"I know it when I see it": employing reflective practice for assessment and feedback of reflective writing in a makerspace classroom

Received 10 September 2020
Revised 2 February 2021
Accepted 4 March 2021

Ofer Chen and Yoav Bergner
Department of Administration, Leadership and Technology, New York University, New York, New York, USA

Abstract

Purpose – In reflective writing, students are encouraged to examine their own setbacks and progress. With a shortage of guidance in how to provide feedback to students on this type of writing, teachers are often left to figure it out on the job. The central hypothesis in this paper is that the lens of reflective practice can help focus teacher efforts and ultimately improve both feedback and instruction. The purpose of this paper is not to produce a universal prescription for assessing reflective writing but rather a protocol for teacher reflective practice that can apply to challenging grading and feedback-giving situations.

Design/methodology/approach – Student assessment is a chance for teachers to learn about their students' abilities and challenges and to provide feedback for improvement. Assessment and grading sessions can also become opportunities for teachers to examine their own instructional and assessment practices. This self-examination process, a cornerstone of reflective practice (Schön, 1984), is challenging, but it may be especially valuable when guidelines for feedback and assessment are hard to come by. Such may be said to be the case in student-centered learning environments such as school Fablabs and makerspaces, where stated goals commonly include cultivating learner self-regulation and resilience. These hard-to-measure constructs are typically assessed through analysis of student reflective journals. This in-depth case study uses mixed-methods to examine how a semester-long intervention affected the grading, feedback and instructional practices of a teacher in a hands-on design classroom. The intervention involved 10 grade-aloud sessions using a computer-based rubric tool (Gradescope) and a culminating card-sorting task. The lens of reflective practice was applied to understanding the teacher's development of their own reflective capabilities.

Findings – During the intervention, the participating teacher grappled with grading and feedback-giving dilemmas which led to clarifications of assessment objectives; changes to instruction; and improved feedback-giving practices, many of which persisted after the intervention. The teacher perceived the intervention as adding both rigor and productive "soul-searching" to their professional practice. Lasting changes in feedback behaviors included a comprehensive rubric and an increase in the frequency, specificity and depth of feedback given to student written work.

Originality/value – Significant prior efforts have been directed separately at the use of reflective practice for teachers, in general, and on the feedback and grading of student process journals. This work combines these lines of inquiry in the reflective classroom assessment protocol, a novel on-the-job professional development opportunity that fosters reflective practice in times of assessment to improve instructional and feedback practices.

Keywords Case study, Makerspaces, Reflective practice, Mixed-methods, Classroom assessment, Formative feedback

Paper type Case study

Information and Learning Sciences
Vol. 122 No. 3/4, 2021
pp. 199-222

2398-5348
DOI 10.1108/ILS-09-2020-0209

Background

Reflective practice for teachers

Teacher educators Cornish and Jenkins (2012) identify reflection activities as one of three effective strategies (along with modeling and applying theory to practice) in preparing teachers for the challenges of formative classroom assessment and transforming students of education into proficient practitioners. Other education researchers have also recommended that reflection activities be embedded in teacher training programs (Nagro *et al.*, 2017) and in the daily work of in-service teachers (Allan and Driscoll, 2014; Farrell, 2012).

There are many benefits associated with applying reflective practice to teaching such as promoting informed decision-making, supporting teachers in developing a rationale for their practice, keeping teachers growing and engaged in their work and cultivating trust with students (Brookfield, 2017). Interventions that promote teacher reflective practice have claimed to raise teacher awareness of the various elements of teaching, and develop critical analysis of how to meet students' needs (Nagro *et al.*, 2017). One such study on reflective practice in informal science education programs led to the production and articulation of design principles and learning objectives (Bevan *et al.*, 2015). Reflective practice was also shown to support novice teachers in "surviving" their first year of teaching (Farrell, 2016) and assisting female pre-service teachers to develop their STEM identity in teaching a makerspace classroom (Blackley *et al.*, 2017). In informal learning environments, reflective practice has been recognized as an effective professional development tool for museum makerspace educators, who point to improved practices and increased confidence in facilitating maker activities (Moore *et al.*, 2020; Grabman *et al.*, 2019).

What exactly is meant by reflective practice? To better explain the design of interventions, including the present one, it is helpful to provide some theoretical foundations from the work of Dewey (1933), Schön (1984) and Mezirow (1998).

Key concepts in reflective practice

Dewey traced the process of inquiry to triggering events which he called perplexities. Schön connected these events to uncertainty, doubt and complexity and Mezirow termed them "disorienting dilemmas." When a process of inquiry is carried out successfully, it involves observation, reasoning and reflection, which leads to an informed decision (Dewey, 1933). This process leads to a critical examination of one's beliefs and assumptions about a given situation which may result in changing one's "habits of mind" (Mezirow, 1998). In reflecting, one is able to connect previous knowledge and experiences with new challenges and information. Reflection also facilitates one's ability to draw on tacit knowledge or know-how, in facing a problem or uncertainty. This ability is considered to be fundamental for all expert practitioners, teachers included (Schön, 1984; Thorsen and DeVore, 2013).

Teachers routinely face uncertainty and complexity in classroom management, instruction and assessment. Indeed, Shavelson (1973) identified decision-making as the most fundamental skill required for teaching and posited that any act of teaching is a result of a decision-making process. During the formative assessment, in particular, teachers are likely to encounter grading and feedback-giving dilemmas. For example, a teacher may debate internally about whether to push a high-achieving student more critically, to respond leniently to a low-achieving student or to treat all students equally. These dilemmas [see Brookhart (1994) and Green *et al.* (2007) for other assessment vignettes] are contextual to the teacher, the assessment task, the teaching environment and the student being assessed (Bell and Cowie, 2001).

While formal pedagogical knowledge serves as an important basis to contend with such issues, it is rarely enough. Teachers tend to rely on their experience and intuition to analyze and respond to a variety of unique challenges and often use a trial-and-error strategy when

it comes to their assessment practices (McMillan and Nash, 2000; Schafer, 1993). In a 1991 survey (Wise *et al.*, 1991), 55% of teachers reported that trial-and-error was the most important source of their knowledge in testing and measurement, with small differences between those who had taken at least one assessment course and those who had not (48% and 59%, respectively). It follows that teachers rely on their experience when it comes to assessing students' performance and abilities. This experience in assessing student assignments accrues over time and is integrated into teachers' professional knowledge, which supports them in decision-making and in navigating uncertainty. Building on Dewey's work 1933, Schön (1984) describes the use of both "reflection-in-action" (in the moment) and "reflection-on-action" (after the fact) as strategies used by teachers and other expert practitioners to access their tacit knowledge in dealing with complex problems.

For Dewey and Schön, reflection is not an isolated process of introspection, it is carried out in a community and relies on a systematic collection of evidence and the use of such evidence in decision-making. For teachers, it is the process of examining what they know about their instruction and their students, through observation, assessment or other means, to guide them in creating learning opportunities for students and supporting them in their learning (Farrell, 2012).

Reflective practice during assessment events as PD on-the-job

Opportunities for self-examination and reflective practice arise naturally when teachers assess student work after class. With evidence of learning laid out in front of them and removed from the hubbub of the classroom, teachers can enter an analytic mindset. They may review how students performed on the assessment, how performances measure up to expectations, what can be improved in students' work, how such improvements might be achieved and why students performed in the way they did. This last question is especially potent, as it can lead the teacher to engage in self-examination of their own teaching practices and on how the instructional time or the assessment were designed and implemented.

During class, teachers may not be aware that they are engaging in formative assessment through discussion and observation (Bell and Cowie, 2001). However, during these sessions, evidence is observed of students' improvement and struggles, strengths and weaknesses. When providing students with written feedback, teachers are better positioned to access this tacit knowledge, accumulated during non-conscious formative assessment events and reconcile it with the evidence laid out in front of them.

The present work aims to develop a method of incorporating reflective practice into routine grading and feedback sessions. We build upon the theories of Dewey and Schön to support teachers in improving their grading and feedback-giving skills and to empower them to communicate better, with students, as well as with administrators, peers and parents. This on-the-job approach is aligned with teachers' expressed preference for receiving professional development in assessment (DeLuca *et al.*, 2018).

Formative assessment in makerspaces and other student-centered learning classrooms

As Blikstein and Worsley (2016) point out, enthusiasm for the educational value of Fab Labs and Makerspaces (FLMs) continues in a tradition of progressive efforts (Dewey, Froebel, Montessori, etc). to make children's education authentic and student-centered. The term student-centered learning (SCL) has been used to span several distinct environments including problem- and project-based classrooms as well as FLMs The common thread in all of these is that learning activities place the student at the heart of the process and acknowledge the active construction of knowledge and skills (Piaget, 1976). Teachers in SCL classrooms express interest in a broad set of process-oriented skills (or soft-skills) such as

agency, metacognition and collaboration, in addition to the acquisition of procedural and content knowledge (Bergner *et al.*, 2019). Educators are also typically concerned with the cultivation of positive dispositions such as lifelong learning and growth mindset (Bell, 2010; Hmelo-Silver, 2004; Hmelo-Silver *et al.*, 2007; Martin, 2015; Bergner *et al.*, 2019).

When it comes to assessment and feedback, these ambitious learning objectives create challenges unique to SCL environments. Process-oriented skills and dispositional shifts are not readily assessed using a written exam and typically demand more than snapshot, end-of-term assessments (Bell, 2010; Hmelo-Silver, 2004; Papert, 1991). In response to this challenge, SCL teachers use a variety of assessments in their classrooms (Wiggins and McTighe, 1998) such as portfolios, design journals, performance assessments and student reflections. Second, SCL environments allow students more freedom to choose the focus of their work and the path to bring that work to completion. This variance in process and product requires flexibility on the side of teachers when assessing student progress and achievement (Bergeron *et al.*, 2019; Murai *et al.*, 2020; Waters and McCracken, 1997). Third, the dynamic nature of the work in these environments can make it harder for teachers to methodically collect evidence of learning during class, which might be easier in frontal instruction (Murai *et al.*, 2020). Finally, there are no gold-standard measurement techniques to assess many of the soft-skills highlighted in SCL classrooms (Duckworth and Yeager, 2015). These challenges are not mutually exclusive with the challenges facing teachers in traditional classrooms but further compound the assessment challenges of SCL teachers.

Assessing student reflections

Student-written reflections and reflective journals are commonly used in educational makerspaces. Moreover, a survey conducted by Peppler *et al.* (2017) indicated that makerspace teachers and instructors view self-reflection as the primary reason for requiring students to create portfolios. This raises the question of how student self-reflection is assessed.

The literature on assessing student reflections, unfortunately, provides little guidance to teachers attempting to use student reflections in formative contexts. Plack *et al.* (2005) suggest a checklist method to assess nine elements of reflection in journals (divided into time, content and stage elements) to determine whether and to what extent, reflection occurs in them. In an experiment conducted by Cheng and Chan (2019), a holistic rubric was used to place student reflections on a five-point scale. The Student Assessment of Reflection Scoring (StARS) rubric (Koole *et al.*, 2012) focuses on three goals of the reflection: awareness of self and the situation; critical analysis and understanding of both self and the situation; and development of new perspectives to inform future actions (p. 12). Though all three instruments demonstrate desirable properties such as interrater reliability, they are more useful for rating and grading student reflections than for giving students quality written feedback.

Some researchers have taken a computational approach to assessing and providing feedback for reflective writing. One such example comes from Gibson *et al.* (2017), who report on the development of an automated system that provides formative feedback on college students' reflections by analyzing the rhetorical moves they use in their writing. While automated systems can provide students with feedback at scale, there are some drawbacks to this approach for classroom assessments, perhaps, particularly so in makerspace and other student-centered learning classrooms. According to Dewey (1933), the reflective process should be done with others to support student sense-making. Moreover, automated systems may be able to give feedback about the depth of the reflection, however, they are unable to incorporate classroom experiences and knowledge of the content domain as a teacher might. Mislevy (2013) anticipates that in assessing students "the strongest evidence will be most contextualized and

the weakest what can be gleaned from external assessments that are not connected with students' instructional contexts or histories" (p. 8). None of the above instruments leave room for teachers to incorporate contextual factors into the assessment process. The present work focuses on the potential benefits teachers can realize, beyond psychometrically sound grades, by engaging deeply in their students' reflections.

Teachers in makerspaces and SCL environments who wish to assess student reflections are, thus, often left to their own devices (McMillan, 2001). It is well-established that, when faced with challenging judgment tasks, people use heuristics to simplify the decision-making process (Kahneman *et al.*, 1982; Gigerenzer and Gaissmaier, 2011). Teachers are not exempt from this. Thus, a teacher may take a holistic view and, based on their experience, assign a rating or grade. Such heuristic/holistic assessment techniques keep the "black box" of classroom assessment (Black and Wiliam, 2010) closed, providing no transparency into assessment criteria. When teachers are unable to articulate the criteria by which they assess their students, it may leave students with a feeling that the entire process is subjective and arbitrary.

This work tries to empower teachers to grapple with complex dilemmas. By promoting a self-examination of assessment and instructional practices, objectives and criteria may be refined and more clearly communicated to students. Grounded in the works of Dewey and Schön, the method presented here includes a combination of using dedicated grading software to offload some of the cognitive strain from participating teachers and a think-aloud protocol. More details about the intervention are provided in the next section.

Methods

The case

Brooklyn Catholic High School (BCHS; pseudonym) is a private, all-girls, Catholic high-school that serves predominantly African-American and Latina students in an urban setting. A few years prior to the present study, BCHS began to scale up efforts to promote science, technology, engineering and math (STEM) education. This initiative included a new school-based makerspace, as well as STEM-related classes and technology-focused after-school programs. To support the transition, the school leadership formed connections and partnerships with several organizations, including the authors' home institution. Partnership activities included regular meetings with school faculty, specifically the newly-formed tech team, mainly to support their curricular design efforts.

Michael has been a teacher for 11 years, following a decades-long career as a cabinet-maker. During the school's transition, he was tasked with developing and teaching a hands-on Fundamentals of Design (FoD) class and becoming the school's technology integration specialist. Michael is a profoundly thoughtful teacher who invests a lot of time in planning classroom activities and trying to connect math concepts to real-life situations. His disposition toward the FoD class is, perhaps, best captured in his own words: "I spent my life making things, for the most part, and giving me the opportunity to share that information, that knowledge with [my students] is an incredible gift to me at this point in my life." Our partnership with Michael included supporting him in designing classroom activities and offering advice about how to implement student-centered learning.

In one of our first planning meetings, Michael stated that his higher-order goals for the FoD class were collaboration, communication and iteration. It is worth noting this focus on "soft-skills" and student resilience is commonly found in learning objectives for Fablab and Makerspace educators (Bergner *et al.*, 2019).

Communication in the FoD classroom was envisioned by Michael as spanning both visual and verbal components. Students would develop and hone those sub-skills through

technical drawing and writing research reports and student reflections. The FoD class was designed to prepare students for the term project of designing a public space, in this case, creating a technical floor plan for a green roof accompanied by a research document detailing each student's decision-making process. The term project was preceded by short-term introductory projects (between 1 and 2 weeks for each project) which focused on developing tools such as orthographic projection, one-point perspective drawing and technical measurement and drawing of large spaces (i.e. the school cafeteria). Every other week, FoD students submitted a written reflection focusing on what was easy for them during that time, what was hard and how they could improve on those aspects.

The topic of this case study emerged from a discussion about how he evaluated student reflections, a task that Michael felt did not play to his content expertise in math or design. He described his approach to evaluating and grading student reflections with critical self-awareness: "I've looked at [the assignments] and kind of[. . .] batched them together by class and said 'this is the best and this is the worst' and then you come up with some sort of scale that relates those facts without it being really objective. It was a very subjective approach to everything."

Michael acknowledged that grading student reflections based on rank-ordering also lacked consistency and that his comments rarely provided actionable feedback. We discussed with him the option of an intervention to support him in improving his grading and feedback-giving practices and in creating a designated reflection assessment instrument based on his own beliefs. The initial design of the Reflective Classroom Assessment Protocol (ReCAP) included two central components: using a specialized grading program and prompting reflective practice during grading sessions using a think-aloud (grade-aloud) protocol. A third component, a card sorting task, was later incorporated to compensate for a shortcoming of the grading software. The intervention was carried out as part of the school's ongoing partnership with the research lab during Michael's work hours and based on his availability. He did not receive any additional compensation for participating in this study.

Student-written reflections were seen as a good target assessment task for grade-aloud sessions for three reasons. First, reflective writing is notoriously difficult to assess. In high-school-age students, process reflections are meant to reveal self-regulation and metacognition skills, which are recognized as hard-to-measure constructs. Second, process reflections are commonly and routinely used in educational makerspaces and other SCL classrooms (Peppler *et al.*, 2017). Reflections were regularly assigned to students for formative purposes in the FoD class and high-quality feedback was, therefore, viewed as a powerful lever for helping students improve their work. Finally, Michael acknowledged that he used a heuristic/holistic assessment strategy in reviewing reflective writing assignments and appreciated an opportunity to delve deeper into the process.

The case is limited to the effects of ReCAP on the teacher's practices and does not include any effects the assessment or its corresponding grade and feedback may have had on students.

The study design

This study focuses on three research questions:

RQ1. To what extent did ReCAP sessions help the participating teacher expose and solve grading and feedback-giving dilemmas?

RQ2. To what extent did the participating teacher clarify his assessment criteria and goals during ReCAP sessions?

RQ3. What are the perceived and observed benefits of engaging in ReCAP sessions for one term to the participating teacher's instruction, assessment design, grading and feedback-giving practices?

To answer the research questions posed in this study, we conducted an exploratory mixed-methods case study (Creswell and Clark, 2017) to investigate the effects of computer-supported reflective grading sessions on teacher assessment and feedback-giving practices. In-depth single case-studies are useful in creating a detailed and rich description of the phenomenon of interest (Bleijenbergh, 2009). In this exploratory study, ReCAP was implemented with one section of Michael's FoD classrooms (10 students) over the 2019 Spring term.

The reflective classroom assessment protocol (ReCAP)

ReCAP was designed to support teachers in improving their assessment and feedback-giving practices through computer-supported reflective grading sessions. There are three elements to ReCAP, designed to work conjointly: think-aloud grading sessions to prompt a reflective mindset; using a grading software to track, document and communicate decisions to students; and participating in a card-sorting task to organize insights into a coherent, reusable, grading-and-feedback instrument.

Think-aloud has been identified as one method of promoting deep processing and stimulating critical reflection on current practice (Ericsson and Simon, 1998; Osmond and Darlington, 2005; Sasaki, 2008). In this study, think-aloud played two roles. As an element of the intervention protocol, think-aloud was used to elicit reflective practice. As a data collection method, think-aloud played a secondary role in exposing the inner-workings of Michael's practices for documentation and analysis.

Think-aloud reports expose information held in the subject's short-term memory and are considered to be direct representations of subjects' cognitive processes (Sasaki, 2008). In providing such an account, the subject engages in sense-making, theory building and interpretation, all of which promote deeper reflection (Ericsson and Simon, 1998). We hypothesized that when teachers engage in deep reflection while grading students' work, they will be better able to communicate their thoughts to students and identify gaps and issues in instruction and assessment.

Gradescope (https://www.gradescope.com) is a web-based application designed to support teachers in grading student assignments. The three main benefits of Gradescope, according to its creators, are speed, consistency and flexibility. When using Gradescope, teachers add feedback items, which can be graded or ungraded, to an item pool and refer back to them instead of writing the same comment twice. If the teacher decides to change the point value associated with a feedback item, the change will apply both backward and forward, supporting the teacher in maintaining consistency across assignments. Gradescope also allows the teacher to write specific comments for each student's submission. Feedback from the item pool and specific feedback is digitally communicated back to students when grading is concluded (Atwood and Singh, 2018; Singh *et al.*, 2017).

We decided to use Gradescope as a second element in the intervention to complement the reflective practice with systematic documentation of decisions, creating a digital record of the feedback items Michael created during grading sessions. Gradescope was also useful in organizing and archiving student assignments for future analysis and review. While Gradescope was specifically chosen for this study because of the aforementioned benefits,

other software or documentation methods could have provided Michael with similar support.

Card sorting is a technique taken from user experience research that is commonly used to cluster items into groups and to create informational hierarchies (Card Sorting, 2013). In an open card sorting task, the subject organizes a set of items into groups in any way they see fit and names each group. In this study, Michael was also given the opportunity to review the items in each group and decide whether some were redundant and whether additional items should be added. Including the card sorting task as a third element of the intervention arose due to two central shortcomings of an earlier design. First, at the time of the intervention, Gradescope did not allow for the grouping of similar items together, a function that is now available. As feedback items accumulate, finding the desired one required cumbersome scrolling and repeated reading of the items in the pool. Second, it is natural to get lost in the details during a grade-aloud session, which fosters a bottom-up process, using specific examples to create a larger and more robust understanding of the assessment criteria and goals. We believed that incorporating card sorting would enable Michael to take a step back and organize the item pool into a coherent feedback instrument while promoting higher-order reflection-on-action at the end of the intervention.

This process enabled us to collect a variety of materials to provide us with an in-depth understanding of how ReCAP sessions may have affected Michael's grading and feedback-giving practices.

Data sources

Case study research draws on multiple sources of evidence to inform about the case (Creswell and Poth, 2016; Yin, 2017). In this study, several sources of data were used to draw inferences: recorded grade-aloud sessions, pre- and post-intervention interviews, a recorded card sorting task, graded student reflections from three academic terms (before, during and after the intervention) and the assessment instrument created by the teacher during the intervention.

Grade-aloud transcriptions. Grade-aloud sessions were conducted every time Michael graded students' reflections. Students were required to submit reflections on a bi-weekly basis, which amounted to 7 reflections submitted per student. Altogether, 57 student reflections were graded by Michael in 10 grade-aloud sessions, each between 30 and 75 minutes, spread across the academic term. This amounted to roughly 10 hours of recorded material. In each session, Michael graded the assignments submitted to him by students. During grading and feedback-giving, Michael followed a think-aloud protocol (Van Someren *et al.*, 1994) which required him to verbalize his thought process throughout. As a data collection method, think-aloud is a way to expose one's internal thought process or "inner speech" (Vygotsky, 1962).

Charters (2003) lists the characteristics of tasks that are most likely to elicit accurate think-aloud responses. First, the task should require the subject to exert an intermediate level of cognitive effort. If the task is too challenging, the additional load created by the need to verbalize thoughts can burden the subject and damage the quality of responses. If the task is too easy, the subject might automate portions of the task, a situation in which utterances are less likely to accurately represent the subject's inner speech. Second, it should be possible for subjects to break down the task into smaller parts to help prevent cognitive overloading. Third, naturally verbal tasks are better suited to be investigated using think-aloud methods. As such, the task of assessing and providing feedback for students' assignments is a good fit for the think-aloud method: it is verbal in nature, can be broken down to smaller parts, is unlikely to be easily automated by teachers and – as teachers are

typically experienced in grading – the task should not present them with a cognitively overwhelming challenge.

During each session, one of the authors was present to observe the process, to provide technical support and to prompt Michael to keep verbalizing his thoughts. The researcher avoided providing any pedagogical advice and for the most part acted as a passive sounding board. At times, clarifying questions about specific decisions were asked by the researcher, usually at the end of the session. Occasionally, a 5-minutes exit interview was conducted to check in on how Michael felt after grade-aloud sessions. These were included in the recordings.

Pre- and post-interviews. We conducted two semi-structured interviews with Michael, each following a designated interview protocol (Rubin and Rubin, 2011). The pre-interview was short (35 minutes) and focused on Michael's expectations from assigning students with the task of writing reflections periodically. This included what he hoped to learn from the assignments about students, what constitutes a successful reflection and what skills he intended for students to develop. The post-interview was longer (90 minutes) and focused on four topics: reflections as a source of information, reflections as a driver of student improvement and change, changes to Michael's assessment and feedback-giving practices and an evaluation of how those changes were supported by following ReCAP. Interviews were recorded and transcribed for analysis.

Graded student reflections. Throughout the grade-aloud process, Michael used Gradescope to document and communicate his grading and feedback decisions. All 57 student reflections graded during the ReCAP sessions were used in the analysis for a high-level understanding of the effects of ReCAP on Michael's grading practices. In addition, 91 graded students' reflections from the term before the intervention were analyzed to provide a baseline for how ReCAP affected Michael's practice and 63 graded students' reflections from the term following the intervention were used to gaini insight into whether changes to Michael's practices persisted.

Gradescope item pool. The different iterations of the item pool, created on Gradescope by Michael during the ReCAP sessions, were used to further inform the case.

Card sorting task. The task was recorded and transcribed for analysis. The final grouping created by Michael during the task was also documented for analysis.

The data sources that informed this case study are summarized in Table 1.

Data processing and analysis

All in-person encounters, including interviews, grade-aloud sessions and the card sorting task were transcribed. We, the two authors of this paper, then conducted two rounds of coding of all transcriptions, along with the observation and interview notes. In the first coding round, we engaged in topical coding, in which the coder organizes text passages into the topics that best fit them. Little interpretation is involved in this process, and it mostly serves to group together excerpts relating to the same topic (Richards, 2014). When each of the authors completed coding the data, we compared the coding schemes and discussed areas of overlap or disagreement.

To sort through and better understand how our code systems fit together, we first mapped related codes one onto another and then discussed the merits of codes that did not directly overlap. We then used a card sorting task (using the research codes as items) to obtain a holistic picture of the insights that emerged from the data. We first classified codes as central or peripheral to our research questions. For example, the code "challenging or borderline judgments" were classified as central while "previous grading habits" was classified as peripheral or contextual. We then sorted the central codes according to

Table 1. Summary of data sources

Timing	Source	Materials	Amount/duration
Fall 2018	Google classroom	Graded reflections	91 student reflections
Spring 2019	Pre-interview	Interview recording	35 min
		Interview notes	
	Grade-aloud sessions	Session recordings	10 grading sessions
		Observation notes	7 reflection tasks graded
			9 h and 50 min
			30–75 min per session
	Gradescope	Graded reflections	57 student reflections
		Assessment instrument	
	Card sorting task	Session recording	55 min
		Observation notes	
		Assessment instrument	
	Post-interview	Interview recording	1 h and 30 min
		Interview notes	
Fall 2019	Gradescope	Graded reflections	63 student reflections

similarity which resulted in seven categories. Those were then further collapsed into three larger groups. In doing so, a clear procedural structure of the findings emerged.

The main code groups that emerged were sources/triggers for reflection, dilemmas, and actions/outcomes of reflection. Sources/triggers of reflection included teacher dispositions, assessment objectives and classroom realities. Actions/outcomes of the reflection included changing the assessment, changes in the teacher's feedback-giving practices, changing instruction and reconsidering/re-understanding the assessment objectives.

From this process, it became apparent that the focus of the second round of coding should be on reflective events. These events were characterized by both a source or trigger and action (including the possibility of "no action"). In addition, each author coded reflective events for depth, based on definitions from Hatton and Smith (1995): descriptive information, descriptive reflection and dialogic/critical reflection.

We compared the results of the second round of coding and discussed any disagreement until consensus was reached for all reflective events that differed on at least one code. In conducting the second round of coding in this manner, we were able to cross-tabulate the codes and examine which sources of perplexity were more likely to instigate shallower or deeper levels of reflections and how depth of reflection was tied to subsequent actions taken.

In addition, we conducted rudimentary quantitative analysis on the student reflections Michael graded during the term and compared those statistics to reflections graded in the terms previous to and following the intervention. This allowed us to quantify Michael's grading and feedback-giving practices before, during and after a semester of ReCAP sessions.

Findings

We begin the findings section with a recurring theme that spanned the entire term and helps to frame the challenge facing Michael and likely other teachers in similar situations. During the first few weeks, the ability to clearly define a good process reflection was elusive. Michael explained: "I want my comments to reflect what it is that I want them to do but part of the problem is I'm not sure myself at this stage all of the things that I want them to be doing[. . .] it's that kind of Supreme Court definition of [obscenity], I kind of know it when I

see it sort of thing and I know it when I don't see it [. . .] " In this quote, Michael voiced his struggle in specifying the criteria for a good reflection.

Prior to the intervention, Michael struggled with what makes a good reflection. He relied on his experience and used a heuristic/holistic approach to assessment, captured by the phrase: "I know it when I see it." While this approach is not necessarily unreliable or inaccurate, Michael acknowledged that it does little to serve formative purposes. By using "gut feelings" in grading, Michael had not been communicating to students how to improve their reflective writing skills. He contended with this central challenge throughout the grade-aloud sessions.

The remainder of the findings section is organized in correspondence with the three research questions. The section will cover a selection of grading and feedback-giving dilemmas encountered in the process and how they were managed, an analysis of the way in which assessment criteria and objectives evolved throughout and a summary of the perceived and observed benefits of ReCAP. We conclude the section with a summary of the quantitative findings.

Grading and feedback-giving dilemmas

In assessing his students' work, Michael was confronted with a variety of grading and feedback-giving dilemmas, indicated by different types of triggering events. Those triggers included: confusing or vague remarks from students; unexpected or novel student behaviors; measures of student work against assessment objectives; considerations of classroom management; other classroom realities (i.e., time constraints); and his own dispositions about teaching. Each time a dilemma was encountered, Michael would either use heuristic decision-making or engage in reflection to disambiguate or negotiate the complexities until reaching a satisfactory solution. Resolution of dilemmas led to various potential actions. Those actions included: reconsidering/re-understanding the assessment objectives; changing the assessment criteria and description; changing grading and feedback-giving practices; changing classroom instruction; changing dispositions; negotiating classroom realities; and "no action."

While in some cases there was a clear through-line between triggers, reflective events and subsequent actions, in many cases or even most, a direct connection was hard to draw. Michael would be confronted with student work that would give him pause, triggering a reflective event. Those reflective events often included stream-of-consciousness memories, ideas and diversions. In the following sections, we include quotes of student work when the reader would be able to draw the line between trigger, event and action. However, in other cases, though we include Michael's reaction, we do not include quotes from the student reflection if they do not inform an outside observer about the reflective event.

Reflection as (means to) an end

Michael viewed the reflective writing assignments as serving dual and interconnected purposes. They were designed to help students improve their design skills. However, the reflections would achieve this aim by focusing the students on their own challenges and strategies for overcoming these challenges. Michael recognized that this intermediate effect of self-reflection on self-regulation was something of an end in itself, not just a means to an end. He said: "the essential purpose of the reflection is for them to think about what each of them individually needs to do [to get better] because it's gonna be different in [each] case." In theoretic terms, the assignments would improve self-regulation and metacognition, which would then mediate improvements in design skills. However, because both design skills (e.g. technical drawing) and reflective writing were new tasks for the students, a tension occasionally arose when Michael perceived a student as "incorrectly" identifying their own

challenges. That is, the student was evidencing an ability to reflect. However, from Michael's expert vantage point, they were not accurate about what they were doing wrong.

An example of this tension came to light in the case of a student who wrote that to improve her technical drawing she should "take [her] time and not overthink it." Michael's perspective on this student's classroom performance was the she "seem[ed] to be slow with everything, taking her time." He deemed her improvement solution to be incongruent with his own observations. While a stated requirement of the reflection task is for students to identify ways to improve their practices, Michael did not anticipate situations in which a student's proposed strategy for improvement would not fall within the set of practices he saw as appropriate for that student to use. This dilemma prompted him to think about how to address the situation and as a response, he ultimately created a new feedback item, "your proposed solution does not address the problem."

This episode merits attention because Michael may be seen as mirroring the student's process on his own reflective journey. Reflection-in-action through grade-alouds is also a means to an end for Michael, inasmuch as it can help him improve his own feedback-giving, grading and instruction practices. Becoming a more reflective teacher-practitioner is something of an end in itself. Michael's feedback here communicates to students that the burden of identifying a solution that corresponds accurately to the issues they are facing is part of a successful reflection. Some readers may think that this kind of corrective-only feedback might work against Michael's learning objectives. Rather than acknowledge the student's self-inquiry as a step in the right direction, the student's self-assessment is judged as inadequate, even wrong. By analogy with the student's reflection, it could be argued that Michael's solution (corrective feedback) does not address the problem (of student inexperience and difficulties with reflective writing). The goal of ReCAP, however and our role within the process did not include questioning Michael's feedback and grading decisions. Rather we set out to foster an environment in which those decisions could be externally communicated.

Assumptions and default positions

Effective critical reflection may lead one to identify assumptions that govern the way they view and interpret situations. During grade-aloud sessions, for example, Michael realized that he was "operating under really erroneous assumptions" about some students' math ability. Referring to some students' stated difficulty with fraction arithmetic and unit conversion, he said: "I couldn't wrap my head around the fact that they would have difficulty with these tasks [that] to my mind, were middle school tasks." Michael reported changing his instruction during classroom tasks to give students "what they need" to succeed, including support on those foundational math procedures. This particular mismatch between expectations and observations may be straightforward, but reflective practice can also uncover more problematic assumptions.

For example, Michael revealed a set of conflicting/conditional dispositions with respect to student effort as a non-achievement factor in grading. He indicated it was important to him that a student "makes the effort to do the work, even if she can't do the work because of extenuating circumstances." By extenuating circumstances, Michael was referring to personal and emotional factors students may contend with such as family problems, depression or anxiety. However, at other times he explained: "Sometimes students will just play a game and try to play a teacher and that's always like my first fallback or default position. You know, like 'you're trying to play me and I'm not gonna let you try to play me.' And that's kind of like the default thing." When Michael assumes his default position, he doesn't cut students any slack and becomes more rigid in how he assesses their

assignments. The dilemma revealed during this reflective process concerned the reciprocity between the sincerity of effort and benefit-of-the-doubt or cutting students slack in grading. Following the grade-aloud sessions, he said, "I had to learn how to reset the default, so to speak, for myself and the reflections gave me that opportunity to see where I can [. . .] where I should reset my default position."

Our understanding of this event is that Michael believes that as a teacher he should be "pulling for students" and consider non-achievement factors in assessment (Chen and Bonner, 2017; McMillan and Nash, 2000). However, when students try to play him, he stops pulling for them and reverts back to his default position of treating all students according to the predetermined standards. He sees applying the objective grading rules as something of a punitive measure. In reflecting on this matter, Michael fortified his belief that he should ward off his default position to enable his students' success.

Refinement of assessment and learning goals

Assessment refinement and clarification

Throughout the term, Michael's reflection assessment criteria and goals evolved and developed. Starting from a blank item pool, Michael created 20 feedback items during the 10 grade-aloud sessions. Item creation slowed as the item pool grew more robust. As shown in Figure 1, the majority of the items in the pool were created during the first three reflection assignments, with no new items created during the last two reflections. During the card-sorting task, five more items were created and one was eliminated, resulting in a pool of 24 repeating feedback items (21 graded and 3 ungraded).

During the card-sorting task, Michael began by grouping the 20 feedback items into six categories that he named: the purpose of reflections, writing style, technical aspects of the process, general comments, grammatical errors and negative behavior. In grouping items and examining the contents of each category, Michael noticed that so far he had created only one positive feedback item. The rest pointed out weaknesses in student reflections. He then created five new positive items, asserting that students "should be rewarded for having a good insight into a particular situation [. . .] [and] for expressing whatever issue they are talking about in a particularly good manner." Although ReCAP is not designed to nudge

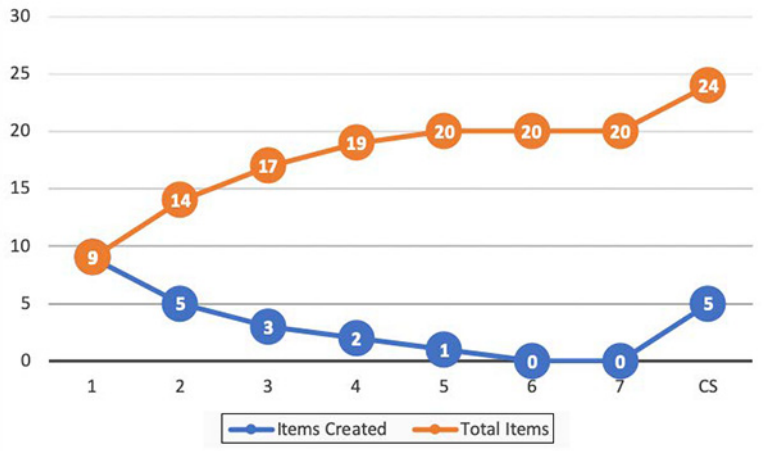

Figure 1.
Items created during each iteration and total items in the pool

teachers deliberately toward practices, we note that pointing out strengths in students' work and not just weaknesses is recognized to be good feedback-giving practice (Shute, 2008).

A second refinement that occurred during the card-sorting task pertained to re-weighting the point values of the feedback items. When Michael created the "purpose of the reflection" item group, he realized that the minor grade deductions given to items in this category did not "give [students] a reason to want to change." He continued that "because it's the heart of it right here, I think it has to be weighted heavier to get them to make the adjustments that I want them to make." As a result, the weight of most of the items in this category was changed from 4% of the overall score to 20%, which presumably signaled importance more clearly to students.

Finally, in grouping items, Michael placed items unrelated to writing quality – having to do with lateness, incompleteness or in one case, making excuses – in a "negative behavior" category. Enforcing grade penalties for late or incomplete work is certainly common practice, although not unchallenged. Measurement experts have historically argued that combining non-achievement factors in grades that supposedly signify achievement is at odds with measurement principles (Brookhart, 1994). What is notable here is that Michael's category label, *negative behavior*, signals quite boldly to students. A student who loses 10% credit for late work might in principle know that the penalty is a disincentive rather than a judgment of quality. However, a student who sees points deducted for negative behavior will understand that they are indeed being evaluated on multiple dimensions, of which reflective writing quality is only one. Interestingly, some of the same critics of grades that collapse multiple dimensions into one have recently come around to recognize the predictive validity, at least in regard to college performance, of this classroom practice (Brookhart *et al.*, 2016).

The final grouped item pool is attached in the Appendix. Grades for reflections were given on a continuous five-point scale.

Learning goal refinement and clarification

In addition to the development and refinement of assessment criteria, the stated goals for the reflection task were negotiated and developed in the process. During the pre-intervention interview, Michael mentioned three distinct, yet inter-related, learning goals of reflective writing. The first two pertained to self-regulation ("I want them to think about what they are doing [. . .] to be able then to analyze what they need to do to be better at what they're doing.") and metacognition: ("self-examination is one skill that I want them to get. I want them to examine their learning process"). In addition to direct learning objectives, Michael felt that the assignment had advisory value as "a personal kind of communication" which would allow him to provide students with individualized supports. These goals were expressed repeatedly over the course of the grade-aloud sessions, but new goals were added as the process progressed.

One goal that Michael articulated for his students was re-evaluating their own understanding and experience of class work: "the thing that reflections will help you with will be whether or not you are making some unexamined assumptions [. . .] And in light of new information or in light of a new experience do you have to go back and reexamine your assumptions." Michael's statement resonates with Mezirow's (1998) approach to critical reflection on assumptions. Indeed, the student work example that precipitated this comment led Michael to the positively-scored rubric item "experiencing an 'a-ha' moment." The student reflected that "To be quite honest, [she] wasn't sure It was possible [to draw the classroom] at first, with all of the tools and equipment that took up the actual room itself. However, in the end, [she had] gotten a good amount of it done before the break began." Michael interpreted this as a process of the student learning to see the classroom differently from its cluttered physical state through the process of technical drawing. In creating a

"more like a platonic sort of shape [rather] than a room," Michael explained, he felt that the student saw "another way to look at something [. . .] to perceive the environment and the reality that she's in."

Another learning goal Michael identified during the ReCAP process was cementing new conceptual knowledge through writing. He elaborated that students are "learning a new vocabulary, they're learning new interactions and how this vocabulary fits together." He thus recognized the role reflective writing can play in facilitating this kind of semantic knowledge restructuring (Vosniadou and Brewer, 1987) and created several rubric items pertaining to the use of terminology. Furthermore, Michael refined and expanded his ideas about the self-regulation goal to specifically include "com[ing] up with a plan to measurably improve their results." The notion of measurable improvement emerged during the grade-aloud session and was attached to an ungraded feedback item ("Going forward, I will ask you to demonstrate those things you've identified as 'things to do to get better' and tell me how I can measure your progress").

The perceived and observed benefits of the reflective classroom assessment protocol

Michael expressed his perceptions of the ReCAP intervention, after using it for one full term, in terms of his consistency/reliability, his creation of "deeper" feedback items and an overall feeling of more intentional engagement with grading, feedback and instruction. Planned and/or enacted changes to instructional practice were also revealed in the grade-aloud sessions and final interview.

"We are going now in a much more rigorous fashion than the other three classes that I've done before [. . .] There was nothing of the depth and the consistency that we have now." Michael attributed part of this transformation to the Gradescope platform itself: "It's forcing me to get better [. . .] I would write superficial comments and didn't have any real point value attached to those comments. And it was just a gut feeling on my part. So it was very inconsistent. Very superficial and very inconsistent, what I was doing in the way I was grading and now it's a lot stronger." Easing the burden of grading while increasing its consistency is doubtless baked into the design of Gradescope. The principal affordance of the application is the creation on-the-fly and reuse of feedback items. Notably, Michael felt that he had created an item pool of high quality.

Feedback was not only more consistent but of a deeper kind, in Michael's perception. He attributed this less to the Gradescope platform than to the ReCAP process: "our whole process, that we've been going through, you and I together, has caused me to really really dig deep in terms of what it is that I'm doing and how I'm doing it in ways that I never did dig deep into. So it's been a definite plus as far as that goes. You know, it's tedious, it's time-consuming, it's agonizing at points, but it's definitely made me better at what I'm doing." The sometimes arduous process of reflecting on grading and feedback-giving dilemmas led Michael to clarification of what students need to support their learning and mastery of new skills.

Alignment with instruction

On several occasions, Michael went into a state of serial questioning when encountering unexpected or novel student behavior.

> [Engaging in grade-aloud sessions] made me want to present valid and honest comments to what it was that they were doing [. . .] That resulted in a whole lot of soul searching in terms of what I've asked them to do, what I expect them to do and the separation between my expectations and their performance. [It] caused me to think a lot about that. They're not doing what I want, Whose fault is that? Is it my fault or is it their fault? Is it, therefore, because they're just not doing it? Or is

> it my fault because I didn't make clear enough to them what it was that I wanted? because maybe I wasn't clear enough myself about what it was that I wanted.

In this process, Michael took a step back to look at the reasons for student performance in the context of what he knew about them, the instructions given to them and his own classroom instruction. In asking such questions, Michael was able to better understand his students' needs and identify potential issues of alignment between assessment design and classroom instruction. Through this process of soul searching, Michael reached several realizations regarding what changes he could make to his instruction that could be useful for his students.

While assessing one student's reflection, Michael commented that he "want[s] her to develop a greater sense of independence." At first, he created an ungraded feedback item which told the student he expected her to develop her independence in class. He later changed his mind about this solution as he continued to think about his own responsibility in supporting his students in becoming more independent. Michael decided that he was looking to "change what [he] was doing during the course of the term as a result of this, because prior to this [process, he] was giving too much individual instruction to the students." Instead of asking students to become more independent, he decided to create a model drawing for students to use as a reference. That way, students "need to become more independent in figuring it out on your own." He extended this notion of modeling to other tasks as well.

In a related realization, Michael theorized that students may benefit from examining good examples of reflection created by their peers. He said that, perhaps, going forward he should "giv[e] them actual examples of what other students have done and my comments to [those] students. We can have a discussion about that [. . .] And how they feel about [my comments]. Do they think [they are] valid [. . .]" In this statement, Michael also implicitly brought up the notion of developing students' assessment literacy (Stiggins, 1995). By being involved in the assessment process, students learn about the goals of assessment and how to differentiate between good and bad responses. In the process, students may develop greater agency and self-regulation skills, while also mastering the requirements of the reflection task (Rust *et al.*, 2003).

Quantitative findings

In this section, we present two brief quantitative analyses from the case. The first set of findings addresses RQ3 using data from graded reflections. Benefits of the intervention may be inferred through direct observation of Michael's grading behaviors, suitably summarized. A second analysis was motivated not by any of our initial research questions, but by a post-hoc inquiry into the possible relationships between depth of reflective event, the type of triggering event and subsequent actions taken. The data for the second analysis were the coded transcripts.

Quantitative changes in grading and feedback

The perceived benefits reported by Michael are supported by the quantitative information collected about the feedback and grades Michael gave before, during and after the intervention. The comparison between the Fall 2018 term (pre-intervention) and the Spring 2019 term (intervention) reveals an overall increase in the amount and frequency of specific (i.e., unique) or repeating feedback given to students. The number of repeating items Michael used prior to the intervention was 10. During the intervention, he created a final pool of 24 items, pointing to either more granular or more comprehensive feedback or both.

Further results of the quantitative comparison are summarized in Table 2 and described next.

During the Fall 2018 term, Michael used an average of 0.90 repeating feedback items per student submission ($SD = 0.42$) compared to 2.39 ($SD = 1.08$) during the intervention [1]. Further, the percentage of student assignments that received specific written feedback rose from 12% to 93%. The length of specific written feedback also increased from a pre-intervention average of 18.8 words ($SD = 12.4$) to 47.2 ($SD = 23.2$) words during the intervention. Perhaps, surprisingly, the mean, range and standard deviation of grades remained remarkably consistent between these two terms.

In trying to observe whether the intervention had any lasting effects on feedback-giving and grading practices, we also analyzed the feedback given by Michael in the post-intervention Fall 2019 term. The comparison revealed a noticeable drop in the average number of repeating feedback items used per reflection from 2.39 to 1.78 ($n = 63$, $SD = 0.87$) and in the percentage of assignments that received specific feedback from 93% to 46% (Those numbers were still notably higher than before the intervention). The average length of specific written feedback, however, remained approximately the same, at 48.2 words ($SD = 29.9$).

In the post-assessment term, the mean reflection grade remained stable, but the variance increased. We believe this can be explained by a confluence of two changes to the rubric made during the card-sorting task. First, Michael increased the point penalty associated with core ("purpose of the reflection") rubric items. At the same time, he added positively-scored items for good observations, use of terminology and well-written reflections. Taken together, these items would increase variance without changing the mean.

Depth of reflection

After coding the grade-aloud transcripts, we examined a cross-tabulation of the (three-level) depth of Michael's reflection with both triggering events and actions he took subsequently. The depth-action analysis (Table 3) revealed that in 48% of shallow reflective events (descriptive information), no action was taken. The second most likely outcome following shallow reflections was "changing the assessment" (17% of the time). This action code was used for the creation of rubric items. In deeper levels of reflection, the most likely action that followed was reconsidering/re-understanding assessment objectives (28% of the time). We also observed that changes to assessment and feedback-giving practices were twice as likely for deep reflective events than shallow (20% versus 10% of the time). Even though we detected only 5 instances of dispositional shifts, all of these occurred following deep reflective events. These findings are consistent with the theoretical framing of this work. Deep reflection is understood to be a prerequisite for formulating understanding (Dewey,

Metric	Pre-intervention (F18)	Intervention (S19)	Post-intervention (F19)
Number of reflections graded	91	57	63
Mean grade (SD)	4.47 (0.38)	4.55 (0.34)	4.49 (0.67)
Grade range	3.3 to 5	3.4 to 5	3 to 5.3
% specific written feedback	11%	95%	46%
Words/feedback (SD)	18.8 (12.4)	47.2 (23.2)	48.2 (29.9)
Avg. repeating feedback items/reflection (SD)	0.9 (0.42)	2.39 (1.08)	1.78 (0.87)

Table 2. Summary quantitative comparison

1933), improving practice (Schön, 1984) and changing dispositions and thinking habits (Mezirow, 1998).

A cross-tabulation of triggers for reflection with the depth of the reflection (Table 4) is also suggestive. When a reflective event was triggered by dispositional conflicts or novel/unexpected challenges it was overwhelmingly likely to lead to a deeper level of reflection (97% and 93% of the time, respectively). Though only 12% of the reflective events recorded were coded as shallow-level, they were more likely to be triggered by classroom management considerations or classroom realities (33% and 29% of the time correspondingly). We report these findings here, even though we feel they are provisional and in need of further substantiation.

Discussion

While the many potential benefits of ReCAP have been described in the findings section, ReCAP is not without drawbacks, the main one being time and energy consumption. In the post-interview, Michael said with a smile, "It's your fault, you're the one that did this to me. You made

Table 3.
Cross-tabulation of depth-of-reflection by action taken (including no action). Proportions are shown such that row totals add up to 100%, with total counts in the last column

Depth of reflection	Reconsidering assessment objectives (%)	Changing grading and feedback-giving practices (%)	Changing the assessment (%)	Negotiating classroom realities (%)	Changing instruction (%)	Changing dispositions (%)	Other actions taken (%)	No action (%)	Total count
Descriptive information	0.07	0.01	0.17	0.01	0.03	0	0.03	0.48	(29)
Descriptive reflection	0.24	0.21	0.15	0.06	0.09	0.01	0.04	0.20	(89)
Dialogic/critical reflection	0.32	0.19	0.11	0.11	0.09	0.04	0.05	0.09	(91)

Table 4.
Cross-tabulation of triggering event by depth-of-reflection. Proportions are shown such that row totals add up to 100%, with total counts in the last column

Triggers	Descriptive information (%)	Descriptive reflection (%)	Dialogic/critical reflection (%)	Total count
External prompts (e.g. researcher)	0.05	0.37	0.58	(59)
New challenges or unmet expectations	0.03	0.01	0.46	(37)
Assessment objectives	0.19	0.43	0.38	(37)
Confusing remarks from students	0.16	0.56	0.28	(32)
Other logistic factors	0.11	0.44	0.44	(18)
Dispositions	0.07	0.36	0.57	(14)
Classroom management	0.33	0.25	0.42	(12)
Classroom realities or pragmatics	0.29	0.29	0.43	(7)

me reflect on what it was that I was doing. You made me aware of the importance of giving them useful feedback. Not superficial comments. It took a lot longer to do these reflections because of that. Much longer than in years past. I mean I could just blow through a whole bunch of them before but [I] just [gave] really superficial, stupid little comments." He goes further to say that for "some of [the reflections], I just really agonized over," pointing out the cognitive demands of a deeply reflective process. However, as described in the findings, it was clear that in Michael's perspective, the effort was worth the gain. We acknowledge that this will not necessarily be the case with any teacher, during every intervention or at any point in time. The selection of Gradescope as the supporting technology was intentional in offsetting a portion of the cognitive and time demands of ReCAP. In addition, it might not be necessary to engage with ReCAP for an entire term if the goals of the teacher are limited to developing a grading-and-feedback instrument. The majority of the items were created while assessing the first three reflections. On the other hand, if a teacher wants to develop their reflective practice, we believe that sustained or periodic engagement with ReCAP would be more beneficial.

We identify several natural extensions of this work. First, as a single in-depth case study, it lacks generalization power. Replicating this process with additional teachers, in different learning environments and for various assessment tasks will form a base of evidence that will increase the validity of any general claims made about ReCAP. Further, ReCAP can be investigated using methods of design-based research (Bakker, 2018) to improve the design of the intervention and make the experience of using it more effective and palatable.

We also see value in going beyond the boundaries of this case and looking into student outcomes (e.g. on-task performance, assessment literacy, perception of feedback) or whether engaging with ReCAP has long-term effects for teachers and students. Finally, for ReCAP to be scalable, we feel there is a need to explore whether the presence of a researcher in the room could be replaced with that of a peer teacher. We believe that if teachers take turns in grading-aloud their students' work, they may both benefit from the process, whether by modeling or by observing.

Conclusion

We have presented the findings from an in-depth case study into the use of ReCAP by one teacher over a full academic term. ReCAP offers a specialized kind of on-the-job professional development in assessment and feedback-giving for teachers. While ReCAP is particularly well-suited for the assessment of hard-to-measure constructs in student-centered learning classrooms, the results could potentially generalize to other types of written reports, not only ones stemming from design activities. ReCAP does not impose any set of assessment criteria or objectives, nor does it ensure that every decision made by teachers will be optimal or even sound. However, ReCAP fosters an environment of reflective practice which empowers teachers to make informed decisions and document those decisions in the form of feedback provided to students and an emergent grading-and-feedback instrument. As such, ReCAP helps teachers externalize their internal decision-making process and open the black box of classroom assessment. We designed the intervention in the hope that such exposure of the internal process would facilitate conversations about the instruments teachers create in the process, whether those take place with peer educators, students or parents. Such conversations are expected to help teachers further improve how they assess their students' assignments.

Note

1. Gradescope requires teachers to assign at least one item from the item pool to consider the assignment as graded. To provide a fair assessment of the change in terms of repeating items

used, we excluded from this analysis the ungraded items "good reflection" and "good effort," which were used in cases where Michael had no material feedback he wanted to give.

References

Allan, E.G. and Driscoll, D.L. (2014), "The three-fold benefit of reflective writing: improving program assessment, student learning, and faculty professional development", *Assessing Writing*, Vol. 21, pp. 37-55.

Atwood, S.A. and Singh, A. (2018), "Improved pedagogy enabled by assessment using gradescope", American society for engineering education annual conference and exposition, salt lake city ut, pp. 24-27.

Bakker, A. (2018), *Design Research in Education: A Practical Guide for Early Career Researchers*, Routledge.

Bell, B. and Cowie, B. (2001), *Formative Assessment and Science Education (Vol. 12)*, Springer Science and Business Media.

Bell, S. (2010), "Project-based learning for the 21st century: skills for the future", *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, Vol. 83 No. 2, pp. 39-43.

Bergeron, L., Schrader, D. and Williams, K. (2019), "Guest editors' introduction: Unpacking the role of assessment in problem-and project-based learning", *Interdisciplinary Journal of Problem-Based Learning*, Vol. 13 No. 2.

Bergner, Y., Abramovich, S., Worsley, M. and Chen, O. (2019), "What are the learning and assessment objectives in educational fab labs and makerspaces?", *Proceedings of fabLearn 2019*, pp. 42-49.

Bevan, B., Gutwill, J.P., Petrich, M. and Wilkinson, K. (2015), "Learning through stem-rich tinkering: findings from a jointly negotiated research project taken up in practice", *Science Education*, Vol. 99 No. 1, pp. 98-120.

Black, P. and Wiliam, D. (2010), "Inside the black box: Raising standards through classroom assessment", *Phi Delta Kappan*, Vol. 92 No. 1, pp. 81-90.

Blackley, S., Sheffield, R., Maynard, N., Koul, R. and Walker, R. (2017), "Makerspace and reflective practice: advancing pre-service teachers in stem education", *Australian Journal of Teacher Education (Education)*, Vol. 42 No. 3, p. 22.

Bleijenbergh, I. (2009), "Case selection", in Mills, A.J., Durepos, G. and Wiebe, E. (Eds), *Encyclopedia of Case Study Research*, Sage Publications, pp. 61-63.

Blikstein, P. and Worsley, M. (2016), "Children are not hackers", *Makerology*, Vol. 1, pp. 64-79.

Brookfield, S.D. (2017), *Becoming a Critically Reflective Teacher*, John Wiley and Sons.

Brookhart, S.M. (1994), "Teachers' grading: Practice and theory", *Applied Measurement in Education*, Vol. 7 No. 4, pp. 279-301.

Brookhart, S.M., Guskey, T.R., Bowers, A.J., Mcmillan, J.H., Smith, J.K., Smith, L.F. and Welsh, M.E. (2016), "A century of grading research: Meaning and value in the most common educational measure", *Review of Educational Research*, Vol. 86 No. 4, pp. 803-848.

Card Sorting (2013), "Department of health and human services", available at: www.usability.gov/how-to-and-tools/methods/card-sorting.html

Charters, E. (2003), "The use of think-aloud methods in qualitative research an introduction to think-aloud methods", *Brock Education: A Journal of Educational Research and Practice*, Vol. 12 No. 2.

Chen, P.P. and Bonner, S.M. (2017), "Teachers' beliefs about grading practices and a constructivist approach to teaching", *Educational Assessment*, Vol. 22 No. 1, pp. 18-34.

Cheng, M.W. and Chan, C.K. (2019), "An experimental test: Using rubrics for reflective writing to develop reflection", *Studies in Educational Evaluation*, Vol. 61, pp. 176-182.

Cornish, L. and Jenkins, K.A. (2012), "Encouraging teacher development through embedding reflective practice in assessment", *Asia-Pacific Journal of Teacher Education*, Vol. 40 No. 2, pp. 159-170.

Creswell, J.W. and Clark, V.L.P. (2017), *Designing and Conducting Mixed Methods Research*, Sage publications.

Creswell, J.W. and Poth, C.N. (2016), *Qualitative Inquiry and Research Design: Choosing among Five Approaches*, Sage publications.

DeLuca, C., Valiquette, A., Coombs, A., LaPointe-McEwan, D. and Luhanga, U. (2018), "Teachers' approaches to classroom assessment: a large-scale survey", *Assessment in Education: Principles, Policy and Practice*, Vol. 25 No. 4, pp. 355-375.

Dewey, J. (1933), "How we think: a restatement of the relation of reflective thinking to the educative process", Vol. 8.

Duckworth, A.L. and Yeager, D.S. (2015), "Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes", *Educational Researcher*, Vol. 44 No. 4, pp. 237-251.

Ericsson, K.A. and Simon, H.A. (1998), "How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking", *Mind, Culture, and Activity*, Vol. 5 No. 3, pp. 178-186.

Farrell, T.S. (2012), "Reflecting on reflective practice:(re) visiting dewey and schon", *Tesol Journal*, Vol. 3 No. 1, pp. 7-16.

Farrell, T.S. (2016), "Tesol, a profession that eats its young! the importance of reflective practice in language teacher education", *Iranian Journal of Language Teaching Research*, Vol. 4 No. 3, pp. 97-107.

Gibson, A., Aitken, A., Sándor, Á Buckingham Shum, S., Tsingos-Lucas, C. and Knight, S. (2017), "Reflective writing analytics for actionable feedback", *Proceedings of the seventh international learning analytics and knowledge conference*, pp. 153-162.

Gigerenzer, G. and Gaissmaier, W. (2011), "Heuristic decision making", *Annual Review of Psychology*, Vol. 62 No. 1, pp. 451-482.

Grabman, R., Stol, T., McNamara, A. and Brahms, L. (2019), "Creating and sustaining a culture of reflective practice: professional development by and for museum-based maker educators", *Journal of Museum Education*, Vol. 44 No. 2, pp. 155-167.

Green, S.K., Johnson, R.L., Kim, D.H. and Pope, N.S. (2007), "Ethics in classroom assessment practices: issues and attitudes", *Teaching and Teacher Education*, Vol. 23 No. 7, pp. 999-1011.

Hatton, N. and Smith, D. (1995), "Reflection in teacher education: towards definition and implementation", *Teaching and Teacher Education*, Vol. 11 No. 1, pp. 33-49.

Hmelo-Silver, C.E. (2004), "Problem-based learning: What and how do students learn?", *Educational Psychology Review*, Vol. 16 No. 3, pp. 235-266.

Hmelo-Silver, C.E., Duncan, R.G. and Chinn, C.A. (2007), "Scaffolding and achievement in problem-based and inquiry learning: a response to kirschner, sweller, and", *Educational Psychologist*, Vol. 42 No. 2, pp. 99-107.

Kahneman, D., Slovic, S.P., Slovic, P. and Tversky, A. (1982), *Judgment under Uncertainty: Heuristics and Biases*, Cambridge university press.

Koole, S., Dornan, T., Aper, L., De Wever, B., Scherpbier, A., Valcke, M., . . . Derese, A. (2012), "Using video-cases to assess student reflection: development and validation of an instrument", *BMC Medical Education*, Vol. 12 No. 1, p. 22.

Martin, L. (2015), "The promise of the maker movement for education", *Journal of Pre-College Engineering Education Research (J-PEER)*, Vol. 5 No. 1.

McMillan, J.H. (2001), "Secondary teachers' classroom assessment and grading practices", *Educational Measurement: Issues and Practice*, Vol. 20 No. 1, pp. 20-32.

McMillan, J.H. and Nash, S. (2000), "Teacher classroom assessment and grading practices decision making", *Presented at the annual meeting of the National Council on Measurement in Education*. ERIC.

Mezirow, J. (1998), "On critical reflection", *Adult Education Quarterly*, Vol. 48 No. 3, pp. 185-198.

Mislevy, R.J. (2013), *Four Metaphors we Need to Understand Assessment*, *The Gordon Commission*, *Princeton*.

Moore, S., Roche, J., Bell, L. and Neenan, E.E. (2020), "Supporting facilitators of maker activities through reflective practice", *Journal of Museum Education*, Vol. 45 No. 1, pp. 99-107.

Murai, Y., Kim, Y., Chang, S. and Reich, J. (2020), "Principles of embedded assessment in school-based making".

Nagro, S.A., DeBettencourt, L.U., Rosenberg, M.S., Carran, D.T. and Weiss, M.P. (2017), "The effects of guided video analysis on teacher candidates' reflective ability and instructional skills", *Teacher Education and Special Education: The Journal of the Teacher Education Division of the Council for Exceptional Children*, Vol. 40 No. 1, pp. 7-25.

Osmond, J. and Darlington, Y. (2005), "Reflective analysis: Techniques for facilitating reflection", *Australian Social Work*, Vol. 58 No. 1, pp. 3-14.

Papert, S. (1991), "Situating constructionism", in Papert, S. and Harel, I. (Eds), *Constructionism*, Ablex Publishing Corporation, Norwood, NJ.

Peppler, K., Keune, A., Xia, F. and Chang, S. (2017), "Survey of assessment in makerspaces", Open Portfolio Project, available at: https://makered.org/wp-content/uploads/2018/02/MakerEdOPP_RB17_Survey_of_Assessments_in_Makerspaces.Pdf

Piaget, J. (1976), "Piaget's theory", *Piaget and His School*, Springer, pp. 11-23.

Plack, M.M., Driscoll, M., Blissett, S., McKenna, R. and Plack, T.P. (2005), "A method for assessing reflective journal writing", *Journal of Allied Health*, Vol. 34 No. 4, pp. 199-208.

Richards, L. (2014), *Handling Qualitative Data: A Practical Guide*, Sage.

Rubin, H.J. and Rubin, I.S. (2011), *Qualitative Interviewing: The Art of Hearing Data*, Sage.

Rust, C., Price, M. and O'Donovan, B. (2003), "Improving students' learning by developing their understanding of assessment criteria and processes", *Assessment and Evaluation in Higher Education*, Vol. 28 No. 2, pp. 147-164.

Sasaki, T. (2008), "Concurrent think-aloud protocol as a socially situated construct", *International Review of Applied Linguistics in Language Teaching*, Vol. 46 No. 4, pp. 349-374.

Schafer, W.D. (1993), "Assessment literacy for teachers", *Theory into Practice*, Vol. 32 No. 2, pp. 118-126.

Schön, D.A. (1984), *The Reflective Practitioner: How Professionals Think in Action (Vol. 5126)*, Basic books.

Shavelson, R.J. (1973), "What is the basic teaching skill", ? *Journal of Teacher Education*, Vol. 24 No. 2, pp. 144-151.

Shute, V.J. (2008), "Focus on formative feedback", *Review of Educational Research*, Vol. 78 No. 1, pp. 153-189.

Singh, A., Karayev, S., Gutowski, K. and Abbeel, P. (2017), "Gradescope: a fast, flexible, and fair system for scalable assessment of handwritten work", *Proceedings of the Fourth (2017) Acm Conference on Learning@ Scale*, pp. 81-88.

Stiggins, R.J. (1995), "Assessment literacy for the 21st century", *Phi Delta Kappan*, Vol. 77 No. 3, pp. 238.

Thorsen, C.A. and DeVore, S. (2013), "Analyzing reflection on/for action: a new approach", *Reflective Practice*, Vol. 14 No. 1, pp. 88-103.

Van Someren, M., Barnard, Y. and Sandberg, J. (1994), *The Think Aloud Method: A Practical Approach to Modelling Cognitive*, *AcademicPress*, *London*.

Vosniadou, S. and Brewer, W.F. (1987), "Theories of knowledge restructuring in development", *Review of Educational Research*, Vol. 57 No. 1, pp. 51-67.

Vygotsky, L. (1962), *Thought and Language*, in Hanfman, E. and Backer, G. (Eds), MIT press, Cambridge, MA.

Waters, R. and McCracken, M. (1997), "Assessment and evaluation in problem-based learning", *Proceedings Rrontiers in Education 1997 27th Annual Conference. Teaching and Learning in an Era of Change*, Vol. 2, pp. 689-693.

Wiggins, G. and McTighe, J. (1998), "What is backward design", *Understanding by Design*, Vol. 1, pp. 7-19.

Wise, S.L., Lukin, L.E. and Roos, L.L. (1991), "Teacher beliefs about training in testing and measurement", *Journal of Teacher Education*, Vol. 42 No. 1, pp. 37-42.

Yin, R.K. (2017), *Case Study Research and Applications: Design and Methods*, Sage publications.

Appendix

Category	Item	Grade
The purpose of reflections	This is more a description of what you did than a reflection of how you did it	−1
	This reflection was more about how you were feeling than what you were doing. I want this reflective process to focus on how to get better at what you are doing in class	−1
	This reflection does not discuss the ways in which you can improve	−1
	Your proposed solution does not address the problem	−1
	An excellent observation	0.3
	Going forward, I will ask you to demonstrate those things you've identified as "things to do to get better" and tell me how I can measure your progress	0
Writing style	Lacks detail	−0.2
	Lack clarity of information	−0.2
	Your experience in class is well presented	0.2
Technical aspects of the process	Incorrect use of terminology	−0.1
	One or more statements is inaccurate	−0.2
	Correct use of terminology	0.1
	Exceptional use of terminology	0.2
General comments	Superficial	−0.3
	Experiencing an "aha" moment	0.3
	Good reflection	0
	Good effort	0
Grammatical errors	Sentence structure/flow is awkward	−0.1
	A few typos that should be corrected before the submission	−0.1
	Reflections should be proofread before submission. There are numerous grammar and sentence structure in this reflection	−0.3
Negative behavior	Nothing submitted = 0 grade	−5
	Late submission	−0.5
	Incomplete; stops in the middle without completing the thought or sentence	−0.5
	Obligations to class supersede obligations to extra-curricular activities	−0.5
Dropped	I'm looking for you to develop the ability to work independently and without frequent direct instruction from me	0

Table A1.
Final item pool

Note: Final item pool generated during the grade-aloud sessions and card sorting task

Corresponding author
Ofer Chen can be contacted at: oc587@nyu.edu