Journal of Mathematics Teacher Education
https://doi.org/10.1007/s10857-020-09485-x

Bridging the gap between observation protocols and formative feedback

Sean Yee[1] · Jessica Deshler[2] · Kimberly Cervello Rogers[3] · Robert Petrulis[4] · Christopher D. Potvin[5] · James Sweeney[6]

Accepted: 28 December 2020

Abstract
In this study, we sought to identify how feedback about classroom observations affected novice university mathematics instructors' (UMIs) teaching practices. Specifically, we examined how a Red–Yellow–Green feedback system (RYG feedback) affected graduate student instructor (GSI) scores on an observation protocol (GSIOP). The protocol was developed specifically for this population, and both the GSIOP and RYG feedback were used within a peer mentoring program for GSIs, wherein novice GSIs were mentored by more experienced GSIs. Mentors observed novices' classrooms using the GSIOP and provided RYG feedback as part of observation–feedback cycles. We analyzed 100 sets of scores, each collected over the course of a semester containing on average three observation–feedback cycles. Analyzing the semester-long datasets longitudinally provided insight into what types of feedback informed and influenced observed teaching. After qualitatively coding the feedback provided to the GSIs by their mentors along multiple dimensions, we found certain forms of feedback were more influential for observable changes in GSIs' teaching. For example, pedagogical feedback that included contextualization (context and focal events) demonstrated a more positive change in GSIOP score than feedback that lacked contextualization. Our results suggest that contextual formative feedback has a positive change to student-focused and teacher-focused observations.

Keywords Graduate student instructors · Feedback · Observation · Mentoring · Observation protocol · Student-centered instruction

Introduction

In 2001, the International Commission on Mathematical Instruction published an important study entitled *The Teaching and Learning of Mathematics at The University Level* (Holton and Artigue 2001). In this study, there was a call for a new paradigm of teaching

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10857-020-09485-x.

✉ Sean Yee
yee@math.sc.edu

Extended author information available on the last page of the article

Published online: 21 January 2021

wherein university mathematics instructors (UMIs) focus on student learning in teaching and design practices. The authors argued that "Mathematical content of courses needs continuous reform, renewal and a close link between what students learn and what they will need in the future" (p. 8). This call for student-centered instruction has grown louder internationally with the popularity of inquiry-based mathematics education (IBME, Laursen and Rasmussen 2019) where the focus centers upon four pillars: (1) students engage deeply with coherent and meaningful mathematics, (2) students collaboratively process mathematical ideas, (3) instructors inquire into student thinking, and (4) instructors foster equity in their design and facilitation choices. The intent of these four pillars is to "account for student learning and thus offer guidance to instructors seeking to develop their teaching practice" (p. 138). Therefore, these four pillars provide guidance toward evidence-based effective teaching strategies that actively engage students. A natural question arises: How are we to support UMIs in this professional growth? More specifically, what components of professional development (PD) help UMIs incorporate these teaching practices (i.e., IBME pillars) into their teaching? To answer such queries, studies of PD programs need to include a method for measuring professional growth.

To provide such PD while measuring UMIs' development in using student-centered teaching techniques, we focused on a specific subset of novice UMIs in the USA, mathematics graduate student instructors (GSIs).[1] In the USA, it is common for UMIs to begin learning about teaching while in mathematics graduate programs. The two primary roles for which graduate students are hired to support the teaching mission of a mathematics department in the US are as an instructor of record for a course (GSI) or to lead a discussion or recitation section (Teaching Assistant, TA) in support of another instructor's course (Ellis et al. 2016b). The presence of PD programs for GSIs was also found to be a characteristic of successful instructional programs in college calculus for undergraduate students (Ellis 2015). A recent survey of all graduate-degree granting mathematics departments in the US found that 66% of responding PhD-granting and 33% of Master's-granting departments provide PD for graduate students specific to teaching mathematics (Ellis et al. 2016b). For the PhD-granting programs, the majority of PD programs were for all graduate students regardless of their specified teaching role. These occurred before the graduate students taught for the first time or during their first semester they were teaching with almost half of them consisting of a semester-long course or seminar. Moreover, what is done in each of these programs can vary widely but may include developing lesson plans, delivering mini lessons, being observed by experienced instructors, and learning about assessment practices (Ellis et al. 2016a). Thus, in the USA, GSIs are taking on the role of many novice UMIs in other countries with respect to teaching expectations.

Researchers have also found that GSIs are receptive to learning about and using student-centered teaching practices (Ellis 2014; Seymour 2005). There are multiple means of providing student-centered pedagogical support for GSIs (e.g., professional development, mentoring, pedagogically focused courses; Speer et al. 2005; Yee and Rogers 2017), but one that transfers internationally is observations and observational feedback. To determine if a novice UMI needs support within their teaching, observations are a natural way to provide direct, informed support. There is currently limited research on UMI teaching observation protocols and even less research on post-observation feedback (Reinholz 2017) at the university level. Multiple observation protocols exist to examine undergraduate

[1] GSI was used instead of TA (Teaching Assistant) because GSI references graduate students who are UMIs with the responsibilities of instructors of record.

classrooms (e.g., MCOP^2, RTOP, C-LASS, etc.), but many do not discuss how to make that observation formative so that feedback can be beneficial for the instructor and students through repeated use (Shute 2008). Specifically, many observation protocols are purposefully descriptive about how they are intended to be used to measure teaching behaviors, but there is little discussion of how to use them for ongoing support and professional growth.

To this end, we created an observation protocol (GSIOP, Rogers et al. 2019) and a post-observation Red-Yellow-Green feedback (RYG feedback)[2] structure at two universities to provide ongoing support for novices emphasizing student-centered teaching practices. The purpose of this paper is to help bridge the research gap between observations and post-observation feedback by identifying how feedback within a peer mentoring program[3] (Rogers and Yee 2018a; Yee and Rogers 2017) informed and influenced future observations. Our research questions for this study are:

(1) In what ways (if any) was feedback provided through the RYG feedback structure associated with changes in observed teaching practices throughout a semester, as shown by changes in GSIOP scores?
(2) How do those changes inform (if at all) methods for providing formative feedback to influence observed teaching?

Related literature

For over a century, psychologists have examined the importance of feedback as a means to change performance, cognition, and understanding in many professions (Kluger and DeNisi 1996). Hattie and Timperley (2007) define feedback as "information provided by an agent (e.g., teacher, peer, book, parent, experience) regarding aspects of one's performance or understanding…that seeks to provide knowledge and skills or to develop particular attributes" (p. 102). In this study, mentors are providing feedback to novices via knowledge and skills for teaching.

Research suggests feedback has two purposes from the recipient's view: directive and facilitative (Black and Wiliam 1998; Gamlem and Smith 2013; Shute 2008). Directive feedback tells the recipient what can be changed or fixed and is embedded in our use of red comments which mandate a direct suggestion on how to address an issue identified by the mentor. Facilitative feedback guides the recipient on their own revisions or conceptualizations and aligns with our yellow comments where the focus is on asking questions of the novice to support them in resolving issues for themselves instead of having the mentor make direct suggestions (Appendices F and G in ESM). In addition to the directive and facilitative feedback, we provide encouraging feedback as green comments to help frame our feedback to the novice as supportive and to recognize improvements in their use of student-centered teaching strategies.

[2] GSIOP is copyrighted by Bowling Green State University (2018). All rights reserved. Rogers and Yee (2018b).

[3] Supported by Collaborative National Science Foundation Grants (NSF DUE 1544342, 1544346, 1725295, 1725230 and 1725264).

Primary and secondary teacher feedback literature

In the primary grades, mathematics coaches observe and provide feedback as part of coaching cycles, where a mathematics specialist works with an elementary school to provide support and feedback to elementary teachers regarding their teaching (Gibbons and Cobb 2017; Gibbons et al. 2017). Gibbons and Cobb's (2017) meta-analysis of highly cited (> 700 citations) coaching literature defines the role of coaches who emphasize student-centered instruction:

> Effective coaching is not a one-way process in which coaches impart technical skills to teachers. Instead, coaches support teachers in addressing problems of practice by engaging them in activities that focus on key disciplinary ideas, how students learn those ideas, and pedagogical principles to support students' learning. (p. 413)

Gibbons and Cobb generated five characteristics of high-quality professional learning: ongoing and intensive, a focus on teacher problems, focus teachers on student thinking, foster communities of practice, and involve pedagogical investigation or enactment. The goals of our peer mentoring program for novices aligns with these five characteristics. For example, the GSIOP was designed to have a student-focused section and teacher-focused section with significant consideration given to student collaboration. Gibbons and Cobb also found that these characteristics provided a way to identify six productive coaching strategies (analyzing classroom video, engaging in the discipline with other teachers, examining student work, engaging in lesson study, co-teaching, and observing instruction) which we also find overlap with the roles of the mentors in our program, as they analyze classroom video, engage other teachers (novices), look at student work, and this is all done through observing instruction.

At the secondary level, observation–feedback cycles are common among novice and pre-service teachers. It has become widely accepted that mentoring and induction methods for teaching are beneficial to teacher retention, success, and self-efficacy (Portner 2005). Moir's (2005) research has generated novice instructors' individual learning plans and regular observation protocol structures for secondary schools (orchestrated by a mentor teacher) to help novice instructors work with mentor teachers to have realistic, tailored, and sustainable induction programs. Consequentially, teachers are retained and provided opportunities to grow (Johnson and Kardos 2002; Kastberg et al. 2018). Hollingsworth and Clarke (2017) research found secondary teachers in Australia valued methods of feedback that allowed them to take ownership of the feedback through discussion and collaboration with the observer.

When comparing primary teacher education's use of coaches with the secondary teacher induction/mentoring program, there are significant overlaps. Gibbons and Cobb's (2017) need to support teachers through regular feedback, and observation is also discussed in Portner's (2005) description of critical factors to quality mentor programs. Although primary and secondary teacher education research advises that such observation feedback is valuable, multiple studies indicate that this feedback is specific to the grade-level, content, and school structure (Kastberg et al. 2018; Portner 2005). Thus, it is important to look to the higher education environment to better tailor instruments for observation feedback for the UMI community.

Importance, complexities, and limited research of GSI feedback

Studies in mathematics education in the USA provide a current picture of UMIs' preparation and the critical value GSI professional growth plays within teaching (Speer and Murphy 2009). With over 200,000 undergraduate students taking courses from mathematics GSIs per semester in the USA, there can be little doubt that mathematics GSIs significantly impact undergraduate courses and students (Belnap and Allred 2009; Lutzer et al. 2007). GSIs have been identified as a key component of success for collegiate mathematics departments for teaching undergraduate mathematics (Bressoud et al. 2015). Similar to international UMI teacher training, GSI education on teaching varies significantly in the USA from a three-day orientation to three semester-long courses depending on the university (Speer et al. 2005). As a result, mathematics departments and research in undergraduate mathematics education continue to focus on methods for supporting and improving GSIs' professional growth in general, and student-centered instructional practices in particular (Rogers and Yee 2018a; Speer and Murphy 2009; Yee and Rogers 2017).

Although mathematics education research has robustly studied the use of feedback within coaching, mentoring, and induction methods at the primary and secondary education levels, our review of the literature found few studies focusing on UMIs' feedback on teaching, specifically GSI peer feedback (Reinholz 2017; Rogers and Steele 2016; Yee and Rogers 2017; Rogers and Yee 2018a). One exception is a recent study by Reinholz (2017) that explores peer feedback with mathematics GSIs observing one another. Reinholz found that feedback not only helped the novice, but enhanced teacher noticing and reflection in the observer, aligning with Reinholz's previous work (2016) where peer assessment led to improved self-assessment. Rogers and Steele (2016) concluded that novice instructors struggle to discuss teaching practices, which Reinholz (2017) argues could be aided by peer feedback. Results from these two studies endorse post-observation feedback as a means of improving GSIs' teaching through discourse and reflection.

Reinholz (2017) reminds us that "how instructors engage with peer feedback is complicated" (p. 7) due to GSIs' beliefs about mathematics and their often-assumed association between mathematics and intelligence. Kluger and DeNisi's (1998) meta-analysis of studies on feedback interventions showed that while overall feedback improves performance, it can also sometimes reduce performance, depending on the type of feedback and means by which it is delivered. Certain feedback was helpful for improving performance as long as attention was directed toward task motivation and task learning rather than praise, negative criticism, or focus on the person. Fundamentally, feedback was found to be most effective when it focused on the task, which in this study would be student learning. This research is helpful for novice UMIs because it suggests the focus should remain centered on student learning within the task instead of judging the performance of the UMI. This provides support for a student-centered observational protocol to help identify feedback for student learning.

Formative feedback

Just as research has demonstrated the value of formative assessment of student learning (Wiliam and Black 1996), a similar value is recognized for formative feedback that provides formative assessment of teaching. Hattie and Timperley (2007) found that summative assessments were often "devoid of effective feedback to students or to teachers" (p. 102).

Table 1 Formative feedback for teacher–student dynamic compared to mentor–novice dynamic

Guidelines for teacher–student formative feedback (Shute 2008, pp. 177–178)	Paralleled guidelines for mentor–novice teacher formative feedback
Focus feedback on the task, not the learner	Focus feedback on the novice's teaching, not the novice
Provide feedback after learners have attempted a solution	Provide feedback after novices have taught
Be specific and clear with feedback message	Provide specifics from the observation to clarify and justify suggestions
Reduce uncertainty between performance and goals	Ask novice teacher before observation what they should watch for and share observation protocol
Give unbiased, objective feedback, written or via computer	Make sure to write and deliver a physical copy of the RYG feedback with supporting evidence
Promote a "learning" goal orientation via feedback	Ask the novice for the measurable goals they are teaching toward
Provide elaborated feedback to enhance learning	Mentors complete GSIOP and additional comments collected from observation, but focus on sharing RYG feedback
Keep feedback as simple as possible but no simpler	Parse out feedback via focused and clear red–yellow–green comments
Present elaborated feedback in manageable units	Limit red–yellow–green comments to sizes that can be worked on for the next observation

To be effective, feedback on teaching should not be a summative judgment based only one observation. Hattie and Timperley's conclusions encourage the need for iterative observation–feedback cycles where the emphasis is on growth and learning shared by both participants. One of the founding pillars for our peer mentoring program was that novices need support to continually grow and improve.

Shute's (2008) research defines formative feedback as "information communicated to the learner that is intended to modify his or her thinking or behavior to improve learning" (p. 154) and identifies multiple formative feedback guidelines to enhance learning while recognizing that a multidimensional view of feedback is necessary due to situational and individual characteristics. These guidelines for formative feedback that demonstrate improvement in student learning align well with our independently generated observation–feedback cycles. Table 1 shows parallels between Shute's (2008) guidelines and our (formative) feedback structure.

The first four parallels in Table 1 demonstrate how our observation feedback can smoothly align with student–teacher formative feedback. The last five parallels emphasize how our RYG feedback is formative because it distills the observation data down to useable feedback, limits the number of suggested red, yellow, and green comments to a manageable number and emphasizes evidence-based suggestions for clarification.

Specificity, focal events, and contextualization of formative feedback

From Shute's (2008) guidelines on feedback, a critical dimension of Shute (2008) significant to our research questions is *specificity*. Although Table 1 supports the need for clear feedback, how specific one should be with details is complicated. Shute states, "providing

feedback that is specific and clear, for conceptual and procedural learning tasks, is a reasonable, general guideline. However, this may depend on other variables, such as learner characteristics and different learning outcomes" (p. 158). If a mentor provides feedback to a novice that focuses on a specific method of specific content taught that day, then the novice may not be able to do anything productive with that feedback if it will not apply to subsequent lessons. For example, a mentor might say:

> The next time you introduce logarithms, you may want to provide opportunities for students to explore the additive property before the exponential property. I have found that this ordering of the content helps my students better retain the information.

Despite the specificity of this feedback with respect to the lesson and the content, as the novice has already taught this content (and most likely teaches only one class a semester), such feedback is not helpful for the novice improving their teaching during the rest of this semester.

If a mentor discusses *why* the additive property works better for the learner before the exponential property, and *how* the novice can integrate this suggestion into other upcoming class topic, the feedback can be more actionable (Cannon and Witherspoon 2005), providing the novice with insight into upcoming instruction. Nilsson and Ryve (2010) emphasize that contextualization that addresses the *why* and *how* must include *focal events*. A focal event aims to steer the observer's attention toward the central issue (p. 245). In linguistics, a focal event is the part of the reasoning that stands out as salient.

> One key way in which context and focal event differ is in their perceptual salience. Generally the focal event is regarded as the official focus of the participants' [observers'] attention, while features of the context are not highlighted in this way, but instead treated as background phenomena. The focal event is placed on center stage, while context constitutes the stage itself. In line with this, the boundaries, outlines, and structure of the focal event are characteristically delimited with far more explicitness and clarity than are contextual phenomena. (Goodwin and Duranti 1992, p. 9)

Thus, in our logarithmic example, the focal event is not the logarithmic properties, but the ordering of properties for student understanding. The focal event provides a way of organizing, conceptualizing, and describing the observer's perception and intention of the context to others. To discuss the "why" illustrated in the logarithmic example, it is important to communicate the observer's focal event relative to the context of the novice's classroom. A mentor could instead suggest:

> When I have taught logarithms, I found students associated the additive property of logarithms with adding exponents. In your upcoming lesson on trigonometric identities, I would suggest identifying prior properties students seem to understand, such as the unit circle, to anchor your lesson content.

The focal event, prior understanding of students, is more explicit and takes what Goodwin and Duranti reference as "center stage." This feedback example clarifies the need for not only context but also a focal event. Thus, we posit that using the linguistic framework of contextualization with focal events (Nilsson and Ryve 2010; Goodwin and Duranti 1992) is a viable qualitative means of coding specific feedback. To this end, we analyzed mentor feedback through the lens of broad and specific (Shute 2008) with the specific feedback identifying a context and focal event, and broad feedback lacking either context or a focal event.

Table 2 Mixed-method triangulation design convergence model for GSI observational study

Research question	Data collected	Data analysis
In what ways (if any) was feedback provided through the RYG feedback structure associated with changes in observed teaching practices throughout a semester, as shown by changes in GSIOP scores?	**GSIOP Scores** Three GSIOPs per novice per semester Each GSIOP had a student-focused and teacher-focused section and score	**Coded** longitudinal trends of GSIOP scores through a full semester as Decrease, Steady, Moderate increase, Substantial increase, Hill, Valley **Triangulated** coding of RYG Feedback with GSIOP longitudinal trends
How do those changes inform (if at all) methods for providing formative feedback to influence observed teaching?	**RYG Feedback** **GSIOP Scores** Three GSIOPs per novice per semester Each GSIOP had a student-focused and teacher-focused section and score	**Coded** RYG Feedback along three dimensions Student/Teacher-focused Broad/Specific Advice/Improvement **Triangulated** coding of RYG Feedback with GSIOP Scores by comparing GSIOP scores with varying types of feedback

Methods

To answer our research questions, we used a mixed-methods triangulation design with a convergence model (Creswell and Clark 2017). This mixed-method research design is appropriate because qualitative and quantitative data were used together (data was not sequential or embedded but convergent in the analysis) for the purpose of triangulating the results of our feedback. To answer the first research question, we quantitatively analyzed changes in GSIOP scores over the course of a semester and qualitatively coded their longitudinal trends to answer our first research question about how the feedback structure was associated with changes in classroom observations. To answer our second research question, we qualitatively coded the RYG feedback for types of formative feedback and used the results with the changes in GSIOP scores to triangulate how feedback influenced observed teaching practices. Table 2 illustrates how data were collected and analyzed relative to each research question.

Context of study

The initial goal of our peer mentoring program was to provide feedback and facilitate discussions among novices around student-centered teaching strategies to improve undergraduate mathematics instruction (Rogers et al. 2019). The GSIOP and RYG feedback were implemented as part of a peer mentoring program (Rogers and Yee 2018a; Yee and Rogers 2017) where novice GSIs (novices) were mentored by experienced GSIs (mentors). The program had multiple components, but the work in this paper focuses on data collected through classroom observations (by using the GSIOP) and RYG feedback meetings to specifically look at observed teaching practices. Scenarios, role playing, video recordings, and live observations helped prepare mentors to provide feedback in the mentor PD (e.g., Appendix G in ESM). Although the context for this study was a program developed

specifically for GSIs, we believe the observation protocol and post-observation feedback structure are applicable to any UMI, especially novice UMIs.

Participants

The participants for this study were 10 mentors and 32 novices from two universities in the USA over two semesters. The mentors were experienced GSIs (GSIs who had previously taught as full instructor of record for at least a year) who applied to be part of the mentoring program. They had a vested interest in helping others improve their teaching and participated in a semester-long professional development (PD) program before becoming mentors. The PD program included training on how to use the GSIOP, how to use the GSIOP data to develop RYG feedback, and how to facilitate post-observation conversations using the RYG feedback. The novices were in their first or second year as instructor of record, and new novices were added between semesters while other novices transitioned to other responsibilities after one semester. Therefore, some novices were participants in both semesters, while others were only participants for one semester, and the result was a total of 50 mentor–novice pairs during the two semesters.

Data sources

For this study, we used RYG feedback, GSIOP scores, GSIOP comments, and observation summaries collected from the mentors. Most novices were observed three times, but certain restrictions (e.g., changes in instructor assignments during the semester) resulted in two novices being observed only twice each, and three novices were observed four times for extra guidance and help. GSIOP comments and observation summaries were used to verify and justify coding. Below we describe the GSIOP data and RYG feedback in detail for purposes of coding.

GSIOP data

The GSIOP was created by tailoring previously developed mathematics observation protocols that focused on student-centered instruction for novice UMIs (e.g., GSIs). Specifically, we modified the already-established Mathematics Classroom Observation Protocol for Practices (MCOP2, Gleason et al. 2017). The MCOP2 was designed for K-16, originating from the STEM-based Reformed Teaching Observation Protocol (RTOP, Sawada et al. 2002), but unlike the RTOP, the MCOP2 includes a means to observe student-centered investigations and collaborative learning environments focusing on mathematics (aligning smoothly with IBME). Thus, we modified the MCOP2 to be applicable for use when observing GSIs and to be used with both native and non-native English-speaking GSIs. Similar to the MCOP2, the GSIOP contains items that are rated on a scale from 0 to 3 in four sections: cover page (necessary communication skills for domestic and international GSIs within the US), student engagement (observer focused on students), teacher facilitation (observer focused on teacher), and lesson design practices (observer focused on lesson design). A thorough explanation of the GSIOP design can be found in Rogers's validation study (2019, see Appendix E in ESM for the full GSIOP). We refer to the student engagement section as *student-focused* and the teacher facilitation section as *teacher-focused* for the remainder of this paper. As our study emphasized student-centered instruction and RYG feedback, we

Table 3 Graduate student instructor observation protocol student- and teacher-focused items

Student-focused items	Teacher-focused items
A. Students engaged in exploration/investigation/problem solving	E. The teacher promoted precision of mathematical language
B. Students used a variety of means (modeling, drawings, concrete materials, manipulatives, etc.) to represent concepts	F. The teacher's questions encouraged student thinking
C. Students critically assessed mathematical strategies	G. In general, the teacher provided wait time
D. Students were involved in the communication of mathematical ideas to others (peer-to-peer)	H. The teacher uses student questions/comments to enhance conceptual mathematical understanding
	I. The teacher incorporates formative assessments (e.g., polling class, exits slips, quick check-in problems) to gauge student understanding during the lesson
Total possible score 0–12	Total possible score 0–15

focused on these two sections of the GSIOP and omitted the cover page and lesson-focused section because they were lesson dependent (See Appendix E in ESM) and did not provide as much insight into methods of student-centered instruction as the student-focused and teacher-focused sections. We summed the questions on the GSIOP student-focused section (4 questions) and the GSIOP teacher-focused section (5 questions) separately. Thus, for each observation of each novice, there were a teacher-focused GSIOP score and a student-focused GSIOP score. The questions from each section are shown in Table 3, and each item in each section could be scored from 0 to 3.

RYG feedback data

The feedback structure for this program dictated that mentors identify key points from the GSIOP that they could summarize for the novice in three categories: teaching practices the novice is doing well (green), teaching practices the novice could work on (yellow), and teaching practices the novice needs to address (red). The choice for a three-tier framework for feedback parallels Roller's (2016) research study that used three types of feedback for prospective secondary mathematics teachers and separated the feedback by time (encouraging, quick fixes, larger issues that take time, p. 482). Mentors were trained to provide manageable feedback by providing limited (2–3) yellow and red comments per observation and they were encouraged to ask questions and collect information about the classroom for purposes of post-observation discussion, but it was not mandated that mentors needed to reference prior observations in RYG feedback. In addition, mentors treated each semester of RYG feedback as independent from prior semesters' because novices may have taught different courses and classes from semester to semester. Post-observation meetings occurred within a week of the observation. Mentors printed out physical copies of the RYG feedback (not just digital) and gave them to the novices so that (1) the mentors had to discuss all RYG feedback and (2) the novices had this copy for their records for their next observation.

Table 4 Categories of changes in observation scores

Category	Description
Decrease	Each observation was at least 2 points lower than the previous one
Steady	Each observation was within one point of the previous one
Moderate increase	Each observation increased and the final was at least two point higher than the first
Substantial increase	Each observation increased and the final was at least three points higher than the first
Hill	Middle score was higher than both other scores and at least two points higher than one of the other scores
Valley	Middle score was lower than both other scores and at least two points lower than one of the other scores

Data analysis

To answer our first research question, we analyzed semester-long changes by finding the differences between GSIOP scores at different points in the semester as well as the overall difference between beginning and end of semester observations. We categorized semester-long changes to GSIOP scores as shown in Table 4.

To answer the second research question, we looked at RYG feedback, GSIOP comments, and observation summaries for advice on teaching through the whole semester that focused on student learning or teacher facilitation to align with the student-focused and teacher-focused sections of the GSIOP. We looked for feedback relevant to the current observation and comments that referenced a noted change based on feedback given in a previous observation that semester. Feedback related to the current observation was coded as *advice*, if it included suggestions of what a novice could do differently, not just indication of something they did poorly. Comments that referenced changes and growth over multiple observations were coded as *improvement*. It is important to note that comments were only coded as improvement if the mentor mentioned changes from prior advice. The focus of these codes is on mentors' feedback, not actual improvements.

Finally, we coded each piece of advice and each noted improvement as *broad* or *specific*. To frame broad versus specific objectively, we used Nilsson and Ryve's (2010) linguistic frame to define *specific* to include a context and focal event. If either a context, or a focal event were not included in the comment, then comment was coded as broad for our coding scheme. We then coded each semester-long set of RYG feedback as having broad advice if it contained broad advice but no examples of specific advice, and specific advice if there was at least one example of specific advice. Similarly, we defined broad improvements and specific improvements for semester-long RYG feedback. Coding advice and improvement as broad or specific provides a categorization of the semester-long feedback described with examples in Table 5.

Advice Without Improvement (AWI) implied advice was given, but improvement was not coded in subsequent observations. AWI could include advice that was coded as broad, specific, or both. No Advice Nor Improvement (NANI) was used when a mentor's feedback lacked advice, and therefore, no improvement could be noted in subsequent observations.

To verify the qualitative coding of advice and improvement as broad or specific, after each research assistant qualitatively coded the results according to Table 5, two additional researchers went back and corroborated the coding by comparing 50% of the

Table 5 Qualitative coding scheme for feedback across an entire semester

Code	Description	Example
SASI	**Specific Advice Specific Improvement**: Feedback included at least one contextualized suggestion with a focal event the novice could take to improve their teaching. In subsequent observations, the mentor noted that the novice had addressed the issues through particular context and focal events	"Elaborate with the material and explain the importance of the concept. For example, one instance in which you could give a little more insight and explanation was when the student used $P(A \cup B) = P(A) + P(B) - P(A \text{ cap } B)$"… (*later observation*) "You elaborated more than last time.. I felt that this was the perfect amount of elaboration. Also, you asked well thought out questions, and you rarely missed good opportunities to ask further questions."
BASI	**Broad Advice Specific Improvement**: Feedback included suggestions without context or a focal event to improve the novice's teaching. In subsequent observations, the mentor noted that the novice had addressed the issues through particular context and focal events	"Have tiny bits of student involvement through to keep students engaged" … *(later observation)* "Student questioning chosen was very effective in engaging students [with 2^x and log_2(x)]"
SABI	**Specific Advice Broad Improvement**: Feedback included at least one contextualized suggestion with a focal event the novice could take to improve their teaching. In subsequent observations, the mentor noted that the novice had improved upon previous issues, but without referencing specific context or focal event	"I encourage you to give more wait time before answering the questions yourself, this can have them participate more" … (*later observation*) "I saw great improvement since last time with student engagement… (*later observation*) "Great student interaction with the activity"
BABI	**Broad Advice Broad Improvement**: Feedback included suggestions without context or a focal event to improve the novice's teaching. In subsequent observations, the mentor noted that the novice had improved upon previous issues, but without referencing specific context or focal event	"Student engagement should be addressed" … (*later observation*)"Even though she ask[ed] many questions, students are not really active in this particular class"… (*later observation*). "She did not just answer but encourage[d] students to respond"
AWI	**Advice Without Improvement**: Feedback included suggestions, but the suggestions did not appear to be noted throughout the subsequent observations	"For the next time, I hope that he can get more active participation during his lecture portions" *No follow up or reference to previous observations*
NANI	**Neither Advice Nor Improvement**: Feedback was either statements extolling the novice's instruction or platitudes on teaching. Mentor did not provide advice nor improvement in subsequent observations	"He did a great job in his lesson of engaging the students, explaining material adequately and also giving his students problems to work on at the end of class." *No advice*

Table 6 Summary of vignettes

Novice	Mentor	Course	Student-focused or teacher-focused	GSIOP score observation 1	GSIOP score observation 2	GSIOP score observation 3	Change category	Change in GSIOP Score (final–initial)	Feedback
Indigo	Jason	Business Calculus	Student	5	6	5	Steady	0	AWI
Indigo	Jason	Business Calculus	Teacher	8	9	8	Steady	0	NANI
David	Hugo	Pre-Calculus	Student	1	4	6	Sub. Inc.	5	SABI
David	Hugo	Pre-Calculus	Teacher	8	10	8	Hill	0	BABI
Sarah	Mickey	Introduction to Statistics	Student	4	5	0	Hill	−4	AWI
Sarah	Mickey	Introduction to Statistics	Teacher	7	10	13	Sub. Inc.	6	BASI
Wendy	Roberto	Introduction to Statistics	Student	3	9	11	Sub. Inc.	8	SASI
Wendy	Roberto	Introduction to Statistics	Teacher	6	10	12	Sub. Inc.	6	SASI

observations and feedback artifacts. Interrater agreement was initially 94% and after discussion of the coding discrepancies, researchers agreed on the appropriate coding for the remaining 6%.

Vignettes

To illustrate the qualitative coding in more detail, we provide four vignettes with the complete RYG feedback for each vignette in Appendices A–D in ESM. In each vignette, we see the RYG feedback provided by a mentor to a novice, and how each piece of advice and noted improvement, if present, as *broad* or *specific*. We look at four mentors' (Jason, Hugo, Mickey, and Roberto) RYG feedback to four novices (Indigo, David, Sarah, and Wendy, respectively). Table 6 summarizes the RYG feedback and the analysis of the RYG feedback as student-focused and teacher-focused for each of the four vignette's GSIOP data and RYG feedback data.

To speak clearly and directly to comments provided by mentors, we will reference specific feedback found in Appendices A–D in ESM using the abbreviation Obs#ColorLetter. For example, Obs3GreenB of Appendix A in ESM reads "Presented work is very clear and easy to follow" and represents the second green comment from Observation 3.

Jason's feedback for Indigo (Appendix A in ESM)

Jason's student-focused feedback for Indigo We identified Jason's student-focused feedback as advice without improvement (AWI). Jason provides broad advice in Obs1RedA, where he suggests having students do part of the problem. In fact, all of Jason's advice is broad because it lacks context. The emphasis on 'having more student interaction' is seen again in Obs2YellowA (without reference to the first observation), but omits any specific context from the observation, keeping it broad and hard for Indigo to implement because Jason does not indicate how the students could "do more." We see the same comment again in Obs3RedA where the emphasis is again on interaction, but does not reference specific context, nor reference previous observations where this could be discussed. Additionally, we see Jason makes comments that could be interpreted as advice with Obs1YellowA and Obs1YellowB so that the instructor is making content accessible to the student with respect to the edge of the board and content consistency, but Jason omits context for these statements. We see more of the same advice in Obs3YellowA and Obs3YellowB, but Jason again does not reference prior observations for improvement.

Jason's teacher-focused feedback for Indigo We identified Jason's teacher-focused feedback as Neither Advice Nor Improvement (NANI). Jason's green comments in all three observations provide compliments that are broad and focused on presentation. However, these do not illustrate advice, nor are any comments referencing prior observations, which would lead to discussion of improvement. Moreover, the comments are broad compliments about a job well done, but with few details to let Indigo know what specific situations are valuable to repeat. Additionally, as described in the data analysis section, suggestions to "avoid" (Obs1YellowA) or "Don't write" (Obs3YellowA) were not followed up with what the novice *should* do and thus were not coded as advice.

Jason's feedback was consistent in emphasizing the same points, broadly, in each observation. Jason's red feedback consistently suggested a focus on student feedback and interaction (student-focused page of the GSIOP) but lacked guidance on how to implement them directly. Moreover, Jason's yellow comments were often suggestions (i.e., Try, Do,

Don't) without questions and openness that would provide space for Indigo to grow into those suggestions and understand where the suggestions are originating, other than Jason's teaching preferences. If we consider Indigo's perspective and look at the feedback from any single observation, it is intimidating to meet the expectations Jason is putting forward, even with the manageable size of the yellow and red comments, due to the lack of direction and open-ended questions in the feedback. Indigo's student-focused and teacher-focused part of the GSIOP both remained steady and had no increase or decrease in score from the beginning to the end of the semester.

Hugo's feedback for David (Appendix B in ESM)

Hugo's student-focused feedback for David We identified Hugo's student-focused feedback as specific advice with broad improvement (SABI). Hugo provides specific advice in the first observation emphasizing student engagement (Obs1YellowA, Obs1YellowB) and formative assessment for student feedback (Obs1RedA) by giving context and focal events via examples of methods of engagement and formative assessment that David can use in the future. Hugo reinforces engagement through exploration and more formative assessment in the second observation (Obs2YellowA, Obs2YellowC). When looking at student-focused feedback for improvement from prior observations, we see Hugo compliments David on learning students' names (Obs2GreenE). We also see that Hugo continues to encourage student-focused methods explicitly from the GSIOP (Obs3YellowA). However, all three of these improvements do not speak to specific contexts from the observation, but rather that they were observed by Hugo somewhere in his observation and Hugo recognizes the improvements have been accomplished or are continuing to be attempted.

Hugo's teacher-focused RYG feedback for David We identified Hugo's teacher-focused feedback as broad advice with broad improvement (BABI). In Obs2YellowB, Hugo suggests motivating the material and he speaks broadly about how it is useful in group work or discovery but does not give an example that demonstrates to David how to motivate the content directly. Similarly, in Obs3RedA, Hugo advises to challenge students before the final exam, but suggests this through a broad, vague statement by stating "let them test their own understanding of the subject" without any examples or focal events. Thus, Hugo's advice is broad because it lacks focal events. Hugo's teacher-focused comments that indicate improvement can be seen with encouragement with David's board work (Obs2GreenD), but how or why his boardwork has become "better" demonstrates a lack of a focal event. Hugo's suggestion on how to continue improving (Obs3YellowA) provides focal events but lacks context for David.

Obs2GreenB and Obs3GreenA indicate that David was focusing on students, but David was still emphasizing his presentation of the content over student involvement. We notice that Hugo pushed hard for formative assessment for student involvement with red comments in the first observation, but then reduced the suggestion about formative assessment to yellow in later observations with more specific and curated variations on formative assessment. This transition to yellow comments is a good example of facilitative feedback because in Obs2YellowC Hugo is leaving the suggestion open for David to determine the best way to include formative assessment. Specifically, Hugo says, "If you don't like the rigidity of that [suggestion] another idea…" to facilitate for David to continue to grow. David's student-focused part of the GSIOP substantially increased by five points over the semester while his teacher-focused GISOP score had a higher middle score (hill) but had an overall change of zero points.

Mickey's feedback for Sarah (Appendix C in ESM)

Mickey's student-focused feedback for Sarah We identified Mickey's student-focused feedback as advice without improvement (AWI). Mickey provided Sarah with many comments focusing on her teaching, but not as many suggestions for student engagement. Obs2YellowA provides broad advice suggesting a means to receive student feedback and engagement for formative assessment. Mickey also provided student-focused advice in Obs3YellowA which directed Sarah to consider the student's view of the material and her role in adjusting that view. No improvement with student-focused feedback was coded.

Mickey's teacher-focused RYG feedback for Sarah We identified Mickey's teacher-focused feedback as broad advice with specific improvement (BASI). Mickey provides broad advice in Obs1YellowA because he does not provide context to clarify the terms "jumping around" and "disorganized," which is especially important for non-native English-speaking novice GSIs. Similarly, Obs1YellowB is broad advice because Mickey does not clarify the context for Sarah to understand how to modify the use of inflection in her voice. We see specific improvement referenced in Obs1GreenA and Obs2GreenA where Mickey complements Sarah on moving away from presentation slides with respect to the context of the previous year.

It is important to note that if this had been a different novice (e.g., Hugo and David), some of the comments on student engagement could have been moved to red. Specifically, Mickey could have made Obs2YellowA a red comment, but declined after considering the background of this novice. We want to emphasize that while we provide general guidelines for when something is a red, yellow, or green comment, Sarah provides a good example why such RYG codes are relative to the novice and mentor. Sarah's student-focused GSIOP score grew by one point between the first two observations but dropped significantly (hill) ending in a deficit while her teacher-focused GISOP score had a substantial increase with an overall point change of six. It is important to note the feedback type and GSIOP score for Sarah because the GSIOP score and feedback type were so different when it came to teacher-focused versus student-focused. Sarah's growth in her teacher-focused GSIOP score may have been associated with Mickey's BASI teacher-focused feedback, while Sarah's decrease in her student-focused GSIOP score may have been associated with Mickey's AWI student-focused feedback.

Roberto's feedback for Wendy (Appendix D in ESM)

Roberto's student-focused feedback for Wendy We identified Roberto's student-focused feedback as specific advice with specific improvement (SASI). Roberto provided specific student advice in Obs1YellowA and Obs2YellowB. In both of these cases Roberto provides context and focal events in which his comments are forward-thinking to Wendy's next class. Roberto also provides student-focused broad advice with Obs3YelowB and providing ideas for how to help Wendy in her next semester of teaching. Roberto's refers to previous observations for student improvement in Obs2GreenC. Notice in this case how Roberto provides encouragement and then references yellow comments (Obs2YellowB) for context for further growth, encouraging and challenging Wendy simultaneously.

Roberto's teacher-focused feedback for Wendy We identified Roberto's teacher-focused feedback as specific advice with specific improvement (SASI) as well. Roberto does provide broad advice where context is not shared in Obs1YellowB and Obs1YellowC, but then

provides specific advice Obs2YellowA with context and the focal event to provide "insight and explanation." Roberto speaks to broad improvements in both Obs2GreenB and Obs-3GreenA with the reinforcement in confidence without context. However, Roberto then provides teacher-focused specific improvement with Obs3GreenB. Roberto notices Wendy's attempt at elaboration and gives specific context (confidence intervals) where he saw improvement and how that improvement is helpful (focal event).

Roberto regularly provided encouraging feedback that referenced improvements and then referenced yellow comments to continue to encourage Wendy and provide deeper feedback than just complements. Roberto consistently applauded and then provided further avenues for improvement on topics such as confidence, elaboration, and student engagement/participation. Roberto recognized Wendy's growth. Wendy's student-focused GSIOP score substantially increased by eight points and her teacher-focused GSIOP score grew by six points. Both GSIOP scores continued to grow through all three observation. Finally, Roberto's Obs3YellowB comment looks to Wendy's future, which provided Wendy with the encouragement that aligns with Roberto's emphasis on confidence throughout all three observations.

Results

We provide a quantitative analysis for longitudinal changes in GSIOP score from the student-focused section and the teacher-focused section to answer our first research question. We did not analyze the number of Red vs. Yellow vs. Green comments because the color coding was dependent upon the mentor. Table 7 shows how many of these semester-long changes fell into each category (substantial increase, moderate increase, steady, decrease, hill, valley) for both the student-focused and teacher-focused sections of the GSIOP.

Table 7 shows nine novice GSIs had a substantial increase in GSIOP score in the student-focused section and twelve students had a moderate increase in GSIOP score in the student-focused section, while nine novice GSIs had a substantial increase in GSIOP score in the teacher-focused section and 14 novice GSIs had a moderate increase in GSIOP score in the teacher-focused section. Looking more closely at the disaggregated data, the student-focused section increased 0.72 points (0.72 out of 12 points, which is an increase of 6.00%) while the teacher-focused section increased an average of 1.30 points (1.30 out of 15 points, which is an increase of 8.67%). Together these two average GSIOP score changes of student-focused (1.30 points) and teacher-focused (0.72 points) gives a total average change of 2.02 points per GSIOP, which averages to 1.01 points per section per novice per semester. Although there was an overall GSIOP score growth of only 1.01 points per section, the larger amount of the feedback growth stemmed from teacher-focused section (8.67% growth) versus the student-focused section (6.00% growth).

We see that the number of datasets that fell into substantial increase, moderate increase, steady, hill, and valley GSIOP change categories had a fairly equal distribution between student-focused and teacher-focused sections, while there were twice as many decreases (10) for the student-focused sections than there was for the teacher-focused Sections (5). Although many of the GSIOP scores remained steady (33 out of 100), there were significantly more novices whose scores increased moderately or substantially (44 out of 100) than decreased (15 out of 100) over a semester. Thus, our results indicated that there was an observed change in teaching throughout a semester

Table 7 Longitudinal semester-long changes in GSIOP scores by student-focused and teacher-focused sections

GSIOP change categories	Substantial increase	Moderate increase	Steady	Decrease	Hill	Valley	Total
Number of student-focused sections	9	12	15	10	2	2	50
Average GSIOP change per student-focused section	5.00	2.50	0.20	− 3.90	− 1.00	− 0.50	0.72
Standard deviation per student-focused section	1.63	0.87	1.11	2.02	3	1.5	3.3
Number of teacher-focused sections	9	14	18	5	3	1	50
Average GSIOP change per teacher-focused section	5.11	2.21	0.28	− 3.60	0.67	− 1.00	1.30
Standard deviation per teacher-focused section	1.29	1.47	0.87	2.58	0.94	0	2.77
Number of student- and teacher-focused sections	18	26	33	15	5	3	100
Average change per student- and teacher-focused sections	5.06	2.35	0.24	− 3.80	0.00	− 0.67	1.01
Standard deviation per student- and teacher-focused section	1.47	1.24	0.99	2.23	2.19	1.25	3.06

via the GSIOP score showing an overall average increase in 1.01 points value, with a majority of that growth occurring from the teacher-focused sections.

To answer our second research question, we wanted to understand the feedback at a more contextual (Nilsson and Ryve 2010) level to determine how the feedback was formative. We computed the total change in score for all novices during a semester by subtracting the initial GSIOP score from the final GSIOP score, and summing those up, for each section. We then divided that total change by the number of novices to get the average change per novice.

Table 8 shows the 50 sets of GSIOP semester-long data analyzed separately as student-focused and teacher -focused, and then analyzed in total. The total highest average change in GSIOP score occurred when mentors provided and noticed Specific Advice and Specific Improvement (SASI, $M = 3.71$, $SD = 2.25$). SASI feedback also resulted in the highest change in GSIOP scores for both student-focused and teacher-focused sections. More specifically, the student-focused SASI feedback from mentors had the absolute largest growth in GSIOP score with an average change of 4.50 points per student-focused section ($M = 4.50$, $SD = 2.29$) generating a growth of 37.5% (4.5 points out of 12 points) in GSIOP score. Concomitantly, the largest change in teacher-focused sections with respect to GSIOP score also occurred with SASI feedback, averaging 3.40 points per teacher-focused section ($M = 3.40$, $SD = 2.15$) generating a growth of 22.67% (3.40 out of 15 points). Advice Without Improvement (AWI, $M = -0.48$, $SD = 2.68$) feedback and No Advice and No Improvement feedback (NANI, $M = -0.25$, $SD = 2.66$) had the least change in both GSIOP scores.

BASI feedback provided high changes as well ($M = 3.17$, $SD = 1.77$), but with fewer student-focused ($N = 2$) and teacher-focused ($N = 4$) feedback instances. SABI feedback influenced the student-focused section more ($M = 3.57$, $SD = 1.76$) than the teacher-focused section ($M = -0.25$, $SD = 2.86$) while BABI feedback influenced the teacher-focused section ($M = 2.38$, $SD = 2.00$) more than the student section ($M = 0.58$, $SD = 3.04$). We also note that there was more teacher-focused SASI and BASI feedback ($N = 10 + 4$) than student-focused SASI and BASI feedback ($N = 4 + 2$). Conversely, there was more student-focused SABI and BABI ($N = 7 + 12$) feedback than there was teacher-focused SABI and BABI ($N = 4 + 8$) feedback.

When looking at teacher-focused feedback with broad improvement, there may in fact be more value to broad advice ($M = 2.38$, $SD = 2.00$) than specific advice ($M = -0.25$, $SD = 2.86$) according to GSIOP scores. This may suggest that broad advice may be sufficient when broad improvements are provided when looking at beginning and final GSIOP scores. However, this was limited to only teacher-focused advice with broad improvements. Once teacher-focused GSIOP scores were combined with student-focused scores, SABI had a higher GSIOP score ($M = 2.18$, $SD = 2.89$) than the BABI GSIOP score ($M = 1.30$, $SD = 1.30$) on average.

It is natural to look for connections between longitudinal trends and feedback types. Figure 1 tallies all longitudinal changes (student-focused and teacher-focused) and compares them to each dataset's respective feedback types.

Figure 1 illuminates many connections between Tables 7 and 8. Specifically, notice how the AWI and NANI feedback often remained steady or had a decrease in changes to GSIOP score. All but one of the novices who received SASI feedback had a substantial increase or moderate increase in change to GSIOP score.

Table 8 Inductive analysis of feedback types cross-referenced with change in GSIOP score

Feedback types	SASI	BASI	SABI	BABI	NANI	AWI	Total
Number of student-focused feedback	4	2	7	12	11	14	50
Average GSIOP change per student-focused section	4.50	3.50	3.57	0.58	−0.73	−0.93	0.72
Standard deviation per student-focused section	2.29	1.50	1.76	3.04	3.05	2.58	3.30
Number of teacher-focused feedback	10	4	4	8	5	19	50
Average GSIOP change per teacher-focused section	3.40	3.00	−0.25	2.38	0.80	−0.16	1.30
Standard deviation per teacher-focused section	2.15	1.87	2.86	2.00	0.75	2.7	2.77
Number of student and teacher feedback	14	6	11	20	16	33	100
Average GSIOP change per student- and teacher-focused feedback	3.71	3.17	2.18	1.30	−0.25	−0.48	1.01
Standard deviation per student- and teacher-focused section	2.25	1.77	2.89	2.81	2.66	2.68	3.06

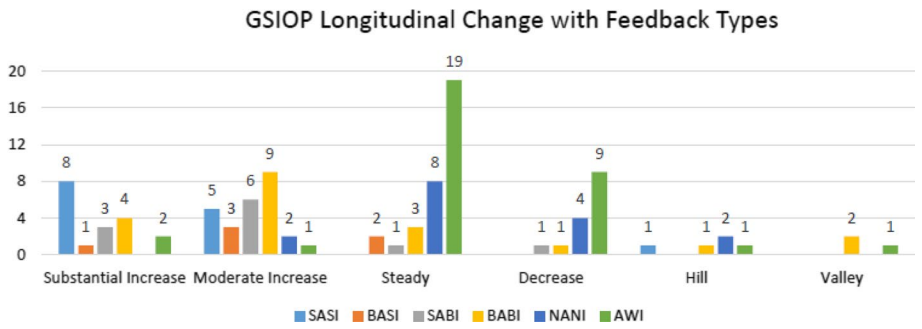

Fig. 1 Comparing feedback types with longitudinal changes

Discussion

We analyzed mentors' feedback and novices' GSIOP scores to investigate (1) In what ways (if any) was the feedback structure associated with changes in observed teaching throughout a semester using the student-centered GSIOP observation protocol, and (2) How do those changes inform (if at all) methods for providing formative feedback to influence observed teaching? We unpack implications of the results of this study by discussing how the qualitative coding, vignettes, and quantitative results are woven together. Then, we consider how the context of this study suggests limitations in scope of application and suggestions for future work. Finally, we include some implications of this work in practice and in research.

Table 8 and Fig. 1 show that AWI and NANI feedback had the most instances of decreased scores or lack of change in scores in the student-focused and teacher-focused sections of the GSIOP. Jason's broad student-focused AWI feedback to Indigo was limited by not providing more context on how to have student's "do more." Similarly, his teacher-focused NANI feedback provided only negative comments about what Indigo was doing wrong. Indigo's GSIOP scores barely changed (5, 6, 5 and 8, 9, 8, respectively) showing no overall change in GSIOP scores for either section. Jason's feedback to Indigo resembled other NANI and AWI feedbacks which provided broad general comments without context and purpose to the referenced context (see Table 5).

SABI and BABI showed some change to student-focused and teacher-focused GSIOP scores but had mixed results, with SABI having greater change in student-focused GSIOP scores and BABI having greater change in teacher-focused GSIOP scores. Hugo's student-focused SABI feedback to David illustrated the former, where specific advice was helpful with the topic of formative assessment. The continued reinforcement of collecting student feedback seemed to make a difference with David and was reflected in his substantial increase and growth to the student-focused GSIOP score (+6). Similar discussion and focus on specific advice showed up in other student-focused SABI feedbacks. Hugo's teacher-focused BABI feedback did not align with the general trend of growth in teaching as indicated by the data, but this may be because Hugo's feedback lacked specific context. Hugo's teacher-focused yellow feedback in the second observation advocated "student exploration" and "self-discovery" admitting it can be difficult but worth giving a try. Such comments lack advice about how and in what context to apply such teaching techniques, making the feedback difficult to use in practice. When student-focused and teacher-focused

scores of the GSIOP were combined, SABI was still higher than BABI overall showing a greater gain in the student-focused section with specific advice offset the loss in the teacher-focused section. Nonetheless, a future study with a larger set of data may want to consider how, given broad improvements in feedback, how teacher-focused feedback and student-focus feedback varied in advice.

Figure 1 illustrates how SASI and BASI feedback had the most moderate and substantial increases in both the teacher-focused and student-focused GSIOP sections. Concomitantly, Table 8 illustrates that SASI and BASI feedback was associated with the largest positive change over an entire semester. Referencing Mickey's teacher-focused BASI feedback to Sarah, there was a lack of clear context when saying things such as the lesson was disorganized or jumped around without giving specifics or purposeful contexts for the statement. However, Mickey then applauds Sarah for moving away from depending on the slides with specific reference to how she has changed over the past year, allowing her to be more flexible in her teaching practices and choices.

For student-focused and teacher-focused SASI feedback (the greatest change in GSIOP scores and the category with the most substantial increases), Roberto's RYG feedback represents the four student-centered SASI feedbacks and the ten teacher-centered SASI feedbacks. Three characteristics distinguish Roberto's feedback, compared to Mickey, Hugo, and Jason's. First, notice how Roberto's novice, Wendy, has similar issues to David regarding engaging with students. Unlike Hugo's responses to David, Roberto's feedback provides green comments and then expands upon them as yellow comments such as shown in his first and second observation. This provides a more thorough explanation of how broad compliments reference contextual advice and improvement. Consistent with Nilsson and Ryve (2010), Roberto's vignette illustrates how the combination of context and focal events is important for clarifying purpose and making feedback actionable (Cannon and Witherspoon 2005; Shute 2008). This result suggests the improved GSIOP scores may be the result of better mentor–novice communication, agreeing with prior linguistic research on the value of focal events (Goodwin and Duranti 1992).

Second, Roberto clearly reviewed and emphasized comments from the previous time and how he saw specific improvements in Wendy's teaching practices. These perceived changes were often mentioned in green, but then pushed Wendy to continue to improve. In his second observation of her, Roberto made sure to emphasize the specifics of the probability of the union of two sets but stayed focused on being clear on how the context was relevant to student confusion, emphasizing the student-focused specific advice. Third, Roberto wrote he would work with Wendy to provide examples in person on how to move away from PowerPoint lectures toward actively engaging with students. Roberto is demonstrating that he is willing to help Wendy to continue to improve her teaching. This is a critical shift because Roberto is not seeing feedback as the final process of the observation, but the starting point for future observations. Indeed, Roberto's third observation's second yellow comment contains "Prepare for next semester by doing something that shows your content knowledge on day one, and that shows that you care for student learning. This will get the semester off to a great start." Roberto sees his role beyond one observation, which is the goal of formative feedback.

The first two characteristics of Roberto's feedback illustrate the need for contextually specific feedback in both advice and improvement. At the core of the third characteristic in Roberto's vignette is the belief that Wendy is continually growing in her teaching practices. It is here we see the connection with formative feedback. Roberto did not judge Wendy, but instead tried to modify her thinking or behavior to improve her teaching, which is Shute's (2008) definition of formative feedback applied to teaching. Roberto's vignette illustrates

the use of specific advice to engage through questioning, followed by specific improvement that promoted continued development demonstrates formative feedback that can positively frame post-observation feedback.

When considering the coding, vignettes, and quantitative data together, our results agreed with Kluger and DeNisi's general result (1998) that certain types of feedback were more effective than other types of feedback. Our results grow the fields' understanding of observational feedback by discerning between advice and improvement, looking at broad and specific types of feedback, and how these types affected teaching practices of novice UMIs, specifically GSIs in this study. Our results illuminated that effective feedback (with respect to change in GSIOP score) combined ways to build on novice teaching practices with contextualized feedback that included focal events (such as Roberto's feedback to Wendy). Conversely, our results showed less effective feedback (with respect to the GSIOP score) provided suggestions that lacked a focal event, lacked context, and were disjoint from prior observations (such as Jason's feedback to Indigo).

Limitations and potential future directions

Although this study focused on GSIs, we conjecture that the observation protocol, feedback structure, and results are applicable to other novice UMIs. A possible mitigating factor is the use of the peer mentoring program. The structure of the post-observation feedback and the overall design of the peer mentoring program, including the training of mentors and the use of the peer mentoring program, could have influenced our results. This in no way voids the results but illustrates that we have not tested for similar results with a different observation protocol and/or feedback structure. To aid in clarifying the training, Appendix G in ESM provides the three handouts we use in the training of the RYG feedback. Moreover, as the vignettes from specific mentor–novice pairs indicate (Appendices A–D in ESM), the peer mentoring interactions are a source of rich data and can be unpacked further in future papers from this project or other research projects about novice instructors' professional development. In particular, one could consider the role of GSIs' identity, agency, or beliefs within peer–mentor interactions. Although these questions are outside the scope of this study, they are important to consider to understand additional factors that can contribute to success in professional development programs.

One may be concerned that the mentors' biases toward a certain teaching style may have influenced their choice of GISOP scores. However, mentors were trained in its use to properly align the observation and score with the detailed box-text following every rating (See full GSIOP in Appendix E in ESM). Moreover, the validation study of the GSIOP (Rogers et al. 2019) shows how the instrument, when properly used, is a reliable and valid instrument for observation of novices teaching undergraduate mathematics courses. A final limitation to consider is the sample size. Tables 7 and 8 show that our sample size for any single longitudinal change or feedback variable was less than twenty in any single student-focused or teacher-focused code, limiting the possibility for regression analyses or further quantitative analyses. This is a valid concern as we had 151 individual data points (observation cycles) but analyzed the data in this study for semester-long changes. Because only three observations of each novice were carried out throughout the semester by each mentor, these are only snapshots of each novice. Our research method aggregated these snapshots to look for trends among the 151 data points as a representative sample of providing feedback by using mixed-methods triangulation. Future studies that incorporate longitudinal

analyses of formative feedback would be an excellent expansion of this study to identify significant variations in specific items of the observation protocol over time.

Implications for practice

For teacher educators who supervise novice UMIs, our results suggest there is a need for making observations and observation protocols formative rather than summative. Specifically, there is a need to articulate growth and change as a teacher rather than just generating an evaluative score for the observation. To do this, novice UMIs require regular feedback and the opportunity to see observation cycles as a critical aspect of professional development. Implementing such cycles in practice is challenging depending on the resources and time available to complete multiple observations and provide formative feedback; thus, we have provided support to implement formative feedback. Appendix G in ESM provides a means to train others in the use of RYG feedback, and we suggest looking into other resources, such as experienced UMIs, who may be incentivized to help in novice UMIs' grow as teachers. If human resources are not available for help in generating multiple observations, one can still consider formative feedback that focuses on contextually specific means of improvement to provide UMIs with a framework for continuing teacher growth and reflection.

Implications for research

Our results echo Kluger and DeNisi's (1998) theory of feedback being "a double-edged sword" where feedback can have a positive influence but may have a negative influence as well, depending on the type of feedback. Table 7 demonstrates overall growth to both the student and teacher sections, but it varies according to the type of feedback (Table 8). In answering our first research question, Table 7 shows that RYG feedback in our study could be connected to increases, decreases, hills, and valleys as trends in GSIOP scores associated with student engagement and teacher facilitation. There were more increases than decreases in GSIOP scores over semester-long observation–feedback iterations, with 6% positive growth occurring with student-focused sections and 8.67% growth occurring with teacher-focused sections. This demonstrates growth in both sections of GSIOP scoring, with more growth in the teacher-focused sections.

In answering our second research question, our coding of feedback (advice/improvement and broad/specific) illustrated how GSIOP scores on the teacher and student sections could be related to the type of feedback. Feedback that included specific improvement (SASI and BASI) resulted in the largest changes in GSIOP scores for both student-focused and teacher-focused sections, with the largest increase stemming from student-focused SASI feedback averaging 37.5% growth in GSIOP score followed by teacher-focused SASI feedback averaging 22.67% growth in GSIOP score. Broad improvement (SABI and BABI) had mixed results depending on whether they were student-focused or teacher-focused, while feedback that lacked improvement (AWI and NANI) had the largest negative and only small positive change with student-focused and teacher-focused GSIOP scores. Thus, when looking over a semester with three observations, identifying and emphasizing improvement was important with specific (context with focal events) improvement receiving the highest change in score.

We hypothesized that Shute's (2008) framework of formative feedback for student learning from the student–teacher dynamic was appropriate for use as formative feedback for

teacher observations in the novice–mentor dynamic (Table 3) with GSIs. Our results verified this and showed specific feedback provided the greatest changes in GSIOP scores and had the most consistent growth. This reinforces the value of formative feedback because the improvement code focused on referencing prior feedback and identifying growth and change in the novice, not feedback on their current observation alone. This is central to the definition of formative feedback. Thus, our results illustrate that specific improvements, a type of formative feedback, was associated with changes in our novices GSIOP scores.

Our study illuminates that if we expect UMIs to grow in their teaching (specifically with student-centered teaching practices, such as IBME), observation protocols should include at least the option of formative feedback. With respect to student learning, if we deconstruct the dichotomy of summative and formative assessment via a spectrum (Harlen and James 1997), we have a spectrum between assessment OF learning (Summative) and assessment FOR learning (Formative). Shute's (2008) definition of formative feedback applied to teaching provides a similar spectrum with observations and observation protocols. Specifically, this study highlights the relationship between assessment OF teaching (Summative) and assessment FOR teaching (Formative). We are not arguing to remove assessment of teaching, but rather bring awareness to the research community to the spectrum so that the community can discuss the use of an observation protocols and feedback's purpose as *more* summative or *more* formative. It is critical that we consider what expectations we have for our UMIs so that the purpose of observation protocols are properly discussed within research. Our results have found that feedback alone is not sufficient to expect growth in an instructor (or their measured scores on an observation protocol). Specifically, our study showed different types of feedback were associated with different effects on changes to observation protocol scores. Contextualized feedback that references improvement resulted in largest change in GSIOP scores, which aligns with Shute's definition of formative feedback. This opens up a research gap that needs to be studied more thoroughly within teacher education.

Our results' connection to formative feedback is aligned with research from primary and secondary schools because both levels often have action plans for their novice teachers (Portner 2005; Harbour and Livers 2018). Action plans provide an opportunity for novice teachers to have specific teaching goals and plans to reach those goals. These plans are often modified and adapted to the classroom, teacher, and culture and are reviewed by the mentor teacher (secondary) or the mathematics coach (primary) along with the novice. As such, the observation feedback is formative with a regular focus on contextualized feedback because that is how the observations reference the specific teaching goals. In White's study (2007) of novice student–teachers in New Zealand, he asked what types of feedback were helpful and found:

> Supervisors giving specific feedback to student-teachers can make a difference (Brawdy and Byra 1995). Specific feedback, containing information relevant to the behaviour of the student–teacher, in contrast to general feedback, which supports the behaviour but provides no information on its 'technical qualities' is an important variable in providing positive changes (Siedentop 1981) (p. 302)

As discussed in the literature review, there is a need for specificity, but previous literature has not studied how to tailor the specificity to support novice UMIs who lack an action plan. Novice UMIs' disciplinary focus is generally not in education, and thus, they may lack necessary educational theories that would give a viable trajectory for teaching goals. They are also often limited in their exposure to teaching experience, seminars, courses, or professional development (Speer et al. 2005). However, our results show that even with

these limitations, focusing on contextualized improvements may be a possible bridge to help post-observation formative feedback be useful and connected to the observation for teaching.

Conclusions

This study found that feedback was associated with positive change in GSIOP score in a subset of novice UMIs, GSIs, along both dimensions of student- and teacher-focused feedback. Although the student-focused change in GSIOP score (6.00%) was less than the teacher-focused GSIOP score (8.67%), they both showed growth. The type of feedback was found to have a meaningful impact. Overall, specific (contextualized) advice and improvements generated the largest growth in both student- and teacher-focused GSIOP scores, with the student-focused GSIOP score having the largest growth (37.5%) followed closely by the teacher-focused GSIOP score (22.67%). Advice with broad improvements had lesser impact, but the results illustrated that advice without improvements or feedback that lacked both advice and improvements had mostly negative growth. The idea of specific improvement emphasizes contextual feedback on how improvement can happen, highlighting feedback *for* teaching, in addition to feedback *of* teaching. To this end, the results of this study found that formative teaching feedback (Shute 2008), not summative teaching feedback, was associated with student-focused and teacher-focused growth in novice UMIs. If we are to fully embrace the paradigm shift articulated in the 2001 International Congress on Mathematical Instruction study that university teaching needs continuous reform, renewal, and attention to student learning (Alsina 2001), then we must provide the same continuous reform, renewal, and attention to teacher growth in our teacher feedback and observational design.

Acknowledgments This work was supported by Collaborative IUSE NSF grants (#1544342 & 1544346; #1725264, 1725295, & 1725230). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We also want to thank the University of South Carolina Reading and Research Group in Mathematics Education for providing insightful feedback.

References

Alsina, C. (2001). Why the professor must be a stimulating teacher. In D. Holton & M. Artigue (Eds.), *The teaching and learning of mathematics at university level: An ICMI study* (Vol. 7). Berlin: Springer.

Belnap, J. K., & Allred, K. (2009). Mathematics teaching assistants: Their instructional involvement and preparation opportunities. In L. L. B. Border (Ed.), *Studies in graduate and professional student development* (pp. 11–38). Stillwater, OK: New Forums Press, Inc.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Brawdy, P., & Byra, M. (1995). Supervision of preservice teachers during an early field teaching experience. *The Physical Educator, 52*(3).

Bressoud, D., Mesa, V., & Rasmussen, C. (Eds.). (2015). *Insights and recommendations from the MAA national study of college calculus*. Washington, DC: Mathematical Association of America Press.

Cannon, M. D., & Witherspoon, R. (2005). Actionable feedback: Unlocking the power of learning and performance improvement. *Academy of Management Perspectives, 19*(2), 120–134.

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research*. London: Sage Publications.

Ellis, J. (2014). Preparing future professors: Highlighting the importance of graduate student professional development programs in calculus instruction. In *Proceedings of the 37th conference of the*

international group for the psychology of mathematics education (Vol. 3, pp. 9–16). Vancouver, BC: PME.

Ellis, J. (2015). Professional development of graduate students involved in the teaching of calculus I. In D. Bressoud, V. Mesa, & C. Rasmussen (Eds.), *Insights and recommendations from the MAA national study of college calculus. MAA notes* (pp 121–128).Washington, DC: Mathematical Association of America.

Ellis, J., Deshler, J. & Speer, N. (2016a). Supporting institutional change: A two-pronged approach related to graduate teaching assistant professional development. In T. Fukawa-Connelly, N. Infante, M. Wawro, and S. Brown (Eds.), *Proceedings of the 19th annual conference on research in undergraduate mathematics education*. Pittsburgh, PA.

Ellis, J., Deshler, J. & Speer, N. (2016b). *How do mathematics departments evaluate their graduate teaching assistant professional development programs?* International Group for the Psychology of Mathematics Education Annual Conference, Szeged, Hungary.

Gamlem, S. M., & Smith, K. (2013). Student perceptions of classroom feedback. *Assessment in Education: Principles, Policy & Practice, 20*(2), 150–169.

Gibbons, L. K., & Cobb, P. (2017). Focusing on teacher learning opportunities to identify potentially productive coaching activities. *Journal of Teacher Education, 68*(4), 411–425.

Gibbons, L. K., Kazemi, E., & Lewis, R. M. (2017). Developing collective capacity to improve mathematics instruction: Coaching as a lever for school-wide improvement. *The Journal of Mathematical Behavior, 46,* 231–250.

Gleason, J., Livers, S., & Zelkowski, J. (2017). Mathematics classroom observation protocol for practices (MCOP2): A validation study. *Investigations in Mathematics Learning, 9*(3), 111–129.

Goodwin, C., & Duranti, A. (1992). *Rethinking context: Language as an interactive phenomenon*. Cambridge: Cambridge University Press.

Harbour, K., & Livers, S. (2018). Using coaching cycles to transfer and sustain effective instructional practices. In *Proceedings from 40th conference of the North American chapter of the psychology of mathematics education (PME-NA)*. Greenville, SC.

Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice, 4*(3), 365–379.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Hollingsworth, H., & Clarke, D. (2017). Video as a tool for focusing teacher self-reflection: Supporting and provoking teacher learning. *Journal of Mathematics Teacher Education, 20*(5), 457–475.

Holton, D., & Artigue, M. (Eds.). (2001). *The teaching and learning of mathematics at university level: An ICMI study* (Vol. 7). Berlin: Springer.

Johnson, S. M., & Kardos, S. M. (2002). Keeping new teachers in mind. *Educational Leadership, 59*(6), 12–16.

Kastberg, S. E., Lischka, A. E., & Hillman, S. L. (2018). Characterizing mathematics teacher educators' written feedback to prospective teachers. *Journal of Mathematics Teacher Education*. https://doi.org/10.1007/s10857-018-9414-6.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254.

Kluger, A. N., & DeNisi, A. (1998). Feedback interventions: Toward the understanding of a double-edged sword. *Current Directions in Psychological Science, 7*(3), 67–72.

Laursen, S. L., & Rasmussen, C. (2019). I on the prize: Inquiry approaches in undergraduate mathematics. *International Journal of Research in Undergraduate Mathematics Education, 5,* 129–146.

Lutzer, D. J., Rodi, S. B., Kirkman, E. E., & Maxwell, J. W. (2007). *Statistical abstract of undergraduate programs in the mathematical sciences in the United States: Fall 2005 CBMS survey*. Providence, MA: American Mathematical Society.

Moir, E. (2005). Launching the next generation of teachers. In H. Portner (Ed.), *Teacher mentoring and induction: The state of the art and beyond* (pp. 59–73). Thousand Oaks: Corwin Press.

Nilsson, P., & Ryve, A. (2010). Focal event, contextualization, and effective communication in the classroom. *Educational Studies in Mathematics, 74*(3), 241–258.

Portner, H. (2005). *Teacher mentoring and induction: The state of the art and beyond*. Thousand Oaks: Corwin Press.

Reinholz, D. (2016). The assessment cycle: A model for learning through peer assessment. *Assessment & Evaluation in Higher Education, 41*(2), 301–315.

Reinholz, D. L. (2017). Not-so-critical friends: Graduate student instructors and peer feedback. *International Journal for the Scholarship of Teaching and Learning, 11*(2), n2.

Rogers, K. C., & Steele, M. D. (2016). Graduate teaching assistants' enactment of reasoning-and-proving tasks in a content course for elementary teachers. *Journal for Research in Mathematics Education, 47*, 372–419.
Rogers, K. C., & Yee, S. P. (2018a). Peer mentoring mathematics graduate student instructors: Discussion topics and concerns. In *Proceedings from 21st conference of the research in undergraduate mathematics education (RUME)*. San Diego, CA.
Rogers, K. C., & Yee, S. (2018b). *GSIOP: Graduate student instructor observation Protocol*: Retrieved from http://personal.bgsu.edu/~kcroger/research.html.
Rogers, K. C., Petrulis R. A., Yee, S. P., & Deshler, J. (2019). Mathematics Graduate Student Instructor Observation Protocol (GSIOP): Development and Validation Study. *International Journal of Research in Undergraduate Mathematics Education (IJRUME)*. https://doi.org/10.1007/s40753-019-00106-4.
Roller, S. A. (2016). What they notice in video: A study of prospective secondary mathematics teachers learning to teach. *Journal of Mathematics Teacher Education, 19*(5), 477–498.
Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., & Bloom, I. (2002). Measuring reform practices in science and mathematics classrooms: The reformed teaching observation protocol. *School Science and Mathematics, 102*(6), 245–253.
Seymour, E. (2005). *Partners in innovation: Teaching assistants in college science courses*. Lanham, MD: Rowman & Littlefield.
Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.
Siedentop, D. (1981). The Ohio State University supervision research program summary report. *Journal of Teaching in Physical Education, 1*(s1), 30–38.
Speer, N. M., & Murphy, T. J. (2009). Research on graduate students as teachers of undergraduate mathematics. In L. L. B. Border (Ed.), *Studies in graduate and professional student development* (pp. xiii–xvi). Stillwater, OK: New Forums Press Inc.
Speer, N. M., Gutmann, T., & Murphy, T. J. (2005). Mathematics teaching assistant preparation and development. *College Teaching, 53*(2), 75–80.
White, S. (2007). Investigating effective feedback practices for pre-service teacher education students on practicum. *Teaching Education, 18*(4), 299–311.
Wiliam, D., & Black, P. (1996). Meanings and consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal, 22*(5), 537–548.
Yee, S.P., & Rogers, K. C. (2017). Mentor professional development for mathematics graduate student instructors. In *Proceedings from 20th conference on research in undergraduate mathematics education* (RUME, pp. 1026–1034). San Diego, CA.

Authors and Affiliations

Sean Yee[1] · Jessica Deshler[2] · Kimberly Cervello Rogers[3] · Robert Petrulis[4] · Christopher D. Potvin[5] · James Sweeney[6]

Jessica Deshler
jmdeshler@mail.wvu.edu

Kimberly Cervello Rogers
kcroger@bgsu.edu

Robert Petrulis
robert.petrulis@epreconsulting.com

Christopher D. Potvin
potvinc1@msu.edu

James Sweeney
jsweeney@coker.edu

[1] University of South Carolina, 1523 Greene Street, Columbia, SC 29208, USA

[2] Department of Mathematics, West Virginia University, PO Box 6310, Morgantown, WV 26506, USA

[3] Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403, USA

[4] EPRE Consulting LLC, 527 Avondale Drive, Columbia, SC 29203, USA

[5] Michigan State University, 619 Red Cedar Road, C212 Wells Hall, East Lansing, MI, USA

[6] Coker College, 300 E. College Ave., Hartsville, SC 29550, USA