

## Effectiveness of guided group work in graduate level quantum mechanics

C. D. Porter<sup>1</sup> and A. F. Heckler<sup>1</sup>

*Department of Physics, The Ohio State University, 191 West Woodruff Avenue,  
Columbus, Ohio 43210, USA*



(Received 22 May 2020; accepted 3 September 2020; published 23 October 2020)

We investigate the effects of guided group work sessions on graduate student performance on a quantum mechanics assessment. Data from a single large Midwestern university were taken over a five-year period, during which guided group work sessions were offered to accompany the graduate-level quantum mechanics course. Students were pre- and post-tested using a set of mostly conceptual items that we call the graduate quantum mechanics assessment. The reliability and validity of this assessment are addressed. A mixed linear model is used to analyze the dependence of post-test scores on factors such as group work attendance, pretest scores, GRE Physics scores, and others. We find a statistically significant effect of group work attendance on post-pre gains, specifically that attendance of one 60-min group work session improves performance on a related post-test item by 6.4%, administered 2–10 weeks after the session. We discuss the lack of a randomized control group and address possible confounding effects such as student self-selection, and attitudinal and motivational factors. Overall, the results of this study indicate that guided group work sessions at the graduate level can be feasible and effective. We note preliminary observations of differences in group interactions and classroom logistics compared to group work at the undergraduate level.

DOI: [10.1103/PhysRevPhysEducRes.16.020127](https://doi.org/10.1103/PhysRevPhysEducRes.16.020127)

### I. INTRODUCTION

According to a 2008 report by the Council of Graduate Schools [1] the 10-year completion rate for students enrolling in a U.S. physics Ph.D. program is 55%. The rate is lower still (37%) for African American students. The lack of gender, racial, and ethnic diversity in physics graduate programs across the country is well known (see, for example, Ref. [2]), with women still earning around 20% of physics Ph.D. degrees.

At the Ohio State University (OSU), a large, public research university in the United States, between the years 2000 and 2010, 25% of incoming physics Ph.D. students left without a Ph.D. and 15% left with no degree at all. Further, 24% of those physics graduate students who leave without a Ph.D. do so in the first year, and another 17% do so in the second year. Graduate core coursework is typically completed in the first year to 18 months of physics graduate school. Since graduate students' time is dominated by core coursework in the first year, it is likely that experiences and outcomes from the core courses influence students' decision or eligibility to remain in a program. Even for those who remain beyond the second year, the core courses are the students' first exposure to graduate school, and may set

the tone for the students' experiences in the department and in the field of physics. At OSU, core course GPAs are used in lieu of a qualifying exam, and thus have great significance for students' progression toward a Ph.D. At OSU, physics graduate students with GPAs below 3.3 after their first attempt at core courses (they may be repeated) are more likely to leave with no Ph.D. than students with GPAs above 3.3 by a factor of 4.3 (95% CI: 2.0 to 8.8). Nationally, core courses may or may not be the principle cause for the loss of 45% of beginning physics Ph.D. students, but it is reasonable to suppose that core courses significantly contribute to this loss. It is therefore compelling to study and improve our physics graduate core courses in order to address issues of student retention and to improve the student experience. Of course, another reason to improve physics graduate core courses is to improve student learning.

In this study we are specifically focused on the implementation and assessment of guided group work sessions (sometimes similar to tutorials). Broadly speaking, the learning benefits of group work, tutorials, and other active learning techniques have been well studied, and have been firmly established for decades (see, for example, Refs. [3,4]). Best practices in the design of tutorials have similarly been long established (see, for example, Refs. [5,6]). Well-designed tutorials have resulted in consistent gains in conceptual understanding and reasoning in many areas of physics, including kinematics [7], mechanics and electromagnetism [8], relativity [9], quantum mechanics [10–12], and even in materials science [13]. An early

---

*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.*

meta-analysis of studies done on introductory physics courses by Hake [14] found that students in courses with group work and active learning components had higher gains than those in traditional lecture-style courses by  $0.48 \pm 0.14$  standard deviations. A more recent meta-analysis by Freeman *et al.* [15] established similar findings for a broader variety of STEM courses, and importantly at levels beyond introductory or freshman courses. These analyses have not, however, been done at the graduate level.

It is well established that undergraduate students have persistent difficulties with learning physics concepts and skills that are not adequately addressed through traditional instruction (e.g., for a review see Ref. [3]). While the concepts and skills may be at a more advanced level, such observations of difficulties are emerging at the graduate level as well. For example, there are a handful of studies that demonstrate significant student misunderstandings or difficulties in graduate-level quantum mechanics. Singh found in 2008 that only 43% of incoming graduate students could correctly write down the time dependence of a superposition of energy eigenstates in a square well, and only 57% could correctly sketch the ground state wave function in a square well [10]. This is consistent with later findings by Porter and Heckler [16] who found that the fraction of graduate students who correctly drew the ground state wave function in an asymmetric 1D well increased from 44% prior to instruction to 68% after their first-year quantum mechanics course. In that same study, the number who drew an excited state correctly was significantly lower: 7% on the pretest and 5% on the post-test. Similar studies exist for graduate student understanding of quantum mechanical spin [17,18] and hydrogen [19].

There have been a few pioneering attempts to implement active learning techniques at the graduate level in physics. Notably, Quantum Interactive Learning Tutorials (QuILTs) were originally developed for advanced undergraduates, but also allowed researchers to study the effects of scaffolded, computer-based tutorials on graduate student learning. These were shown to be very effective interventions for both populations in content areas such as the double slit experiment [20] and degenerate perturbation theory [21]. This suggests that some of the same active learning techniques, and even the same material, that is beneficial to advanced undergraduate physics majors may also benefit graduate students. However, none of these early studies has systematically examined the feasibility and effectiveness of implementing a collection of weekly guided group work sessions implemented throughout the term(s) of a course at the graduate level.

Graduate-level guided group work (GGW) sessions were implemented at OSU in an attempt to offer weekly and semester-long supplemental support to graduate students, a need that is augmented when students come from a wide variety of backgrounds and undergraduate institutions. Over the course of five years, as these sessions were

introduced, data were collected on student performance, and we report here on the effectiveness of this instructional method in graduate quantum mechanics. Materials for these sessions are available upon request from the corresponding author.

One of the many reasons for the limited number of studies on graduate populations is that, in contrast to introductory subjects, there are no validated, widely adopted conceptual assessments for graduate quantum mechanics. There are several early assessments that may yet become widely adopted as interest grows [10,22]. In this study, we have systematically and iteratively developed and implemented our own sets of assessment items, which we refer to collectively as the graduate quantum mechanics assessment (GQMA). The validity and reliability of this set of items are discussed in Sec. II.

In this study, we explore the effectiveness of group work at the graduate level by asking two principal research questions: (i) To what extent is it feasible? (ii) To what extent does it positively influence student learning of key concepts (here in quantum mechanics)? The first question addresses the logistics including development of engaging material, matching the content highlighted by different faculty in different years, and getting graduate students to participate and to see the benefit of attendance. Answering the second question is done with evidence from pre- and post-testing, but is made more difficult by the absence of a randomized control group, for reasons discussed in Sec. II. Comparisons will be made between students who attended GGW sessions and those who did not, which introduces the possibility of self-selection or other biases that could influence the results. These potential confounding factors are discussed throughout the remaining sections, and are partially addressed by using mixed linear models to control for student factors such as pretest scores, GRE scores, and the pattern of attendance in relevant and non-relevant tutorial sessions.

## II. METHODS AND IMPLEMENTATION

### A. Participation and data collection

This study uses 4 years of student data (after an initial year of observation and development) collected from graduate students enrolled in graduate-level quantum mechanics at OSU. All students were invited to attend voluntary GGW sessions, and an average of 30% of enrolled students attended each week (approximately 10 students). All (140) students were invited to participate in conceptual pre-post testing; 133 (95%) did so. Invitations to the GGW sessions took the form of an advertisement on the first day of lecture, and weekly reminder emails describing what material would be discussed in the coming GGW session. For the first session each semester, food was provided as an extra incentive to participate and learn from first-hand experience more about the potential benefits of

continued attendance of the GGW sessions. The GGW instructor was not the core course instructor, but rather a physics education research (PER) postdoc or an advanced graduate student TA. Additionally, if students performed poorly on an early assessment such as a first midterm, or in-class quizzes, the students were encouraged by instructors or academic advisors to take advantage of the GGW sessions. Students with weaker backgrounds in quantum mechanics were often encouraged from the beginning to attend the GGW sessions. Although struggling students could not be required to attend, these advisory interventions represent a possible sample bias, namely, that weaker students were encouraged more to attend than stronger students. On the other hand, there is a complementary possible source of sample bias in the form of student enthusiasm and interest; students passionate about quantum mechanics may be more likely to attend.

During the study period, instructors of the graduate quantum mechanics course offered credit equivalent to that of one homework assignment (between 1% and 3% of the total points in the course) for taking the GQMA pretest and post-test. Students could complete the assessment and opt out of participation in research with no penalty. Overall, 95% of enrolled students agreed to participate in research.

### B. Guided group work

Over the past 5 years, we have been iteratively developing GGW sessions to accompany graduate core courses in physics: Quantum Mechanics (QM), Electricity and Magnetism, Classical Mechanics, and Statistical

Mechanics. Hereafter, we will focus on group work sessions in Quantum Mechanics.

There are many types of group work and peer-led learning (for a useful discussion of these, see Ref. [23]). The type used here is closest to what Hodges calls “Peer-led team learning (PLTL)”, in that students self-select into groups, and it is external to the regularly scheduled lecture. Our GGW sessions are designed to take place once per week for each core course, and each session lasts 1 h. The group work itself consists of questions that range from conceptual to calculational. They are designed to be relevant to a given week’s homework and lecture content, which is especially necessary since attendance of these group work sessions is optional in the current format. In-session observations and student feedback were used to iterate the group work questions. This included dropping questions that failed to generate good discussion, and clarifying early versions of questions. In some cases, the resulting sessions might rightly be called “tutorials” in the sense that they are single topic and heavily scaffolded. In other cases, an instructor might cover a variety of topics in a given week, and the corresponding group work would be better described as simply “guided group work.” Examples of topics treated in quantum mechanics GGW sessions and related example questions are shown in Table I.

In the five years of development, four different instructors have taught the course, and they have used three different textbooks. Because topics have been taught slightly differently and sometimes in different orders, a corpus of questions has been developed that should accommodate most choices of book and topic ordering. Specific

TABLE I. Some topics emphasized in guided group work sessions and an example question for each topic. This is not an exhaustive list of topics or questions related to these topics.

Topic	Example group work item
Wave functions	In the square well below <finite potential well shown>, qualitatively sketch: (a) The ground state ( $n = 1$ ), (b) the 2nd excited state ( $n = 3$ ), and (c) the 5th excited state ( $n = 6$ ), assuming all these states exist. Compare with your neighbor and resolve any differences.
Math or linear algebra	You are working on a quantum mechanics problem with a friend, and the problem involves an operator $\hat{\Omega}$ . You are very pleased with your choice of basis, in which the matrix corresponding to $\hat{\Omega}$ is diagonal: $\hat{\Omega} = c \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 4 \end{pmatrix}.$ (a) Find a way of representing your basis states $ 1a\rangle,  1b\rangle,  2\rangle,  4\rangle$ . (b) Your friend insists he has used a different basis than you, but he also has a diagonal matrix. How is this possible? Convince your skeptical friend.
Expectation values	Without doing a direct calculation, explain to your partner which values of $n$ yield nonzero results in the following expressions in the context of a quantum harmonic oscillator: (a) $\langle n   \hat{x}^2   0 \rangle$ , (b) $\langle 3   \hat{x}^3   n \rangle$ , (c) $\langle n   \hat{p}   0 \rangle$ .
Operators vs eigenvalues	When can we make the replacement $e^{-i\hat{H}t/\hbar} \rightarrow e^{-iEt/\hbar}$ ? Explain this to your neighbor.
Spin	Explain how you can experimentally produce a spin state $ \Psi\rangle = \frac{1}{\sqrt{3}} \begin{pmatrix} 1+i \\ 5 \end{pmatrix}$ .

sets of questions are chosen from this larger body each week to match the pace or choices of the core course instructor. Although the GGW instructor is usually kept informed about the content of the week's lectures and the corresponding homework, they are not privy to any information about what will be covered on tests or quizzes. Attendance of these GGW sessions has varied from year to year, ranging from as little as 20% of students enrolled in the course to as much as 60%. Based on anecdotal student feedback, this appears to depend heavily on the perceived difficulty of the core course. The fact that attendance of these GGW sessions is optional means that students are not randomly assigned, and no randomized control group exists. Even within the group of students who often attend GGW sessions, they do not generally attend all sessions. One might reasonably question whether any differences between attendees and nonattendees could be attributed entirely to selection effects. This is addressed further using a mixed linear model in Sec. III.

Field notes have been recorded during GGW sessions, which to date constitute more than 100 h of observation. These notes cover a number of conceptual difficulties which became the bases for future group work questions, future GQMA item development, and a number of papers on graduate student misunderstandings in quantum mechanics [16,18,19]. These difficulties are not the focus of this work. Instead, here we summarize general observations about the GGW sessions and the classroom dynamics, to address the questions of feasibility and student engagement. Here we will present four key observations. (i) Simply put, the GGW sessions work. Students show up voluntarily, engage with each other and with the material, and ask productive questions. This is not yet meant as a claim about learning gains or cohort building, merely a statement of the functionality of GGW as a structure in graduate education. (ii) Once graduate students are engaged, their discussions are more expertlike, and require less external prompting than might be required in undergraduate group work sessions. For example, a common practice in introductory group work is to withhold answers until near the end of class, and withhold full solutions entirely. This is done for good reasons, but these reasons do not seem as relevant for graduate students. Graduate students were not observed waiting complacently for answers to be given, knowledge that an answer would be shared did not seem to diminish their discussions, and giving an answer did not end debate, but often rekindled a discussion. As a result, answers were often shared as the session progressed, rather than strictly at the end. (iii) Graduate students express a need for group autonomy in the group work sessions. This includes the ability to check answers as needed, as described above in (ii), and also the ability to skip some questions. Groups would skip questions about which the members were very confident; this may be problematic when materials have been tuned to address known common graduate student

misunderstandings. However, meeting students halfway is critical if attendance is optional. One effective measure against students skipping over a problematic area is the sharing of answers throughout the session, described above, and verbally emphasizing tricky problems or often-unexpected answers. This gives students the freedom to pursue the questions most interesting to their group, but ensures that they give additional consideration before skipping a question entirely. (iv) Graduate students vary in their willingness to discuss problems and interact with each other. Some students are excited to work on physics problems in a discussion format, while others are very shy, and in some cases fearful of exposing knowledge gaps or misunderstandings. Although this was not studied rigorously, one should keep in mind that this fear may be more pronounced among graduate students than among undergraduates, given how entangled physics expertise may be with the identity of a physics graduate student. Tension was effectively diffused in GGW sessions in several ways, including providing snacks, which initiated discussion naturally. Another way of validating students' initial understanding was the explicit inclusion of common student misunderstandings in problem stems in the group work material—a technique commonly used in undergraduate tutorial materials [24]. Consider the following example taken from an early lesson on linear algebra, and choice of basis:

“Some physics graduate students are disagreeing about Pauli matrices, and their use in the context of spins. Friend A insists that  $\hat{S}_x = \frac{\hbar}{2} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ . Friend B insists that  $\hat{S}_x = \frac{\hbar}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ . Friend C insists  $\hat{S}_x = \frac{\hbar}{2} \begin{pmatrix} i & 0 \\ 0 & -i \end{pmatrix}$ . Help the friends reconcile their differences... careful here.”

This type of presentation is aimed at normalizing the occurrence of intelligent people disagreeing, and even missing subtle details, such as being able to choose a basis that diagonalizes a matrix of interest, or overgeneralizing the freedom introduced by a choice of basis to incorrectly allow non-Hermitian matrices to correspond to physical observables.

### C. Graduate quantum mechanics assessment

In four of the years that GGW has been developed, we have been iteratively developing and using sets of conceptual items designed to assess graduate student understanding of topics central to quantum mechanics. Such assessments have not yet been used for the other core course areas. These assessments were designed to put more emphasis on conceptual understanding (as opposed to calculations) than the archetypical graduate quantum mechanics exam, although there are still some calculations required. One assessment was designed for each semester of graduate QM, and these were given as both pretests (in the first 10 days of the semester) and post-tests (in the last ten days of the semester). One hour was allotted for each

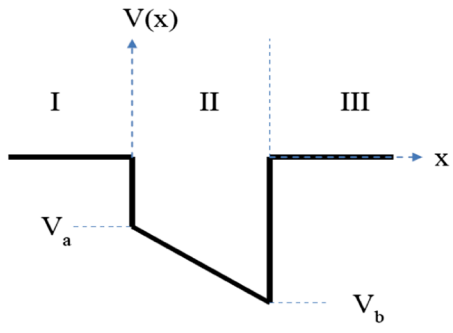
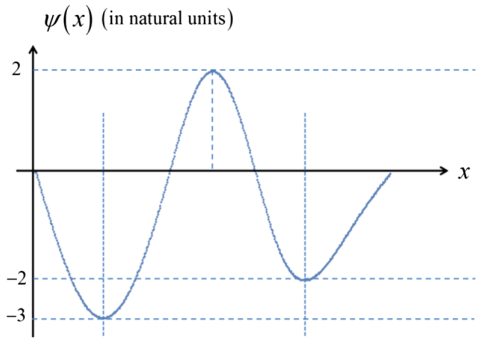
test, with most students finishing in less than 40 min. We call these assessments, collectively, the graduate quantum mechanics assessment (GQMA).

Because of variation in the book used by the instructor and coverage of topics, occasionally an item or two will be removed from or added to the GQMA to ensure that the coverage of the assessment matches the coverage by the class. This means that if an instructor chooses not to cover, for example, coherent states of the harmonic oscillator, any items on coherent states would be removed from the

assessment and the lack of coverage would not be reflected in lower GQMA scores. Low scores on GQMA items reflect poor performance on a covered topic, not variation in choice of topic coverage. In this work, some items are included in analysis that have only been used in a subset of years; but all items here were used on both the pretest and post-test of at least 1 yr, and more typically 3 yr.

Because only 1 h could reasonably be allotted for the assessment, and due to the complexity of material at this level, it was not feasible to include multiple items on every

TABLE II. Some topics emphasized on the GQMA, number of items covering the topic, and an example item for each topic. This is not an exhaustive list of GQMA items. In total, there were 14 topics and 34 items in the GQMA used in this study. 24 were scored out of 10 points, such that the scores are approximately a continuum; 10 were graded as simply “right” or “wrong.”

Topic	No. of items	Paraphrased example
Wave functions (continuum score)	1	Make a qualitative sketch of the ground state wave function in the potential well shown.
		
Math or linear algebra (continuum score)	2	Let $\hat{A} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ , and the vector $ a\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ . Calculate $\langle a \hat{A} a\rangle$ .
Expectation values (continuum score)	3	On the wave function diagram, mark with an “x” the approximate expectation value of position. Mark with a circle the approximate position at which the particle is most likely to be found.
		
Operator action vs measurement (binary item—right or wrong)	2	Consider the statement “Acting with an operator $\hat{H}$ is the same thing as making a measurement in a lab of the physical observable associated with that operator.” Do you agree or disagree with this? Explain.
Spin (continuum score)	3	A system consists of two spin-1/2 particles. One particle is “spin-up” and one is “spin-down.” If a measurement is made of the total spin of the system, what might you find?

topic assessed. This means that treating the assessment as a single-topic scale (or even collection of subscales) in the validation process is not strictly appropriate. Indeed, we are currently undertaking a project to utilize misunderstandings identified from the GQMA to develop subscales on specific QM topics including wave functions, spin, and several others. A few example GQMA items are shown in Table II, and the validity and reliability of the GQMA items used is discussed in the next subsection.

Many of the GQMA items were scored pseudocontinuously, as in assignment of a whole-number score between 0 and 10 (we will refer to these 24 items as *continuum-scored* items). An additional 10 items were *binary* (right or wrong). These two outcome types require different models, and are analyzed separately in Sec. III.

#### D. Assessment reliability and validity

GQMA items were written based on observations of graduate students in GGW sessions, and their discussions of key fundamental topics in graduate QM. Three of the 34 items used on the GQMA were modifications of items developed by other researchers [10] (on measurement, operator action, and wave functions); all other items emerged from student observation. Some GGW questions were written with the intent of confronting suspected or known misunderstandings in quantum mechanics, and notes were taken on the effectiveness of these. Other times, unexpected confusion would emerge, and questions were crafted later to address these.

The GQMA was given to six instructors from four different physics departments. Feedback was solicited through discussions and email on the importance of the topics being covered, and on the appropriateness of the difficulty level. Adjustments were made in accordance with instructors' advice. All items used in this analysis were deemed topically important and of appropriate difficulty by instructors. After the first cohort of students took the GQMA, student responses were reviewed. Modifications were made to some items to correct unclear wording or to refocus student attention on the desired skill or content. Any items to which substantive changes were subsequently made have been omitted from this analysis.

Mean scores on individual GQMA items were between 36% and 77% on the pretest and between 38% and 83% on the post-test. Mean scores are shown in Fig. 1. It is possible that some items are near ceiling on the post-test.

To determine the reliability of a scale, one might consider its Cronbach's alpha score. A set of 11 items used on one first-semester assessment had a Cronbach's alpha of 0.74 (acceptable) despite wide variation in the skills and topics addressed by each item. There are caveats to using this as a descriptor of GQMA reliability. Any assessment of a topic as broad as quantum mechanics is bound to be heterogenous, such that high values of alpha would likely be difficult to achieve. Conversely, scales with a large number of items can

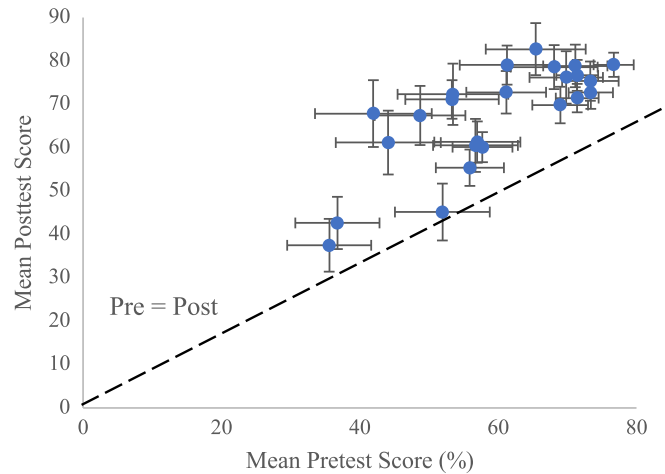


FIG. 1. Mean pretest and post-test scores on GQMA, by item for the 24 continuum-scored items used in this study. Scores are averaged over the 133 student participants.

yield moderate to high values of alpha even when the items are dissimilar or only weakly related [25]. A more appropriate use of Cronbach's alpha would be to confirm unidimensionality of items that fall into one factor [26]; this is not practical here, since the exploratory GQMA was not designed to consist of subscales.

Alternatively, one might consider repeated-measure reliability. That is, is there a correlation between student performance on the pretest and on the post-test? Pearson correlation coefficients for pre and post-testing ranged by item from 0.15 to 0.9. This is an imperfect descriptor of item reliability, since there are numerous interventions between the pretest and post-test, some of which are given to all students (such as lecture and homework) and some of which are given only to some (GGW). The analysis in this paper has been repeated with the removal of all items for which the correlation between pre- and post-test scores was not statistically significantly positive; the results were qualitatively unchanged.

Another way of checking repeated-measure reliability would be to look at the performance of different cohorts on the same item, and see if a repeated measures ANOVA reveals a significant dependence on cohort. Only one item in our set showed a statistically significant dependence on cohort, and that was not significant after a *post hoc* Bonferroni correction.

Finally, discrimination indices were calculated for all items. Those with very low or negative discrimination indices (fewer than 10% of items) were removed from analysis. Remaining items' discrimination indices ranged from 0.2 to 0.7.

### III. RESULTS AND DISCUSSION

#### A. Student performance

Before analyzing the data in a way that more carefully considers the complexity of the dataset, to get a general

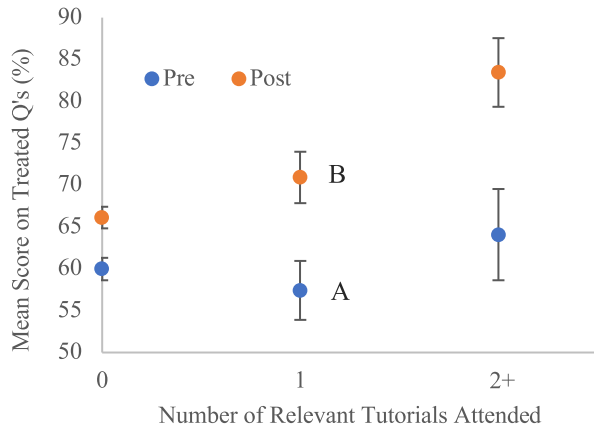


FIG. 2. Mean student pretest and post-test scores on the set of all continuum-scored GQMA items for which they attended  $N$  relevant GGW sessions, for  $N = 0, 1, 2$  or more. Error bars represent standard error. The numbers of student items in the three treatment categories are 607, 107, and 37, respectively, and these originate from 133 students' work.

sense of the results, it is instructive to look at the average of individual pretest and post-test item scores for all items that were relevant to the GGW sessions as a function of relevant GGW treatments. Figure 2 shows the average pre- and postscores over all students and all continuum-scored items for which the student attended 0, 1, or 2 or more GGW sessions related to that item's content. For example, the point labeled "A" in Fig. 2 indicates that the average pretest score over all instances in which a student attended one GGW session relevant to a particular continuum-scored item was 57%. The point labeled "B" indicates that the post-test score averaged over all such instances was 71%. It must be noted here that most topics are covered in one week's GGW session, and only a few topics are revisited in a second or third session. The number of items covered in 0, 1, 2, and 3 GGW sessions were 5, 20, 5, and 4, respectively.

Figure 2 provides an indication that the GGW sessions were effective. Specifically, student average scores over all

items for which no relevant GGW session was attended did improve by about 5% from pre to post. Students who attended a single GGW session related to item  $j$  did slightly worse on item  $j$  on the pretest than those who attended none, but had much higher gains ( $\sim 13.5\%$ ). This slightly worse performance on the pretest could be interpreted at least partly as a selection effect: students who performed poorly early on made it a point to attend GGW. This is a hypothesis that needs more investigation, although it is known that students who received low scores on other measures early on (like midterms) did begin attending GGW in response. Students who attended GGW so often that they were treated on some items more than once did better on those items on both the pretest and the post-test. Gains in this category were higher still (at 19.4%). The increased pretest score compared to other categories may indicate a different kind of selection effect: students who are very interested in quantum mechanics, who perform well initially, may be more likely to attend the majority of GGW sessions.

Given that there is not a randomized control group in this study, it is important to consider the factors that might affect performance on the GQMA. Here, we consider a mixed linear model of student performance given by

$$S_{ij} = S_0 + \alpha_j + \beta_i + \delta \times \text{GRE}_i + \theta \times T_{ij} + \nu \times N_{ij} + \gamma \times P_{ij} + \epsilon_{ij}. \quad (1)$$

The symbols are explained in Table III, and resulting values of the fit parameters are shown in Table IV. This model assumes continuous variables and is applied only to the subset of GQMA items that were scored on a continuum. Binary (right or wrong) items are discussed separately.

The principal outcome of interest from this model is  $\theta$ , the estimate of gain in score on a given post-test item for every GGW session attended relevant to that item. Essentially, this is a measure of the effectiveness of the

TABLE III. Parameters used in the mixed linear model of student post-test performance on continuum-scored GQMA items.

Symbol	Factor type	Meaning
$S_{ij}$	Dependent var.	Post-test score of student $i$ on item $j$
$S_0$	Intercept	Post-test score intercept
$\alpha_j$	Random	Normally distributed, zero-mean, random effect on item $j$ ( <i>nominally associated with item difficulty</i> )
$\beta_i$	Random	Normally distributed, zero-mean, random effect on student $i$ ( <i>nominally associated with student achievement</i> )
$\text{GRE}_i$	Fixed	GRE Physics score for student $i$ (grand mean centered, scaled from 0 to 100)
$\delta$	Fit parameter	Increase of GQMA post-test score (0 to 100) per 1% increase in GRE Physics score
$T_{ij}$	Fixed	GGW treatments of student $i$ relevant to item $j$ (number: 0, 1, 2...)
$\theta$	Fit parameter	Increase of GQMA post-test score (0 to 100) per relevant GGW attendance
$N_{ij}$	Fixed	GGW treatments of student $i$ NOT relevant to item $j$ (number)
$\nu$	Fit parameter	Increase of GQMA post-test score (0 to 100) per irrelevant GGW attendance
$P_{ij}$	Fixed	GQMA pretest score of student $i$ on item $j$
$\gamma$	Fit parameter	Increase of GQMA post-test score (0 to 100) per increase in pretest score (0 to 100)
$\epsilon_{ij}$	Residual	Residual random (unexplained) variation

TABLE IV. Estimates of the best fit parameters in model (1). The estimate of each parameter is in units of post-test score (1–100) per increment of the parameter, as described in Table III. Here \* means  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ . The percentage of variance explained was calculated by the iterative removal of the fixed parameter that explained the most variance, and recalculation of the variance. The data include scores from 133 student participants.

Parameter	Estimate	95% C.I.	% variance
$S_0$	46.5***	19.3 to 73.8	...
$\gamma$	0.334***	0.269 to 0.400	17.0
$\delta$	0.277**	0.063 to 0.490	4.6
$\theta$	6.42***	2.54 to 10.3	2.5
$\nu$	-0.123	-0.970 to 0.724	0.22
Parameter	St. Dev.		% variance
$\alpha_j$	8.6		30.9
$\beta_i$	7.8		18.4
$\epsilon_{ij}$	24.2		28.3

group work. Note that this model accounts for normal variation between students' performance ( $\beta$ ) nominally associated with student achievement, normal variation in average score on each item ( $\alpha$ ) nominally associated with item difficulty, the number of nonrelevant GGW sessions attended ( $N_{ij}$ ), the pretest score ( $P$ ), and the GRE Physics score.

The model estimate for the parameter  $\theta$  is positive and statistically different from zero, with a value of 6.42. One might interpret this as meaning that post-test scores on items went up on average by 6.42% for each GGW session a student attended that was relevant to that item (though here it must be noted that multiple GGW treatments on the same topic were relatively rare, occurring in only 9 of the items). The standard deviation of the set  $\beta_i$  (normally distributed variation in student performance) is 7.8. In other words, attending one relevant GGW session will increase an average student's performance by about 0.82 standard deviations of typical between-student variation. As a caveat, it must be noted that the standard deviation of the residual (which is variation in performance unexplained by the model) is 24.2, such that this increased performance accounts for only 0.25 standard deviations of this unexplained variation.

GRE Physics scores as well as pretest and post-test performance have all been expressed as a percentage of the maximum points to facilitate direct comparisons. As expected, for a given GQMA item, on average the post-test score depends on the pretest score (i.e.,  $\gamma > 0$ ), and interestingly, the post-test score also depends on the GRE physics score ( $\delta > 0$ ), even controlling for the pretest score. The estimated magnitudes of  $\gamma$  and  $\delta$  indicate that attending one GGW session results in an average increase in the post-test score on items relevant to that session by the same amount as increasing the GRE Physics score by 23% or the pretest score for that same item by 19%.

The only parameter from Eq. (1) that is not statistically different from zero is  $\nu$ , the coefficient of GGW sessions unrelated to item content. This is to be expected, and it is important when considering self-selection effects. It would be possible for GGW attendees to be those students who are most enthusiastic about QM and most driven to succeed at it, or for them to be especially motivated on the post-test, because they invested time throughout the semester. Both of those possibilities would result in students of higher motivation or interest correlating with overall GGW attendance, regardless of the topic. If the driving factor behind higher post-test scores for attendees were of this sort, the post-test score on a particular item would correlate with total attendance of GGW sessions, including those irrelevant to that item (i.e., we would have found that  $\nu > 0$ ). Then this increase in score due to GGW participation may not necessarily be attributed to student learning during group work, and might alternatively be better explained by some motivational or self-selection factors. This was not the case. What was observed, instead, is that attendance of irrelevant GGW sessions was consistent with having no effect on post-test performance, providing more evidence that the mean improvement on post test scores could indeed be due to student learning in relevant GGW sessions. Many variations of Eq. (1) were explored. Terms were kept in the linear model only if their addition significantly lowered the Akaike's information criterion. Factors that were examined and rejected include student score on the first midterm in quantum mechanics, and different dependencies on GGW attendance (such as quadratic). In all models employed, the coefficient of relevant GGW attendance was always

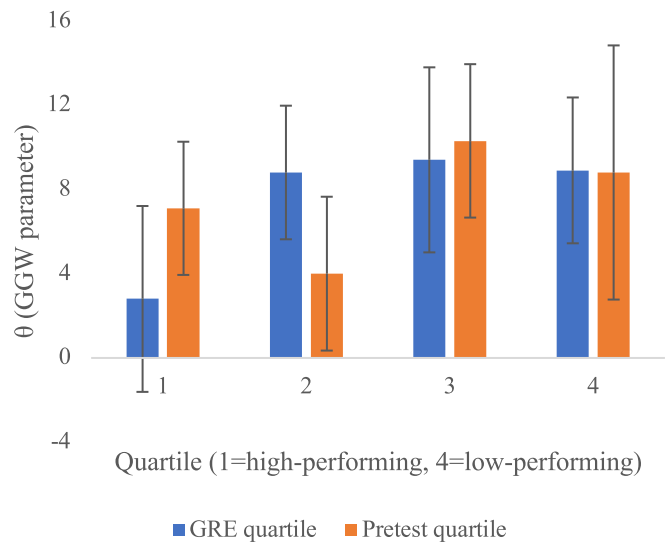


FIG. 3.  $\theta$  resulting from the mixed linear model (1) for quartiles of the 133 total students. Here, the 4th GRE Physics quartile is the 25% of students with the lowest GRE Physics scores, and so on. The error bars indicate standard error.



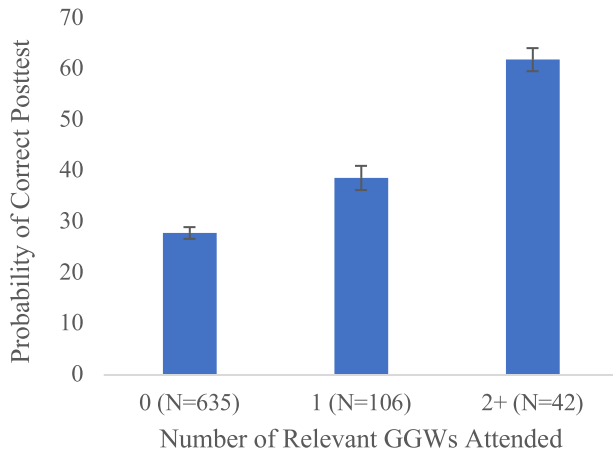


FIG. 4. Probability of correctly answering a binary (right or wrong) item on the post-test, given that students answered incorrectly on the pretest. Results are averaged over different students and items, and are separated by the number of item-relevant GGW treatments students received.  $N$  is the number of student items in each category. Error bars represent standard error, although there are correlations within the data.

positive and significant. It is worth noting that all of these estimates were not substantively changed upon removal of items with poor retest reliability. For example, removal of the four items with the lowest retest reliability changed  $\theta$  to 7.06.

One characterization of the goodness of fit is  $R^2$  as described by Nakagawa and Schielzetch [27]; two variations are described in that work and both are calculated here:  $R^2_{\text{GLLM}(m)} = 0.24$ , and  $R^2_{\text{GLLM}(c)} = 0.74$ . The difference is that  $R^2_{\text{GLLM}(c)}$  includes the variance explained by the random variables in the total variance explained, whereas  $R^2_{\text{GLLM}(m)}$  does not.

It is important to know how this apparent utility of GGW instruction varies for students of different incoming preparation: is this most helpful for struggling students, or for

students who are already high performing? The mixed linear model described by Eq. (1) was therefore applied to quartiles of students, with the quartiles determined first by GQMA pretest score, and then separately by GRE Physics score. There were very few participants in the top quartile by GRE, resulting in statistically insignificant results. In all other cases,  $\theta$  is positive, and significant. The results are summarized in Fig. 3.

Based on this rough treatment, the effect appears to be positive and statistically different from zero for all but one quartile (the highest-achieving quartile by GRE score). In general, there is significant overlap in error bars, and the values of  $\theta$  are not statistically different for the different quartiles.

## B. Student performance on binary items

We conducted a separate analysis of binary-scored items because the properties and logical analysis of these items are different from that of continuum-graded items. The fourth item in Table II: “Operator action vs measurement” is an example of such a binary-graded item. We describe a full binary logistic model below, but first to get a general sense of the effectiveness of the GGW sessions, let us consider cases in which a given binary item is answered incorrectly on the pretest and determine the probability that it was answered correctly on the post-test, as a function of number of relevant GGW sessions attended. The results are presented in Fig. 4, which are consistent with the findings from the continuum-graded items and indicate a clear, additive benefit of attending multiple relevant GGW sessions. Specifically, before controlling for any variables like pretest score or GRE scores, the probability of answering binary items correctly on the post-test is 1.4 times higher if students attend a single related GGW session than if they attend none; it is 2.2 times more likely if they attend two related GGW sessions.

Similar to the continuum-graded items, we conducted a regression analysis on the binary items, though in this case

TABLE V. Parameters used in the mixed linear model of student post-test performance on binary-scored GQMA items.

Symbol	Factor type	Meaning
$p_{ij}$	Dependent var.	Probability of answering correctly on the post-test
$S'_0$	Intercept	Intercept related to post-test probability
$\alpha'_j$	Random	Zero-mean, random effect on item $j$ (nominally associated with item difficulty)
$\beta'_i$	Random	Zero-mean, random effect on student $i$ (nominally associated with student achievement)
$\text{GRE}_i$	Fixed	GRE Physics score of student $i$ (grand mean-centered, scaled 0 to 100)
$\delta'$	Fit parameter	Weight of GRE Physics score
$T_{ij}$	Fixed	GGW sessions attended (number: 0, 1, 2...) by student $i$ , related to item $j$
$\theta'$	Fit parameter	Weight of relevant GGW treatments
$N_{ij}$	Fixed	GGW sessions attended (number: 0, 1, 2...) by student $i$ , irrelevant to item $j$
$\nu'$	Fit parameter	Weight of irrelevant GGW treatments
$\text{Pre}_{ij}$	Fixed	GQMA pretest score on item $j$ by student $i$ (right or wrong 1 or 0)
$\gamma'$	Fit parameter	Weight of pretest score

TABLE VI. Values of the best fit parameters in model (2). Bootstrapping was used to determine 95% CIs. \* indicates  $p < 0.05$ , \*\*= $p < 0.01$ , \*\*\*= $p < 0.001$ . Data from 133 students are included in the analysis.

Parameter	Estimate	Standard error	Exp(value)	95% C.I. on Exp(value)
$S'_0$	0.29	1.7	1.33	0.05 to 39.4
$\delta'$	0.053***	0.015	1.05	1.02 to 1.09
$\theta'$	1.13**	0.36	3.11	1.52 to 6.36
$\nu'$	-0.061	0.058	0.94	0.84 to 1.06
$\gamma'$	1.62***	0.40	5.06	2.30 to 11.1
Parameter	St. Dev.			
$\alpha'_j$	0.12			
$\beta'_i$	0.40			

we used a binary logistic regression. The model used was as follows:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = S'_0 + \alpha'_j + \beta'_i + \delta' \times \text{GRE}_i + \theta' \times T_{ij} + \nu' \times N_{ij} + \gamma' \times \text{Pre}_{ij}. \quad (2)$$

This binary logistic regression is clearly different from the mixed linear model analysis done on continuum-scored questions in that the variable being predicted is the probability of a correct answer, rather than an average score, and one input (pretest score) is binary. Accordingly, the same fixed factors are considered in this analysis, as in Eq. (1), but the parameters have slightly different mathematical meanings (hence the primed variables). These are described in Table V. Best fit values for the parameters are shown in Table VI.

The resulting estimates for the binary items are qualitatively similar to the continuum items and to this extent are a confirmation of those findings. In particular, the nonzero value of  $\theta'$  indicates that attending one relevant GGW session increases the probability of answering correctly on the relevant post-test item by  $76 \pm 7\%$ , and attending 2 GGW sessions increases it by  $90 \pm 8\%$ . Further, the probability of answering correctly on a post-test item increases, as expected if the pretest was answered correctly, and increases with GRE score. Finally, since  $\nu'$  does not differ significantly from unity, there is no statistically significant dependence on attendance of unrelated GGW sessions.

#### IV. CONCLUSIONS

We have designed and implemented guided group work sessions for graduate students in quantum mechanics. Sufficiently many students attended and engaged consistently and voluntarily for group work and discussions to be feasible and productive, though we have observed that there are some important differences in how to run these sessions

compared to sessions with novice undergraduate students. We have collected significant evidence that guided group work sessions improve performance on a mixture of conceptual and calculational assessment items in graduate level quantum mechanics. Although there is not a randomized control group in this study, the independence of performance on attendance of unrelated GGW sessions is strong evidence that alternative explanations for the gain in performance, such as selection effects or other attitudinal or motivational considerations do not account for the observed positive effects of GGW attendance. The coefficient describing post-test score dependence on relevant GGW attendance was large and positive (6.42% increase per GGW attendance). Comparing this to the coefficient related to the GRE Physics test, a student who attends one GGW session on a topic would do as well on related items as a student who scored 23% higher on the GRE Physics test, but did not attend a relevant GGW session.

These findings are robust against modifications of the model, including the removal of variables, and restriction to subsets of students or items. The benefit is present across all quartiles of student performance on the Physics GRE, and GQMA pretest.

Qualitatively, the observed positive benefit of GGW at the graduate level is consistent with similar benefits of GGW observed at the undergraduate level, and indicates the broad effectiveness of GGW in physics education. The results of this study indicate that implementing GGW in a graduate level QM course may be beneficial. However, there are a number of factors to be mindful of, and which warrant further study. For example, the optional attendance of GGW sessions not only removes a randomized control group, but also means that nonattendees were not ensured equal time on some alternative form of instruction. The claim in this work is that GGW sessions add to student understanding achieved during lecture; it would be very useful to know whether GGW adds to student understanding more than alternative instructional modes. Additionally, the findings in this work are based on a specific implementation of GGW at one institution; the ratio of conceptual to calculational items, the duration of the weekly sessions, and the voluntary attendance policy were all roughly constant across the multiple years examined in these data. This leaves a wide parameter space to be explored in the incorporation of GGW into graduate core courses. In particular, including GGW into traditional lecture might involve shorter 15-20-min spurts of group work, preceded and/or followed by lecture. It is not clear how the effectiveness of this would compare to the 60-min sessions we have studied. Further, the implicit expectation of full attendance in a traditional lecture session may negatively impact the dynamic of students freely sharing misunderstandings of the topic to the entire class, as opposed to with a smaller, committed group of voluntary participants, as was the case in our study. GGW sessions in

lecture might constitute a higher-stakes environment in which to air such misunderstandings, and steps must be taken to ensure it is a supportive environment.

### ACKNOWLEDGMENTS

This work was supported by the Innovations in Graduate Education National Science Foundation's Research Traineeship Award under Grant No. 1735027. This work was also supported in part by the OSU Center for Emergent Material, an NSF MRSEC under Grant No. DMR-1420451, and by the OSU M.S.-Ph.D. Bridge Program, an APS Bridge Program site. The authors wish to thank the

graduate students of The Ohio State University Physics Program for participation in this study. We thank the Physics faculty from OSU and collaborators for their useful discussions, suggestions, and endorsement of our study. We wish to acknowledge the contribution of Abigail Bogdan, who helped review and analyze the first year of data and helped with early drafts of the assessment. We also wish to thank The Ohio State University Physics M.S. to Ph.D. Bridge Program and the American Physical Society Bridge Program, for supporting the role of our Physics Education Research group in advanced physics and graduate physics courses at The Ohio State University.

- 
- [1] Council of Graduate Schools PhD Completion Project, Ph.D. Completion and Attrition: Analysis of Baseline Demographic Data from the Ph.D. Completion Project (Council of Graduate Schools, Washington, DC, 2008), [https://cgsnet.org/sites/default/files/phd\\_completion\\_attrition\\_baseline\\_program\\_data.pdf](https://cgsnet.org/sites/default/files/phd_completion_attrition_baseline_program_data.pdf).
  - [2] S. Malcom, Diversity in physics, *Phys. Today* **59**, No. 6, 44 (2006).
  - [3] J. L. Docktor and J. P. Mestre, Synthesis of discipline-based education research in physics, *Phys. Rev. ST Phys. Educ. Res.* **10**, 020119 (2014).
  - [4] L. Springer, M. E. Stanne, and S. S. Donovan, Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis, *Rev. Educ. Res.* **69**, 21 (1999).
  - [5] L. C. McDermott, Millikan Lecture 1990: What we teach and what is learned—Closing the gap, *Am. J. Phys.* **59**, 301 (1990).
  - [6] L. C. McDermott, Physics education research—the key to student learning, *Am. J. Phys.* **69**, 1127 (2001).
  - [7] P. S. Shaffer and L. C. McDermott, A research-based approach to improving student understanding of the vector nature of kinematical concepts, *Am. J. Phys.* **73**, 921 (2005).
  - [8] S. J. Pollock and N. D. Finkelstein, Sustaining educational reforms in introductory physics, *Phys. Rev. ST Phys. Educ. Res.* **4**, 010110 (2008).
  - [9] R. E. Scherr, P. S. Shaffer, and S. Vokos, The challenge of changing deeply held student beliefs about the relativity of simultaneity, *Am. J. Phys.* **70**, 1238 (2002).
  - [10] C. Singh, Student understanding of quantum mechanics at the beginning of graduate instruction, *Am. J. Phys.* **76**, 277 (2008).
  - [11] M. C. Wittmann, J. T. Morgan, and L. Bao, Addressing student models of energy loss in quantum tunneling, *Eur. J. Phys.* **26**, 939 (2005).
  - [12] L. Bao and E. F. Redish, Understanding probabilistic interpretations of physical systems: A prerequisite to learning quantum physics, *Am. J. Phys.* **70**, 210 (2002).
  - [13] R. Rosenblatt, A. F. Heckler, and K. Flores, A tutorial design process applied to an introductory materials engineering course, *Adv. Engineer. Educ.* **3** (2013).
  - [14] R. R. Hake, Interactive-engagement vs. traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses, *Am. J. Phys.* **66**, 64 (1998).
  - [15] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, and M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8410 (2014).
  - [16] C. D. Porter and A. F. Heckler, Graduate student misunderstandings of wave functions in an asymmetric well, *Phys. Rev. ST Phys. Educ. Res.* **15**, 010139 (2019).
  - [17] E. Marshman and C. Singh, Validation and administration of a conceptual survey on the formalism and postulates of quantum mechanics, *Phys. Rev. ST Phys. Educ. Res.* **15**, 020128 (2019).
  - [18] C. D. Porter and A. F. Heckler (to be published).
  - [19] C. D. Porter, A. Bogdan, and A. F. Heckler, Student understanding of potential, wavefunctions and the Jacobian in hydrogen in graduate-level quantum mechanics, in *Proceedings of the 2016 Physics Education Research Conference, Sacramento, CA* (AIP, New York, 2016), pp. 244–247.
  - [20] R. Sayer, A. Maries, and C. Singh, Quantum interactive learning tutorial on the double-slit experiment to improve student understanding of quantum mechanics, *Phys. Rev. ST Phys. Educ. Res.* **13**, 010123 (2017).
  - [21] C. Keebaugh, E. Marshman, and C. Singh, Developing and evaluating an interactive tutorial on degenerate perturbation theory, in *Proceedings of the 2016 Physics Education Research Conference, Sacramento, CA* (AIP, New York, 2016), pp. 184–187.
  - [22] L. D. Carr and S. B. McKagan, Graduate quantum mechanics reform, *Am. J. Phys.* **77**, 308 (2009).
  - [23] L. C. Hodges, Contemporary issues in group learning in undergraduate science classrooms: A perspective from student engagement, *Life Sci. Educ.* **17** (2018).
  - [24] L. C. McDermott, P. S. Shaffer, and University of Washington, Physics Education Group, *Tutorials in In-*

- troductory Physics* (Prentice Hall, Upper Saddle River, NJ, 2002), Vol. 2.
- [25] D. Streiner, Starting at the beginning: an introduction to coefficient alpha and internal consistency, *J. Personality Assess.* **80**, 99 (2003).
- [26] M. Tavakol and R. Dennick, Making sense of Cronbach's alpha, *Int. J. Med. Educ.* **2**, 53 (2011).
- [27] S. Nakagawa and H. Schielzeth, A general and simple method for obtaining R<sup>2</sup> from generalized and linear mixed-effect models, *Methods Ecol. Evol.* **4**, 133 (2013).