International Journal of STEM Education

RESEARCH

Open Access

Mental simulations to facilitate teacher learning of ambitious mathematics instruction in coaching interactions

Marguerite E. Walsh*, Eben B. Witherspoon, Christian D. Schunn and Lindsay Clare Matsumura

Abstract

Background Many studies have shown that ambitious, "student centered" approaches to STEM instruction benefit K-12 student learning. However, relatively little research has systematically investigated the learning processes that support teachers to skillfully enact these challenging pedagogies. In this study, we used a mixed-methods, case-comparison design to examine one kind of teacher learning routine, Mental Simulations for Teacher Reflection (MSTR), for advancing robust teacher learning in the context of one mathematics-focused instructional coaching intervention. Specifically, this study draws from a large, state-wide representative dataset to select carefully matched, contrasting cases to analyze the quality of coach–teacher conversations for teachers who showed very similar baseline instructional quality but then large differences in levels of improvement. We began by qualitatively coding detailed transcripts from selected coach–teacher pairs as they reflected on lesson artifacts (i.e., lesson plans, student work, and coach observations) using MSTR as an analytical lens. Next, quantitative analyses were conducted to determine the extent to which mental simulations characterized significant differences in the conversations of high- vs. low-instructional growth pairs. Lastly, additional qualitative analyses explored finer-grain distinctions in the quality of mental simulation talk in high- vs. low-growth pairs.

Results Quantitative analyses showed high-growth pairs were significantly more likely to engage in mental simulation talk compared to their low-growth counterparts. Moreover, the high-growth pairs were much more likely to initiate (i.e., raise an instructional ambiguity or problem for discussion) as well as complete (i.e., generate and weigh alternative instructional strategies) a MSTR routine. Qualitative analyses further revealed that engaging teachers' in-depth pedagogical reasoning to connect specific teaching moves to conceptual learning goals in mental simulations was a key distinction of the high-growth coaches.

Conclusions These findings indicate MSTR captured meaningful variation in coaching quality in this context. Notably, all coaches discussed the same instructional topics with teachers (i.e., teaching–learning goals and dimensions) and engaged in the same training that did not explicitly include MSTR, suggesting the possibility that MSTR captured a more implicit process of effective coaches. This study thus offers insight into the 'black box' of teacher learning and how it can be supported in similar professional learning contexts.

Keywords Teacher professional learning, Mental simulations, Instructional coaching, Adaptive expertise

Introduction

Designing and facilitating professional learning experiences that enable STEM teachers to transform entrenched practices and skillfully enact new kinds of instruction is a long-standing challenge (Kennedy, 2016; Tharp & Gallimore, 1991). This challenge has accrued

*Correspondence:
Marguerite E. Walsh
m.walsh@pitt.edu
Learning Research and Development Center, University of Pittsburgh, 3420 Forbes Ave, Pittsburgh, PA 15260, USA

renewed significance as STEM learning reform efforts to implement student-centered instructional approaches have gained momentum across international contexts (Garrett et al., 2019; Resnitskaya & Gregory, 2013; Wilkinson et al., 2015). These instructional approaches emphasize engaging students in authentic, inquiry-focused dialogue to collectively construct meaning and build robust conceptual knowledge. Such instruction is complex because it requires teachers to elicit and respond to students' emergent thinking, which makes teaching a dynamic process rather than a scripted or formulaic act (Kavanagh et al., 2020b).

However, the bulk of STEM classroom activity is still characterized by teacher-centered practices that many believe undermines students' opportunities to deeply engage with academic concepts and tasks (Tekkumru-Kisa et al., 2018). Not only are teacher-centered practices and routines strongly embedded in mainstream STEM education and therefore resistant to change, but student-centered models also require a sophisticated level of knowledge and adaptive skill that many teachers struggle to develop (Sun & van Es, 2015; Wells & Arauz, 2006). A number of teacher professional learning approaches have emerged to address this challenge, such as video clubs (e.g., Sherin & van Es, 2009), instructional coaching (e.g., Fishman et al., 2017), and professional learning communities (e.g., Prenger et al., 2019), all built on the premise that teachers learn best when they are supported to reflect, discuss, and experiment in authentic and collaborative contexts.

Such interventions can be effective for improving instruction, but there is also substantial variation in program designs (Garrett et al., 2019) such that the implementation of even well-specified professional development protocols vary significantly across facilitators (see, e.g., Downer et al., 2009). Importantly, variation in teaching and learning outcomes of teacher professional learning interventions is often underexplored, limiting our knowledge of how and why some programs and features are effective for transforming STEM teaching practice and others are not (Kraft et al., 2018; Osborne et al., 2019).

One underlying issue is insufficient attention to teacher learning mechanisms—i.e., the interactional processes driving teachers' learning—that connect teaching outcomes to a larger theory of change (Kavanagh et al., 2020a; Kennedy, 2016). Especially missing is research to investigate links between specific features of teacher discourse processes during professional learning experiences (e.g., coach–teacher reflective dialogues) and evidence of individual change in teacher knowledge or practice (Lefstein et al., 2020). If the goal is to develop STEM teachers' adaptive skill for noticing and responding to varying student thinking, how do we conceptualize the cognitive processes that support teachers in developing this kind of expertise? What specific yet flexible routines can advance this learning in professional development practice? The present study explores these questions by applying cognitive perspectives that have been extensively studied and supported in other domains (e.g., student learning and expert-novice comparative research) to the study of STEM teacher learning and development. Specifically, we aim to contribute to a theory of teacher learning by identifying and examining one theory-based interactional routine in instructional coaching (i.e., Mental Simulations for Teacher Reflection) to support teachers' learning of ambitious, student-centered math instruction. We propose that theory-guided empirical investigation into the 'black box' of teacher learning and development can provide new insights into what makes for more effective teacher learning interventions.

Theoretical background

Challenges to facilitating ambitious, student-centered instruction

Approaches to classroom education are often characterized in broad terms as a dichotomy between teacher-centered practices that cast teachers as purveyors and students as recipients of knowledge, and student-centered approaches that emphasize students and teachers as co-constructors of knowledge (Gallimore & Tharp, 1990). Some debates remain about the relative merits of teacher- vs student-centered practices on more nuanced levels (e.g., linked to specific kinds of learning tasks or content, see Kirschner et al., 2006; Zhang et al., 2022). Most recently, however, the bulk of instructional quality research and student achievement standards emphasize the advantages of student-centered instructional approaches for supporting robust STEM learning in classrooms (Stein et al., 2015; Windschitl et al., 2008). Moreover, teachers are typically well-versed in teacher-centered forms of instruction (Kennedy, 2016). Thus, the present study is focused specifically on teacher learning mechanisms and routines that support teachers' transition to student-centered approaches to STEM teaching.

Rooted in constructivist learning theory, student-centered models are characterized by teaching practices that aim to elicit and leverage student ideas as the primary basis for lesson planning and enactment, and foster students' conceptual knowledge development by engaging and scaffolding their thinking and reasoning around complex concepts and ideas (Resnick et al., 2015). Across STEM teaching, there is a common focus on supporting students to develop the conceptual knowledge and reasoning and argumentation skills to productively analyze and engage with core disciplinary concepts (Sun & van

Es, 2015). In math education, for example, standards for ambitious teaching include a greater focus on developing students' conceptual understanding and mathematic reasoning skills and less focus on procedural accuracy (NCTM, 2000; Sun & van Es, 2015). Similarly, for ambitious science and engineering teaching, advocates stress engaging students in the authentic inquiry practices such as generating hypotheses, collecting, interpreting, and assessing data, and constructing scientific explanations (Windschitl et al., 2008).

Optimally adapting one's teaching moves and decisions in response to live dialogue between multiple actors presents a steep challenge for novice and experienced teachers alike (Wells & Arauz, 2006). One aspect of this challenge is the degree of uncertainty or ambiguity inherent to an instructional approach that primarily hinges on eliciting and responding to students' in-the-moment thinking. Student-centered pedagogies in STEM are thus 'ambitious' in that they require teachers to quickly identify, interpret, and address significant instructional moments and dilemmas within a highly complex and volatile environment that many teachers (understandably) struggle to navigate effectively (Lampert, 1985). The scope of this challenge is evidenced by research showing that teachers often assimilate new student-centered practices within existing classroom routines and practices (e.g., I-R-E patterns of classroom talk), with the result being that traditional, teacher-centered approaches are functionally maintained (Lefstein et al., 2015).

One problem is that traditional approaches to teacher learning and professional development, long characterized by sporadic and decontextualized lecture-style workshops, were ineffective for meeting these new demands for teacher learning (Kennedy, 2016). In response, recent decades have seen a surge in research and development around 'practice-based' teacher learning programs that provide more robust and meaningful contexts for STEM teacher learning (Ball & Cohen, 1999; Hill & Papay, 2022).

Practice-based learning to develop adaptive teaching expertise

To support teachers' transition from teacher-centered to student-centered approaches, teacher learning research has emphasized situating teachers' learning in authentic contexts (Borko et al., 2008). This research has led to a wide range of 'practice-based' programs that often feature collaborative learning environments with teacher groups or cohorts (see e.g., Borko, et al., 2011; van Es & Sherin, 2008) or individualized learning formats where teachers meet one-on-one with an expert coach or facilitator (see e.g., Murphy et al., 2018; Sedova et al., 2016). These approaches are organized around the central principle that effective teacher learning is fundamentally situated in the artifacts, practices, and discourses that comprise the professional knowledge and on-the-ground work of teaching (Greeno et al., 1996). Much research has explored 'high leverage' practices and features of such programs associated with improved teaching and learning outcomes, including for example a strong focus on instructional content (e.g., particular science activities and concepts); opportunities for ongoing collaboration and practice; and the specification of teaching moves and routines linked to learning goals (Desimone & Pak, 2017).

However, a swell of research has also emerged in response to what some view as an over-emphasis on the 'visible' aspects of teaching (i.e., skills, routines, and behaviors) in practice-based programs at the expense of a greater focus on the goals, assumptions, and principles that give meaning to those behaviors in practice (Kavanagh et al., 2020a; Lefstein & Snell, 2013; Philip et al., 2019). When conceptualizations of effective teaching become distilled into a set of 'high-leverage' practices, they can lead to an over-emphasis on teaching procedures that undermines the foundational principles of responsive teaching (Kennedy, 2016). This could in turn beget a view of expert teaching as akin to 'routine expertise', characterized by high efficiency to perform a standardized set of procedures but low ability to adapt to novel or volatile situations (Hatano & Inagaki, 1986). While procedural knowledge is important for adaptive teaching expertise, it is insufficient for flexibly and skillfully deploying teaching moves in response to shifting contexts for their function—the hallmark of student-centered teaching (Kavanagh et al., 2020a; Shulman, 1986).

The concept of adaptive expertise has therefore been invoked to describe expert performance for the kind of flexible and responsive professional practice necessary for effective student-centered teaching (Hatano & Inagaki, 1986). Adaptive teaching experts "understand when and why to use particular procedures and can associate them with a set of underlying goals that guide their use" (Ghousseini et al., 2015, p. 464). Specifically, hinging one's teaching moves and decisions on student learning (e.g., eliciting and leveraging student ideas; interpreting and scaffolding students' conceptual knowledge development) requires the ability to flexibly and efficiently marshal the declarative ('knowing what'), procedural ('knowing how') and conditional ('knowing why and when') knowledge relevant to a discipline (Bransford et al., 2005; Männikkö & Husu, 2019). Without incorporating conditional thinking based on student input, teacher learning that is only focused primarily on enactment (e.g., implementing 'best practices') may limit teachers' opportunities to develop the professional judgement needed to successfully navigate the inherent

'messiness' or complexity of student-centered teaching (Lefstein & Snell, 2013).

Thus, central to a vision of adaptive teaching expertise in STEM are the specialized thinking and reasoning processes that enable expert teachers to perceive, interpret, and make well-informed in-the-moment decisions responsive to students (Kavanagh et al., 2020b; Sherin & van Es, 2009). These pedagogical reasoning processes, also described as the "thinking that underpins informed professional practice" (Loughran, 2019, p. 4), constitute the core 'mechanism' through which teachers make professional judgements and decisions. Viewed this way, pedagogical reasoning processes are inextricable to expert teaching performance, as they are the invisible cognitive 'work' that bind specific teaching actions and behaviors to more abstract pedagogical goals and meanings (Kavanagh et al., 2020b). The significance of pedagogical reasoning also draws on the notion that STEM teaching practice always involves solving novel problems in dynamic contexts, where teachers face a constant stream of ambiguous situations or "pedagogical dilemmas" (Lampert, 1985) where they must "consistently choose between alternative courses of action, all of which create new pedagogical dilemmas" (Kavanagh et al., 2020a, p. 3). To meet such demands, STEM teachers must rely on well-practiced pedagogical thinking and reasoning skills to make choices that best support student learning in uncertain or ambiguous situations.

Reflection as a mechanism for developing adaptive teaching expertise

Expertise research across domains (e.g., sports, music, chess) consistently emphasizes the critical role of professional learning activities and resources that promote focused deliberation and analysis keyed to specific, core disciplinary goals and skills that underlie expert practice (e.g., Anders Ericsson, 2008). Thus, if the central challenge of adaptive teaching is defined in terms of responding to students in ways optimize their learning, implications for teacher learning include activities and routines designed to: (1) explicitly target the pedagogical reasoning processes involved in achieving specific learning goals; and (2) provide opportunities for teachers to engage socially in the processes that underlie productive decision-making in the face of uncertain or novel pedagogical dilemmas (Kavanagh et al., 2020a).

Decades of research on expert problem solving and deliberate practice further suggests the critical role of planning and reasoning through alternative scenarios and solution strategies for developing and sustaining growth in disciplinary skill and performance (Ericsson, 2006; Mosier et al., 2018). Applied to teacher learning, these activities should thus also engage teachers' hypothetical and counterfactual thinking to, for example, hypothesize alternative strategies that could have been used to achieve better outcomes in past lessons (retrospective), as well as simulating ways to enact these alternatives when similar dilemmas arise in future lessons (prospective) (Bransford et al., 2005). But what are the more specific social and cognitive processes that facilitate this desired type of teacher learning (i.e., understanding the 'why' and 'when' of student-centered teaching moves and routines) and how can they be supported in professional learning contexts?

Reflection on practice is often cited as a core mechanism through which professional learning activities support teacher knowledge and skill development (Tannebaum et al., 2013; Zeichner & Liston, 1996). A rich body of scholarship has evolved to describe the processes through which reflection, particularly expert-guided reflection, is hypothesized to support teacher growth and change (Rodgers, 2002a; Schön, 1983). Similar to cognitive perspectives on expertise, this research has stressed that as a mechanism for professional learning, reflection should be strongly tied to clear purposes (i.e., be goal driven) and concrete plans for future action (Beauchamp, 2015; Rodgers, 2002b). That is, reflective inquiry is most robust when the goal is to discover new insights linked to a specific disciplinary issues, concepts, or practices and has explicit implications for change in behavior or performance.

In this sense, reflection is often seen as a bridge between theory and practice. In Schön's (1983) influential work, the relationship between reflection and practice was further articulated on three levels that each accord with a particular type of deliberative thinking about classroom interactions: retroactive (reflection-on-action), anticipatory (reflection-for-action), or contemporaneous (reflection-in-action). During reflection-on-action, teachers can systematically decompose and study instructional moments and features they may have overlooked in the moment of activity and re-assess their teaching choices in light of shifting pedagogical goals or perspectives (Harlin, 2014). During reflection-for-action, teachers can engage in counterfactual thinking and reasoning about the hypothesized impact of alternative teaching actions "based on their understanding of (co-occurring) cause–effect relationships as well as any mediating processes" (Loughran, 1996; van der Linden & McKenney, 2020, p. 709). In tandem, these reflective processes can therefore enable teachers to mentally rehearse and analyze planned future actions based on prior reflective insights and an informed understanding of the available courses of action and their potential trajectories and impacts in practice.

Notably, causal reasoning to infer and clarify potential causes and effects (what happened and why?) and meanings (why is this significant?) is a central component of all levels of reflection that has important implications for expert judgement and decision-making. Research on expert chess players, for example, has shown that chess masters consistently engage in deliberative study of how and why certain move sequences led to better outcomes given key contextual details (de Groot, 1978; Ericsson, 2006). Similarly, research on problem solving in dynamic and uncertain contexts has shown that experts regularly hypothesize potential causes and corresponding implications for best or 'better' alternative actions (Price et al., 2021).

In the context of teacher learning, significant challenges exist across all levels of reflection. Much research has shown, for example, teachers often struggle to notice the content of student thinking and ideas during reflection on past lessons and tend to superficially evaluate their teaching actions rather than reason in-depth about the impact of their choices on student thinking opportunities (Sherin & van Es, 2009; Sun & van Es, 2015; Tekkumru-Kisa & Stein, 2014). Beyond interpreting past interactions, it can be even more difficult to marshal these reflective processes to plan for future instructional situations and problems (i.e., reflection-for-action). These challenges thus call for a learning context that engages and scaffolds teachers' reflective thinking about how to interpret, respond, and organize their teaching around student thinking and ideas.

Instructional coaching as a context for teacher reflection

Instructional coaching is one kind of model that can scaffold these reflective practices in teacher professional learning and has been shown to be effective for improving teaching and learning outcomes (Correnti et al., 2021; Fishman et al., 2017; Kraft et al., 2018; Matsumura et al., 2019; Resnitskaya & Wilkinson, 2015; Sedova et al., 2016). In such coaching models, collaborative and expert-guided reflection on teaching practice is widely cited as a core mechanism for supporting robust teacher knowledge and skill development (Borko et al., 2008; Correnti et al., 2021; Sedova, 2017). Specifically, coaching interactions can, in theory, provide the opportunity for teachers to engage in reflection-on-and-for-action, building the kind of pedagogical reasoning and problem-solving skills needed to make responsive and informed classroom decisions (i.e., reflection-in-action).

Although a rich body of research has developed around the design, implementation, and evaluation of coaching-based professional learning programs, we still have little understanding of the specific routines and activities that distinguish coaching interactions that directly and consistently encourage this kind of teacher reflection. More specifically, relatively little is known about the kinds of interactional routines in coaching that develop teachers' capacity to make responsive and well-reasoned instructional choices as opposed to a more surface-level, 'pro-forma' adoption of new instructional techniques and routines (Lefstein et al., 2015). Moreover, despite being near-universal in professional learning practice, what makes for productive reflection is often not well-defined or rigorously studied as an empirical context for advancing specific teacher learning principles and change processes (Gaudin & Chaliès, 2015). These problems are underscored by the multitude of research showing disappointing or mixed results across professional learning programs and instantiations (Major & Watson, 2018; van der Linden et al., 2022). Finally, in contrast to the wealth of research on student learning, research on teacher learning more generally has lacked theorization and empirical study around the cognitive mechanisms that support robust teacher learning outcomes.

Mental simulations to build adaptive expertise in coaching

Here we propose a routine based on mental simulations that instructional coaches can use to directly and consistently facilitate the kind of reflection that supports teachers to develop adaptive student-centered teaching expertise. Grounded in cognitive research, mental simulations are a kind of 'what-if' reasoning engaged to solve problems and hypothesize possible alternatives for future action (Christensen & Schunn, 2009; Landriscina, 2015). Mental simulations operate as a form of reasoning based in the systematic manipulation of 'mental models', defined as representations of the causal relationships, dynamics, and processes embodied in a particular situation or domain. Mental models play a key role in human cognition and learning because they support understanding, reasoning, and prediction (Gentner, 2002; Price et al., 2021). For example, one's representation of a particular situation (e.g., the nature of a student's confusion about a topic) can guide future action (revisiting that topic) to attain desired outcomes (student comprehension).

The importance of simulation-based thinking in problem solving is evident in expertise research across disciplines. Ericsson's influential work on deliberate practice, for example, asserts that a key mechanism for continued growth in expert performance is experts' ability to generate and reason through the hypothetical impacts of alternative decisions and evaluate their relative merits (Ericsson, 2006). Similarly, in their recent review of STEM experts' naturalistic problem-solving processes, Price et al. (2021) found that all experts they studied engaged in some form of mental simulations thinking to

"make predictions for dependencies and observables and interpret new information" (p. 12).

Critically, as pointed out by Price et al. (2021), the kinds of 'authentic' problems faced by experts in complex disciplines such as science and engineering are by nature ill-structured and ambiguous. Thus, a core competency in experts' practice is how to solve problems in situations of high uncertainty and incomplete information. For example, one's existing conception of a problem (mental model) might be too inaccurate or imprecise to enable effective judgement and action. Similarly, novel problems and dilemmas might arise in ways that challenge prior assumptions and problem-solving routines. In these cases, mental simulations can be used to reduce uncertainty and build knowledge by systematically thinking through counterfactual alternatives and generating conclusions about likely causes and effects.

We argue that mental simulations, as a means for clarifying and solving problems in situations of high volatility or uncertainty, is a particularly useful frame for characterizing the pedagogical reasoning and decision-making process involved in facilitating student-centered teaching. Other teacher learning researchers have similarly emphasized causal reasoning and inference in the context of teacher rehearsal and debrief discussions (e.g., thought experiments) that enables teachers to reason productively with partial knowledge (whether incomplete or imprecise) or counterfactuals to generate hypotheses that can be tested in future lessons (Keller et al., 2022; Munson et al., 2021). Building these kind of reasoning skills is important because it supports teachers to productively navigate ambiguous or open-ended situations (Forbus, 2002) that are inherent to complex domains, including student-centered teaching. Moreover, the kind of 'what if' reasoning in mental simulation also theoretically supports the development of more abstract or higher-level knowledge, as it requires the mind to infer information that is missing or not precisely known (Chen et al., 2019; Trickett & Trafton, 2007). Simulating the impacts of multiple alternatives by iteratively connecting specific moves and procedures to larger goals increases knowledge of the 'what' (declarative) the 'why' (conceptual) the 'how' (procedural) and the 'when' (conditional)—precisely the kind of knowledge organization and development implicated in adaptive expertise (Carbonell et al., 2014).

As a teacher learning routine, mental simulation is located at the nexus of reflection-on and reflection-for practice. For example, in trying to understand how a classroom interaction went awry during reflection-on-action, a teacher can be supported to consider and make explicit how they are interpreting the dynamics of that situation, including the causal and normative assumptions driving their inferences (i.e., elicit their existing 'mental model'). Once elicited, these assumptions can be critically interrogated, and counterfactual thinking about the causes and effects of alternative teaching assumptions and decisions can be hypothesized (i.e., generate alternative mental models) and systematically weighed against one another to inform decision-making (i.e., reflection-for-action). By constructing multiple alternative scenarios, teachers are also supported to generate more informed predictions and explanations when they encounter similar situations in the future (Trickett & Trafton, 2007).

Mental simulations for teacher reflection (MSTR)

Integrating these cognitive learning theory concepts with research on effective instructional coaching designs and interactional routines, we have developed a framework, termed Mental Simulations for Teacher Reflection or MSTR, that specifies the components of a mental simulation routine aligned with mechanisms for developing adaptive teaching expertise (Walsh, 2021). MSTR especially builds on expertise research on the cognitive processes involved in interpreting, clarifying, and resolving discipline-specific situations and problems, particularly in the context of dynamic spaces where no one perspective or solution strategy readily applies. For example, MSTR incorporates key components and processes associated with naturalistic decision-making in expert practice (e.g., framing and interpreting problem types and sources; selecting and evaluating alternatives, see Price et al., 2021).

MSTR extends this research to the context coaching routines that support adaptive teaching expertise. Specifically, MSTR outlines three basic components needed for a mental simulation routine to support teachers to develop an integrated system of conceptual and procedural knowledge for student-centered teaching and the pedagogical and counterfactual reasoning skills to recognize, interpret, and resolve discrepancies between learning goals and outcomes. These include: (1) establishing ambiguities; (2) proposing alternatives; and (3) weighing alternatives. The first component problematizes an aspect of teaching, establishing a kind of ambiguity related to past (enacted) or future (planned) instructional decisions: no teacher moves (alternatives) yet established for a situation; it is unclear whether the currently proposed alternative will be successful; or that a variety of possible alternatives exist. The second component adds at least one possible teaching move or strategy to address the ambiguity, but often more than one. The third component is iteratively applied to each alternative that is proposed, engaging in-depth reasoning about the relative merits of each proposed alternative relative to targeted

Table 1 MSTR components and descriptions

MSTR component	Description
(1) Establishing ambiguities	'Establishing Ambiguities' situates the context or 'problem space' for the ensuing simulation discussion. In early stages of a coaching conference, this is typically achieved by the coach offering an initial problem statement or interpretation of the larger pedagogical issue or 'dilemma' represented by a particular lesson scenario, either planned (in the case of a Pre-lesson conference) or actualized (in the case of a Post-lesson conference). This component lays the requisite foundation for a simulation discussion to even occur at all, as counterfactual thinking and problem-solving processes can only be engaged when classroom interactions are problematized and recognized as carrying some degree of uncertainty. A new ambiguity 'statement' marks the beginning of a potential new simulation
(2) Proposing alternatives	'Proposing Alternatives' refers to the specification of potential options for teaching moves that could be used to address an established ambiguity. Proposed alternatives can draw on a wide variety of possibilities that range in terms of specificity (e.g., question 'types' vs. specific phrasings) and temporality (i.e., planning moves for an upcoming lesson or hypothesizing alternatives based past events). In a coaching conference, this component is typically initiated by the coach prompting the teacher to offer specific ideas for how to approach an ambiguity or problem that can be subsequently raised for further discussion and inquiry
(3) Weighing alternatives	In 'Weighing Alternatives,' the coach and teacher systematically consider the relative merits of the proposed options for alternative moves, including discussing the ways in which outcomes could vary based on differing student responses or solution strategies. This could involve discussion of: (1) Which alternatives are more or less viable or valuable given particular lesson context (i.e., student learning goals, potential student responses and learning progressions, and learning tasks); (2) Reasons why (or for non-selected alternatives, why not) selected alternatives are useful for advancing student learning goals; and (3) How selected alternatives will be specifically enacted and utilized in subsequent lesson(s)

goals for students' conceptual learning (see Table 1 for further description).

Research questions

In the current study, we empirically test the usefulness of the MSTR framework in accounting for variation in teacher learning outcomes in one instructional coaching context focused on implementing a particular model for ambitious math instruction. Coaching conversations in this context were focused on two specific dimensions of ambitious math instruction: (1) maintenance of cognitive demand in high-level tasks; and (2) exploring and facilitating the public display of student thinking (described in greater detail below). Importantly, we first established that coaches in the study consistently discuss these dimensions of math instruction quality with fidelity during the coaching sessions. Nonetheless, we found varying degrees of teacher 'uptake', as evidenced by coding of their classroom videos along these two dimensions before, during, and after their participation in the coaching. For the present study, we analyzed transcripts of coaching dialogues, contrasting coach–teacher pairs in which teachers showed very little vs. high levels of uptake of the focal instructional practices. We explore whether there were systematic differences in the kinds of talk that occurred between 'low' vs 'high' growth pairs according to our MSTR framework (i.e., differed in the frequency and quality of observed mental simulation talk). Our research questions specifically ask:

RQ1: Are mental simulations more frequent in the coaching dialogues of high-growth teachers relative to low-growth teachers?

RQ2: Are particular components of mental simulations (establishing ambiguities, proposing alternatives, weighing alternatives) more frequent in the coaching dialogues of high-growth teachers relative to low-growth teachers?

RQ3: Are there qualitative differences in the mental simulation talk that distinguish high- vs. low-growth coach–teacher pairs?

Methods

Study context

The current study was part of a larger project examining the impact of instructional coaching on the practice of grades 3–8 mathematics teachers throughout a multi-year, state-wide coaching project in collaboration with the Tennessee Department of Education. A primary goal of the larger project was to design, test, and iteratively refine a model for mathematics instructional coaching developed by the Institute for Learning, a practitioner-focused research program located at the University of Pittsburgh (Russell et al., 2020). The coaching model involved repetitions of a Coaching Cycle that included four main stages (see Fig. 1). Coaches received extensive training in the coaching model before and during implementation. The primary aims of the coach training focused on facilitating teachers' use of complex mathematical tasks, surfacing student thinking about the mathematics in the task, and orchestrating student-centered small-group and whole-class discussions to advance students' mathematical thinking (Stein et al., 2015).

Each coach was partnered with two to four teachers for the study. The coaches engaged in five coaching cycles, labeled Cycle A–E, with each teacher they were coaching. Each cycle was documented through audio recordings

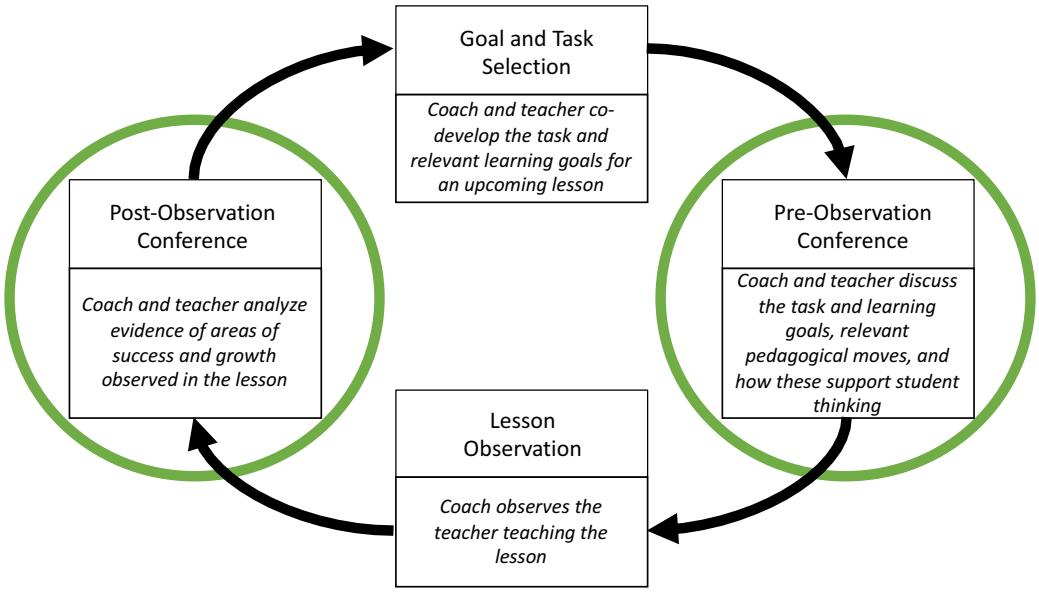

Fig. 1 An overview of the coaching cycle (circles indicate the analytic focus of the current study)

of the pre-observation conference, video-recordings of observed lesson, and audio recordings of post-observation conference. The teachers also video-recorded two uncoached lessons: a lesson that represented their typical mathematics instruction prior to working with their coach (i.e., before Cycle A), and a lesson that represented typical mathematics instruction after working with their coach (i.e., after Cycle E). The seven video-recorded lessons for each teacher (pre, cycles A–E, post) were used to measure teacher uptake of the focal instructional practices (as described in a later section).

While opportunities for teacher learning could exist at all four stages in the coaching cycle, we focused our analyses of frequency and quality of mental simulations on the Pre- and Post-Observation Conferences because those involved the most in-depth interactions between coach and teacher. For the Pre-Observation Conference (often approximately 45 min), conducted one to two days prior to the lesson, the coach and teacher carefully analyze and discuss the teachers' selected task, focusing on the conceptual learning goals and the specific teaching moves and routines that can be used to support those goals. For the Post-Observation Conference (of similar length), conducted within two days of the lesson, the coach and teacher compare and discuss evidence of student learning (gathered during the Lesson Observation) relative to established goals.

Study design

To answer our research questions, we made several strategic sampling and analytic decisions. Broadly speaking, two kinds of approaches are commonly used for examining predictors of learner growth: light experience sampling in many participants (e.g., surveys; infrequent observations) and in-depth experience coding for a small number of participants (e.g., single case studies, microgenetic analyses). Both approaches have advantages and disadvantages (e.g., concerns about confounds, generalizability, and power). Applied to our situation, the choice was between coding a few randomly selected units of coaching dialogue from many teachers in our sample versus coding full transcripts of dialogue from only a few teachers.[1]

For our study purposes, we opted for the second approach, specifically implementing an extreme groups design that analyzes in depth the full transcript data from carefully selected cases. Coding full transcripts was required by the nature of the coaching sessions and the mechanics of the coding process. In particular, coaching sessions were only loosely structured and coding for mental simulations turned out to be an iterative process that required broader knowledge of how coaching conversations unfolded across the entire transcript. Extreme groups sampling supported the study aims and research questions. Specifically, we were interested in whether and to what extent mental simulations characterized the conferring conversations of highly effective coaches compared to their below-average counterparts. Thus, as

[1] The intensive nature of the mental simulations coding, further detailed below, precluded the possibility of incorporating full transcript data from all teachers (n = 420 transcripts) in analysis.

is detailed in the 'case selection' section, we purposefully sampled from a reduced set of cases (i.e., from 'high' and 'low' distribution regions of a more representative sample) to support inferences about characterizing coaching talk patterns linked to more or less robust teacher learning.

Participants

Available sample

The full dataset from the larger project included 32 coaches working with 105 partner teachers. Beginning with such a large dataset allowed us to strategically select teachers for in-depth analysis that entered the coaching program at very similar levels of teaching expertise (i.e., baseline instructional quality), but then demonstrated significantly varied rates of growth. Specifically, we selected four teachers from the pool who began the study at the overall mean but varying in subsequent growth: two teachers showing high growth and two teachers showing low growth (see Case Selection for details). These four teachers interacted with four partner coaches (i.e., each teacher always with the same coach and each teacher had a unique coach). All teachers and coaches consented to participation in the study.

Case selection

Case selection was based on ratings of teachers' classroom instruction by experts in this form of mathematics instruction who were trained for this ratings task. The overall instructional quality rating was based on the extent to which the cognitive demand of high-level instructional tasks was maintained during the lesson (rubric score from 1 to 4) and the degree to which teachers explored and facilitated the public display of student thinking (rubric score from 1 to 4; for further information, see Stein & Kaufman, 2010). These topics constituted the primary focus of the coaching sessions; pilot work established that all of the coaches did indeed focus on these topics throughout all the coaching sessions. The ratings were combined to yield a composite on a scale from 2 to 8. A higher score on this composite therefore represents not only maintenance of cognitive demand of the task during the lesson, but also whether students had the opportunity to engage in and make public their (conceptual) thinking. For this overall measure, an intra-class correlation (ICC) of 0.62 was calculated (similar to previous applications, Russell et al., 2020). ICC scores between 0.6 and 0.74 are commonly considered 'good' for observational data (Cicchetti, 1994; Hallgren, 2012;), indicating adequate inter-rater reliability for this measure. Average scores across raters were used for analysis; it is assumed that conflicting ratings reflected cases in which the classroom involved instruction at mixed levels of cognitive demand or student authority. Finally, to estimate growth on this measure across the pre-test, five coached sessions, and post-test, a model-based cubic growth curve was fit to the full seven points, and pre-values and post-values were then taken from the curve (Russell et al., 2020). This approach reduced the impact of variation in teaching ratings due to extraneous factors (coding noise, lesson details, time of day effects, etc.), providing a more rigorous estimate of instructional growth.

In the overall sample, teachers showed a range of scores at both pre-observation ($M = 5.3$, $SD = 1.8$) and post-observation ($M = 6.9$, $SD = 0.5$). Some teachers had scores on our outcome that were already at the ceiling and had little room to grow; other teachers had scores near the bottom and their coaches needed to focus on basics of selecting appropriate tasks rather than on how to teach using appropriate tasks. Fortunately, the distribution was roughly normal, such that most teachers began in the middle of the instructional quality scale. Therefore, the 63 teachers who started near the mid-point of the teaching quality scale ($M = 5.3$, $SD = 0.1$) were considered for selection.

These 63 coach–teacher pairs on average showed statistically significant growth in teaching quality over time ($M = 1.6$), speaking to the value of the coaching model (Russell et al., 2020). However, there was also variation in growth ($SD = 0.4$), with some teachers making only modest improvements while other teachers made strong improvement. From this subset of mid-point starting teachers, we selected two teachers from the top third of the sample in terms of growth on the cognitive demand scale ($M = 2.0$, $SD = 0.3$), and another two teachers from the bottom third in terms of growth ($M = 1.2$, $SD = 0.4$). We found that selecting four coach–teacher pairs from this subset provided a sufficiently large amount of data to detect large effects (as described in the following sections) while also providing a practically manageable corpus for in-depth coding. Moreover, by randomly selecting coach–teacher pairs within the higher and lower growth regions (rather than the most extreme cases), generalizability to typical more successful and less successful coaching is better supported. This selection resulted in four unique selected coaches (i.e., a different coach for each teacher) that we categorize as either 'high-growth' or 'low-growth' coach–teacher pairs. Both high- and low-growth cases involve urban and non-urban schools. The two high-growth coaches had 5 and 13 years of prior experience teaching math and the two low-growth coaches had 13 and 20 years of such prior experience.

For selection of coaching conferences to systematically code within these four coach–teacher pairs, we omitted the very first round of coach–teacher pre- and post-conferences which tended to be more introductory (i.e.,

Table 2 Mental simulation component codes, inter-rater reliability (Cohen's Kappa, % agreement), coding rules, and brief summaries of typical examples from dataset

Code	Rule	Examples
Ambiguity ($\kappa = 0.70$, 99%)	When either the coach or teacher identified a situation where there was uncertainty about what the teacher would do, including potential outcomes of a particular teaching move or decision relative to the lesson context or concerns about a common problem of practice the teacher is struggling with	*Pre-conference*: Coach asks teacher to consider how the specific choices in her lesson plan might be expected to advance student learning goals *Post-conference*: Teacher expresses uncertainty about what alternative strategies she could have used to achieve better outcomes, while reflecting on what occurred during a previous lesson
Alternative ($\kappa = 0.68$, 98%)	When a teacher or coach raised a possible alternative scenario that could have taken place, as the result of either a different teacher move or a different student contribution	*In pre- or post-conference*: Coach prompts the teacher to brainstorm different moves she might use to help students link ideas; or how she would respond if a student came up with an unanticipated solution strategy
Weighing ($\kappa = 0.66$, 96%)	When the coach and teacher engaged in discussion of the affordances and/or constraints of potential alternative strategies relative to student learning goals, including at least some pedagogically relevant reasons to explain their choices (i.e., alternatives selected and/or evaluated without reasoning did not count as weighing)	*In pre- or post-conference*: Coach and teacher might discuss how the phrasing of a particular question shaped (or could shape) students' thinking opportunities; or what aspects of the mathematics concept would be highlighted if the teacher selected particular student answers to share

Coaching Cycle A) and the last rounds which tended to be more administrative (i.e., Coaching Cycle D and E) and instead focused on the second and third rounds of conversations (i.e., Coaching Cycle B and C) where much of the growth in both years was observed. This resulted in 16 transcripts across our high- vs. low-growth case comparison sample for qualitative coding.

Measures

Our primary independent variables (mental simulations and their components) were operationalized through qualitative coding of the Pre- and Post-lesson coaching conference transcripts using the MSTR framework (see Table 1). When a particular discussion segment included all three components, it was counted as a mental simulation. When a particular discussion segment included only establishing an ambiguity (component 1) but did not also involve both raising and weighing of alternatives (components 2 & 3), it was called a mental simulation initiated but not pursued (see Table 2 for coding rules and examples).

Procedure

The coding process was iterative and highly contextualized across entire transcripts of coach–teacher dialogue. Mental simulations varied considerably in length, at times unfolding across multiple pages of transcript or, conversely, be contained within a few talk turns. This means it was not possible to sample transcripts by a priori units of time or a pre-specified number of transcript lines to code mental simulations: if the selected transcript is mid-simulation, it would not be possible to know whether it was indeed a full simulation (e.g., with an established ambiguity, proposed and weighed alternatives) or just part of a simulation (e.g., an ambiguity that was raised but never resolved). Similarly, sometimes statements initially sounding like alternatives turned out to be simple restatements of a prior idea, which would not be known without coding the larger transcript. Further, mental simulations were sometimes split by various other conversations and then resumed.

For these reasons, coding for mental simulations proceeded in several stages. The first stage involved segmenting transcripts that comprised a mental simulation discussion (as opposed to other kinds of coach–teacher talk). For a segment of dialogue to 'count' as a mental simulation, it had to contain all three basic components. Hence, coders identified all areas of a transcript where an ambiguity was initiated and contained at least some alternatives or weighing in the ensuing dialogue. These codes were used to set the 'boundary' for the beginning and end of each simulation segment, through the application of a binary segmentation code (Simulation = 1, Non-Simulation = 0) to each coach and teacher utterance ($\kappa = 0.88$, 96% agreement). Once we broadly identified the proportion of simulation vs. non-simulation talk for each transcript, the first and second authors then proceeded to code all instances of alternatives and weighing for all mental simulation segments. These codes were applied deductively to the remaining transcripts; this final Cohen's kappa and raw percent agreement are reported for each code as a measure of reliability (see Table 2). It was at this stage of coding that previously identified mental simulation segments and components were often revisited and subsequently revised (e.g., an ambiguity statement indicating the beginning of a new mental simulation or the resumption of a prior one). Finally, these codes were discussed between raters until 100% agreement was reached; this final set of codes were used as data for our quantitative analyses.

The realities of the coding process also made a blind coding approach infeasible. Longitudinal coding of full transcripts meant there were clear differences in the quality of the coaching talk between high- and low-growth coaching pairs. However, several steps were implemented to minimize bias as is typically done in this kind of coding methodology. First, together with a larger group of researchers, we developed and refined the MSTR framework and coding rules over the span of several months using subsets of transcripts drawn from the larger sample (see Tables 1 and 2 for final versions). Second, as noted, the first and second authors independently coded each transcript in the final sample and achieved adequate inter-rater reliability for each mental simulation segment and sub-component (see Table 2). Third, the coding process included weekly meetings with the larger group who would challenge assumptions and offer critiques and suggestions for improving the MSTR framework and its application. Fourth, though we hypothesized that mental simulations would be an important feature distinguishing high- vs. low-growth coaches, we did not have any a priori expectations about how mental simulations might emerge or unfold differently by teacher growth levels. Finally, as part of the results for our third research question, we provide readers with detailed examples (i.e., four in-depth transcript excerpts), thereby increasing transparency. Collectively, these steps minimize the potential for bias and bolster confidence in the validity of our methodological approach and findings.

Analyses

Quantitative analysis

Once all 16 transcripts had been coded and 100% agreement reached on the final set of codes, inferential tests were performed on these codes to see if there were statistically significant differences in the number (per

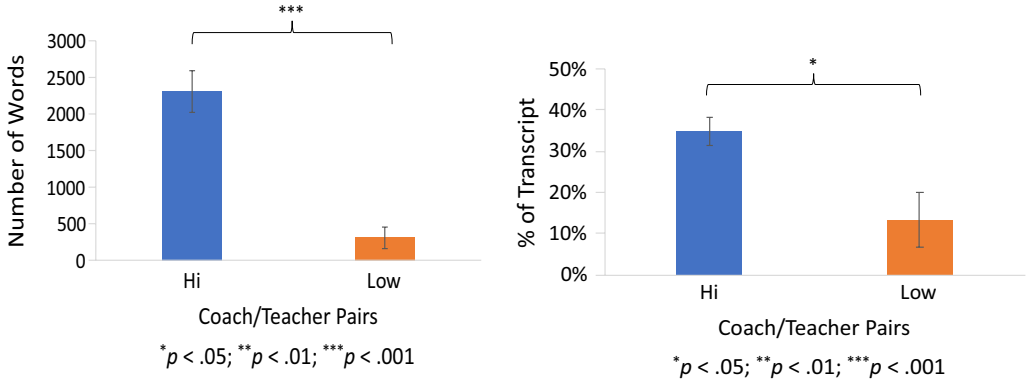

Fig. 2 Mean (and SE bars) per coach–teacher conference by high- vs. low- growth pairs (along with statistical significance of the group contrast) in the **A** total amount of teacher and coach talk (in words) and **B** percent of each conference transcript, that is contained within a Simulation

conference) of each of these components that occurred during the Pre- and Post-Conference conversations between coaches and teacher pairs who showed high growth on our outcome measure, and those pairs who showed low growth on our outcome measure. For differences in means, a *t*-test was performed; for raw counts where the distribution was non-normal, the non-parametric Mann–Whitney test was performed. As discussed in the Results section below, while these analyses only applied to four coach–teacher pairs, the amount of data they contained ($n=16$ transcripts) provided sufficient power to detect statistically significant differences.

Qualitative analysis

Lastly, to answer our third research question, we conducted a final series of thematic and inductive coding (Miles & Huberman, 1994) to more closely examine whether and to what extent notable distinctions emerged between the high- and low-growth coach–teacher pairs in terms of the quality of their Simulation discussions. This process was characterized by several rounds of the first and second authors independently re-analyzing each mental simulation segment for each coach–teacher pair, creating analytic memos (Miles & Huberman, 1994) to note recurring themes and observations both within and across coach–teacher pairs, with a particular focus on key differences between high- vs. low-growth pairs. Once completed, the coders engaged the larger research team in several collaborative discussions and joint analysis of key areas of interest in the transcripts (i.e., discussion segments that either exemplified or potentially challenged emerging themes and patterns) to reach shared conclusions about what we viewed to be the most significant points of contrast in the simulation talk of high- vs. low-growth pairs.

Results

Quantitative differences

Our data showed very large quantitative differences in how high- and low-growth coach–teacher pairs spent their time during Pre- and Post-Conference conversations. In terms of the amount of the words in each Conference conversation that was spent engaged in a Simulation, there was more Simulation discussion between the coach and teachers who showed high growth ($M=2308$ words) than those who showed low growth ($M=318$ words). These differences were shown to be statistically significant ($t(13)=5.97$, $p<0.001$); see Fig. 2a) and very consistent across both of the coach–teacher pairs within each group.

Because the high-growth coach and teacher pairs were more likely to have longer conversations in general, we also wanted to ensure that these differences held when correcting for the length of the transcripts. We therefore conducted another test comparing the amount of time spent in a discussion of a Simulation as a percentage of each transcript. These differences were also significant, with the high-growth pairs spending about 35% of the total discussion engaged in Simulation talk, while the low-growth pairs only spent about 13% of their time engaged in Simulation talk ($t(13)=2.93$, $p<0.05$; see Fig. 2b). These results suggest that the differences we observed in amount of Simulation talk between high- vs. low-growth pairs were not simply driven by more talk overall in the high-growth pairs. Rather, they suggest the high-growth coach–teacher pairs devoted a significantly greater proportion of their time on focused Simulation talk relative to the low-growth pairs.

For the next phase of analysis, we then began to examine differences between each component of a Simulation (i.e., Ambiguities, Alternatives, and Weighing) between the high- and low-growth pairs. First, we

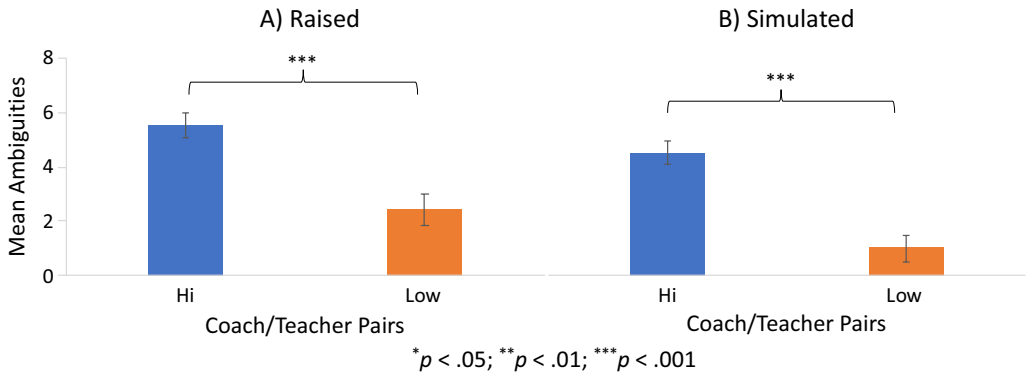

Fig. 3 The mean number (with SE bars) per coach–teacher conference by high- vs. low-growth pairs (along with statistical significance of the group contrast) of **A** all Ambiguities, and **B** all Ambiguities that are raised and then Simulated

tested to see if there was a difference in the number of Ambiguities. While, by definition, Simulations could only occur if an Ambiguity had been raised, it was also possible that Ambiguities were raised that did not turn into Simulations (i.e., never led to alternatives being raised or weighed). We were therefore interested to see if there were differences in the number of Ambiguities, as well as differences in the "take up" and Simulation of those Ambiguities.

As illustrated in the left-hand columns of Fig. 3, the data showed that the high-growth coach–teacher pairs were significantly more likely to raise more ambiguities overall ($t(13) = 4.22$, $p < 0.001$), suggesting these conversations were characterized by significantly more time spent problematizing teaching–learning situations relative to the low-growth pairs. This disparity is notable because it might indicate a greater overall tendency for the high-growth pairs to frame teaching decisions and outcomes (planned or executed) as inherently open for interpretation and further inquiry, and ill-suited for routinization—a key tenet of student-centered instructional practice.

Importantly, of those Ambiguities that were raised, the high-growth pairs were also more likely to follow up and Simulate those Ambiguities by suggesting Alternatives and Weighing them ($t(13) = 5.45$, $p < 0.001$), as illustrated by the right-hand columns of Fig. 2. However, this analysis also revealed that in one of the low-growth pairs, the coach and teacher either never raised Alternatives, and/or never Weighed them, meaning that in this low growth case, there were no Simulations identified. Therefore, because this meant that only one of our low growth cases contained Simulations, additional analyses looking at the relative presence of Alternatives or Weighing within Simulations by the high- and low-growth pairs were underpowered. As such, we continued to examine qualitatively key patterns of coach–teacher talk that distinguished the quality of Simulation talk for high- vs. low-growth pairs (i.e., more or less effective coaching conferences).

Qualitative differences

Our qualitative analyses revealed four themes that characterized diverging patterns of mental simulation talk between the high- vs. low-growth pairs. Two major themes characterized the Simulation talk of the high-growth pairs, both of which involve complete simulations (i.e., contains all three components). Two major themes also characterized the Simulation talk of the low-growth pairs, but are further differentiated by whether or not a complete simulation emerged. Specifically, the first theme involves incomplete simulations (i.e., missing one or more MSTR components), and the second theme involves complete but less sophisticated simulations. Below, we describe each theme and provide illustrative transcript excerpt examples, with key bits of text highlighted in bold.

High-growth theme 1: problematizing lesson components (planned or executed) and proposing specific alternatives linked to learning goals

The first high-growth theme is characterized by the coach consistently encouraging to the teacher identify ambiguities in a planned or enacted lesson scenario, and ensuring that each ambiguity is tied to specific, actionable alternatives. In a pre-lesson conference, for example, a teacher's anticipated lesson tasks and interactions are not treated as 'givens' and left unquestioned but are rather problematized and connected to specific teaching moves (i.e., establishing a clear vision of the 'what' and the 'how'), as illustrated in the excerpt below (Fig. 4). In the discussion leading up this excerpt, the coach had asked the teacher what kinds of strategies she expected students to use in a task focused on identifying equivalent fractions, to

Turn (T)	Speaker	Transcript Excerpt
1	Coach:	So are there any particular strategies that you – so you said you are waiting to see what comes up today. And I think that's fine…**But is there a specific one that you're like it has to come up**? If nobody brings it up, how can we get it come up**.**
2	Teacher:	Let's see. Well, I expect the one half [to come up].
3	Coach:	So one half has to come up. **So if nobody in your class says one half today –how are we gonna bring that up?**
4	Teacher:	We can take our manipulatives out and take out 12 and turn 6 over to different color. And then we can see the relationship between there are 6 that are red, 6 are yellow, **so then hopefully somebody would say half are red or one out of every two, maybe – however they wanna do it.**
5	Coach:	So another thing we could do is say: "**Earlier one of my other classes somebody said this same picture represented one half. How do you think they saw that?"** Do you see one half of this picture?" I mean, you got the picture up there already from the six twelfths. **So do they see that one half as a fraction?**

Fig. 4 Excerpt illustrating high-growth theme 1

which the teacher had initially replied she was just going to "wait and see" what the students come up with. The excerpt below shows the exchange that followed:

Here, the coach responds by problematizing what the teacher had previously framed as an unambiguous situation (T1), pressing the teacher to think more intentionally about advancing specific learning goals. The teacher identifies one strategy—dividing by one-half (T2)—as an important goal of the lesson, and the coach then presses her to specify how they might introduce it if it's not spontaneously brought up by students (T3). Notably, the teacher's suggestion to bring out manipulatives and "hope" that students will identify one-half (T4) does not amount to a specific strategy or question that could be used to advance student thinking. Rather than leaving it there and moving on, the coach offers another alternative (T5) that includes specific questions to assess and advance students' thinking about equivalent fractions. The difference in coaching of leaving non-specific plans as sufficient vs. digging further was a key distinction separating high- vs. low-growth pairs.

High-growth theme 2: iteratively interrogating the relationship between proposed alternatives (what, how) and student learning goals (why, for whom) in Weighing discussions

The second high-growth theme pertains to the quality of the Weighing Alternatives component of MSTR. Specifically, this theme is characterized by the coach iteratively pressing teachers to explain their thinking behind instructional choices and how proposed alternatives are valuable relative to conceptual learning goals (i.e., the 'why') as well as describing how an abstract goal or principle applies in a particular lesson context and how this would be achieved (i.e., the 'what' and 'how'). Similarly, the coach couples any of her own suggested alternatives or claims with reasons and explanations anchored in robust pedagogical goals and concepts. In both cases, teacher and coach are obliged to make explicit their assumptions about how and why selected alternatives do (or perhaps do not) advance learning goals, a including questioning and explicating the math 'behind' a given task. Figure 5a, b, which shows exchanges that followed soon after Fig. 4, illustrates these patterns. This excerpt (below) begins with the coach querying the teacher to consider another strategy that may come up in the lesson, continuing the discussion from above:

In her response, the teacher recalls a previous lesson where a student had tried to divide by thirds, indicating that she had told the student they could not do that because "we don't have equal groups" (T2). In the exchange that follows, the coach presses her to explain her reasoning behind this (Ts 3 & 5) and encourages her to think about an advancing question that could help students make sense of equivalent fractions (T5). This questioning continues as the coach elicits and interrogates the mathematical assumptions implicit in the teacher's lesson plan (T9). Interestingly, as is revealed by her responses in Ts 8, 10, and 12, the teacher seems to reach an impasse as she struggles to explain her reasoning or come up with specific questions to ask students, and in her concluding remarks (T12) appears to confront a limit to her understanding of the mathematics. Notably, had the coach not continued to push the teacher to explain her reasoning and propose specifics for advancing student thinking, they may not have arrived at this juncture.

This line of inquiry continues in the following excerpt (Fig. 5b), the latter half of which also highlights how the coach modeled productive pedagogical reasoning in explicating and justifying her claims and suggestions:

(a)

Turn (T)	Speaker	Transcript Excerpt
1	Coach:	Are there any other ones? […] What about threes, if we had thirds?
2	Teacher:	Somebody tried that– wait, one thirds, let me think. So I think we tried this one before, and **I said we can't do that because we don't have equal groups**. So I told them they either all have to be red, all have to be yellow, or they all have to match.
3	Coach:	**And why is that?**
4	Teacher:	To be equivalent or to show the ratio.
5	Coach:	And so what sense are kids making of that? **Or is there a question we could ask them to help them think about that?**
6	Teacher:	Probably would be better to ask them a question.
7	Coach:	**And what would the question be?**
8	Teacher:	**I'd say it has to be…equal?** *[Laughs]*
9	Coach:	So why would it have to be that? What do want kids to understand about that? **Why do the groups have to be what you're saying**?
10	Teacher:	They have to be equal because... *[Trails off]*
11	Coach:	But they are equal: four, four, and four.
12	Teacher:	Well, they do have to be equal number. But they also have to be equal in makeup. Like if they're – **because you can't have a fraction and a fraction**. So here we see two fourths. Here we see two fourths. Here we see two fourths. Or they all have to be red, or they all have to be yellow or they all have to – if they're mixed they have to match. **I don't know how to describe why- nobody has [ever] asked me why. I just said that's the rule.**

(b)

Turn (T)	Speaker	Transcript Excerpt
13	Coach:	And so I wrote down here that **one of your goals was for students to own the learning**.
14	Teacher:	**They should be telling me why**. *[Laughs]*
15	Coach:	…So let's say that a kid does this and says – **what fraction would he say or she say for this?**
16	Teacher:	I would think they would see one and a half thirds. And then say wait a minute, we can have a fraction and a fraction. And so then they're either stuck or...
17	Coach	So you're saying like this one and a half over three?
18	Teacher	Yeah. Like one –
19	Coach:	**So could you have one and a half thirds?**
20	Teacher:	No.
21	Coach:	**—ever?**
22	Teacher:	Um…
23	Coach:	So there is a world of math where kids do use...
24	Teacher:	Yeah. So if we thought about one and a half thirds, what did we really talk about? Like two thirds and a sixth. Oh my gosh, should I start getting new materials? *[Laughs]*
25	Coach:	No, no. But I just want us to be careful. Remember the article we read about the rules that expire? **That's not a rule that will hold true that you can't write a fraction on top of a fraction**. **So if we talk about one and a half thirds, we're really talking about three sixths.** That would be equivalent to it. And you can write a fraction like this.
26	Teacher	You can.
27	Coach	You can.
28	Teacher	Yeah, so I would think – again, I'm having – I told them the groups have to match. But I really – I'm not sure why.
29	Coach	So let's look. If we were to look at the six twelfths here, and if a kid was to say "I could split it up into three groups", and I go "one, two, three, so my denominator is three." **And so my numerator is one and a half– So we wouldn't want to say "you can't write a fraction like that."** But then we might wanna talk about "**So here we have four eggs that are eaten, and here we have two eggs that are eaten**."
30	Coach	**[Because] it's harder to compare – to talk about fractions when they're not the same size pieces. So I think if we were to look at some of these other representations, kids would see that.** I mean, I think that's something for us to really think about today and watch for our kids doing that and what is that that they're doing? And I do think somebody will do that.

Fig. 5 **a** Excerpt highlighting high-growth theme 2. **b** Excerpt illustrating high-growth theme 2

Here, the coach begins by offering a slight redirect, taking up the teacher's final statement in Fig. 4a indicating she had just told students "that's the rule" (T12) and connecting to a previously stated goal for her students to "own the learning" (T13). This pivot to the teacher's personal goals for her students, followed by prompts for her to consider how a student might specifically think and respond (e.g., T15) leads back to what the teacher was struggling with (improper fractions) from a different perspective (through the lens of possible student solution strategies). Once the teacher's misconception is contextualized (Ts 19–24), the coach then connects back up to a larger principle in teaching math ("rules that expire") and applied to the math goals of the current lesson (T25). The coach then returns to what students might do in the upcoming lesson and how the teacher can help advance their learning in this task, backing her assertions with strong conceptual and pedagogical and reasons (Ts 29–30). These sequences where the coach both presses the teacher to link lesson specifics to larger goals and concepts, as well as models these processes in her own thinking and explanations, was a defining feature of the mental simulation talk of high-growth coaches.

Low-growth theme 1: ambiguity raised, but no alternatives meaningfully proposed or weighed

The first low-growth theme is characterized by an initial statement or question establishing that a teaching–learning situation involves some level of uncertainty or requires further discussion (i.e., ambiguity raised), but alternatives for addressing the situation are not meaningfully proposed or considered in connection to learning goals. One common example of this begins with the coach asking a teacher to describe their lesson plan, targeted learning goals, and the ways in which students might respond to a particular question or task, including strategies that could be used to assess and advance their learning. In the ensuing discussion, the teacher might offer a relatively abstract articulation the targeted learning goals without connecting them to specific teaching moves. Conversely, coach and teacher might discuss a procedural summary of the lesson plan, often in the form of simply recapitulating each planned step in a relatively rote fashion. In both cases, the tendency is for the coach to enjoin the teacher's ideas with evaluative statements or immediately move on to another topic or phase of the lesson. As a result, lesson specifics and alternatives are either superficially linked to learning goals or not at all, as is highlighted in the following excerpt (Fig. 6) concerning a teacher's upcoming lesson about inverse relationships:

Here, the coach begins by raising a potential ambiguity about how students might respond and be supported in the context of the instructional task and goals (Ts 3 & 5). The teacher then iterates the learning goal (T8), to which the coach responds with a relatively superficial evaluative statement (T7). Notably, the coach declines to press the teacher for any specifics about how the task advances this learning goal, instead pivoting to query the teacher about her general plan for the lesson (T10). In what follows, even though the teacher is voicing specific teaching moves and questions to use during this phase of the lesson (e.g., Ts 4 & 6), these are not considered 'true' alternatives because they were not taken up as meaningfully different strategies to be considered relative to student learning goals or progressions (i.e., Weighed). Notably, the coach could have responded by raising up one or more proposed alternatives for further discussion. Instead, she continued to ask the teacher direct questions about the procedures of her planned lesson (e.g., Ts 3 & 5) and moved on without treating the anticipated effects of these moves as an area of legitimate inquiry—i.e., asking 'what' and (to a lesser extent) 'how', but not 'why' (e.g., T9)—and offering suggestions that elicited one-word evaluative answers from the teacher without deeper discussion (e.g., Ts 5–6; 7–8). These kinds of sequences stand in contrast to what was typical in the high-growth pairs, where the coach would press the teacher to explain their reasoning behind a proposed move (rather than accept it as 'given') and, conversely, identify the specifics of teaching moves that could be used to further a learning goal.

Low-growth theme 2: alternatives weighed superficially

The final low-growth theme specifically pertains to the quality of the Weighing Alternatives talk. Specifically, in contrast to the high-growth coaches, when Weighing did occur in the low-growth coaching dialogues, it tended to be relatively superficial, characterized by the coach advancing the conversation past establishing an ambiguity and proposing alternatives, but only just so. The coach might, for example, press the teacher to identify a specific kind of student misconception that might arise during a task and consider how proposed alternatives connect back to a larger learning goal, but the conversation does not advance beyond a relatively vague or surface-level rendering of short answers, direct feedback, or evaluations. As a result, the initial ambiguity is reduced to a low-inference procedural problem with a relatively straightforward (i.e., unambiguous) answer. This theme is illustrated in the following excerpt (Fig. 7) that begins where the teacher is asked to consider what kinds of student misconceptions or struggles she anticipated for an upcoming lesson focused on fractions and varied representation:

Here, the coach the prompts the teacher to specify what kinds of misconceptions students might have and

Turn (T)	Speaker	Transcript Excerpt
1	Coach:	Good. Um… Let's talk about your lesson… that you're gonna teach tomorrow. And… I saw that you picked out the birthday party task.
2	Teacher:	Yeah.
3	Coach:	And… I've read through that, so let's talk about your lesson tomorrow, your goals for the lesson, the specifics of the lesson— and then talk about, um, **how students might respond and strategies we can use to support their learning in this task.**
4	Teacher:	OK, great. That sounds great.
5	Coach:	OK. As you would implement the birthday party task, um… **What… What is your mathematical goal for the lesson?**
6	Teacher:	Well, I think the biggest goal I have is that I want the kids to get to realize, or… to, um… **reinforce that the multiplication and the division are inverse**, um… relationships, and that **you can use either one of those to find division or multiplication facts.**
7	Coach:	**And that's a good task, I think, to, um… to really… for that to be your goal**. I think you've… you selected a good task. True.
8	Coach:	OK, so at the beginning of the class, um… **I guess you're— planning on presenting the past to the kids.**
9	Teacher:	Right.
10	Coach:	**Right. Are you just gonna do that on the Prometheum board?**
11	Teacher:	Yes. I thought **I would put it up and then we would write it together and talk about it, make sure everybody understood what the, um… problem was** that they were gonna work on, right?
12	Coach:	And then, um… **Then, are you gonna give them some self- think time?**
13	Teacher:	Yes. I thought we'd do the private think time for a little bit and then, um… I'm thinking…
14	Coach:	**Do you think we will do, like, group time after that**? Like, they could share with the people at their table what they… what they think privately?
15	Teacher:	**Yes, I think that would be good.**
9	Coach:	OK. Um… **Then after they do that, and they've shared with their group, then you're gonna give them some group time**—
10	Teacher:	Yes.
11	Coach:	—to maybe think about how everybody has… how each one of the people at that table has, um… represented that.
12	Teacher:	Right.
14	Teacher:	**Yes. I think that's really good**, **when they can see each others' ideas and maybe, make, y'know… They'll think about it a different way**.
15	Coach:	**I agree, good.**

Fig. 6 Excerpt illustrating low-growth theme 1

how to address them (T1). After the teacher identifies one way students might struggle to get started with the task (T2) and the coach prompts her for specific ways to address this (T3) the teacher proposes some questions to pose to students (T4) without explanation of why they would be effective. Rather than taking up these proposed alternatives or pressing the teacher to explain how they will advance student thinking, the coach offers another alternative without explanation (T5), which in turn elicits a brief evaluation from the teacher that does not link to the initial ambiguity or conceptual learning goals (T6). The coach then challenges her own suggestion (T7), a move that could have provided a productive context for weighing the merits of this activity relative to learning goals. Instead, she immediately undermines this suggestion before again pivoting to it as an option but without any pedagogically relevant reasons (T7). Notably, the teacher then raises another ambiguity related to whether students should be given a choice (T9), which the coach affirms without elaboration and offers another option (T10) which also elicits a brief affirmation from the teacher (T11). In the final phase of the discussion, the coach pivots to press the teacher to consider how the proposed activities connect back to the larger goal (T14), to which the teacher gives a short response and seeks input from the coach (T15). Similar to T7 above, the coach initially sets up what may have been a productive weighing discussion by not answering directly but

Turn (T)	Speaker	Transcript Excerpt
1	Coach:	**We could figure out some strategies…maybe some misconceptions we think the kids might have and how we might try to… to work through that.**
2	Teacher:	Yes**, I think some of them may have trouble getting started, um… I'm just not sure**. They might **not realize that there's four pans of brownies and each has 16 brownies**. Sometimes, a **few of them still wanna add instead of multiply.**
3	Coach:	OK. Yes. So, what… if they had those misconceptions…**What do you think you could do at that point to help them?**
4	Teacher:	Well, I think I would just try to ask them some questions. Um… I might ask, "**What's the question that you're trying to answer**?" Y'know, for them to actually think about what they're trying to find out. Or, "**How could you find out about how many brownies are in four pans**?"
5	Coach:	OK. Yeah. I think so. Another idea that I had… I wondered if… **we could cut these [brown pieces of paper] into brownie pans**. And then they could actually draw their [brownies]
6	Teacher:	**I think that would be cute.**
7	Coach:	**But, um… is that really gonna help them when they start dividing**? I… I've thought about it. 'Cause I think your idea of the tiles is much better, but we could also use these if they would like to do this.
9	Teacher:	I think that'd be a good thing. It would be something different. **Should we give them a choice**?
10	Coach:	**We could do a choice. Or we could have each group do it**— They could have multiple representations.
11	Teacher:	Yes, 'cause **I think that's really important, that they get more than one way**—— to do it. So I think that's a great idea.
12	Coach:	And then if we go… **If you think back to your goal for the lesson,** is that you really wanted them **to… know the inverse relationship between multiplication and division.**
13	Teacher:	Right.
14	Coach:	So when we think about that, um… **and what they're doing to solve the problem… Do you think… the way that we have been solving problem is really going back to your goal?**
15	Teacher:	**Yes. I do. I think it does. What do you think?**
16	Coach:	[pause] Well, OK. **Let's talk about this.**
17	Teacher:	OK.
18	Coach:	When they make their tiles—
19	Teacher:	Yes.
20	Coach:	—and they divide that out, and then… I think… **I think it would be good for you to… after they do that, for you to point out that relationship**—
21	Teacher:	Yes. Yeah, I think they'll definitely need some, um…guiding for that.

Fig. 7 Excerpt illustrating low-growth theme 2

opening the question up for further discussion (T16). However, the ensuing discussion once again amounts to a fairly superficial sequence offering direct feedback (T20) and an evaluative statement (T21). In contrast, as shown above, the high-growth coaches would more often respond to teachers' suggestions or direct questions with consistent pressing and scaffolding moves and ensuring the pedagogical rationale associated with various alternatives was made explicit.

Discussion

Contributions

In this study, we applied a novel framework, Mental Simulations for Teacher Reflection (MSTR), to conceptualize and study the micro-level processes in coaching interactions that link to robust outcomes in STEM teaching practice. Through a mixed-methods approach, we explored the relationship between MSTR and effective coaching interactions at multiple levels of complexity and specificity—important steps towards de-mystifying the 'black box' of teachers' professional learning (Goldsmith et al., 2014).

MSTR proved to be a useful framework for studying instructional coaching in two ways. First, we found that mental simulations comprised a substantial proportion of the instructional coaching talk, particularly for the high-growth coach–teacher pairs (approximately 30% of the overall talk of high-growth coach–teacher pairs). Notably, this finding was not one of fidelity of implementation where coaches did as they were asked. Mental simulation routines were not an explicit component of the coaching model or coach training (consistent with not being a majority of the coaching talk), suggesting that mental simulations were a substantive but perhaps implicit feature of the highly effective coaching interactions. This finding is especially interesting against the backdrop of research that has often described mental simulations or similar forms of 'what if' reasoning as a key feature of expert problem solving and practice (e.g., Ericsson, 2006; Klein, 2008; Price et al., 2021) but without clear delineation or specificity of the component processes. By identifying the component pieces of a mental simulations routine in a teacher learning context (i.e., Establishing Ambiguities, Proposing Alternatives, Weighing Alternatives), MSTR provides a means to formalize and study in detail a core part of these coaches do with their time spent with teachers. In other words, MSTR provided an explicit language for an otherwise implicit aspect of coaching and provides a framework for understanding the work of highly effective coaches.

Second, MSTR successfully provided an explanation for how coaching interactions can lead to significant variation in teacher outcomes even when the basic coaching model is implemented with fidelity. Specifically, by closely studying the interactional routines of coaches trained and supported to implement the same coaching model, this study contributes insight into the mechanisms that support more or less robust teacher learning and instructional growth. Results from our quantitative analyses, for example, suggest that the high-growth coach–teacher pairs spent a significantly greater amount of time problematizing teaching–learning situations (i.e., raising and exploring ambiguities) relative to the low-growth pairs. Our qualitative analyses added further texture to these results by showing, for example, how the quality of the Weighing component of MSTR—i.e., the pedagogical and counterfactual reasoning processes to simulate the causes and effects of proposed alternatives—was a key distinction of the highly effective coaches. Moreover, these analyses suggested that even when ambiguities were raised in the low-growth coaching dialogues (signaling the start of a potential simulation), they frequently failed to emerge as full simulations, often because viable alternatives would never be specified or if they were, not meaningfully interrogated.

Taken together, these findings suggest that MSTR provides a productive lens for conceptualizing and studying one mechanism underlying STEM teacher learning, particularly one that applies to coach-guided reflection. There already exists a wealth of research on teachers' reflective sensemaking and pedagogical reasoning as key for building STEM teaching knowledge and practice (e.g., Kavanagh et al., 2020a; Sherin & van Es, 2009). The primary contribution of the present study is that it provides empirical support for one conceptually aligned routine for concretely guiding the collaborative work of coaches and teachers in reflection. Extensively studied in other domains of cognitive research, there is strong evidence to suggest mental simulations play a fundamental role in basic inquiry and knowledge-building processes (Landriscina, 2015). For example, research suggests mental simulations are a key process through which causal and counterfactual reasoning can support high-level inference to connect specific instances or 'particulars' to larger ideas and principles (Trickett & Trafton, 2007), reducing uncertainty and clarifying cause–effect relations in engineering and design (Christensen & Schunn, 2009) and generating and evaluating hypotheses in scientific inquiry (Trickett et al., 2009). Adapting this theoretical framework to teacher learning adds clarity and specificity to the kinds of pedagogical thinking and reasoning processes implicated in the development adaptive teaching expertise.

Efforts of this kind are especially needed if we are to better understand how to design and disseminate effective professional learning experiences for STEM teachers. Teacher reflection, for example, while richly rooted in learning theory and nearly ubiquitous in practice-based teacher learning programs, often lacks conceptual clarity or clear learning objectives in practice (Beauchamp, 2015; Lefstein et al., 2020). Missing greater clarity on the 'what' and 'how' of effective teacher reflection, many researchers have raised concerns that reflection as a learning context is often bereft of meaning in practice (Gaudin & Chaliès, 2015). Indeed, the very structured coaching program that was the basis of our study (see Russell et al., 2020) included reflection as a core component, but nonetheless had large variation in outcomes. More specifically, the coaching model require coaches to have teachers reflect on: (1) specific learning goals linked to student's conceptual math understanding, and (2) teaching practices for furthering those goals (e.g., asking students assessing and advancing questions to elicit and further their thinking). It also had routines for doing that reflection before and after teaching using classroom artifacts. But nonetheless, as highlighted in our findings, there were clear differences in the nature and quality of

the reflection practices depending on coach assignment, and MSTR appeared to capture a significant amount of this variation. By analyzing these coaching dialogues through the lens of a well-supported conceptual framework, this study shed some light on why programs such as these might be highly effective for some teachers and less so for others.

Practical implications

MSTR can also be used to inform the work of instructional coaches and other facilitators engaged in similar professional learning contexts. The two ways that low-growth coaches struggled with mental simulation talk each have implications for practice. First, there was evidence of significantly fewer ambiguities overall in the low-growth coaching interactions. That is, these coaching pairs were less likely to raise up teaching–learning situations as open for interpretation and further inquiry. Interestingly, this might also be thought of as violating a core premise of student-centered teaching, but here applied to coaching aimed at teacher learning. Future implementations might benefit from more explicit coach training around how to identify and frame ambiguities in teachers' planned or enacted lessons. This might include, for example, developing protocols and resources around the kinds of ambiguities associated with a particular instructional topic (e.g., common ways in which student misconceptions may emerge and how to address them).

The second primary way in which the low-growth coaches appeared to struggle is that they less frequently leveraged an established ambiguity into a productive conversation that included both proposing and weighing alternatives. This suggests that coach training could be advanced by teaching coaches how to work productively with teachers to reason through their pedagogical challenges, including supporting coaches to recognize and respond to teachers' individual learning needs.

Limitations and future directions

While we have argued that mental simulation offers a robust learning routine in coaching practice, more research is needed to understand whether and to what extent MSTR is effective for instructional coaching in other contexts (e.g., other coaching models, other grades, other pedagogical topics, disciplines other than mathematics). To our knowledge, no other research to date has explicitly applied a mental simulations lens to understand and specify what a robust teacher learning mechanism 'looks like' in the context of instructional coaching. While we believe there is strong rationale to believe MSTR will be conceptually and empirically robust in application to other teachers and coaching contexts, it is still a relatively novel and untested framework. The aim of the present study is to provide a foundation for future research to further explore, extend, and refine the concepts embodied by MSTR.

Similarly, while we believe MSTR contributes empirical clarity and guidance to the largely under-specified context of teacher reflection, we acknowledge certain limitations of our present framing and approach. Most notably, though MSTR as a learning routine is socially situated (i.e., in coach–teacher interactions), we adopted a largely cognitive perspective in how we framed MSTR as a tool for advancing teacher learning in coach-guided reflection (i.e., scaffolding individual teachers' adaptive knowledge and skill development). However, reflection in education has important social, philosophical, and political dimensions (e.g., ideological notions of race, power, and academic ability) that intersect with individual-level cognitive and experiential factors in ways that can profoundly shape teacher learning and practice (Louie, 2020; Philip, 2011). As such, we hope future iterations of MSTR could include more explicit focus on engaging teachers' critical reflection around how societal-level constructs (e.g., normative assumptions about learning and learners) influence their own pedagogical reasoning, decision-making, and actions.

Moreover, our qualitative analyses and findings in particular raise questions about individual differences across coaches and mental simulations discussions. For example, why were the low-growth coaches less inclined to problematize and engage with instructional ambiguities? Are there dispositional, experiential, or knowledge-related factors that could explain their coaching practices? Is it perhaps the case that different kinds of belief systems or lay theories might have tacitly influenced how they framed instructional situations in their conversations with teachers? Similarly, though our findings suggest the importance of pedagogical and counterfactual reasoning in coaching reflections, more research is needed to understand the specifics of what makes these processes effective for change in practice. For example, is it the depth and/or breadth of reasons offered for or against a few alternatives that's important in the weighing process? Or, alternatively, is the variety of alternatives proposed and simulated most important for supporting teachers to flexibly apply this knowledge in practice? More research is needed to better understand the potential influence of these coach and discussion quality factors.

Finally, because this study focused on the combined content of coach–teacher dialogues, we cannot draw any strong inferences about the respective contributions of coaches and teachers in mental simulation talk. For example, is it important that teachers participate heavily in the Weighing discussion or, as suggested by some researchers, is it more important that

this reasoning occurs at all, regardless of who's voicing it? (Haneda et al., 2017; Witherspoon et al., 2021). On a related point, to what extent were differences within teachers driving differences between high- and low-growth pairs? Nested models applied to the full study dataset showed that coaches accounted for between 41 and 48% of variance in teacher outcomes (depending on which outcome measure was used and which covariates were included), suggesting that much of the variation in outcomes were driven by differences within coaches. However, more research is needed to examine the specific factors underlying coach and teacher-level contributions.

Conclusions

Developing teachers' proficiency for student-centered instruction is a challenging yet vital enterprise for improving the learning opportunities and outcomes for a wider range of students in K-12 classrooms. Decades of research have significantly advanced our understanding of how to design and implement high-leverage professional development practices such as instructional coaching (Kraft et al., 2018). However, teacher and student learning outcomes remain highly variable both within and across intervention efforts despite significant research expenditure and wide-scale professional development investment by schools and districts (Jacob & McGovern, 2015). One issue is that professional development research has generally focused less on theorizing and systematically investigating the teacher learning processes and mechanisms that shape differential outcomes across contexts (Lefstein et al., 2020). By integrating key insights from cognitive learning perspectives, this study brings a new perspective to recent, practice-based teacher learning research that primarily draws from situated and socio-cultural learning theory perspectives. This echoes recent calls that education research in general, and teacher learning research in specific, can greatly benefit from a 'de-siloed' approach to advance forward, rather than lateral progress in the field (van der Linden & McKenney, 2020). Specifically, by weaving together research on adaptive expertise, mental simulation, and reflection-based coaching, this study brings additional clarity to the 'what' and the 'how' of effective coaching to promote student-centered teaching practices across schools and contexts.

Acknowledgements
We would like to thank Dena Zook-Howell for her valuable insights and contributions to our analytic framework.

Author contributions
All authors contributed to the paper. MW and EW conducted primary data analyses and writing of the manuscript. CS helped conceptualize the study design and research questions, review, edit, and organize all sections of the paper, and collaborate in the quantitative portion of the analysis. LCM contributed to conceptualization of the analytic framework and collaborated in the qualitative portions of the analysis. All authors participated in the formation of the coding scheme and code reconciliation. All authors read and approved the final manuscript.

Funding
This work was supported by the James S. McDonnell Foundation [Grant Number 220020525]. The views and conclusions contained herein are those of the authors and should not be represented as official policies of the James S. McDonnell Foundation.

Availability of data and materials
Data for quantitative analyses can be included as a supplementary file in published article. Due to IRB constraints, transcript data cannot be publicly shared.

Declarations

Ethics approval and consent to participate
This study was approved by The University of Pittsburgh (IRB # STUDY20010062).

Competing interests
The authors have no conflicts of interest to declare.

Received: 4 July 2022 Accepted: 20 January 2023
Published online: 03 February 2023

References

Anders Ericsson, K. (2008). Deliberate practice and acquisition of expert performance: A general overview. *Academic Emergency Medicine, 15*(11), 988–994.
Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. In G. Sykes & L. Darling-Hammond (Eds.), *Teaching as the learning profession: Handbook of policy and practice* (pp. 3–32). Jossey Bass.
Beauchamp, C. (2015). Reflection in teacher education: Issues emerging from a review of current literature. *Reflective Practice, 16*(1), 123–141.
Borko, H., Jacobs, J., Eiteljorg, E., & Pittman, M. E. (2008). Video as a tool for fostering productive discussions in mathematics professional development. *Teaching and Teacher Education, 24*(2), 417–436.
Borko, H., Koellner, K., Jacobs, J., & Seago, N. (2011). Using video representations of teaching professional development programs. *Mathematics Education, 43*, 175–187.
Bransford, J., Derry, S., Berliner, D., Hammerness, K., & Beckett, K. L. (2005). Theories of learning and their roles in teaching. *Preparing teachers for a changing world: What teachers should learn and be able to do, 40*, 87.
Carbonell, K. B., Stalmeijer, R. E., Könings, K. D., Segers, M., & van Merriënboer, J. J. (2014). How experts deal with novel situations: A review of adaptive expertise. *Educational Research Review, 12*, 14–29.
Chen, Y. C., Benus, M. J., & Hernandez, J. (2019). Managing uncertainty in scientific argumentation. *Science Education, 103*(5), 1235–1276.
Christensen, B. T., & Schunn, C. D. (2009). The role and impact of mental simulation in design. *Applied Cognitive Psychology: THe Official Journal of the Society for Applied Research in Memory and Cognition, 23*(3), 327–344.
Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284.
Correnti, R., Matsumura, L. C., Walsh, M., Zook-Howell, D., Bickel, D. D., & Yu, B. (2021). Effects of online content-focused coaching on discussion quality and reading achievement: Building theory for how coaching develops teachers' adaptive expertise. *Reading Research Quarterly, 56*(3), 519–558.
de Groot, A. D. (1978). *Thought and choice in chess* (2nd ed.). Mouton.
Desimone, L. M., & Pak, K. (2017). Instructional coaching as high-quality professional development. *Theory into Practice, 56*(1), 3–12.

Downer, J. T., Locasale-Crouch, J., Hamre, B., & Pianta, R. (2009). Teacher characteristics associated with responsiveness and exposure to consultation and online professional development resources. *Early Education and Development, 20*(3), 431–455.
Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. *The Cambridge Handbook of Expertise and Expert Performance, 38*(685–705), 2–2.
Fishman, E. J., Borko, H., Osborne, J., Gomez, F., Rafanelli, S., Reigh, E., & Berson, E. (2017). A practice-based professional development program to support scientific argumentation from evidence in the elementary classroom. *Journal of Science Teacher Education, 28*(3), 222–249.
Forbus, K. (2002). Qualitative modeling of common sense understanding. *Cognitive Science Society Virtual Colloquium Series*. Retrieved from: http://cognitivesciencesociety.org/colloquium/archive.html.
Gallimore, R., & Tharp, R. (1990). Teaching mind in society: A theory of education and schooling. In L. Moll (Ed.), *Vygotsky and education: Instructional implications and applications of sociohistorical psychology* (pp. 175–205). Cambridge, UK: Cambridge University Press.
Garrett, R., Citkowicz, M., & Williams, R. (2019). How responsive is a teacher's classroom practice to intervention? A meta-analysis of randomized field studies. *Review of Research in Education, 43*(1), 106–137.
Gaudin, C., & Chaliès, S. (2015). Video viewing in teacher education and professional development: A literature review. *Educational Research Review, 16*, 41–67.
Gentner, D. (2002). Psychology of mental models. *International encyclopedia of the social and behavioral sciences* (pp. 9683–9687). Amsterdam: Elsevier Science.
Ghousseini, H., Beasley, H., & Lord, S. (2015). Investigating the potential of guided practice with an enactment tool for supporting adaptive performance. *Journal of the Learning Sciences, 24*(3), 461–497.
Goldsmith, L. T., Doerr, H. M., & Lewis, C. C. (2014). Mathematics teachers' learning: A conceptual framework and synthesis of research. *Journal of Mathematics Teacher Education, 17*(1), 5–36.
Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. *Handbook of Educational Psychology, 77*, 15–46.
Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23.
Haneda, M., Teemant, A., & Sherman, B. (2017). Instructional coaching through dialogic interaction: Helping a teacher to become agentive in her practice. *Language and Education, 31*(1), 46–64.
Harlin, E. M. (2014). Watching oneself teach–long-term effects of teachers' reflections on their video-recorded teaching. *Technology, Pedagogy and Education, 23*(4), 507–521.
Hatano, G., & Inagaki, K. (1986). Two courses of expertise. In H. Stevenson, H. Azuma, & K. Hakuta (Eds.), *Child development and education in Japan* (pp. 262–272). Freeman.
Hill, H. C., & Papay, J. P. (2022). Building better PL: How to strengthen teacher learning. Retrieved from https://annenberg.brown.edu.
Jacob, A., & McGovern, K. (2015). The mirage: Confronting the hard truth about our quest for teacher development. *TNTP*. Retrieved from https://tntp.org.
Kavanagh, S. S., Conrad, J., & Dagogo-Jack, S. (2020a). From rote to reasoned: Examining the role of pedagogical reasoning in practice-based teacher education. *Teaching and Teacher Education, 89*, 102991.
Kavanagh, S. S., Metz, M., Hauser, M., Fogo, B., Taylor, M. W., & Carlson, J. (2020b). Practicing responsiveness: Using approximations of teaching to develop teachers' responsiveness to students' ideas. *Journal of Teacher Education, 71*(1), 94–107.
Keller, L., Cortina, K. S., Müller, K., & Miller, K. F. (2022). Noticing and weighing alternatives in the reflection of regular classroom teaching: Evidence of expertise using mobile eye-tracking. *Instructional Science, 50*(2), 251–272.
Kennedy, M. M. (2016). How does professional development improve teaching? *Review of Educational Research, 86*(4), 945–980.
Kirschner, P., Sweller, J., & Clark, R. E. (2006). Why unguided learning does not work: An analysis of the failure of discovery learning, problem-based learning, experiential learning and inquiry-based learning. *Educational Psychologist, 41*(2), 75–86.
Klein, G. (2008). Naturalistic decision making. *Human Factors, 50*(3), 456–460.
Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research, 88*(4), 547–588.
Lampert, M. (1985). How do teachers manage to teach? Perspectives on problems in practice. *Harvard Educational Review, 55*(2), 178–195.
Landriscina, F. (2015). The role of mental simulation in understanding and in creating scientific concepts. *Innovazione nella didattica delle scienze nella scuola primaria e dell'infanzia: al crocevia fra discipline scientifiche e umanistiche*, *141*. Retrieved from: https://iris.unimore.it/bitstream/11380/1081802/6/ALTIERO_st2015_PUP_INTERNO.pdf#page=141.
Lefstein, A., Louie, N., Segal, A., & Becher, A. (2020). Taking stock of research on teacher collaborative discourse: Theory and method in a nascent field. *Teaching and Teacher Education, 88*, 102954.
Lefstein, A., & Snell, J. (2013). *Better than best practice: Developing teaching and learning through dialogue*. Routledge.
Lefstein, A., Snell, J., & Israeli, M. (2015). From moves to sequences: Expanding the unit of analysis in the study of classroom discourse. *British Educational Research Journal, 41*(5), 866–885.
Loughran, J. J. (1996). *Developing reflective practice: Learning about teaching and learning through modelling*. Routledge Falmer.
Loughran, J. (2019). Pedagogical reasoning: The foundation of the professional knowledge of teaching. *Teachers and Teaching, 25*(5), 523–535.
Louie, N. (2020). Agency discourse and the reproduction of hierarchy in mathematics instruction. *Cognition and Instruction, 38*(1), 1–26.
Major, L., & Watson, S. (2018). Using video to support in-service teacher professional development: The state of the field, limitations, and possibilities. *Technology, Pedagogy, & Education, 27*(1), 49–68.
Männikkö, I., & Husu, J. (2019). Examining teachers' adaptive expertise through personal practical theories. *Teaching and Teacher Education, 77*, 126–137.
Matsumura, L. C., Correnti, R., Walsh, M., Bickel, D. D., & Zook-Howell, D. (2019). Online content-focused coaching to improve classroom discussion quality. *Technology, Pedagogy and Education, 28*(2), 191–215.
Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Sage.
Mosier, K., Fischer, U., Hoffman, R. R., & Klein, G. (2018). Expert professional judgments and "naturalistic decision making." In K. A. Ericcson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *The Cambridge handbook of expertise and expert performance* (2nd ed., pp. 453–475). Cambridge University Press.
Munson, J., Baldinger, E. E., & Larison, S. (2021). What if…? Exploring thought experiments and non-rehearsing teachers' development of adaptive expertise in rehearsal debriefs. *Teaching and Teacher Education, 97*, 103222.
Murphy, P. K., Greene, J. A., Firetto, C. M., Hendrick, B. D., Li, M., Montalbano, C., & Wei, L. (2018). Quality talk: Developing students' discourse to promote high-level comprehension. *American Educational Research Journal, 55*(5), 1113–1160.
National Council of Teachers of Mathematics. (2000). *Principles and standards for school mathematics*. Reston, VA: Author.
Osborne, J. F., Borko, H., Fishman, E., Gomez Zaccarelli, F., Berson, E., Busch, K. C., et al. (2019). Impacts of a practice-based professional development program on elementary teachers' facilitation of and student engagement with scientific argumentation. *American Educational Research Journal, 56*(4), 1067–1112.
Philip, T. M. (2011). An "ideology in pieces" approach to studying change in teachers' sensemaking about race, racism, and racial justice. *Cognition and Instruction, 29*(3), 297–329.
Philip, T. M., Souto-Manning, M., Anderson, L., Horn, I., Carter Andrews, J., & StillmanVarghese, J. M. (2019). Making justice peripheral by constructing practice as "core": How the increasing prominence of core practices challenges teacher education. *Journal of Teacher Education, 70*(3), 251–264.
Prenger, R., Poortman, C. L., & Handelzalts, A. (2019). The effects of networked professional learning communities. *Journal of Teacher Education, 70*(5), 441–452.
Price, A. M., Kim, C. J., Burkholder, E. W., Fritz, A. V., & Wieman, C. E. (2021). A detailed characterization of the expert problem-solving process in science and engineering: Guidance for teaching and assessment. *CBE—Life Sciences Education, 20*(3), 43.
Resnick, L., Asterhan, C., & Clarke, S. (2015). *Socializing intelligence through academic talk and dialogue*. American Educational Research Association.
Resnitskaya, A., & Gregory, M. (2013). Student thought and classroom language: Examining the mechanisms of change in dialogic teaching. *Educational Psychologist, 48*(2), 114–133.

Resnitskaya, A., & Wilkinson, I. A. (2015). Positively transforming classroom practice through dialogic teaching. In S. Joseph (Ed.), *Positive psychology in practice: Promoting human flourishing in work, health, education, and everyday life* (pp. 279–296). Hoboken, NJ: John Wiley & Sons Inc.
Rodgers, C. R. (2002a). Seeing student learning: Teacher change and the role of reflection. *Harvard Educational Review, 72*, 230–253.
Rodgers, C. (2002b). Defining reflection: Another look at John Dewey and reflective thinking. *Teachers College Record, 104*(4), 842–866.
Russell, J. L., Correnti, R., Stein, M. K., Thomas, A., Bill, V., & Speranzo, L. (2020). Mathematics coaching for conceptual understanding: Promising evidence regarding the Tennessee math coaching model. *Educational Evaluation and Policy Analysis, 42*(3), 439–466.
Schön, D. A. (1983). *The reflective practitioner*. Basic Books.
Sedova, K. (2017). A case study of a transition to dialogic teaching as a process of gradual change. *Teaching and Teacher Education, 67*, 278–290.
Sedova, K., Sedlacek, M., & Svaricek, R. (2016). Teacher professional development as a means of transforming student classroom talk. *Teaching and Teacher Education, 57*, 14–25.
Sherin, M. G., & Van Es, E. A. (2009). Effects of video club participation on teachers' professional vision. *Journal of Teacher Education, 60*(1), 20–37.
Shulman, L. (1986). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*(1), 1–23.
Stein, M. K., Engle, R. A., Smith, M. S., & Hughes, E. K. (2015). Orchestrating productive mathematical discussion: Helping teachers learn to better incorporate student thinking. In *Socializing intelligence through academic talk and dialogue* (pp. 357–388). Washington, DC: American Educational Research Association.
Stein, M. K., & Kaufman, J. H. (2010). Selecting and supporting the use of mathematics curricula at scale. *American Educational Research Journal, 47*(3), 663–693.
Sun, J., & Van Es, E. A. (2015). An exploratory study of the influence that analyzing teaching has on preservice teachers' classroom practice. *Journal of Teacher Education, 66*(3), 201–214.
Tannebaum, R. P., Hall, A. H., & Deaton, C. M. (2013). The development of reflective practice in American education. *American Educational History Journal, 40*(1/2), 241.
Tekkumru-Kisa, M., & Stein, M. K. (2014). Learning to see teaching in new ways: A foundation for maintaining cognitive demand. *American Educational Research Journal, 52*(1), 105–136.
Tekkumru-Kisa, M., Stein, M. K., & Coker, R. (2018). Teachers' learning to facilitate high-level student thinking: Impact of a video-based professional development. *Journal of Research in Science Teaching, 55*(4), 479–502.
Tharp, R. G., & Gallimore, R. (1991). *Rousing minds to life: Teaching, learning, and schooling in social context*. Cambridge University Press.
Trickett, S. B., & Trafton, J. G. (2007). "What if…": The use of conceptual simulations in scientific reasoning. *Cognitive Science, 31*(5), 843–875.
Trickett, S. B., Trafton, J. G., & Schunn, C. D. (2009). How do scientists respond to anomalies? Different strategies used in basic and applied science. *Topics in Cognitive Science, 1*(4), 711–729.
van der Linden, S., & McKenney, S. (2020). Uniting epistemological perspectives to support contextualized knowledge development. *Educational Technology Research and Development, 68*(2), 703–727.
van der Linden, S., van der Meij, J., & McKenney, S. (2022). Teacher video coaching, from design features to student impacts: A systematic literature review. *Review of Educational Research, 92*(1), 114–165.
van Es, E. A., & Sherin, M. G. (2008). Mathematics teachers'"learning to notice" in the context of a video club. *Teaching and Teacher Education, 24*, 244–276.
Walsh, M. E. (2021). *Building Teacher Learning Theory and Research in the Era of Student-Centered Instructional Reforms*. Doctoral dissertation, University of Pittsburgh.
Wells, G., & Arauz, M. R. (2006). Dialogue in the classroom. *The Journal of the Learning Sciences, 15*(3), 379–428.
Wilkinson, I. A., Murphy, P. K., & Binici, S. (2015). Dialogue-intensive pedagogies for promoting reading comprehension: What we know, what we need to know. In *Socializing intelligence through academic talk and dialogue* (pp. 37–50). Washington, DC: American Educational Research Association.
Windschitl, M., Thompson, J., & Braaten, M. (2008). Beyond the scientific method: Model-based inquiry as a new para-digm of preference for school science investigations. *Science Education, 92*(5), 941–967.
Witherspoon, E. B., Ferrer, N. B., Correnti, R. R., Stein, M. K., & Schunn, C. D. (2021). Coaching that supports teachers' learning to enact ambitious instruction. *Instructional Science, 49*(6), 877–898.
Zeichner, K., & Liston, D. (1996). *Reflective teaching: An introduction*. Lawrence Erlbaum.
Zhang, L., Kirschner, P. A., Cobern, W. W., & Sweller, J. (2022). There is an evidence crisis in science educational policy. *Educational Psychology Review, 34*(2), 1157–1176.

Publisher's Note