is comprised of the positive indirect effect of Form and a positive direct effect of sex on nonexam scores. Here, again controlling for ACT score, Form, and conscientiousness, female students score higher than men on nonexam scores; the indirect effect of Form further boosts the nonexam scores for female students.

## VI. DISCUSSION AND CONCLUSION

This study essentially consisted of two parts. First, we developed and evaluated an approach to characterize and quantify the presentation qualities of student-written solutions to physics show-work problems. Second, we coded show-work solutions from a female-male balanced sample of 225 students and found that some aspects of SWP were moderately to weakly, but reliably, predictive of show-work scores and course grades, even when controlling for ACT scores and conscientiousness. Further, we found differences in SWP mean scores between male and female students, and we found evidence that this difference partially mediates male-female differences in performance on all graded components of the course that we studied. We provide more specific comments below in our responses to each research question, and we also provide further comments on the limitations and implications of the results.

### A. RQ1: What are useful indicators that characterize SWP, and do these indicators cluster into a small number of meaningful categories (i.e., factors)?

For the two show-work problems studied, an emergent coding process revealed that SWP can be meaningfully and reliably quantified into seven indicators, which cluster into 2 orthogonal factors: Form, which is comprised of the indicators straight lines, spacing, layout, and handwriting legibility, and Annotation, which is comprised of the indicators final answer markings, defined variables, and text elaboration. While this coding scheme has not yet been applied to other kinds of questions, for example, from topics beyond mechanics, we attempted to create a coding scheme that was general enough to expect that it may be appropriate for a broader range of questions. Certainly, a topic for future work would be to determine the extent to which the coding rubric is more generally applicable.

In terms of assigning productive meaning to the 2 orthogonal factors, we interpret Form as measuring visual clarity and organization and Annotation as a measure of communication of explanatory additions. While it is not clear whether scoring high in Form requires more time for a student, given prior research discussed in Sec. I, it is reasonable to hypothesize that high levels of Form may reduce extraneous cognitive load by making solutions easier to parse for both the student while creating and ruminating over their work during the problem-solving process and for the grader while evaluating it. The potential connection of high Form with improvements in problem-solving is a line of inquiry worth pursuing in future studies. Annotation, on the other hand, appears to capture explanatory additions to the written solution, although its functional role is less clear from the present study. Annotation may either support cognitive processes through clarifying structures of problem-solving when students are unsure of how to proceed, or it may be increasing extraneous load through students adding nonessential information under time constraints during exams. Taken together, these findings suggest SWP can be operationalized as a structured construct rather than treated as an overall impression of neatness. This distinction between Form and Annotation can provide a useful framework for future studies of students' written work for show-work physics problems.

### B. RQ2: To what extent do the indicators and factors correlate with student performance on the same SW question, other SW questions, multiple-choice questions, exam scores, and nonexam graded components?

The SWP factor Form (and all its component indicators) is positively correlated with all course performance outcomes, typically at the level of $r \sim 0.3$, and controlling for ACT score and conscientiousness, Form predicts most outcomes, such as all show-work scores and course grade, explaining 4%–5% of the variance. However, Form is

more weakly correlated with multiple-choice question scores ($r \sim 0.15$) and explains no extra variance when ACT score and conscientiousness are included. This difference in the predictiveness of Form for SW vs MC questions is especially visible in the linear regression models in Table V, which indicates some level of discriminant validity of this SWP factor. More specifically, Form may be measuring something that is more predictive of show-work question scores than multiple-choice scores, which is consistent with an expectation that Form is measuring some characteristic of answering show-work problems that is not used in answering multiple-choice problems. This format-specific pattern provides some support for the (discriminant) validity of our measure Form as a construct specifically related to SW performance rather than a generic achievement measure.

One possible explanation for the link between Form and performance scores is that students who present their work more clearly are engaging in more effective problem-solving strategies. Well-organized and presented solutions may indicate deliberate planning and self-regulation, both of which are linked to higher academic achievement [63,64]. Prior research on expertise in physics problem-solving suggests that novice learners often jump directly into calculations without a structured approach, whereas intermediate and expert learners tend to plan their solutions [65]. The idea that students who make the effort to clearly present their responses by ensuring adequate spacing, maintaining straight lines, and writing more neatly may be engaging in a process of more effective problem-solving provides a hopeful hypothesis for improving student performance in STEM problem-solving.

A related possibility is that producing well-structured work requires students to adopt the perspective of the grader or audience, which prompts them to organize their reasoning clearly and communicate it effectively. This perspective-taking may encourage students to approach problems from multiple angles, fostering divergent thinking, a process associated with creativity and insight [66]. In this way, Form may not only reflect organization and clarity but also cognitive flexibility that enhances problem-solving.

On the other hand, the association between Form and SW score may at least in part be due to the process of grading itself and not a student-level mechanism. Specifically, graders may favor solutions with better Form, regardless of the solution content or accuracy. However, this explanation cannot fully account for our observations, since Form was also associated with nonexam grades and (more weakly) MC scores.

It is also interesting to note, again especially evident in the regression models, that while Form predicts nonexam scores, even controlling for conscientiousness, ACT scores have negligible (or at best very small) predictive power.

In addition, ACT scores were only weakly correlated ($r \sim 0.15$), while Form is stronger ($r \sim 0.3$), which is another indication that Form is measuring something distinct from what ACT scores are measuring.

In contrast, Annotation has a very weak and possibly inconsistent predictive power on course performance. For example, as shown in Table IV, Annotation scores from SW1 were weakly positively correlated with final exam score, but Annotation scores from SW2 (the more psychometrically consistent SW item) were weakly negatively correlated with the final exam score. Rather than interpreting this as annotation being unimportant for problem-solving, we interpret it as evidence that Annotation may capture multiple sets of behaviors. In some cases, it may reflect productive explanation or clarification of reasoning. In other cases, it may reflect time-consuming, extraneous writing that does not improve solution correctness under timed assessment conditions. Future work should distinguish potentially productive annotation, which may clarify essential reasoning, from extraneous annotation, which may increase time and cognitive load without improving (or negatively affecting) performance.

These findings indicate that Form, rather than Annotation, is the more robust, predictive, and potentially helpful SWP factor. This implies that it may be productive to consider presentation quality not (only) as aesthetic but rather as a possible mechanism underlying how students reason, how they communicate their reasoning, and how that reasoning is interpreted by graders in assessment contexts.

### C. RQ3: To what extent is SWP correlated with conscientiousness? If they are correlated, to what extent does SWP mediate self-reported conscientiousness and student performance?

Contrary to our initial expectations, SWP Form and Annotation are both, at best, very weakly correlated with conscientiousness ($r \sim 0.10$), and the 95% confidence interval of the correlation overlaps with zero. Nonetheless, both Form and conscientiousness independently predict student performance in regression and mediation models. This suggests that Form and conscientiousness measure distinct student behaviors that impact course performance. The fact that there appears to be a distinction between these two seemingly similar constructs provides a starting point for future investigation and perhaps more insights into the nature of each construct.

Since SWP does not significantly correlate with conscientiousness, it also does not mediate the relationship between conscientiousness and performance. This opens up the possibility that SWP is a distinct and potentially teachable set of behaviors or skills that may improve performance regardless of a student's baseline (and perhaps less malleable) level of conscientiousness.

### D. RQ4: To what extent is SWP different between male and female students? If there are differences, to what extent does SWP mediate the relation between sex and course performance?

Female students scored higher than male students on Form by about 0.6 standard deviations and scored especially higher—more than one standard deviation—in handwriting legibility. Female students also scored higher than male students on Annotation, though the difference appears to be smaller, around 0.2–0.4 standard deviations.

Mediation models indicated that Form partially mediated the effect of sex on all six course performance outcomes investigated. While there was no overall difference in almost all scores between female and male students, it appears that this lack of difference may be in part due to competing effects: a positive indirect effect of female students scoring relatively higher on Form, which in turn was associated with better course scores, and a weak, sometimes insignificant, negative direct effect of female students scoring lower, especially for MC questions, when controlling for ACT score, conscientiousness, and Form. Put another way, the mediation analysis indicates that, were it not for the positive mediating effect of Form, female students may tend to score lower than males on several of the course performance outcomes. The exception was for nonexam questions, in which the indirect effect of Form coherently added to the direct effect of sex, allowing for a significant overall difference in nonexam scores, with female students scoring higher.

The findings of this study can also help contextualize the observed differences in male and female student performance [26,27], as discussed in Sec. I C. Prior research has shown that female students often underperform on multiple-choice questions but perform similarly to male students on show-work items, potentially due to differences in exam endurance, stereotype threat, or cognitive resource allocation [27,42–45]. In our study, Form partially mediated the relationship between sex and performance outcomes, with female students exhibiting higher Form scores. These results suggest that the clarity and organization of written solutions may be one mechanism driving differences between male and female performance on some introductory physics assessments and assignments. Notably, because MC items are machine-graded, this compensatory pattern cannot be explained by grader bias affecting MC scoring. However, grader bias mechanisms remain plausible for show-work and some TA-graded nonexam components, and the present dataset cannot disentangle grader-mediated pathways from student-level cognitive pathways. We therefore emphasize that the mediation results should be interpreted as evidence of a correlational pattern rather than a causal claim. One possible explanation for the observed sex differences in Form is that, through broader socialization, females may be more likely to be socialized to value aesthetics and self-presentation [67], which may extend to handwriting legibility and the visual organization of their work [68]. Although the present study cannot test this explanation, it provides one idea for future work on why Form differs between male and female students. These findings suggest that Form may be relevant to understanding assessment differences between male and female students even when overall performance outcomes appear similar, and that it may be important to consider presentation quality in future research on fairness and grading in introductory physics.

### E. Practical implications

There are a number of practical implications arising from the products of this study. First, we have developed and characterized a reasonably valid and reliable protocol and rubric for measuring the construct of show-work presentation quality. There are some caveats attached to this claim: the methods can be somewhat time-consuming, and it has only been validated for a limited topic and population in physics. Nonetheless, with proper implementation and confirmation of validity in a specific context, this method could be used in future research on SWP.

The remaining practical implications are derived from the empirical results. Naturally, since the study was correlational, one must be cautious in making claims about any implications. The first implication is that since we have found an association between SWP and student performance, there is now reason to investigate the extent to which this relationship may be causal, as briefly discussed in the next section. Had we found no association, there would have been less support for doing such an experimental study.

In fact, an experimental study investigating SWP as a causal factor could use the results of this study to provide design insights into specific and important features of a SWP intervention aimed at improving SWP. For example, an intervention could facilitate practice and reflection on all four indicators of Form in worked examples, presenting it as a tool for problem-solving and communication. Students could also practice via low-stakes assignments such as homework, quizzes, or recitation activities, which could incorporate feedback on SWP. Further, the distinction between Form and Annotation may help interventions to provide targeted guidance on problem-solving.

Finally, SWP may be relevant to grading fairness. Instructors can consider how SWP is rewarded in grading rubrics and train graders to separate solution clarity from solution correctness to help reduce unintended bias while still valuing clear communication as a stated learning goal. If presentation is valued by instructors as a way to effectively communicate scientific reasoning, those expectations can be clarified to students by introducing aspects of SWP explicitly rather than applying them implicitly during grading.

### F. Limitations and recommendations for future research

There are a variety of limitations to this work. First, the research data were limited to a calculus-based physics course population at a single public R1 university over the span of one semester. As a result, the SWP measurement rubric and empirical findings from this study may not be generalizable, for example, to algebra-based courses or to other institutional settings such as community colleges, liberal arts colleges, or other universities with populations that substantively differ by academic preparation or other relevant experiences. Expanding this study to include data from different institutions, course levels, and student populations would help determine whether these findings can be generalized across various educational populations.

Second, this study focused exclusively on two Newtonian mechanics-based show-work questions, which may limit its applicability to other physics or other STEM topics. Problem-solving strategies and annotation habits may differ depending on the complexity, conceptual demands, and mathematical structure of a given problem. Investigating how students approach different physics topics, such as thermodynamics or electromagnetism, may yield a more comprehensive understanding of how show-work presentation impacts problem-solving across a larger variety of physics topics.

Third, there were only two coders analyzing the show-work solutions, and they were aware of the study being, at least in part, about potential sex-based differences. While the student names were removed, the coders may have inadvertently looked for perceived sex-based differences. For example, prior research by Burr found that handwriting features such as neatness and roundness were often stereotypically associated with femininity, and participants were able to identify a writer's gender at rates significantly higher than chance [48]. Malouff and Thorsteinsson further demonstrated that graders may unintentionally apply lower expectations or interpret responses more critically for female students [30]. This represents a potential threat to the validity of the study procedure, as coders may have inferred the sex of the student based on handwriting style or other visual cues even though the solutions were anonymized. Additionally, despite only having two coders, while steps were taken to ensure reliability, such as looking at Cohen's kappa values, iterating the coding rubric, and observing within-student SWP consistency between the two SW questions studied, having a larger pool of coders would reduce the potential for individual coder biases.

Fourth, while the physics course graders used a rubric to evaluate the show-work responses, there were no records of which graders assessed which student solutions. This could introduce systematic errors due to potential variability in grader bias in grading the show-work final exam questions. Further studies should incorporate mechanisms to track grader identity and intergrader reliability checks to investigate consistency in physics course grading.

Fifth, the conscientiousness survey consisted of only two items. Adding more items to the conscientiousness survey could potentially increase the validity and reliability, allowing for a more precise measure of conscientiousness levels. This may be one reason why we did not find differences in conscientiousness between males and females, which has been shown in prior research, for example, by Henderson *et al.* [23].

Sixth, there may be a bias in the regression estimates due to the fact that regression and mediation analyses were performed only with students who had ACT scores, which represents 190 students out of the 225 students in the full study sample. Table I indicates slightly higher course grades for these students.

These limitations also suggest several directions for future research. As mentioned above, to provide more evidence for SWP as a causal factor in performance, future studies could implement a control-treatment design with an intervention explicitly teaching students how to improve SWP skills associated with Form and examine differences in performance between the conditions. Such studies could also help determine whether Form influences students' problem-solving processes, grader interpretation of written solutions, or both. Further, since our study indicates excessive annotation may be negatively correlated with some student performance outcomes, if further replication studies find this effect to be valid, it would be valuable to examine potential mechanisms responsible for the effect. For example, future work could investigate whether students who engage in Annotation do so because they have a lack of confidence in solving the problem and resort to algorithmic procedures learned from prior education or if it stems from difficulty managing exam time constraints.

## ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation under DUE IUSE Grant No. 2235595.

## DATA AVAILABILITY

The data that support the findings of this article are not publicly available because they contain sensitive personal information. The data are available from the authors upon reasonable request.

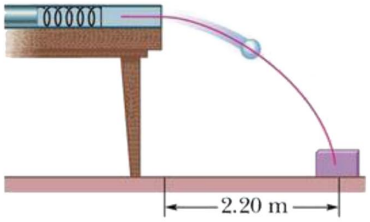
## APPENDIX A: SHOW-WORK QUESTIONS 1 AND 2 (SW1 AND SW2)

Images of the show-work problems 1 and 2 are shown in the exact format presented to students on the final exam.

**UnansweredSW1** 15 pts

Do not upload your SW solutions in Final A. Submit your MC answers before the due time. Then upload and submit your SW solutions in Final B.

Two children are playing a game in which they try to hit a small box on the floor with a marble fired from a spring-loaded gun that is mounted on a table. The target box is 2.2 m horizontally from the edge of the table; see the figure below. Bobby compresses the spring 1.1 cm, but the center of the marble falls 27.0 cm short of the center of the box. How far should Rhoda compress the spring for the marble to hit the center of the box? Show your work. (Ignore all frictional forces and use g=10 m/s$^2$.)



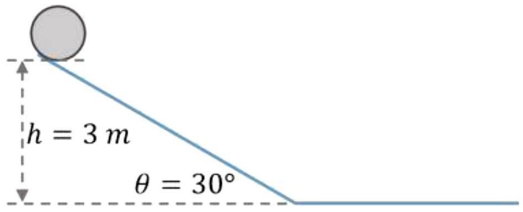

**UnansweredSW2** 15 pts

Do not upload your SW solutions in Final A. Submit your MC answers before the due time. Then upload and submit your SW solutions in Final B.

A uniform cylinder with R=0.2 m and a mass of 2 kg is initially at rest on a ramp at a height of 3 meters above the ground. It then rolls down the ramp without slipping. (Use g=10 m/s$^2$, I=(1/2) MR$^2$)

(a) (10 points) Find the linear acceleration of the cylinder as it rolls down the ramp.

(b) (5 points) Find the rotational kinetic energy of the cylinder when it reaches the bottom of the ramp.



## APPENDIX B: COHEN'S KAPPA VALUES FOR 11 INDICATORS

TABLE IX. Cohen's kappa values to measure interrater reliability for 11 indicators. Crossed-out work, smudging, diagrams, and photographic care were excluded from later analyses due to exploratory factor analysis findings. Holistic neatness was not included in this table because it was removed early on from the analysis.

| Indicator | SW1 | SW2 | Included in analysis |
|---|---|---|---|
| Straight lines | 0.93 | 0.91 | Yes |
| Spacing | 0.91 | 0.82 | Yes |
| Layout | 0.92 | 0.83 | Yes |
| Handwriting legibility | 0.90 | 0.93 | Yes |
| Final answer markings | 0.98 | 0.99 | Yes |
| Defined variables | 0.98 | 0.99 | Yes |
| Text elaboration | 0.96 | 0.99 | Yes |
| Crossed-out work | 0.96 | 0.97 | No |
| Smudging | 0.94 | 0.96 | No |
| Diagrams | 0.97 | 0.93 | No |
| Photographic care | 0.95 | 0.89 | No |
| Overall | 0.94 | 0.92 | ⋯ |

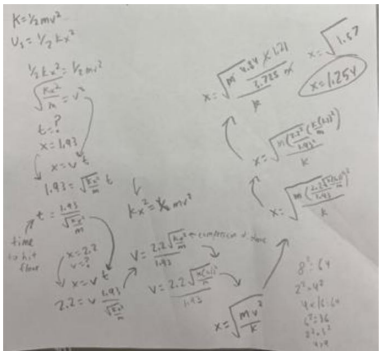
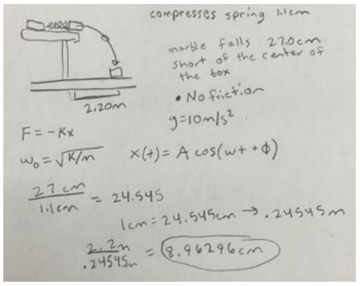
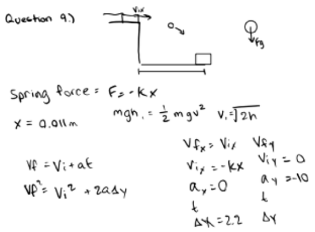
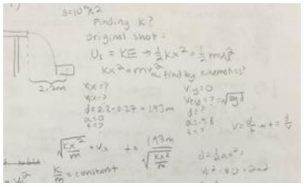
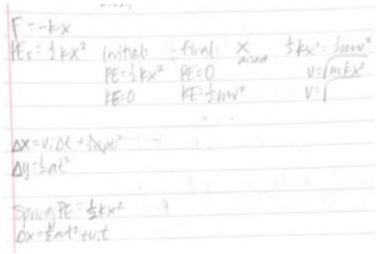
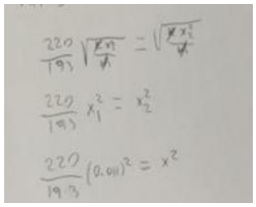
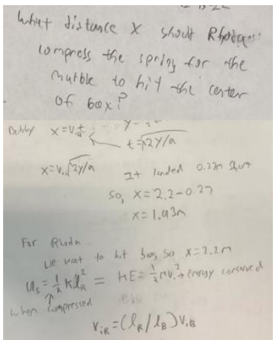
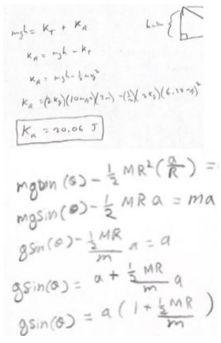
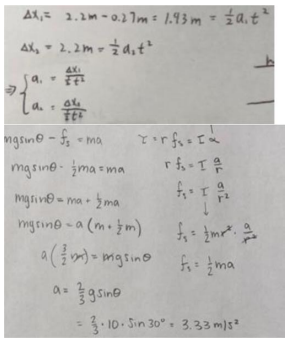
## APPENDIX C: SPEARMAN CORRELATION TABLE FOR QUESTION 2 INDIVIDUAL INDICATORS

TABLE X. Spearman correlation table for all seven individual indicators. Bolded values denote unadjusted $p < 0.05$.

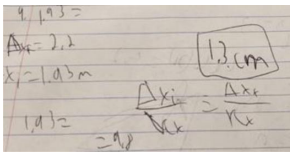
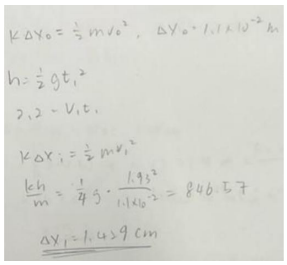
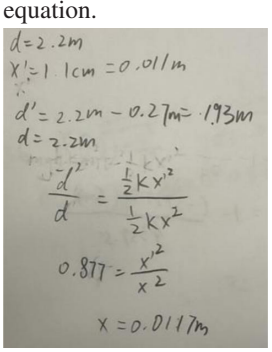
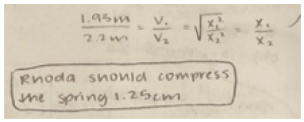
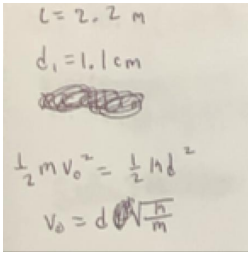
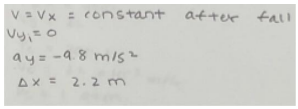
| | Handwriting legibility | Spacing | Layout | Straight lines | Final answer markings | Defined variables | Text elaboration |
|---|---|---|---|---|---|---|---|
| Spacing | **0.34** | | | | | | |
| Layout | **0.29** | **0.5** | | | | | |
| Straight lines | **0.31** | **0.24** | **0.2** | | | | |
| Fin. ans. markings | 0.05 | −0.08 | 0.01 | 0.11 | | | |
| Defined variables | −0.05 | **−0.16** | **−0.17** | 0.1 | 0.05 | | |
| Text elaboration | 0.1 | **−0.15** | −0.02 | 0.09 | **0.24** | **0.2** | |
| Conscientiousness | 0.12 | 0.01 | 0.03 | 0.13 | 0.13 | **0.15** | 0.05 |
| Sex | **0.49** | 0.07 | **0.15** | 0.08 | 0.04 | 0.1 | **0.23** |
| Course grade | **0.2** | **0.22** | **0.25** | **0.29** | 0 | **−0.19** | −0.04 |
| Avg final exam | **0.16** | **0.16** | **0.24** | **0.23** | −0.03 | **−0.22** | −0.1 |
| Avg MC | 0.05 | 0.12 | **0.14** | **0.13** | −0.12 | **−0.18** | −0.11 |
| Avg SW | **0.27** | **0.18** | **0.27** | **0.26** | 0.07 | **−0.15** | −0.05 |
| Nonexam comp. | **0.3** | **0.21** | **0.2** | **0.22** | 0.08 | −0.01 | 0.06 |
| SW1 | −0.04 | 0.02 | 0.08 | 0.06 | −0.06 | −0.12 | −0.02 |
| SW2 | **0.25** | 0.06 | **0.28** | **0.25** | **0.15** | −0.13 | 0 |
| SW1 Form | **0.58** | **0.45** | **0.43** | **0.33** | 0.01 | −0.04 | 0.04 |
| SW1 Annotation | 0.03 | −0.04 | −0.01 | 0.1 | **0.2** | **0.19** | **0.39** |
| SW2 Form | **0.67** | **0.75** | **0.7** | **0.6** | 0.04 | −0.09 | 0.02 |
| SW2 Annotation | 0.03 | **−0.22** | −0.1 | **0.14** | **0.5** | **0.7** | **0.74** |
| ACT math | 0.11 | 0.07 | 0.11 | **0.18** | −0.11 | −0.13 | −0.06 |

## APPENDIX D: SHOW-WORK PRESENTATION RUBRIC

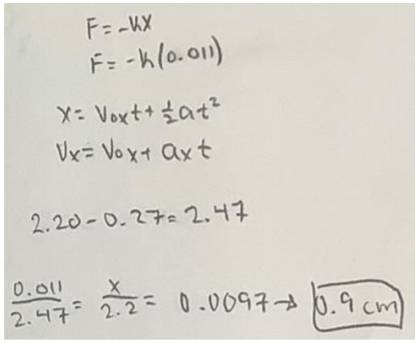
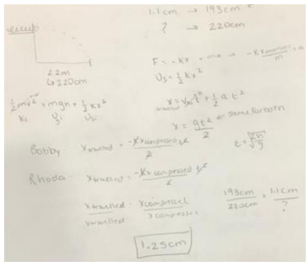
Examples of student work are shown in the table for indicators.

| Indicator name | 0 | 1 | 2 |
|---|---|---|---|
| Straight lines of writing (horizontally) | Lines are not straight. Lines are slanted at multiple angles.  | Lines are somewhat straight. They are all slanted at around the same angle starting from the top of the page.  | Lines are very straight. No slanted lines starting from top of the page. Lines do not get increasingly or decreasingly slanted throughout solution.  |
| Spacing (vertical + horizontal) | Spacing is very inconsistent; some work is difficult to read due to proximity. Work bleeds into lines above or below 3 or more times. Multiple equations are horizontally placed next to each other where they could look like they relate to the same equation.  | There is not enough space between lines to read the work easily, or spacing may be inconsistent for a couple of lines. Work bleeds into lines above or below 1 or 2 times.  | There's enough space between lines to read clearly and spacing between lines is consistent. There is no overlap between lines or work being squished due to proximity to other lines.  |
| Handwriting legibility | Difficult to read either due to size or style of handwriting. Letters may be squished on top of each other.  | Handwriting is somewhat legible. Certain parts of the solution may be less legible than others.  | Handwriting is legible and easy to read.  |

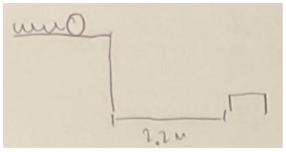
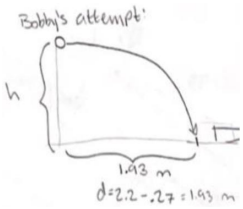
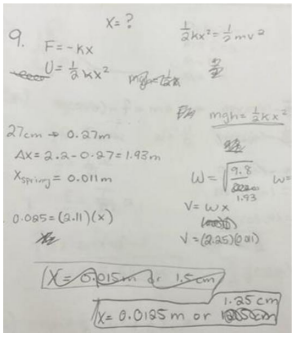
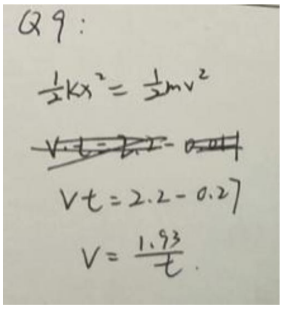
*(Table continued)*

| Indicator name | 0 | 1 | 2 |
|---|---|---|---|
| Layout | No clear progression with items written out of order (either later work written above or to the left of previous work). Work is not written top to bottom, left to right, or in multiple columns going top to bottom.  | Some logical progression to lead the grader, but there are logical jumps between different lines or areas of work.  | Clear logical progression to help lead the grader using spatial organization such as straight lines of work or arrows and symbols. Previous work is used in the next line of work unless there is an arrow connecting disparate areas of work.  |
| Final answer markings *mark "X" for missing data if student does not arrive at a final answer | Not marked at all and/or in the middle of work without any symbols identifying it. | At the end of the solution, written either furthest right or on its own line, but without circling, boxing, or using symbols. Only the final number is required, not necessarily a left hand and right-hand side of equation.  | Circled, starred, or boxed symbol defining answer. Answer can be written in sentence format.  |
| Defined variables *constants, such as g, do not count as variables | No variables are defined. | One to two variables are defined and can be interspersed with work.  | Three or more, or all, variables are defined (including in diagram) using an equal sign and either a number or text in its own area on the page.  |

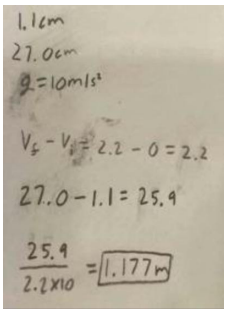
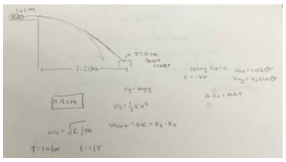
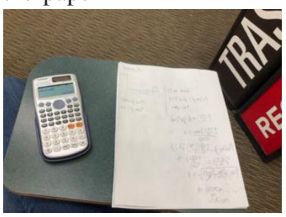
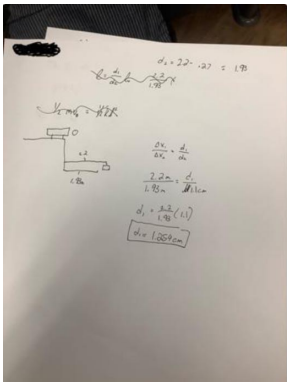
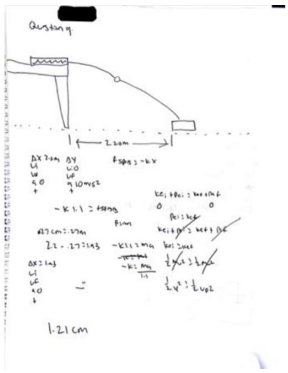
*(Table continued)*

| Indicator name | 0 | 1 | 2 |
|---|---|---|---|
| Text elaboration(not including the diagram text) | No text elaboration provided   | One to two words or phrases explaining their logical progression of work. Answer in sentence or phrase format can count as text elaboration.  | Three or more words or phrases explaining the logical progression of their work or defining variables outside of subscripts. Answer in sentence or phrase format can count as text elaboration.  |

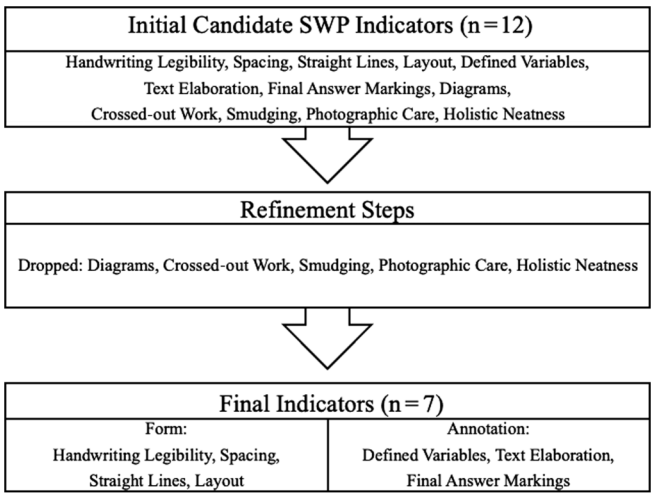
Below are indicators that were not used in the final rubric.

| Indicator name | 0 | 1 | 2 |
|---|---|---|---|
| Diagram | No diagrams drawn or only extraneous drawings not related to the problem. | Diagram drawn but 0 or 1 numbers or variables labeled. Some extraneous details or partially correct drawn diagram (i.e., normal force pointing incorrect way).   | Relevant diagram pertinent to problem-solving (i.e., no extraneous details) with two or more variables or numbers or numbers labeled. A free body diagram can be added if necessary.   |
| Crossed-out work | Significant amount of crossed-out work that makes it difficult to read the overall solution.   | Some lines of work are crossed out, but the overall solution is legible.   | No crossed-out work. |

*(Table continued)*

| Indicator name | 0 | 1 | 2 |
|---|---|---|---|
| Smudging (only for nondigital answers) | Three or more smudges or poorly erased areas, so it is difficult to read. You can easily read the "erased" work for multiple lines underneath newly written work on top.   | Two to three dark smudges from erasing work, or you can read the work underneath new work for one line at most.  | No smudges or one small one. Work is generally erased cleanly. |
| Photographic care (only for nondigital answers) | Extraneous details in the background (two or more colors of different items in background including a separate page in background with other problems' solutions); the background is larger than just a small border surrounding the solution; picture is not cropped to page size, very small, too dark or bright, out of focus, shadows on part of page, transparency; you can see through the page into other work or solutions behind the paper  | A small border around the solution that is one color; contrast is off but is still readable; paper is in focus. There may be a shadow on the page that obscures some work.  | Page fit to size (small amount of background past border); large enough to be legible; contrast is well adjusted. Multiple solutions of problems on the same page are fine.  |
| Holistic neatness (qualitative and intuitive judgment) | Not neat; difficult to grade | Somewhat neat; somewhat easy to grade | Very neat; easy to grade |

## APPENDIX E: SHOW-WORK PRESENTATION INDICATOR FLOW CHART




Flow chart depicting how SWP indicators became grouped and which were dropped.


[1] J. H. Larkin and F. Reif, Understanding and teaching problem-solving in physics, Eur. J. Sci. Educ. **1**, 191 (1979).

[2] J. P. Mestre, J. L. Docktor, N. E. Strand, and B. H. Ross, Chapter Nine—Conceptual problem solving in physics, in *Psychology of Learning and Motivation*, edited by J. P. Mestre and B. H. Ross (Academic Press, New York, 2011), Vol. 55, p. 298.

[3] J. L. Docktor, J. Dornfeld, E. Frodermann, K. Heller, L. Hsu, K. A. Jackson, A. Mason, Q. X. Ryan, and J. Yang, Assessing student written problem solutions: A problem-solving rubric with application to introductory physics, Phys. Rev. Phys. Educ. Res. **12**, 010130 (2016).

[4] F. Reif and J. I. Heller, Knowledge structure and problem solving in physics, Educ. Psychol. **17**, 102 (1982).

[5] L. Hsu, E. Brewe, T. M. Foster, and K. A. Harper, Resource letter RPS-1: Research in problem solving, Am. J. Phys. **72**, 1147 (2004).

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

[7] P. Heller, R. Keith, and S. Anderson, Teaching problem solving through cooperative grouping. Part 1: Group versus individual problem solving, Am. J. Phys. **60**, 627 (1992).

[8] E. Burkholder, L. Blackmon, and C. Wieman, Characterizing the mathematical problem-solving strategies of transitioning novice physics students, Phys. Rev. Phys. Educ. Res. **16**, 020134 (2020).

[9] B. Ibrahim, L. Ding, A. F. Heckler, D. R. White, and R. Badeau, How students process equations in solving quantitative synthesis problems? Role of mathematical complexity in students' mathematical performance, Phys. Rev. Phys. Educ. Res. **13**, 020120 (2017).

[10] M. De Cock, Representation use and strategy choice in physics problem solving, Phys. Rev. ST Phys. Educ. Res. **8**, 020117 (2012).

[11] C. O. Sweedler-Brown, Effect of training on the appearance bias of holistic essay graders, J. Res. Dev. Educ. **26**, 24 (1992).

[12] V. Connelly, J. E. Dockrell, and J. Barnett, The slow handwriting of undergraduate students constrains overall performance in exam essays, Educ. Psychol. **25**, 99 (2005).

[13] J. Stewart and S. Ballard, Effect of written presentation on performance in introductory physics, Phys. Rev. ST Phys. Educ. Res. **6**, 020120 (2010).

[14] S. Graham, C. A. MacArthur, and J. Fitzgerald, *Best Practices in Writing Instruction* (Guilford Publications, New York, USA, 2025).

[15] R. E. Mayer, Rote versus meaningful learning, Theory Pract. **41**, 226 (2002).

[16] J. Sweller, Cognitive load during problem solving: Effects on learning, Cogn. Sci. **12**, 257 (1988).

[17] B. J. Zimmerman, Becoming a self-regulated learner: An overview, Theory Pract. **41**, 64 (2002).

[18] O. P. John and S. Srivastava, The big five trait taxonomy: History, measurement, and theoretical perspectives, in *Handbook of Personality: Theory and Research*, edited by L. A. Pervin and O. P. John (Guilford Press, New York, 1999), p. 138.

[19] S. Mammadov, Big five personality traits and academic performance: A meta-analysis, J. Pers. **90**, 222 (2022).

[20] A. E. Poropat, A meta-analysis of the five-factor model of personality and academic performance, Psychol. Bull. **135**, 322 (2009).
[21] N. Conrad and M. Patry, Conscientiousness and academic performance: A mediational analysis, Int. J. Scholarship Teach. Learn. **6** (2012).
[22] B. W. Roberts, C. Lejuez, R. F. Krueger, J. M. Richards, and P. L. Hill, What is conscientiousness and how can it be assessed?, Dev. Psychol. **50**, 1315 (2014).
[23] R. Henderson, D. Hewagallage, J. Follmer, L. Michaluk, J. Deshler, E. Fuller, and J. Stewart, Mediating role of personality in the relation of gender to self-efficacy in physics and mathematics, Phys. Rev. Phys. Educ. Res. **18**, 010143 (2022).
[24] V. W. Berninger, K. H. Nielsen, R. D. Abbott, E. Wijsman, and W. Raskind, Gender differences in severity of writing and reading disabilities, J. Sch. Psychol. **46**, 151 (2008).
[25] K. P. Feder and A. Majnemer, Handwriting development, competency, and intervention, Dev. Med. Child Neurol. **49**, 312 (2007).
[26] A. B. Simmons and A. F. Heckler, Grades, grade component weighting, and demographic disparities in introductory physics, Phys. Rev. Phys. Educ. Res. **16**, 020125 (2020).
[27] S. Salehi, S. Cotner, S. M. Azarin, E. E. Carlson, M. Driessen, V. E. Ferry, W. Harcombe, S. McGaugh, D. Wassenberg, A. Yonas, and C. J. Ballen, Gender performance gaps across different assessment methods and the underlying mechanisms: The case of incoming preparation and test anxiety, Front. Educ. **4**, 107 (2019).
[28] N. van Drempt, A. McCluskey, and N. A. Lannin, A review of factors that influence adult handwriting performance, Aust. Occup. Ther. J. **58**, 321 (2011).
[29] S. Graham, V. Berninger, N. Weintraub, and W. Schafer, Development of handwriting speed and legibility in grades 1-9, J. Educ. Res. **92**, 42 (1998).
[30] J. M. Malouff and E. B. Thorsteinsson, Bias in grading: A meta-analysis of experimental research findings, Aust. J. Educ. **60**, 245 (2016).
[31] R. Greifeneder, S. Zelt, T. Seele, K. Bottenberg, and A. Alt, Towards a better understanding of the legibility bias in performance assessments: The case of gender-based inferences, Br. J. Educ. Psychol. **82**, 361 (2012).
[32] F. G. W. C. Paas and J. J. G. Van Merriënboer, Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach, J. Educ. Psychol. **86**, 122 (1994).
[33] J. Hattie, *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement* (Routledge, London, 2008).
[34] E. W. Burkholder, J. K. Miles, T. J. Layden, K. D. Wang, A. V. Fritz, and C. E. Wieman, Template for teaching and assessment of problem solving in introductory physics, Phys. Rev. Phys. Educ. Res. **16**, 010123 (2020).
[35] T. Santangelo and S. Graham, A comprehensive meta-analysis of handwriting instruction, Educ. Psychol. Rev. **28**, 225 (2016).
[36] S. L. Hammerschmidt and P. Sudsawad, Teachers' survey on problems with handwriting: Referral, evaluation, and outcomes, Am. J. Occup. Ther. **58**, 185 (2004).
[37] Z. Ivcevic and M. Brackett, Predicting school success: Comparing conscientiousness, grit, and emotion regulation ability, J. Res. Pers. **52**, 29 (2014).
[38] B. Dumfart and A. C. Neubauer, Conscientiousness is the most powerful noncognitive predictor of school achievement in adolescents, J. Individ. Differ. **37**, 8 (2016).
[39] A. L. Duckworth and M. E. P. Seligman, Self-discipline outdoes IQ in predicting academic performance of adolescents, Psychol. Sci. **16**, 939 (2005).
[40] A. Zusho, P. R. Pintrich, and B. Coppola, Skill and will: The role of motivation and cognition in the learning of college chemistry, Int. J. Sci. Educ. **25**, 1081 (2003).
[41] K. G. Rice, F. G. Lopez, and C. M. E. Richardson, Perfectionism and performance among STEM students, J. Vocat. Behav. **82**, 124 (2013).
[42] C. J. Ballen, S. Salehi, and S. Cotner, Exams disadvantage women in introductory biology, PLoS One **12**, e0186419 (2017).
[43] P. Balart and M. Oosterveen, Females show more sustained performance during test-taking than males, Nat. Commun. **10**, 3798 (2019).
[44] S. M. Lindberg, J. S. Hyde, J. L. Petersen, and M. C. Linn, New trends in gender and mathematics performance: A meta-analysis, Psychol. Bull. **136**, 1123 (2010).
[45] K. F. Stanger-Hall and E. H. Chudler, Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes, Life Sci. Educ. **11**, 294 (2012).
[46] C. Hill, C. Corbett, and A. St. Rose, *Why So Few? Women in Science, Technology, Engineering, and Mathematics* (American Association of University Women, 2010).
[47] C. A. Moss-Racusin, J. F. Dovidio, V. L. Brescoll, M. J. Graham, and J. Handelsman, Science faculty's subtle gender biases favor male students, Proc. Natl. Acad. Sci. **109**, 16474 (2012).
[48] V. Burr, Judging gender from samples of adult handwriting: Accuracy and use of cues, J. Soc. Psychol. **142**, 691 (2002).
[49] S. I. Hofer, Studying gender bias in physics grading: the role of teaching experience and country, Int. J. Sci. Educ. **37**, 2879 (2015).
[50] S. D. Schneid, C. Armour, and K. Brandl, Beyond right or wrong: How partial credit scoring on multiple-choice questions improves student performance and assessment perceptions, Br. J. Clin. Pharmacol. **92**, 2542 (2026).
[51] G. W. Procop, T. McCarthy, A. Schlinsog, and M. Ghofrani, A comparison of free-response and multiple-choice questions on the American board of pathology primary certification examinations, Acad. Pathol. **13**, 100248 (2026).
[52] H. Mei, A. Heckler, and Q. Liu, Evaluating conscientiousness in student responses to show-work exam questions (2013), Accessed: Aug. 12, 2025, https://osf.io/ven87.
[53] J. Cohen, A power primer, Psychol. Bull. **112**, 155 (1992).
[54] S. D. Gosling, P. J. Rentfrow, and W. B. Swann, A very brief measure of the big-five personality domains, J. Res. Pers. **37**, 504 (2003).
[55] R. Eisinga, M. te Grotenhuis, and B. Pelzer, The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown?, Int. J. Pub. Health **58**, 637 (2013).
[56] B. G. Glaser, A. L. Strauss, and E. Strutzel, The discovery of grounded theory; strategies for qualitative research, Nurs. Res. **17**, 364 (1968).

[57] K. Charmaz, Constructing grounded theory, in *Introducing Qualitative Methods*, 2nd ed. (SAGE Publications, London, 2014).

[58] R. E. Boyatzis, *Transforming Qualitative Information: Thematic Analysis and Code Development* (Sage Publications, Thousand Oaks, CA, 1998).

[59] K. Sijtsma, J. L. Ellis, and D. Borsboom, Recognize the value of the sum score, psychometrics' greatest accomplishment, Psychometrika **89**, 84 (2024).

[60] K. F. Widaman and W. Revelle, Thinking thrice about sum scores, and then some more about measurement and analysis, Behav. Res. Methods **55**, 788 (2023).

[61] T. Lange, K. W. Hansen, R. Sørensen, and S. Galatius, Applied mediation analyses: A review and tutorial, Epidemiol. Health **39**, e2017035 (2017).

[62] Y. Rosseel, Lavaan: An R package for structural Equation modeling, J. Stat. Softw. **48**, 1 (2012).

[63] P. R. Pintrich and E. V. D. Groot, Motivational and self-regulated learning components of classroom academic performance, J. Educ. Psychol. **82**, 33 (1990).

[64] B. J. Zimmerman, Motivational sources and outcomes of self-regulated learning and performance, in *Handbook of Self-Regulation of Learning and Performance*, Educational Psychology Handbook Series (Routledge/Taylor & Francis Group, New York, NY, US, 2011), pp. 49–64.

[65] J. Larkin, J. McDermott, D. P. Simon, and H. A. Simon, Expert and novice performance in solving physics problems, Science (New York, N.Y.) **208**, 1335 (1980).

[66] M. A. Runco, Divergent thinking, creativity, and ideation, in *The Cambridge Handbook of Creativity*, Cambridge Handbooks in Psychology, edited by J. C. Kaufman and R. J. Sternberg (Cambridge University Press, Cambridge, England, 2010), p. 446

[67] M. Merino, J. F. Tornero-Aguilera, A. Rubio-Zarapuz, C. V. Villanueva-Tobaldo, A. Martín-Rodríguez, and V. J. Clemente-Suárez, Body perceptions and psychological well-being: A review of the impact of social media and physical measurements on self-esteem and mental health with a focus on body image satisfaction and its relationship with cultural and gender factors, Healthcare (Basel) **12**, 1396 (2024).

[68] C. Leaper and C. K. Friedman, The socialization of gender, in *Handbook of Socialization: Theory and Research* (The Guilford Press, New York, NY, US, 2007), pp. 561–587.