Characterizing graduate teaching assistants' teaching practices in physics "mini-studios"

Tong Wan
Department of Physics, University of Central Florida, 4111 Libra Drive, Orlando, FL, 32816

Constance M. Doty,[1] Ashley A. Geraets,[2] Erin K. H. Saitta,[2] and Jacquelyn J. Chini[1]
[1]Department of Physics, University of Central Florida, 4111 Libra Drive, Orlando, FL, 32816
[2]Department of Chemistry, University of Central Florida, 4111 Libra Drive, Orlando, FL, 32816

In this study, we characterized GTAs' teaching practices in algebra-based introductory physics "mini-studios," which combine student-centered recitation and inquiry-based labs. We documented both GTA and student actions using an observation protocol adapted from the Laboratory Observation Protocol for Undergraduate STEM (LOPUS). We observed 72 mini-studio sessions led by 11 GTAs over two semesters. We used an agglomerative hierarchical cluster analysis and identified three clusters that described the similarities and differences between individual sessions. Two clusters contained sessions characterized by more interactive GTAs but they varied in the amount of feedback, lecture and whole class questioning the GTA provided. In the third cluster, GTAs tended to wait for students to call on them before engaging. Student behaviors also varied between the clusters, suggesting correlations between student behaviors and GTA instructional styles, in contrast to previous findings with LOPUS in other contexts. We discuss implications of these findings for future research.

I. INTRODUCTION

Active learning has been proven to be effective at improving student performance in undergraduate science, technology, engineering and mathematics [1]. To support use of active learning, the physics education research community has engaged in curriculum development, often developing curricula for recitation or laboratory environments in which students work in small groups (e.g., *Tutorials in Introductory Physics* [2]). Graduate teaching assistants (GTAs) often lead instruction in these settings, especially at large, research-intensive universities. Research has shown that GTAs' teaching skills are positively correlated with student learning in transformed recitation and lab sections [3,4]. However, GTAs who engage in the same professional development may vary substantially in their teaching actions [5]. These findings suggest that more effective GTA professional development (GTA PD) is needed to support GTAs in high-fidelity implementation of active learning strategies.

As part of a larger project on developing an effective GTA PD program, we have started to characterize GTAs' teaching practices in active learning environments. In this study, we evaluate the extent to which GTAs teaching the same transformed course employ certain pedagogical skills. We observed GTAs' classrooms using an observation protocol adapted from the Laboratory Observation Protocol for Undergraduate STEM (LOPUS) [6]. We performed a cluster analysis to categorize small-group sessions led by GTAs. Our research questions are:

(1) In what ways may GTAs' instructional practices differ when teaching in the same instructional environment?

(2) How do GTAs' instructional styles impact student behaviors?

In this paper, we report three different instructional styles observed among the GTAs, and discuss relationships between GTAs' instructional styles and student behaviors. Furthermore, we provide suggestions for future research.

II. METHODS

A. Conceptual framework: Instructional Capacity

Cohen and Ball describe instructional capacity as the capacity to produce worthwhile and substantial learning [7]. Cohen and Ball argue that efforts in educational reform focus on either improving curriculum *or* enhancing instructors' teaching practices. However, they propose that instructional capacity depends on the *interactions* among instructional materials, instructors, and students. Instructional activities that students and instructors engage with are influenced by instructional materials. In return, instructors' pedagogical content knowledge [8], prior experiences as students, and beliefs about teaching impact how instructors interact with students and instructional materials. Students' backgrounds and past educational experiences also impact their responses to instructors' teaching and their engagement with the instructional activities.

In line with this framework, we documented both GTAs' and students' actions, which allowed us to evaluate how student behaviors may correlate with GTAs' instructional styles. In addition, the framework informed the interpretation of our findings (see Section IV).

B. Instructional context and participants

The instructional environment chosen for this study is the "mini-studios" for an algebra-based introductory physics course at the University of Central Florida. The "mini-studios" combine recitation and laboratory activities [9]. Each mini-studio session is comprised of a 75-minute tutorial based on the University of Maryland Open Source Tutorials [10], a 15-minute quiz, and an 80-minute lab based on the Investigative Science Learning Environment (ISLE) curriculum [11].

GTAs participated in a one-semester pedagogy seminar during their first semester as GTAs and weekly on-going preparation meetings; see Ref. [12] for more details. In addition, the GTAs rehearsed two pedagogical skills, cold calling and error framing, in a mixed-reality simulator [13] at the beginning of one semester; we do not report on the rehearsed skills in this analysis.

A total of 11 GTAs volunteered to participate in the study. Six of the participants were new GTAs, and the other five were experienced GTAs (four previously taught the mini-studios and one taught calculus-based physics labs). We conducted the classroom observations over two consecutive semesters. Seven of the 11 GTAs participated in both semesters. In each semester, every GTA was observed about four times (range of three to five), making up a total of 72 mini-studio session observations.

C. Coding with LOPUS

Our observation protocol was adapted from LOPUS. LOPUS was designed to capture both GTA and student actions in two-minute intervals during an entire class period. The codes describe GTA or student behaviors (e.g., GTA is talking to individual group of students one-on-one) rather than criteria used for evaluating quality of teaching (e.g., "adequate" interaction with students). All actions occurring in a two-minute interval are documented (i.e., the codes are not mutually exclusive).

Codes for GTA actions and student actions are shown in Table I and Table II, respectively. Some of the codes from the original LOPUS were not included in our protocol either because they rarely occurred during our observations or they were found to be highly correlated with other codes. In addition, we added a new code, VF (GTA providing verbal feedback), to evaluate the extent to which GTAs provide clear feedback to student responses.

The observations were conducted by three researchers (T.W., C.M.D., and A.A.G.). During the observations, GTAs

TABLE I. Descriptions and average fractions for GTA action codes adapted from LOPUS.

GTA Code	Abbreviated Definition	Average fraction
Lec	Lecturing to the class or making announcements	0.16±0.11
RtW	Real-time writing on the board, doc cam, etc.	0.12±0.10
FUp	Providing follow-up or feedback on activity	0.01±0.02
D/V	Showing a demonstration or video	0.01±0.02
M	Monitoring class or individual groups	0.25±0.18
PQ	Posing a worksheet- or lab-related question	0.03±0.05
1o1-Talk	Talking to individual student or group of students one-on-one	0.67±0.21
1o1-TPQ	Posing a question to individual students or group of students	0.25±0.17
VF	Providing feedback to student responses	0.10±0.12
VM	Verbal monitoring	0.22±0.13
TI [14]	Initiating one-on-one interaction with students	0.20±0.13
Adm	Performing administrative tasks	0.15±0.10
W	Waiting and generally unavailable to students	0.18±0.17

were asked to wear a microphone so that the observers could hear the GTAs through headphones. In order to establish inter-rater reliability (IRR), three mini-studio sessions (each led by a different GTA) were observed by either pairs of observers or all three observers toward the beginning of the first semester. Each researcher first coded the entire session independently. Then, we discussed disagreements and resolved inconsistencies. We then calculated IRR for each pair of observers. We used Cohen's Kappa for the overall IRR (i.e., IRR for all the codes during each session) and Gwet's AC1 for individual codes (e.g., Adm). We used Gwet's AC1 for individual codes because Cohen's Kappa can be extremely low even when a high percent agreement is achieved if an action is coded very frequently or very infrequently [15]. The three observers achieved an average Cohen's Kappa of 0.77±0.15, and an average Gwet's AC1 of 0.87±0.20.

For each session, we calculated the fraction of occurrence for GTA and student codes. For example, if code SQ occurred in seven 2-min intervals and the session lasted for 140 minutes (70 2-min intervals), then the fraction for code SQ would be 0.1. The average fraction and standard deviation for each code are shown in Table I and Table II.

TABLE II. Descriptions and average fractions for student action codes adapted from LOPUS.

Student Code	Abbreviated Definition	Average fraction
Wks/Lab	Working on worksheet or performing lab activity	0.87±0.09
TQ	Taking a quiz	0.14±0.09
SQ	Asking the GTA a worksheet- or lab-related question with entire class listening	0.02±0.03
1o1-SQ	Individual student or a group of students asking the GTA a worksheet- or lab-related question	0.50±0.17
WC	Engaging in whole class discussion	0±0.01
SI	Initiating one-on-one interaction with the GTA	0.36±0.12

D. Cluster analysis

The variables we used to conduct a cluster analysis were the fractions of occurrence for GTA and student actions observed. We scaled the fractions such that the average fraction for each code is zero and the standard deviation is one. This gives equal weight to all the codes, which allows the follow-up statistical tests to tease out codes that do not occur very frequently but vary substantially between clusters. We then conducted a hierarchical agglomerative cluster analysis on all 72 mini-studio sessions based on the scaled fractions for all the codes.

Agglomerative clustering works in a bottom-up manner. Initially, each session is considered as its own cluster. The two clusters that are most similar (i.e., have the smallest distance) are then combined into a bigger cluster. The dissimilarity (i.e., distance) in this case is calculated in Euclidean space. This process is iterated until all the sessions are combined in a single cluster. We performed the analysis in R using Ward's method [16], which minimizes the total within-cluster variance. We then determined the optimal number of clusters using the elbow method [17] and found three clusters.

III. RESULTS

A. GTAs' instructional styles

In order to identify the ways (i.e., the codes) in which these clusters are different, we used the Kruskal-Wallis rank sum test since the sample size is relatively small [18]. Effect size was calculated using the eta-squared for Kruskal-Wallis test [19]. We then used Dunn's multiple comparison test with Holm-Bonferroni corrections to determine significant differences between pairs of clusters. The analysis was conducted in R [20,21].

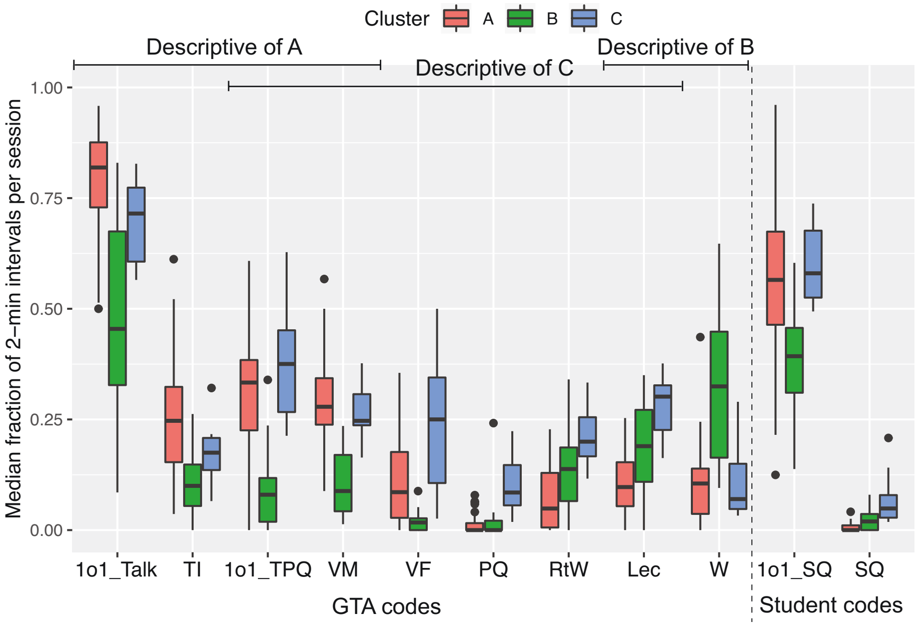

FIG. 1. Boxplot for codes with large effect sizes. For GTA codes, the four leftmost codes are higher for cluster A (than either cluster B or C), the two rightmost codes are higher for cluster B, and the six codes in the middle are higher for cluster C.

We found 11 codes that are associated with large effect sizes ($\eta^2 > 0.14$ [22]), as shown in Fig. 1; nine codes describe GTA actions and two describe student actions. Below we describe GTAs' instructional styles (each associated with a cluster of sessions observed) using these nine GTA codes.

Instructional style A: GTAs who were observed teaching the sessions in cluster A had a significantly higher fraction of talking to individual students or groups of students one-on-one (1o1_Talk, $\chi^2(2) = 31.5127$, $p_{\text{K-W}} < 0.001$). Dunn's test suggested that the difference came from comparing clusters A and B ($p_{\text{H-B}} < 0.001$). In addition, GTAs in cluster A initiated significantly more conversation with students compared to GTAs in cluster B (TI, $\chi^2(2) = 20.0585$, $p_{\text{K-W}} < 0.001$, $p_{\text{H-B}} < 0.001$). Furthermore, GTAs in cluster A (and GTAs in cluster C) posed more questions to individuals or groups of students (1o1_TPQ, $\chi^2(2) = 32.6633$, $p_{\text{K-W}} < 0.001$, $p_{\text{H-B}} < 0.001$) and had a higher fraction of verbal monitoring (VM, $\chi^2(2) = 37.7364$, $p_{\text{K-W}} < 0.001$, $p_{\text{H-B}} < 0.001$) when compared to GTAs in cluster B. In summary, GTAs in cluster A engaged more often with students in small groups.

Instructional style B: GTAs in cluster B spent significantly more time waiting (W, $\chi^2(2) = 27.5757$, $p_{\text{K-W}} < 0.001$) when compared to both cluster A ($p_{\text{H-B}} < 0.001$) and cluster C ($p_{\text{H-B}} = 0.001$). In addition, GTAs in cluster B spent more time lecturing compared to GTAs in cluster A (Lec, $\chi^2(2) = 27.655$, $p_{\text{K-W}} < 0.001$, $p_{\text{H-B}} = 0.003$), but less time compared to cluster C ($p_{\text{H-B}} = 0.008$). As mentioned previously, GTAs in cluster B had lower fractions of 1o1_Talk and TI compared to cluster A. Moreover, they had the lowest fractions of 1o1_TPQ, VM and VF (verbal feedback). To summarize, GTAs in cluster B tended to wait for students to call on them and interacted with students less frequently.

Instructional style C: Similar to instructional style A, GTAs in cluster C had significantly higher fractions of 1o1_TPQ and VM as well as less waiting compared to cluster B. However, GTAs in cluster C posed more questions (PQ, $\chi^2(2) = 27.2269$, $p_{\text{K-W}} < 0.001$) and provided more verbal feedback (VF, $\chi^2(2) = 27.7761$, $p_{\text{K-W}} < 0.001$) compared to both clusters A and B. Furthermore, GTAs in cluster C used the highest fraction of lecturing and real-time writing (RtW, $\chi^2(2) = 16.0461$, $p_{\text{K-W}} < 0.001$). To summarize, GTAs in cluster C engaged students in both whole-class and small-group settings, provided more verbal feedback, but also lectured more frequently.

GTAs may use multiple instructional styles. As shown in Fig. 2, seven out of 11 GTAs were found to use more than one of the instructional styles. Three GTAs used all three instructional styles. The results suggest that the same GTA may (or may not) use different instructional styles when leading physics mini-studios. Although the observations were conducted over two semesters, we did not notice any trend in changes in instructional styles for GTAs who were observed in both semesters. Our results are consistent with the finding from Stains et al. that STEM faculty members vary their instructional styles day to day [23].

New GTAs are more interactive than experienced GTAs. We compared the distributions of new GTAs and experienced GTAs in each cluster. As shown in Fig. 2, 29 out of 35 (83%) sessions in cluster A and nine out of 14 (64%) sessions in cluster C were led by new GTAs, while 17 out of 23 (74%) sessions in cluster B (less interactive) were led by experienced GTAs. Due to a small sample size, we used

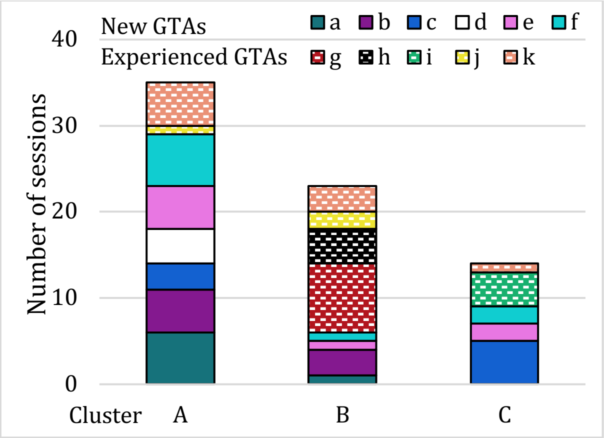

FIG. 2. Number of sessions in each cluster.

Fisher's exact test to determine whether the difference in proportions of cluster B and non-cluster B (i.e., clusters A and C) is significant between new GTAs and experienced GTAs. The results suggest that sessions led by experienced GTAs were less interactive (i.e., higher fraction in cluster B) than sessions led by new GTAs ($p < 0.001$).

B. Student behaviors

Student behaviors were also found to vary across clusters. Two student codes were found to have large effect sizes: 1o1_SQ (individual student asks GTA a question) and SQ (student asks GTA a question with entire class listening). The fraction of 1o1_SQ was found to be higher in both cluster A (1o1_SQ, $\chi^2(2) = 18.8961$, $p_{\text{K-W}} < 0.001$, $p_{\text{H-B}} < 0.001$) and cluster C ($p_{\text{H-B}} < 0.001$) when compared to cluster B. For SQ, cluster C was found to be higher than cluster B (SQ, $\chi^2(2) = 33.8605$, $p_{\text{K-W}} < 0.001$, $p_{\text{H-B}} = 0.011$), and cluster B was found to be higher than cluster A ($p_{\text{H-B}} < 0.001$).

Student behavior is associated with GTA instructional style. For clusters A and C in which GTAs were observed to be more engaged (i.e., posing more questions to individuals or groups of students and verbally monitoring students more frequently), individual students asked more questions in small groups as well. Moreover, in cluster C in which GTAs spent more time lecturing and posed more questions to the whole class, students also asked more questions with the entire class listening. The results suggest an association between GTA instructional style and student behavior.

IV. DISCUSSION

In this paper, we characterized GTAs' instructional practices in the context of physics "mini-studios." We identified three different instructional styles among GTAs who led the same instructional environment. We found that the same GTA may use different instructional styles, suggesting multiple observations are necessary to characterize GTAs' teaching practices. We hypothesize that the same GTA may use different instructional styles due to variations in learning goals between the recitation/lab units [24]. We also found that sessions led by experienced GTAs tend to be less interactive. We conjecture that this may be due to lack of incentives for high-quality teaching and increasing research tasks as graduate students progress through graduate school. Prior research has shown that interactions with students can be another factor that causes GTAs to become less interactive over time. For example, Wheeler et al. [25] found that some GTAs whose teaching beliefs shifted toward "TA as facilitator" during TA PD, reverted back to "TA as disseminator" after teaching. Similarly, Wilcox et al. found that GTAs' behaviors are influenced by their perceptions of student expectations [12].

Our results also showed that student behaviors are associated with GTAs' instructional styles, which supports the conceptual framework that the interaction between instructors and students should be considered when evaluating instructional capacity. However, it contrasts the study by Velasco et al., who found that student behavior is independent of GTA instructional styles in the context of a traditional chemistry laboratory [6]. Velasco et al. argue that the interaction between GTAs' and students' behaviors may be limited in the traditional laboratory. This argument is consistent with the framework that the interaction between students and instructors depends on the curriculum. Our study and Velasco et al. together suggest that in active learning environments, GTAs' behaviors are more likely to influence students' behaviors. Thus, we hypothesize that GTAs who make use of an interactive instructional style support student engagement in active learning. However, as discussed above, prior research shows that student behaviors can cause GTAs to become less interactive. These findings together appear to support the two-way interplay between GTA and student behaviors: instructors may become less interactive in response to student expectations, but instructors who maintain interactive teaching may be able to shift student behaviors. Therefore, we suggest that future research further explores how GTAs' behaviors can influence students' behaviors and in return, how students' behaviors further impact GTAs' behaviors.

V. LIMITATIONS

An important limitation of this study may be that two of the three observers were the facilitators for the weekly GTA preparation meetings, which may have had an impact on GTA behaviors. In addition, only GTAs were wearing microphones during observations, therefore only voices from GTAs and students close to GTAs were captured. Moreover, student group dynamics may also interact with GTA behaviors, but they were not documented in this study. Lastly, we did not report the nature of verbal interaction between GTA and students (e.g., scientific principles, data analysis).

ACKNOWLEDGEMENTS

This work is funded in part by NSF grant DUE 1725554.

[1] 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).

[2] L.C. McDermott and P.S. Shaffer, *Tutorials in Introductory Physics*, Prentice-Hall, Englewood Cliffs, NJ (1998).

[3] K. M. Koenig, R. J. Endorf, and G. A. Braun, Effectiveness of different tutorial recitation teaching methods and its implications for TA training, Phys. Rev. ST Phys. Educ. Res. **3**, 010104 (2007).

[4] J. B. Stang and I. Roll, Interactions between teaching assistants and students boost engagement in physics labs, Phys. Rev. ST Phys. Educ. Res. **10**, 020117 (2014).

[5] E. A. West, C. A. Paul, D. Webb, and W. H. Potter, Variation of instructor-student interactions in an introductory interactive physics course, Phys. Rev. ST Phys. Educ. Res. **9**, 010119 (2013).

[6] J. B. Velasco, A. knedeisen, D. Xue, T. L. Vickrey, M. Abebe, and M. Stains, Characterizing instructional practices in the laboratory: The Laboratory Observation Protocol for Undergraduate STEM, J. Chem. Educ., 93(7) (2016).

[7] D. K. Cohen and D. L. Ball, Instruction, Capacity, and Improvement; CPRE Research Report Series RR-43; Consortium for Policy Research in Education: University of Pennsylvania, Graduate School of Education, 1999.

[8] L. Darling-Hammond, Constructing 21st-century teacher education, J. Teach. Educ. **57**, 300 (2006).

[9] J. Chini and J. Pond, in *Proceedings of the Physics Education Research Conference, Minneapolis, 2014*, (2016), pp. 51–54.

[10] A. Elby et al., Open Source Tutorials in Physics Sense-making (2008).

[11] https://sites.google.com/site/scientificabilities/ISLE-labs.

[12] M. Wilcox, Y. Yang, and J. J. Chini, Quicker method for assessing influences on teaching assistant buy-in and practices in reformed courses, Phys. Rev. Phys. Educ. Res. **12**, 020123 (2016).

[13] J. J. Chini, C. L. Straub, and K. H. Thomas, Learning from avatars: Learning assistants practice physics pedagogy in a classroom simulator, Phys. Rev. Phys. Educ. Res. **12**, 010117 (2016).

[14] TI indicates that it is the GTA who initiates the interaction, but it does indicate whether the GTA asks a question or makes a statement. Therefore, 1o1-TPQ or 1o1-Talk is often co-coded with TI. If the GTA is still talking to the same student after a 2-min interval, TI will not be coded in the following 2-min interval but 1o1-Talk would continue to be coded.

[15] N. Wongpakaran, T. Wongpakaran, D. Wedding and K. L. Gwet, A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples, BMC Medical Research Methodology 13(1), 61 (2013)

[16] J. H. Ward, Jr., Hierarchical grouping to optimize an objective function, J. Am. Stat. Assoc., 58, 236 (1963**)**.

[17] https://cran.r-project.org/web/packages/factoextra/factoextra.pdf.

[18] We also tried Welch's ANOVA as it is robust when the assumption of homogeneity of variance is violated. The results from Welch's ANOVA do not exclude any of the codes shown in Fig. 1.

[19] M. Tomczak and E. Tomczak, The need to report effect size estimates revisited. An overview of some recommended measures of effect size, Trends in Sport Sciences, **1**(21) (2014), pp. 19-25.

[20] https://cran.r-project.org/web/packages/dunn.test/dunn.test.pdf.

[21] https://ggplot2.tidyverse.org/.

[22] J. M. Maher, J. C. Markey, and D. Ebert-May, The other half of the story: Effect size analysis in quantitative research, CBE Life Sci Educ. 12(3) (2013).

[23] M. Stains et al., Anatomy of STEM teaching in North American universities, Science, **359**, 8383.

[24] We note that half of the observations (36 sessions) were conducted in the mechanics course and the other half were in electromagnetism. In each course, six recitation/lab units were used during the observations. Therefore, we did not have enough data to test this hypothesis.

[25] L. B. Wheeler, J. L. Maeng, and B. A. Whitworth, Characterizing teaching assistants' knowledge and beliefs following professional development activities within an inquiry-based general chemistry context, J. Chem. Educ., 94(1) (2017).