CAPTURING INSTRUCTIONAL PRACTICE AT SCALE: CONCEPTUALIZING AND DESCRIBING THE PROFESSIONAL PRACTICE OF TEACHERS AND COACHES

A Dissertation

Presented to

The Faculty of Education and Human Development

University of Virginia

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

by Arielle Boguslav

May 2023

Educational Policy Studies
School of Education and Human Development
University of Virginia
Charlottesville, Virginia

APPROVAL OF THE DISSERTATION

This dissertation, ("Capturing Instructional Practice at Scale: Conceptualizing and Describing the Professional Practice of Teachers and Coaches"), has been approved by the Graduate Faculty of the School of Education and Human Development in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

____
Julie Cohen (Chair)

____
Allison Atteberry

____
Beth Scheuler

____
Vivian Wong

DEDICATION

This dissertation is dedicated to all the teachers, administrators, and other education professionals who have dedicated their careers to supporting student learning and well-being.

ACKNOWLEDGEMENTS

Like raising children, completing a doctoral degree takes a village, without whom I never could have gotten to this point. I cannot express enough my gratitude to everyone in the EdPolicyWorks village. So much of my learning and progress is due to the interactions I've had with faculty, staff, and students in classes, at workshops, in hallways, and at social events. I am especially grateful to those people with whom I've worked on research projects or had the opportunity to discuss my dissertation: Jim Wyckoff, Luke Miller, Daphna Bassock, Sara Rimm-Kaufman and of course the members of my dissertation committee – Julie Cohen, Allison Atteberry, Vivian Wong, and Beth Scheuler. As committee chair and my advisor, Julie Cohen has been instrumental in my development as a scholar. Thank you for seeing my potential and nurturing it over the last five years.

I am also grateful to the many people outside of UVA who have provided feedback and support throughout my doctoral program. To Kylie Anglin, thank you for bringing me on board to help you think about coaching quality and fidelity – that was a pivotal moment that helped me developed my coaching moves work in Chapters 2 and 3. To Josh Goodrich, thank you for sharing your deep understanding of what it's like to be a practitioner of coaching and pushing my thinking. To Nick Kochmanksi, Lynsey Gibbons, Heather Hill, Kavita Matsko, and Abby Reisman, thank you for taking the time to engage with my coaching moves taxonomy and provide feedback.

I also want to acknowledge the many friends that have supported my doctoral journey. To the Ed Policy students, thank you for all the intellectual and emotional support you've provided. I especially want to thank Walter Herring and Dan Rodriguez for their friendship and support – our talks and walks almost made me forget the challenge of being a cohort of one – until you both left a year ahead of me! To Katie Waddell, thank you for holding me accountable to getting my work done, sharing your trials and tribulations, and letting me vent mine. To Emma Weinrich, our weekly check-ins have had a huge role in keeping me sane and motivated through the dark times of the pandemic and isolation. To Paige Linnell, thank you for believing in my capacity to contribute to education research and for helping me to believe in myself.

Finally, thank you to the extended Boguslav village: my parents, Bruce and Linda; my siblings, Louis and Mayla; my husband, Reagan; and all of the extended family that came with him when we got married. Each of you has had a role in encouraging me, supporting me, and distracting me (when needed) over the course of the last five years. Mom and Dad, thank you for always being so excited to brag about my accomplishments, even though it sometimes makes me feel a little awkward. Mayla, thank you for being the trailblazing younger sister who lit the path for me to follow in your footsteps to a PhD. Louis, thank you for the many conversations we've had about my dissertation and your genuine curiosity about my work. To the Merediths, thank you for welcoming me so completely into your family and celebrating my accomplishments as your own. To Reagan, thank you for having the patience to wait for me to find you and for maintaining that patience as we've built our lives in the midst of the intensity, stress, and chaos of finishing a dissertation.

TABLE OF CONTENTS

Dedication..iv

Acknowledgements v

Elements

Dissertation Overview 1

Chapter 1 22

Chapter 2 74

Chapter 3 131

DISSERTATION OVERVIEW

Great teachers change students' lives. Just having a single great primary school teacher can increase a student's lifetime earnings by $16,000 (Chetty et al., 2010). Great teachers increase academic achievement, increase college attendance, and decrease teenage pregnancy, among other outcomes (Chetty et al., 2014; Rivkin et al., 2005). At the same time, teacher effectiveness is highly variable and often the most disadvantaged students are assigned to the least effective teachers (Akiba et al., 2007; Allen et al., 2018; Atteberry et al., 2015; Chetty et al., 2010; Clotfelter et al., 2005; Dolton & Newson, 2003; Grissom, 2011; Kane et al., 2008; Ost, 2014; Rivkin et al., 2005). Improving teacher effectiveness requires, among other things, effective teacher preparation and professional development experiences that support the development of professional expertise and enable teachers to successfully adapt to changing conditions and incorporate new research insights into their practice (Didion et al., 2020; Johnson, 2006; Kraft et al., 2018; Lynch et al., 2019; Pellegrini et al., 2021).

When it comes to providing high quality instruction for students, backwards-planning from learning goals and ongoing, formative assessments of student learning to evaluate progress toward these goals are two bedrock principles supported by strong bodies of evidence and increasingly embedded in school curricula and policies (Bennett, 2011; Graff, 2011; Jones et al., 2009). So too might we expect such principles to be important foundations for providing teachers with high quality preparation and development experiences. Indeed, substantial efforts have been made over the last thirty years to articulate key learning goals for teacher preparation and development and develop tools for assessing the quality of teachers' instruction. Teacher education programs have adopted common professional standards to guide teacher preparation curricula (Council for the Accreditation of Educator Preparation, 2013). Teacher education researchers have developed frameworks for decomposing the complex work of teaching into practices, strategies, and techniques that can further guide the content of teacher preparation courses (Grossman & Dean, 2019; McDonald et al., 2013). At the same time, observational measures of teaching quality are now widely used to guide teacher professional development efforts and evaluate their effects (Gitomer, 2009; Wylie & Lyon, 2013; Wylie, 2020).

This dissertation builds on this foundation of work to address remaining gaps in the tools available to enable backwards-planning and formative assessment for the purposes of teacher learning and development. In doing so, this dissertation especially focuses on two key lessons learned from the last few decades of research. First, conceptual frameworks and formative assessment tools must include attention to the practice of teaching as the conduit through which teachers enact their knowledge, skills, and dispositions (Grossman et al., 2009). Second, common tools that can be applied across contexts are vital for creating a common professional language and facilitating the aggregation and synthesis of knowledge (McDonald et al., 2013).

The first chapter addresses the need for formative assessments of pre-service teachers' instructional practice that teacher educators and teacher preparation programs can use to provide tailored supports and make programmatic improvements over time. While there is a robust literature on measures of in-service teachers' instructional practice, literature focused on the pre-service period is sparse (Bartanen & Kwok, 2021). Yet the pre-service period is pivotal in setting pre-service teachers up for success as they become teachers of record. Furthermore, the pre-service period presents unique measurement challenges, discussed further below, that are not addressed in the literature on in-service measures.

Beyond providing teacher educators and professional development facilitators with instructional tools like formative assessments, we must also attend to knowledge, skills, and dispositions they need to support teacher development. This is a burgeoning area of research that has produced lots of studies. However, the field has not yet turned those studies into coherent frameworks and syntheses, like those available for teaching, that can guide efforts to prepare and support teacher educators and professional development facilitators (Gibbons et al., 2021; Gibbons & Cobb, 2017).

In the second and third chapters, I focus specifically on the needs of coaches who support teacher development through ongoing dialogue with teachers about the day-to-day details of a teacher’s classroom and instruction (Donaldson & Woulfin, 2018; Galey, 2016; Hunter & Springer, 2022; Johnson, 2016; Kraft et al., 2018; Kutsyuruba & Godden, 2019; Lochmiller, 2021; Matsko et al., 2020; Neufeld & Roper, 2003). While not the only form of support provided to teachers, coaching is widespread, highly valued by teachers, and supported by robust evidence of its effectiveness for supporting teacher development and student learning (Domina et al., 2015; Kraft et al., 2018; Lochmiller, 2021; Davis & Higdon, 2008; Hardt et al., 2020; Kraft et al., 2018; Ronfeldt et al., 2018; Ronfeldt & Reininger, 2012; R. Stanulis & Floden, 2009; Clark & Byrnes, 2012; Gross, 2010; Ronfeldt et al., 2020). Together, this makes coaching one of the most promising levers for ensuring that all students have access to high quality instruction (Alston et al., 2018; Hiebert & Morris, 2012; Kloser et al., 2019). Yet, researchers know little about the specific mechanisms and features that make coaching effective, making it difficult to identify goals for coach preparation or develop formative assessment tools that provide information about coaching quality (Gibbons & Cobb, 2016, 2017; Heineke, 2013; Robertson et al., 2020). The second chapter therefore focuses on identifying a range of features that may matter, while the third chapter explores the relationship between specific features and teachers’ instructional practice to generate initial evidence about which features may be most promising.

In doing so, these chapters focus specifically on the discourse strategies that coaches use in their interactions with teachers for reasons I discuss further below.

Chapter 1. Different methods for assessing pre-service teachers' instruction: Why measures matter

We cannot effectively support pre-service teachers in developing their instructional practice without reliable formative assessment tools to guide teacher educators' instruction. Most teacher preparation programs (TPPs) collect observational ratings of teachers' instructional practice during clinical placements (Caughlan & Jiang, 2014). In capturing observable skills these measures are especially vital, but also particularly challenging. This is because, in addition to concerns raised about existing observational measures of in-service teacher practice, the pre-service context is an especially challenging and under-explored area. Unlike in-service teachers, pre-service teachers are not solely responsible for their own classrooms, but instead teach in the context of mentors' classrooms, making it difficult to separate pre-service teachers' own instructional skills from the context in which they are observed. Yet, few studies have explored the measurement properties of observational tools used by teacher preparation programs, and I am not aware of any studies that attend to the issue of mentor effects.

In the first chapter, co-authored with Julie Cohen, we evaluate the strengths and weaknesses of two innovative observational approaches to measuring pre-service teachers' instructional skills and growth over time within a university-based teacher preparation program. The first approach employs the Classroom Assessment Scoring System (Pianta & Hamre, 2009) to provide a holistic picture of teachers' classroom interactions during clinical placements. The second approach employs researcher-created measures of more discrete skills culled from simulation-based rehearsal activities collected across the year (Grossman et al., 2009). For observational measures to improve TPPs, several things must be true. First, differences in scores between teachers must result primarily from consistent differences between teachers rather than differences in the circumstances under which teachers were observed. Second, measures must be sensitive enough to detect differences between teachers. In this chapter, we evaluate the extent to which the two innovative measures used by Lambeth University meet these conditions by answering the following research questions:

1. To what extent does each measure capture consistent differences between teachers?
2. To what extent do raters and mentors influence teacher scores?
3. How well can each measure differentiate between individual teachers, groups of teachers, and facets of instruction?

In answering these questions, we identify both the strengths of these measures and the ongoing challenges for learning about teacher practice. First, we find evidence that Lambeth's use of best practices for rating procedures helped reduce the influence of systematic rater effects on scores. We also find evidence that the standardized nature of the simulation-based measures reduced the influence of other conditions of assessment, ultimately ensuring that a larger proportion of the variation in scores reflected consistent differences between teachers. In terms of measure sensitivity, we find evidence that finer-grained measures, such as the simulation-based measures, are better able to detect differences between teachers than broader tools like the Classroom Assessment Scoring System (CLASS).

At the same time, our results demonstrate that pre-service observational measures suffer from the same challenges that in-service observational measure do. At most, only 20% of the variation in CLASS or simulation scores represents consistent differences between pre-service teachers (PSTs). Drawing conclusions about individual PSTs' skills based on these scores is risky. This issue of score consistency is especially prominent for CLASS scores, where we find evidence that differences in the clinical placement context and mentor instructional skills are likely contributing to differences in scores between teachers. While we recognize the importance of observing how PSTs enact their knowledge and skills in real classrooms, we argue that it's equally important for the measures we use to capture what PSTs can do on their own, instead of what they can do only when assisted by mentors. At a minimum, our work suggests the value of exploring new or modified measures whose indicators focus squarely on PST practices to minimize the potential for mentor effects. Measures focused on classroom quality, like the CLASS, might be less useful in the pre-service context than for in-service teachers.

The pre-service period is both brief and formative. We have only a very short window to provide experiences and supports that position candidates for success as they become teachers of record. Preparation programs and teacher educators stand to reap large benefits when they develop systems that enables data-driven programmatic decision-making. Preparation programs can make better decisions about how best to allocate limited time and personnel to support the individual teachers and areas of instructional practice with the greatest need. These data can also help programs and researchers better evaluate the effects of specific preparation experiences on teachers' learning and skill development. However, the details of the data matter, especially in the pre-service period.

Chapter 2. Parsing Coaching Practice: A Systematic Framework for Describing Coaching Discourse

The administrators and coaches responsible for implementing coaching programs face a dizzying array of coaching models and ideas about what coaches can say and do in their interactions with teachers to support their development (Gibbons & Cobb, 2017; Knight, 2009; Kraft et al., 2018). While studies exploring the effects of specific models abound, few studies make comparisons across different models of coaching to understand what coaching practices are the most helpful, for whom, and under what circumstances. Furthermore, synthesizing across studies is complicated by the lack of a common language for describing the practices coaches use. This makes it difficult to identify patterns across studies in how coaching practice supports teacher development.

Rather than relying on coaching practitioners to “figure things out” on their own, we need a systematic program of research designed to identify effective coaching practice across contexts and program models. To do so, we first need a coherent framework that can provide a common language for describing coaching practices and outline potentially promising strategies that warrant further investigation. This chapter introduces such a framework, focusing on concrete coaching discourse “moves,” or questioning and feedback discourse strategies coaches may use in their interactions with teachers as the foundation for a framework of coaching practice (Boerst, et al., 2011). This framework focuses exclusively on coaching discourse moves because they are under explored in the literature and likely influence teacher learning. While the coaching literature is filled with discussions of high-level coaching practices and purposes, such as building trust and supporting teacher self-reflection, there is a dearth of analogous research on how concrete coaching discourse strategies support teacher development (Heineke, 2013; King et al., 2004; L’Allier et al., 2010; Obara, 2010; Robertson et al., 2020; Sisson & Sisson, 2017; Walpole et al., 2010). Yet, there is good reason to believe that coaching discourse strategies matter for teacher development. Given the substantial evidence of the role of teacher discourse in student learning, it seems unlikely that teacher learning would not also be influenced by coach discourse (Demszky & Hill, 2022; O’Connor & Michaels, 1993; Rowe, 1986; Tobin, 1987). Additionally, the limited available literature provides suggestive evidence of the importance of coach discourse strategies (Heineke, 2013; Hunt, 2016; Robertson et al., 2020; Sims & Fletcher-Wood, 2021).

The final framework highlights 40 questioning and feedback strategies that coaches may use in their conversations with teachers. For example, the *Cause & Effect* move refers to questions that ask the teacher to identify or reflect on a causal relationship between two classroom events. In doing so, the framework makes three primary contributions. First, the framework can guide future empirical research, providing a common language for describing coaching discourse and articulating aspects of coaches' interactions with teachers that warrant further investigation. Second, the framework can provide a practical toolkit and technical vocabulary for coaching practitioners, synthesizing our existing knowledge of coaching discourse into a flexible repertoire of discourse strategies that can be used to reflect on and plan for coaching conversations. Third, in serving as a common language for coaching research and practice, the framework can foster greater integration between coaching research and practice. A shared framework and language for describing coaching discourse will facilitate the systematic accumulation, synthesis, and application of new knowledge about coaching (Boerst, et al., 2011; Charalambous & Praetorius, 2020; Hiebert & Morris, 2012; Kloser et al., 2019; McDonald et al., 2013).

Chapter 3. Identifying Promising Coaching Moves to Support Teacher Development

Quantitative measures that detail variation in coaches' discourse and interactions with teachers offer an important route for understanding how coaching interactions influence teacher improvement at scale. Unfortunately, few such tools exist and those that do typically rely on teacher surveys and/or coach self-report, which may not accurately reflect what happens in coaching conversations (Reddy et al., 2019; Richardson et al., 2020). These tools are often also designed for the specific needs of a single study context to evaluate, for example, the extent to which coaching interactions conform to a particular coaching model (Huff et al., 2013; Powell & Diamond, 2013; Wayne & Coggshall, 2022). This limits the extent to which study findings may be applicable to other models and contexts. Furthermore, such tools have limited utility for conducting research on other models, requiring researchers interested in conducting such studies to invest substantial time and effort to develop their own measurement tool(s) (Anglin et al., 2021).

To advance our understanding of the kinds of coaching interactions that best support teacher development, we need measures that leverage direct observation of coaching interactions and are applicable across a wide range of program models and contexts. This chapter, co-authored with Kylie Anglin, builds on the previous chapter to apply the coaching moves framework to transcriptions of coaching conversations. In doing so, we generate a quantitative picture of the nuances of coaches' discourse that allows us the ask the following research questions:

1. What coaching moves do coaches tend to use?
2. How do the moves that coaches use vary across conversations, teachers, coaches, and contexts?
3. What is the relationship between variation in the coaching moves used and teachers' observed instructional practice after participating in coaching?

This approach allows us to explore coaches' interactions at a larger scale than prior qualitative work, while preserving a level of detail that is typically lost in larger-scale quantitative studies of coaching. Additionally, the concrete nature of the discourse moves included in the taxonomy helps ensure that our findings can be clearly interpreted by researchers and coaching practitioners alike.

In answering these research questions, we make both substantive and methodological contributions. Substantively, we provide a new glimpse into the black box of coaching interactions, highlighting both overall trends in the discourse moves coaches use and how the patterns of moves coaches use vary across conversations, integrating data from the 40 move variables to identify distinct profiles of coaching moves that provide a holistic picture of this variation. In doing so, we gain insight into the ways that coaches implemented and adapted the coaching protocol they were asked to follow. We find, for example, that coaches tended to primarily employ discourse moves that were highlighted in the coaching protocol provided to them, suggesting that coach supports like coaching protocols can be helpful in shaping coaches' interactions with teachers. At the same time, we find variation in the emphasis coaches placed on different components of the coaching protocol. Rather than reflecting differences between coaches, we find that individual coaches emphasized different components of the protocol in different conversations, raising questions about how these different approaches to implementing the protocol might have affected teacher development. In exploring the relationship between this variation and teachers' subsequent instructional practice, we identify several coaching moves that may be particularly effective for supporting teacher development, though more work is needed to confirm that this relationship is causal.

Methodologically, this study introduces a measurement tool and analytic method that other researchers can apply to a variety of other contexts and research questions. Because the coaching moves taxonomy is applicable across many different coaching models, other researchers may apply it to understand coaching practice in other coaching programs.

Researchers may also focus their analysis on a specific subset of moves to investigate coach fidelity to specific coaching protocols in the context of randomized controlled trials (RCTs), thereby avoiding the substantial cost and challenge of creating study-specific measures of fidelity (Anglin et al., 2021). When used on an ongoing basis throughout a study, researchers may also use the coaching moves tool to guide the feedback and support provided to coaches. Additionally, this study serves as the first step in a broader program of work in which we intend to develop an automated Natural Language Processing-based tool for coding coaching transcripts. Once developed, this automatic tool will dramatically reduce the resources required to apply the coaching moves taxonomy to transcriptions of coaching conversations. As more studies using the coaching moves tool are conducted, researchers will be able to aggregate findings through conceptual reviews and quantitative meta-analyses with relative ease, ultimately allowing us to provide clearer, evidence-based guidance to coaching practitioners.

Conclusion

Teacher practice shapes student learning opportunities (Kane & Staiger, 2013). Teacher preparation and professional development opportunities, in turn, support teaching practice (Kraft et al., 2018). However, our ability to support teaching practice is only as robust as the tools we can use to guide this work. Similarly, we cannot effectively prepare and support coaches if we lack formative assessment tools to inform the supports that coaches receive. Unfortunately, we cannot create such formative assessment tools because we do not yet know what high quality coaching practice looks like (Gibbons & Cobb, 2017; Robertson et al., 2020). These are the critical gaps this dissertation addresses.

References

Akiba, M., LeTendre, G. K., & Scribner, J. P. (2007). Teacher Quality, Opportunity Gap, and National Achievement in 46 Countries. *Educational Researcher*, *36*(7), 369–387. https://doi.org/10.3102/0013189X07308739

Allen, R., Burgess, S., & Mayo, J. (2018). The teacher labour market, teacher turnover and disadvantaged schools: New evidence for England. *Education Economics*, *26*(1), 4–23. https://doi.org/10.1080/09645292.2017.1366425

Alston, C. L., Danielson, K. A., Dutro, E., & Cartun, A. (2018). Does a Discussion by Any Other Name Sound the Same? Teaching Discussion in Three ELA Methods Courses. *Journal of Teacher Education*, *69*(3), 225–238. https://doi.org/10.1177/0022487117715227.

Anglin, K. L., Wong, V. C., & Boguslav, A. (2021). A natural language processing approach to measuring treatment adherence and consistency using semantic similarity. *AERA Open*, *7*, 23328584211028616.

Atteberry, A., Loeb, S., & Wyckoff, J. (2015). Do first impressions matter? Predicting early career teacher effectiveness. *AERA Open*, 1(4), 1-23. https://doi.org/10.1177/2332858415607834.

Bartanen, B., & Kwok, A. (2021). Examining Clinical Teaching Observation Scores as a Measure of Preservice Teacher Quality. *American Educational Research Journal*, *58*(5), 887–920. https://doi.org/10.3102/0002831221990359.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25.

Boerst, T., Sleep, L., Ball, D., & Bass, H. (2011). Preparing teachers to lead mathematics discussions. *Teachers College Record*, *113*(12), 2844–2877.

Caughlan, S., & Jiang, H. (2014). Observation and teacher quality: Critical analysis of observational instruments in preservice teacher performance assessment. *Journal of Teacher Education*, *65*(5), 375–388.

Charalambous, C. Y., & Praetorius, A.-K. (2020). Creating a forum for researching teaching and its quality more synergistically. *Studies in Educational Evaluation*, *67*, 100894. https://doi.org/10.1016/j.stueduc.2020.100894

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review*, *104*(9), 2633–2679.

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2010). $320,000 Kindergarten Teachers. *Phi Delta Kappan*, *92*(3), 22–25. https://doi.org/10.1177/003172171009200306

Clark, S. K., & Byrnes, D. (2012). Through the eyes of the novice teacher: Perceptions of mentoring support. *Teacher Development*, *16*(1), 43–54. https://doi.org/10.1080/13664530.2012.666935

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. (2005). Who teaches whom? Race and the distribution of novice teachers. *Economics of Education Review*, *24*(4), 377–392.

Council for the Accreditation of Educator Preparation. (2013). *CAEP Accreditation Standards*.

Davis, B., & Higdon, K. (2008). The Effects of Mentoring/Induction Support on Beginning Teachers' Practices in Early Elementary Classrooms (K-3). *Journal of Research in Childhood Education*, *22*(3), 261–274. https://doi.org/10.1080/02568540809594626

Demszky, D., & Hill, H. (2022). The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts. *ArXiv Preprint ArXiv:2211.11772*.

Didion, L., Toste, J. R., & Filderman, M. J. (2020). Teacher professional development and student reading achievement: A meta-analytic review of the effects. *Journal of Research on Educational Effectiveness*, *13*(1), 29–66.

Dolton, P., & Newson, D. (2003). The relationship between teacher turnover and school performance. *London Review of Education*.

Domina, T., Lewis, R., Agarwal, P., & Hanselman, P. (2015). Professional Sense-Makers: Instructional Specialists in Contemporary Schooling. *Educational Researcher*, *44*(6), 359–364. https://doi.org/10.3102/0013189X15601644

Donaldson, M. L., & Woulfin, S. (2018). From tinkering to going "rogue": How principals use agency when enacting new teacher evaluation systems. *Educational Evaluation and Policy Analysis*, *40*(4), 531–556.

Galey, S. (2016). The evolving role of instructional coaches in US policy contexts. *The William & Mary Educational Review*, *4*(2), 11.

Gibbons, L. K., & Cobb, P. (2016). Content-Focused Coaching: Five Key Practices. *The Elementary School Journal*, *117*(2), 237–260. https://doi.org/10.1086/688906

Gibbons, L. K., & Cobb, P. (2017). Focusing on Teacher Learning Opportunities to Identify Potentially Productive Coaching Activities. *Journal of Teacher Education*, *68*(4), 411–425. https://doi.org/10.1177/0022487117702579

Gibbons, L. K., Lewis, R. M., Nieman, H., & Resnick, A. F. (2021). Conceptualizing the work of facilitating practice-embedded teacher learning. *Teaching and Teacher Education*, *101*, 103304. https://doi.org/10.1016/j.tate.2021.103304

Gitomer, D. (Ed.). (2009). *Measurement Issues and Assessment for Teaching Quality*. SAGE Publications, Inc.

Graff, N. (2011). " An effective and agonizing way to learn": Backwards design and new teachers' preparation for planning curriculum. *Teacher Education Quarterly*, *38*(3), 151–168.

Grissom, J. A. (2011). *Can Good Principals Keep Teachers in Disadvantaged Schools? Linking Principal Effectiveness to Teacher Satisfaction and Turnover in Hard-to-Staff Environments*. 34.

Gross, P. A. (2010). Not Another Trend: Secondary-Level Literacy Coaching. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *83*(4), 133–137. https://doi.org/10.1080/00098651003774844

Grossman, P., & Dean, C. G. (2019). Negotiating a common language and shared understanding about core practices: The case of discussion. *Teaching and Teacher Education*, *80*, 157–166. https://doi.org/10.1016/j.tate.2019.01.009

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record*, *111*(9), 2055–2100.

Hardt, D., Nagler, M., & Rincke, J. (2020). *Can peer mentoring improve online teaching effectiveness? An RCT during the covid-19 pandemic*.

Heineke, S. F. (2013). Coaching Discourse: Supporting Teachers' Professional Learning. *The Elementary School Journal*, *113*(3), 409–433. https://doi.org/10.1086/668767

Hiebert, J., & Morris, A. K. (2012). Teaching, Rather Than Teachers, As a Path Toward Improving Classroom Instruction. *Journal of Teacher Education*, *63*(2), 92–102. https://doi.org/10.1177/0022487111428328

Huff, J., Preston, C., & Goldring, E. (2013). Implementation of a coaching program for school principals: Evaluating coaches' strategies and the results. *Educational Management Administration & Leadership*, *41*(4), 504–526.

Hunt, C. S. (2016). Getting to the heart of the matter: Discursive negotiations of emotions within literacy coaching interactions. *Teaching and Teacher Education*, *60*, 331–343. https://doi.org/10.1016/j.tate.2016.09.004

Hunter, S. B., & Springer, M. G. (2022). Critical Feedback Characteristics, Teacher Human Capital, and Early-Career Teacher Performance: A Mixed-Methods Analysis. *Educational Evaluation and Policy Analysis*, 01623737211062913.

Johnson, K. G. (2016). Instructional Coaching Implementation: Considerations for K-12 Administrators. *Journal of School Administration Research and Development*, *1*(2), 37–40.

Johnson, S. M. (2006). The Workplace Matters Teacher Quality, Retention, and Effectiveness. *Teacher Quality*, 34.

Jones, K. A., Vermette, P. J., & Jones, J. L. (2009). An integration of "backwards planning" unit design with the "two-step" lesson planning framework. Education, 130(2).

Kane, T. J., & Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation (NBER Working Paper No. 14607). http://www.nber.org/ papers/w14607.

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Research Paper. MET Project. *Bill & Melinda Gates Foundation*.

King, D., Neuman, M., Pelchat, J., Potochnik, T., Rao, S., & Thompson, J. (2004). *Instructional Coaching: Professional Development Strategies that Improve Instruction*. Annenberg Institute at Brown University.

Kloser, M., Wilsey, M., Madkins, T. C., & Windschitl, M. (2019). Connecting the dots: Secondary science teacher candidates' uptake of the core practice of facilitating sensemaking discussions from teacher education experiences. *Teaching and Teacher Education*, *80*, 115–127. https://doi.org/10.1016/j.tate.2019.01.006

Knight, J. (2009). *Coaching: Approaches and perspectives*. Corwin Press.

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. https://doi.org/10.3102/0034654318759268

Kutsyuruba, B., & Godden, L. (2019). The role of mentoring and coaching as a means of supporting the well-being of educators and students. *International Journal of Mentoring and Coaching in Education*, *8*(4), 229–234. https://doi.org/10.1108/IJMCE-12-2019-081

L'Allier, S., Elish-Piper, L., & Bean, R. M. (2010). What Matters for Elementary Literacy Coaching? Guiding Principles for Instructional Improvement and Student Achievement. *The Reading Teacher*, *63*(7), 544–554. https://doi.org/10.1598/RT.63.7.2

Lochmiller, C. R. (2021). Guest editorial: Coaching for improvement in education: new insights and enduring questions. *International Journal of Mentoring and Coaching in Education*.

Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis*, *41*(3), 260–293.

Matsko, K. K., Ronfeldt, M., Nolan, H. G., Klugman, J., Reininger, M., & Brockman, S. L. (2020). Cooperating teacher as model and coach: What leads to student teachers' perceptions of preparedness? *Journal of Teacher Education*, *71*(1), 41–62. https://doi.org/10.1177/0022487118791992

McDonald, M., Kazemi, E., & Kavanagh, S. S. (2013). Core practices and pedagogies of teacher education: A call for a common language and collective activity. *Journal of Teacher Education*, *64*(5), 378–386. https://doi.org/10.1177/0022487113493807

Neufeld, B., & Roper, D. (2003). *Coaching: A strategy for developing instructional capacity*.

O'Connor, M. C., & Michaels, S. (1993). Aligning Academic Task and Participation Status through Revoicing: Analysis of a Classroom Discourse Strategy. *Anthropology Education Quarterly*, *24*(4), 318–335. https://doi.org/10.1525/aeq.1993.24.4.04x0063k

Obara, S. (2010). Mathematics coaching: A new kind of professional development. *Teacher Development*, *14*(2), 241–251.

Ost, B. (2014). How do teachers improve? The relative importance of specific and general human capital. *American Economic Journal: Applied Economics*, *6*(2), 127–151. https://doi.org/10.1257/app.6.2.127

Pellegrini, M., Lake, C., Neitzel, A., & Slavin, R. E. (2021). Effective programs in elementary mathematics: A meta-analysis. *AERA Open*, *7*, 2332858420986211.

Pianta, R. C., & Hamre, B. K. (2009). Classroom processes and positive youth development: Conceptualizing, measuring, and improving the capacity of interactions between teachers and students. *New Directions for Youth Development*, *2009*(121), 33–46.

Powell, D. R., & Diamond, K. E. (2013). Implementation fidelity of a coaching-based professional development program for improving Head Start teachers' literacy and language instruction. *Journal of Early Intervention*, *35*(2), 102–128.

Reddy, L. A., Glover, T., Kurz, A., & Elliott, S. N. (2019). Assessing the effectiveness and interactions of instructional coaches: Initial psychometric evidence for the instructional coaching assessments–teacher forms. *Assessment for Effective Intervention*, *44*(2), 104–119.

Richardson, G., Yost, D., Conway, T., Magagnosc, A., & Mellor, A. (2020). Using instructional coaching to support student teacher-cooperating teacher relationships. *Action in Teacher Education*, *42*(3), 271–289.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417–458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

Robertson, D. A., Ford-Connors, E., Frahm, T., Bock, K., & Paratore, J. R. (2020). Unpacking productive coaching interactions: Identifying coaching approaches that support instructional uptake. *Professional Development in Education*, *46*(3), 405–423. https://doi.org/10.1080/19415257.2019.1634628

Ronfeldt, M., & Reininger, M. (2012). More or better student teaching? *Teaching and Teacher Education*, *28*(8), 1091–1106. https://doi.org/10.1016/j.tate.2012.06.003

Ronfeldt, M., Bardelli, E., Truwit, M., Mullman, H., Schaaf, K., & Baker, J. C. (2020). Improving Preservice Teachers' Feelings of Preparedness to Teach Through Recruitment of Instructionally Effective and Experienced Cooperating Teachers: A Randomized

Experiment. *Educational Evaluation and Policy Analysis*, *42*(4), 551–575. https://doi.org/10.3102/0162373720954183

Ronfeldt, M., Brockman, S. L., & Campbell, S. L. (2018). Does cooperating teachers' instructional effectiveness improve preservice teachers' future performance? *Educational Researcher*, *47*(7), 405–418.

Rowe, M. B. (1986). Wait time: Slowing down may be a way of speeding up! *Journal of Teacher Education*, *37*(1), 43–50.

Sims, S., & Fletcher-Wood, H. (2021). Identifying the characteristics of effective teacher professional development: A critical review. *School Effectiveness and School Improvement*, *32*(1), 47–63. https://doi.org/10.1080/09243453.2020.1772841

Sisson, D., & Sisson, B. (2017). *The literacy coaching handbook: Working with teachers to increase student achievement*. Routledge, Taylor & Francis Group.

Stanulis, R., & Floden, R. E. (2009). Intensive Mentoring as a Way to Help Beginning Teachers Develop Balanced Instruction. *Journal of Teacher Education*, *60*(2), 112–122. https://doi.org/10.1177/0022487108330553

Tobin, K. (1987). The role of wait time in higher cognitive level learning. *Review of Educational Research*, *57*(1), 69–95.

Walpole, S., McKenna, M. C., Uribe-Zarain, X., & Lamitina, D. (2010). The relationships between coaching and instruction in the primary grades: Evidence from high-poverty schools. *The Elementary School Journal*, *111*(1), 115–140.

Wayne, A. J., & Coggshall, J. G. (2022). How to ensure high-quality instructional coaching at scale. *Phi Delta Kappan*, *103*(5), 42–46.

Wylie, C., & Lyon, C. (2013). Using the formative assessment rubrics, reflection and observation tools to support professional reflection on practice. *Educational Testing Service*.

Wylie, E. C. (2020). Observing Formative Assessment Practice: Learning Lessons Through Validation. *Educational Assessment*, *25*(4), 251–258.

CHAPTER 1

Different Methods for Assessing Pre-Service Teachers' Instruction: Why Measures Matter

Arielle Boguslav and Julie Cohen

Abstract

Over the last two decades, teacher preparation programs (TPPs) have become increasingly responsible for collecting and analyzing pre-service teacher (PST) data. Inherent in the theory of action is the idea that TPPs can use such data to provide more targeted supports and improve the preparation experiences offered. Most TPPs collect observational ratings of PSTs' instructional practice during clinical placements. In capturing observable skills these measures are especially vital, but also particularly challenging. This is especially true in the pre-service period, an often-overlooked context. For observational measures to improve TPPs, several things must be true. First, differences in scores between PSTs must result from consistent differences between PSTs. Second, measures must be sensitive enough to detect differences between PSTs. In this chapter we investigate one TPP's efforts to gain insight into PSTs' instructional skills using two innovative observational measures. Our results highlight several strengths of these measures alongside remaining challenges.

Introduction

Over the last two decades, teacher preparation programs (TPPs) have become increasingly responsible for the collection and analysis of data on pre-service teachers' (PST) knowledge, skills, and dispositions (Bastian et al., 2016). Several states have implemented formal accountability systems that require TPPs to provide data on their graduates, share these data as a measure of TPP effectiveness, and even levy penalties on programs deemed ineffective (von Hippel et al., 2016). Inherent in the theory of action behind these policies is the idea that TPPs can use PST data to provide more targeted support based on PSTs' needs, as well as adjust programmatic experiences based on the degree to which such experiences promote PSTs' development (Bastian et al., 2018; Davis & Peck, 2020).

Among other measures, most TPPs collect observational ratings of PSTs' instructional practice during clinical placements (Feuer et al., 2013). In capturing observable skills—what PSTs *do* in interactions with students—these measures are especially vital, but also particularly challenging (Gitomer, 2009). Ratings of teaching practices are often biased by inconsistent raters and the contextual characteristics of an observed lesson from the characteristics of the students to the content of the lesson (Bartanen & Kwok, 2021; T. Kane & Staiger, 2012). In addition, unlike in-service teachers, PSTs are not solely responsible for their own classrooms, but instead teach in the context of mentors' classrooms. This raises the possibility that mentor characteristics and their teaching skills could influence PST scores on observational measures of instructional quality. To our knowledge no study has investigated this measurement issue, despite the likelihood of "mentor effects."

For observational measures to improve TPPs, several things must be true. First, differences in scores between PSTs must result primarily from consistent differences between

PSTs, rather than differences, for example, between raters scoring classroom observations or the mentors in whose classrooms PSTs are placed (T. Kane & Staiger, 2012). If this is not the case, then TPPs risk erroneously attributing differences between raters or other contextual factors to differences between PSTs. For example, if scores on an observational rubric are substantially influenced by supervisors' standards, the PSTs with the lowest scores may not be the PSTs most in need of targeted support, but instead, may be those rated by supervisors with harsher standards (Bartanen & Kwok, 2020). Second, to inform TPP decision-making, measures must be sensitive enough to detect relevant differences between PSTs. Third, differences in scores must reflect the specific differences that TPPs wish to understand (M. Kane, 1992; Papay, 2012). For TPPs interested in understanding how PSTs' instructional skills differ, measures that largely reflect differences in PST writing ability or socio-economic background will be unhelpful (Gitomer et al., 2021). Similarly, for TPPs interested in understanding how PSTs' will enact their knowledge and skills in the classroom, measures that observe PSTs' practice in clinical placements but are not reflective of PSTs' practice as teachers of record will also be unhelpful (Diez, 2010; Henry et al., 2013). If the measures TPPs use do not meet these conditions, data analysis and data-driven decision-making are unlikely to lead to desired improvements in PSTs' preparedness for the classroom.

In this paper we investigate one TPP's efforts to generate data on PSTs' instructional skills using two distinct types of observational measures of teaching: Classroom Assessment Scoring System (Pianta & Hamre, 2009) scores collected during clinical placements, and measures of more discrete skills culled from simulation-based Instructional Activities collected across the year (Lampert & Graziani, 2009). We focus on observational measures because they are most prone to measurement issues, and because TPPs are ultimately charged with preparing novices who can engage in productive and supportive interactions with children. Figuring out how to measure instructional quality in valid and reliable ways has been a longstanding puzzle for research on teaching. Few, however, have explored the challenges particular to doing so in the context of pre-service preparation, when time is in short supply and candidates are typically observed in contexts that are not fully "theirs."

We begin by describing the study context and the measures. We then generate hypotheses about how design features of each measure (e.g. rubric indicators and scoring procedures) may influence the measurement issues they exhibit. We then use data collected by the TPP to test these hypotheses by answering the following research questions:

1. To what extent does each measure capture consistent differences between PSTs?
2. To what extent do raters and mentors influence PST scores?
3. How well can each measure differentiate between individual PSTs, groups of PSTs, and facets of instruction?

In this way, we can identify both the strengths of these measures and the ongoing challenges for learning about PST practice. The pre-service period is both brief and formative. We have only a very short window to provide experiences and supports that position candidates for success as they become teachers of record. It is vital that TPPs have data that can inform such efforts.

Background & Conceptual Framework

Why observe PSTs' instruction?

Many have argued that TPPs would benefit from having more insight about the degree to which preparation experiences support PSTs' development (Davis & Peck, 2020; Goldhaber, 2019). In our conceptual framework, shown in Figure 1, we articulate a theory of action for how such data can result in improvements in teaching when PSTs enter the classroom. In the first column of the conceptual framework, we highlight four key mechanisms by which PST data may support TPP improvement. In the remaining three columns, we demonstrate how these mechanisms might be implemented in practice and highlight the proximal outcomes for the TPP and the distal outcomes to which they might contribute.

[insert Figure 1 here]

First, we highlight how measurement tools can help create a shared understanding of and common language for teaching when teacher preparation faculty, supervisors, mentors, and PSTs all use the same tools (Davis & Peck, 2020). This facilitates teacher preparation stakeholders—faculty, mentors, supervisors, and candidates —working toward the same learning goals and fosters coherence across preparation experiences. The remaining three mechanisms highlight how specific analyses can contribute to improvements in the supports and experiences provided to PSTs. First, analysis of PST data can allow TPPs to diagnose and respond to PSTs' individual learning needs (M. Allen & Coble, 2018; Peck et al., 2014).

Second, analysis of PST data can help TPPs identify persistent and program-wide areas for development, which can, in turn, inform changes to course content and program curriculum (M. Allen & Coble, 2018; Peck & McDonald, 2013). By comparing PSTs' mastery of different facets of instruction (i.e., relative strengths in classroom management versus orchestrating discussions), TPPs can tailor learning experiences to meet observed needs (i.e. more practice facilitating student discourse). Moreover, TPPs can compare skill development across cohorts or licensure tracks (e.g., elementary versus secondary) to identify areas of relative weakness and redesign or supplement program content to provide additional support for those areas of development.

Finally, TPPs can compare the scores of PSTs that participated in different preparation experiences (e.g. candidates who tutored struggling readers in a reading development course with those who did not have tutoring experience) or were exposed to different preparation pedagogies (e.g., a mathematics method that incorporated many rehearsals of core teaching practices) (McDonald et al., 2013). In this way, TPPs can identify experiences and pedagogies that are more and less promising for supporting PSTs' development (Hill et al., 2020; Peck & McDonald, 2013).

What makes a measure useful?

Realizing improvements from PST data depends on what data can tell us. Here we highlight three key measurement properties that influence the conclusions that can be drawn: reliability, sensitivity, and validity.

Reliability. Reliability refers to the extent to which differences in scores reflect consistent differences between PSTs (Bell et al., 2012; Ho & Kane, 2013). When reliability is low, differences in scores are influenced by differences in the conditions of assessment, other than who is being assessed. When measures of PSTs' skills are not reliable, conclusions TPPs draw about which PSTs need more support may instead reflect, for example, which PSTs were rated by more stringent supervisors (Bartanen & Kwok, 2021).

While it is not possible to create measures with perfect reliability, measures that require human judgment, as observational measures do, tend to be less reliable than paper and pencil tests with an established correct answer (Bell et al., 2019; Hill et al., 2012). This is because it is very difficult to ensure raters assign scores in the same way (Bell et al., 2015). Prior literature highlights the relatively low reliability of observational measures in both in-service and pre-service contexts (Bartanen & Kwok, 2021; Ho & Kane, 2013).

Observational measures face an additional reliability challenge: the influence of the specific instructional context on scores. Prior work has highlighted the influence of lesson content, time of year, and student demographics on teacher observation scores (Bell et al., 2012; Casabianca et al., 2015). This issue may be further complicated when attempting to measure PST practice, as PSTs are typically observed in their mentor's classroom, where a classroom's instructional and emotional climate may be driven not by the PST's instructional practices, but by the mentor's instructional practices across the school year (Bartanen & Kwok, 2021).

The literature highlights three primary strategies for increasing measure reliability. First, TPPs can focus on reducing the influence of contextual characteristics by standardizing the assessment context. This includes providing raters with extensive training and feedback to ensure consistent ratings, providing PSTs with a standardized lesson plan or learning objective, and observing PST instruction in the context of standardized teaching simulations (Cohen & Goldhaber, 2016; Cohen et al., 2020). Second, TPPs can average PST scores over multiple observations under a variety of contextual conditions to isolate the differences between PSTs that are consistent across observations (T. Kane & Staiger, 2012). In this case, randomly assigning contextual characteristics, such as raters or lesson objectives, for each observation is especially helpful (van der Lans et al., 2016).

Sensitivity. Sensitivity is the extent to which measures can detect statistically significant and practically meaningful differences when making comparisons between individual PSTs, groups of PSTs, or facets of instruction (M. Allen & Coble, 2018). Data analysis and data-driven decision-making are unlikely to lead to TPPs' desired improvements in PSTs preparedness for the classroom if measures are not sufficiently sensitive. Identifying PSTs that may need additional targeted support implicitly requires comparing PSTs' scores against one another.

Understanding the impacts of specific preparation experiences or pedagogies requires comparing scores between the group of PSTs that participated in a particular experience and those who did not. Similarly, identifying specific dimensions of instructional skill where all or some PSTs need more support requires comparing scores from one dimension to another. To accurately interpret these comparisons, the measures TPPs use must be sensitive enough to detect these differences (Cohen & Goldhaber, 2016; Weisberg et al., 2009). When measures are not sufficiently sensitive, TPPs risk concluding that there are no differences, when in fact the measures used may simply be unable to identify them.

The less reliable a measure is, the sensitive it is likely to be because scores on less reliable measures are influenced by both contextual characteristics and differences between PSTs. The contextual characteristics can drown out the between-PST differences we want to detect. Scoring procedures also influence sensitivity. For example, a measure with scores ranging from 1-4 provides coarser differentiation than a measure with scores ranging from 1-10, assuming PSTs receive the full range of scores (T. Kane & Staiger, 2012). Of course, this issue is further exacerbated if PSTs do not receive the full range of scores, as is frequently the case (Kraft & Gilmour, 2017).

The scope and granularity of a measure can also influence sensitivity (Janssen et al., 2015). Measures that provide a broader picture of overall PST skills tend to provide a less detailed and nuanced picture of specific areas of strength or improvement (Hill et al., 2020; Mancenido, 2022). Because of reliability issues and logistical constraints, providing finer-grained information necessarily requires trading off breadth for depth. It is not feasible, for example, to ask raters to individually score hundreds of finer-grained indicators of PST skill (Bell et al., 2015). If the main differences between PSTs, however, are more nuanced, then broader measures will be less sensitive than more granular ones. After completing intensive methods coursework, PSTs may exhibit broadly similar skills when presenting instructional content, leading classroom discussions, and providing feedback. In this case, important differences among PSTs in the more specific skills of metacognitive modelling or lesson pacing, however, will not be detected by broad measures of instructional skill.

Common measures of PSTs' instructional are likely to be limited in sensitivity because of these issues. In addition to suffering from low reliability, these measures generally aim to provide a broad picture of PSTs' skills, providing little information about finer-grained details of PST development (Hill et al., 2020; Mancenido, 2022). Additionally, these measures use a relatively limited score range, often including a maximum of only three to five points (Bartanen & Kwok, 2021; Henry et al., 2013).

Validity. Finally, validity refers to the extent to which differences in scores reflect the specific differences that TPPs wish to understand (M. Kane, 1992; Papay, 2012). These differences may include differences in a) the extent to which PSTs have mastered the skills they've been taught, b) the extent to which PSTs enact what they've learned in the classroom, and c) the extent to which PSTs' instructional practice supports student learning (Diez, 2010). A specific measure of PST practice may provide information about one, multiple, or none of these differences. Furthermore, a measure's validity, or ability to provide information about each difference, often varies across contexts (M. Kane, 1992).

Challenges with validity can occur for several reasons. First, a measure may capture other characteristics of PSTs aside from the specific skills that the measure is designed to capture (Bell et al., 2012). Second, a measure may provide information about PSTs' current instructional skills without providing any information about how PSTs' will enact those skills in the classroom or how student learning will be affected. Third, like sensitivity, low validity can also result from low reliability, where contextual characteristics, including mentor teachers, might mask insight into PSTs' skills (Cohen & Goldhaber, 2016).

Despite the growing interest in practice-based teacher education, many commonly used measures assess PST knowledge outside of the context of classroom practice. Written exams, surveys, and reflection assignments, used commonly for course assessments and licensure exams, often bear little resemblance to the daily interactions teachers have with K-12 students. Such measures are likely not a good proxy for how well teachers are able to *apply* their knowledge, skills, and dispositions in the classroom in ways that support student learning (Diez, 2010). Even performance assessments may not provide a strong reflection of PSTs' instructional skills.

Because of their explicit focus on PSTs' instructional practice in real classrooms, observational measures have the potential to add to the incomplete picture created by other measures of PST skills. Existing literature provides validity evidence demonstrating a relationship between teacher observational scores and measures of student learning (T. Kane & Staiger, 2012). However, this evidence comes from observations of in-service teachers with robust, researcher-designed measures used by highly trained raters. TPPs rarely use such measures and methods (M. Allen & Coble, 2018). Studies that investigate PST observation scores during student teaching find little evidence of a relationship with measures of PSTs' instructional skills as teachers of record (Bartanen & Kwok, 2021; Henry et al., 2013).

Data & Methods

Context

We draw on data from a university-based teacher preparation program at a large, public university in the southeastern United States, which we call Lambeth University. The university offers multiple pathways to licensure for PSTs, enrolling approximately 120 teacher PSTs each year, across the elementary, secondary, and special education licensure areas.

The program primarily takes a practice-based approach to preparing PSTs (Forzani, 2014). Coursework emphasizes the development of robust content and pedagogical knowledge, while also attending to how this knowledge is enacted in practice. In all courses, instructors aim to make explicit links between theory and teaching practice through representations, decompositions, and approximations of practice (Grossman et al., 2009). In addition to role-play exercises during coursework, many instructors also provide opportunities for PSTs to teach digital student avatars in mixed-reality simulations, where PSTs can practice and experiment with specific skills, techniques, and approaches before working with real children (Cohen et al., 2020; Dieker et al., 2014). This allows faculty to observe and assess candidates engaged in teaching practice in ways that are otherwise difficult to replicate in a university classroom.

All PSTs also participate in a clinical experience each semester that allows them to apply their learning. Early clinical experiences vary by program but range from one-on-one tutoring with one child to interning in a classroom for 15 hours a week. Across programs, the final semester features a formal, full-time student teaching placement designed to give candidates experience with all aspects of teaching from setting up a classroom to attending faculty meetings. The university has dedicated staff responsible for coordinating clinical experiences who strategically cultivate relationships with local school divisions to identify mentors with appropriate pedagogical skills and the ability to scaffold novice teacher development. In matching PSTs with specific placements, the staff pay particular attention to providing PSTs with experiences that are varied, are reflective of the kinds of contexts in which PSTs plan to work, and match mentors' strengths with PSTs' developmental needs.

In addition to scaffolding and feedback provided by faculty instructors in coursework, PSTs also participate in multiple coaching cycles during each clinical experience, following a modified version of the evidence-based *My Teaching Partner* (MTP) program (J. Allen et al., 2015). In each cycle, PSTs first record a video of their instruction in their clinical placement. Field supervisors then select targeted video segments for analysis and provide PSTs with several reflective prompts to promote focused self-reflection. The PST, supervisor, and mentor then meet over Zoom to discuss the PST's practice. The process helps cultivate prospective teachers' ability to analyze their own teaching and enhance their capacity to reflect on ways to improve their practice. Additionally, PSTs meet weekly in groups to discuss and reflect on their clinical experiences as part of a seminar course each semester.

Measures

Lambeth University has dedicated substantial resources to developing a robust data collection system for research and program improvement purposes (Cohen, 2015). The university systematically collects data on PST instructional practice at multiple timepoints using two observational measures: the Classroom Assessment Scoring System (CLASS), designed to capture broad features of classroom climate and instructional support (Pianta & Hamre, 2009), and more micro-level measures of teaching assessed in the context of mixed-reality simulations (Cohen et al., 2020).

For both measures, the university employs rating procedures that are largely aligned with what existing literature suggest are best practices for observational measures of teaching. For both measures, raters have no relationship with the PSTs they rate; all raters complete formal training and are required to pass a rater certification test; they also receive ongoing feedback and support at weekly calibration meetings (Park et al., 2015). Finally, observation videos are randomly assigned to different raters within each observation timepoint (T. Kane & Staiger, 2012).

CLASS Scores

PSTs select up to four of videos collected as part of the MTP program to submit for external scoring. Videos are independently scored by the university using the CLASS framework, an observation protocol designed by Robert Pianta and colleagues that articulates the components of high-quality classroom interactions (J. Allen et al., 2011). Developed by a team of psychologists, CLASS foregrounds the value of strong, warm relationships, focusing on the tenor of interactions between a teacher and students and among students and treating the classroom as the unit of analysis. To date, the framework has been used as a measure of classroom quality for the purposes of conducting research, guiding coaching conversations as part of the MTP professional development program, and evaluating the quality of early childcare programs (J. Allen et al., 2015; Araujo et al., 2016; Bassok et al., 2021). CLASS is designed to provide a high-level view of a classroom across three broad interactional domains: Emotional Support, Classroom Organization, and Instructional Support (Hafen et al., 2015). Each domain is also comprised of 3-5 dimensions. For example, one dimension within the domain of Classroom Organization is Productivity and one dimension within Instructional Support is Quality of Feedback. Raters select a score on a scale from 1 to 7 for each dimension, using several dimension-specific indicators to distinguish between teacher-student and student-student interactions that are low quality (1-2 points), mid quality (3-5 points), or high quality (6-7 points).

Decades of research documents the reliability, sensitivity, and validity of CLASS when used to assess classroom quality for in-service teachers. When scores from multiple observations-- each rated by a different rater-- are averaged together, over 60% of the variation in average scores can be attributed to differences between teachers (T. Kane & Staiger, 2012). CLASS scores have also been used to detect differences between individual teachers and document changes in scores over time (Bassok et al., 2021; T. Kane & Staiger, 2012; La Paro et al., 2004). Finally, numerous studies in a variety of classroom contexts have shown that higher CLASS scores are associated with other established measures of classroom quality and stronger student outcomes, including academic performance and behavioral outcomes like student engagement (e.g. J. Allen et al., 2013; Araujo et al., 2016; T. Kane & Staiger, 2012; La Paro et al., 2004). However, more recently a growing body of literature is raising concerns about the influence of contextual factors on CLASS scores and our ability to use these scores to explore change in teacher practice over time (reliability) and the ability for CLASS scores to detect differences between teachers (sensitivity) (Briggs & Alzen, 2019; Casabianca et al., 2015; Gitomer et al., 2014; Wallace et al., 2020).

Simulation-Based Scores

Beginning in 2017-18, all PSTs also participated in two standardized simulations at as part of their general methods courses. In these mixed-reality simulations, PSTs practice teaching virtual students voiced by a trained actor. These simulations provide opportunities for PSTs to engage in "approximations of practice" (Grossman et al., 2009) and also serve as a standardized assessment platform. In the first simulation scenario, referred to as "Redirection," PSTs practice redirecting off-task student behaviors in the context of a discussion about setting classroom norms on the first day of school (Cohen et al., 2020). In the second simulation scenario, referred to as “Text-Focused Instruction,” PSTs practice scaffolding students during a discussion of a fiction text (Cohen et al., in review). In this scenario, PSTs facilitate the text-based discussion using researcher-provided questions and respond to standardized student responses that are more or less supported by textual evidence. PSTs complete each simulation scenario four times over the course of the program.

The measures used to assess PSTs’ instructional skill in the simulations are aligned with the limited focus of each scenario. The rubric for the Redirection simulation was designed based on the Responsive Classroom framework used for management in the local K-12 schools (Charney, 1993; Responsive Classroom, 2014). The rubric for the Text-Focused Instruction simulation was designed to reflect high-quality instructional practices highlighted in the relevant literature on Text-Focused Instruction (Castles et al., 2018; Reznitskaya et al., 2009). The simulation rubrics represent a smaller grain size of practice than the CLASS. Whereas the entire focus of the simulation rubric is capturing Redirection, the CLASS treats “effective Redirection of misbehavior” as only one component of the broader dimension of Behavior Management. Unlike the CLASS, these simulation measures were designed to identify differences between groups of PSTs and have not been used to compare individual PSTs’ skills (Cohen et al., 2020).

As part of a larger research study, PSTs were randomly assigned to a short coaching session or a short self-reflection protocol in between their second and third simulation session. For about half the PSTs, the coaching session was for the Redirection scenario, and for the other half it was for Text-Focused Instruction scenario. For more detailed descriptions of the simulation measures and the coaching and self-reflection protocols see Author 2020, 2021.

Sample

For each measure, we select a sample that maximizes sample size, while also ensuring a similar number and timing of observations, to avoid bias from missing data. For CLASS scores, our sample consists of the 83 PSTs who entered Lambeth in 2017-18 and for whom four observations are available, two for each of two clinical placements. In total, our sample includes 135 unique mentors[1] from 62 unique schools across more than 10 counties in one state. Schools are primarily elementary schools serving mostly students who are white and not eligible for free & reduced-price meals. For simulation scores, our sample consists of the 60 PSTs who entered Lambeth in 2018-19 and for whom eight observations (four for each simulation scenario) are available.

Hypotheses

First, we expect that scores for both measures will be influenced by characteristics of the context in which PSTs are observed. Second, because the simulation context is standardized, we expect those scores will be less influenced by characteristics of the context. Third, we hypothesize that characteristics of the placement context, such as the classroom culture largely created by the mentor, will influence PSTs' CLASS scores. Fourth, because of the high-quality rating procedures employed by Lambeth, we hypothesize that raters will have limited influence on PST scores. Finally, if simulation scores are indeed less influenced by characteristics of the context, then we expect simulation scores to be more sensitive to differences between PSTs.

Methods

Given our focus on cross-measure comparisons, we analyze CLASS scores for the two domains that are conceptually most closely aligned to the constructs measured using the simulation rubrics: Classroom Organization (aligned with the Redirection scenario) and

[1] Our sample includes more mentors than PSTs because PSTs work with one mentor during early clinical experiences and a different mentor for their formal student teaching experience.

Instructional Support (aligned with the Text-Focused Instruction scenario). For each research question, we conduct analyses separately for each simulation scenario and CLASS domain. Below we provide a general overview of our methods. More detailed explanations of the statistical models for each research question are included in Appendix A.

RQ1: To what extent does each measure capture consistent differences between PSTs?

Following prior studies of observation scores, we draw on generalizability theory to answer our first research question (Bartanen & Kwok, 2020; Briggs & Alzen, 2019). Variation in observation scores is decomposed into distinct sources with the goal of distinguishing between 1) variation that reflects consistent differences between PSTs, and 2) measurement error that results from differences in the conditions and context of measurement. We then calculate the proportion of the variation that reflects consistent differences, relative to overall variation in scores. Because our statistical models allow for growth in PST scores over time, we report results from two approaches: 1) combining variation in PSTs' initial scores with variation in PSTs' growth and dividing by overall variation (Raudenbush & Bryk, 2002); 2) calculating a separate proportion for variation in PSTs' initial scores and growth over time (Briggs & Alzen, 2019).

RQ2: To what extent do raters and mentors influence PST scores?

We decompose variation in observation scores into contextual sources we hypothesize might influence scores without contributing to our understanding of PSTs' instruction. For CLASS scores, we separate consistent differences between PSTs from variation between mentors and any remaining measurement error. For simulation scores, we separate consistent differences between PSTs from variation between raters and any remaining measurement error.

RQ3: How well can each measure differentiate between individual PSTs, groups of PSTs, and facets of instruction?

To explore the degree to which CLASS and simulation scores are sensitive to differences between individual PSTs and groups, we explore the magnitude and significance of additional parameters from the models used in RQ1 and RQ2. To explore sensitivity to different facets of instruction, we evaluate the correlation between the raw and predicted Classroom Organization and Instructional Support scores and compare the magnitude of the scores for each domain.

Limitations

There are several limitations of the measures we employ. We recognize that these measures necessarily reflect specific conceptualizations of core components of high-quality instruction. Many crucial aspects of teaching, including culturally and linguistically responsive pedagogies, are not included in these measures (Pacheo, 2009). We also do not yet know the extent to which CLASS scores from clinical placements or simulation-based scores provide meaningful information about PSTs' instruction when they enter the classroom as teachers of record. Though CLASS scores have been used in prior studies to evaluate pre-service candidates' instructional skills during student teaching (e.g. Malmberg et al., 2010), we are not aware of any studies that directly evaluate the predictive validity of CLASS scores for instructional skills down the road. Such evidence is also lacking for simulation scores, though such research is currently under way. Unfortunately, generating this evidence using data from Lambeth is virtually impossible because of the lack of longitudinal data systems for connecting PSTs to their later employment and teaching outcomes in this state. Nonetheless, these kinds of predictive validity analyses are only possible when measures primarily reflect differences between PSTs. When measures are heavily influenced by contextual characteristics, such as raters, they cannot tell us much about differences between PSTs' skills, let alone predict what PSTs will do when they enter the classroom. We therefore argue that our analyses raise important considerations for

TPPs and provide a proof of concept that is relevant to any measure of PSTs' skills, including those that capture other aspects of teaching.

Our analyses are also limited by unavoidable deviations from the ideal design for a generalizability study, which requires that each PST be observed by every rater, in every kind of classroom context, and teaching every kind of lesson (Briggs & Alzen, 2019). Under these conditions, we could directly measure how each contextual characteristic influences scores. Random assignment of rater, classroom context, and lesson provide a more feasible alternative to estimate the average effects of each contextual characteristic. Because mentors are not randomly assigned to PSTs in our data, effects attributed to mentors may instead reflect differences between PSTs. This would be the case if PSTs with weaker instructional skills tend to be assigned to more skilled mentors. The estimated between-mentor variation in scores would then reflect both these initial differences in PSTs' instructional skills *and* the influence of mentors on PST scores. Rather than providing a precise estimate of the proportion of the variation in CLASS scores attributable to consistent differences between PSTs, our results provide a reasonable range for this proportion.

Our analyses are also limited by the lack of access to data on all contextual characteristics of interest. While we have access to rater information for simulation scores, we do not have access to rater information for CLASS scores. While we have access to mentor information as a proxy for classroom context for CLASS scores, we do not have access to information about the (subtle) differences in avatar responses across actors and simulation sessions. Any attempt to decompose the variation in CLASS and simulation scores, therefore, suffers from the problem that not all potential contextual factors are accounted for in the model. This means that our estimates serve as an upper bound, as the true proportions would be lower if any of these unobserved factors influence scores (Briggs & Alzen, 2019).

Results

RQ1: To what extent does each measure capture consistent differences between PSTs?

The proportion of variation in CLASS scores that reflects differences between PSTs is low relative to the overall variation in scores. Specifically, we estimate that 3-4% of the variation in individual Instructional Support and Classroom Organization scores reflects consistent differences between PSTs (Table B1). This means that 96-97% of the variation in scores reflects measurement error. When we separately consider variation in growth over time, we find that 15% of the variation in growth in Instructional Support scores reflects consistent differences between PSTs. The estimate for growth is higher because PST growth is calculated using all four scores, while the proportion for individual scores is calculated only a single score. For Classroom Organization, all PSTs effectively grow at the same rate, so we cannot estimate what proportion of this "growth" variation reflects consistent differences between-PSTs.

Consistent with our hypothesis about the affordances of the standardized context, a greater proportion of the variation in simulation scores (~20%) reflects consistent differences between PSTs. This is approximately five times larger than the proportion for CLASS scores. Like Classroom Organization, all PSTs effectively grow at the same rate (once we account for coaching effects) so we cannot estimate what proportion of this variation reflects consistent differences between PSTs. However, for Text-Focused Instruction, we estimate that about 15% of the variation in growth over time reflects consistent differences between PSTs.

Notably, the estimates for variation in growth over time are similar for CLASS and simulation scores. However, directly comparing these two estimates is misleading since the growth estimates are scaled by the intervals between observations, which are greater for CLASS scores (see Appendix A). To allow more direct comparison, we can instead compare the proportion of variation in growth if CLASS scores and simulation scores were collected at the same intervals. To do so, we recalculate the estimate for simulation scores using the scaling factor for CLASS scores. In this case, about 40% of the variation in growth on Text-Focused Instruction scores would reflect consistent differences between PSTs. This is about two and half times larger than the estimate for CLASS Instructional Support.

Complete results from all statistical models are provided in Tables B2-B5.

RQ2: To what extent do raters and mentors influence PST scores?

Since PSTs are observed within the context of a specific classroom led by a specific mentor, we hypothesize that some of the variation in CLASS scores may result from systematic differences among these classroom/mentor contexts, which will not be accounted for in the analyses above. This means that the proportion of variation in scores attributed to consistent differences between PSTs will be artificially inflated if mentors influence PST scores.

When we separate out variation in scores between mentors, we find that 9-17% of the variation in CLASS scores can be explained by the mentor in whose classroom a PST is observed (Table B6). We also find that once variation between mentors is accounted for, the proportion of variation that reflects differences between PSTs falls to effectively zero.

Unlike mentors, who are not randomly assigned to candidates, raters *were* randomly assigned to observe specific PSTs at specific timepoints, following established best practices (T. Kane & Staiger, 2012). Consistent with our hypothesis, we see little evidence that raters systematically influence simulation scores. When we separate out variation in scores between raters, we find that raters explain only 1-3% of the variation in scores (Table B7). These results provide additional evidence in support of randomly assigning raters, when feasible.

RQ3: How well can each measure differentiate between individual PSTs, groups of PSTs, and facets of instruction?

Differences between PSTs

The lower the proportion of variation in scores that reflects differences between PSTs, the less likely it is for a measure to be able to detect those differences. This is a concern for both CLASS and simulation scores. Figure 2 illustrates the differences that CLASS and simulation scores can detect between PSTs. For each of the four outcomes, we plot the growth trajectories for all PSTs to illustrate how baseline scores and growth over time vary. Here, the y-axis reflects each PSTs' simulation or CLASS score, and the x-axis reflects the observation timepoint. These graphs were created using PSTs' predicted scores, isolating only between-PST differences after accounting for measurement error (the most conservative approach).

[insert Figure 2 here]

Differences between PSTs for CLASS scores are practically small. The difference in baseline CLASS scores between a PST at the 16th percentile and a PST at the 84th percentile corresponds to 0.2-0.3 points (out of 7) for both Instructional Support and Classroom Organization. This difference is only statistically significant for Instructional Support, however. There is also a statistically significant difference in growth rate for Instructional Support, corresponding to a difference of 0.03 points between a PST at the 16th percentile and a PST at the 84th percentile. Relative to the 7-point CLASS scale, these differences represent at most 5% of the maximum possible difference (6 points).

Consistent with our hypothesis, differences between PSTs in simulation scores are statistically significant and larger. The difference in baseline simulation scores between a PST at the 16th percentile and 84th percentile is 1.11 points for Text-Focused Instruction and 1.71 for Redirection. Relative to the 10-point simulation scale, this represents 12-19% of the maximum possible difference in scores (9 points), more than double the estimate for CLASS scores. We acknowledge the possibility that our results stem from PSTs having very similar instructional skills, rather than a lack of measure sensitivity. However, anecdotal evidence from Lambeth teacher educators and prior work documenting large differences between PSTs once they enter the classroom, suggest that this is unlikely to be the case (e.g. Boyd et al., 2008).

We include results from significance tests and detailed estimates of the differences between PSTs in Tables B8-B12.

Differences between Groups

Here, we compare simulation scores between PSTs that participated in coaching versus self-reflection between practice sessions as one way to explore score sensitivity to differences in learning experiences. PSTs that received coaching for the Text-Focused Instruction simulation score 1.5 points higher immediately after coaching and 0.5 points higher when observed 5 months later, though the latter estimate is not significant in some models (Table B4). PSTs that received coaching for the Redirection simulation score 2.5 points higher immediately after coaching and 1.3 points higher two months later (Table B5).

These results highlight that the simulation-based measures are sensitive enough to detect differences between PSTs that participated in different preparation experiences. Additionally, these results reinforce the feasibility of comparing groups of PSTs, even when a large portion of the variation in scores reflects measurement error. If measurement error and the influence of contextual characteristics like raters are the same across groups of PSTs, we can safely make comparisons across groups.

Differences between Facets of Instruction

Here, we compare PSTs' scores between Classroom Organization and Instructional Support as one way of exploring CLASS scores' sensitivity to differences between facets of instruction. In Figure 3, we graph the scores over time for all PSTs using all three statistical approaches. These results provide suggestive evidence that CLASS scores are sufficiently sensitive to identify program-wide patterns in areas of relative strength. Across all three graphs, Classroom Organization scores are consistently higher than Instructional Support scores, with a difference of 1.00-3.25 points. This suggests that PSTs in our sample are, on average, considerably stronger in management skills (consisting of behavior management, productivity, and avoiding negative climate) than providing instructional support (consisting of instructional learning formats, content understanding, analysis and inquiry, quality of feedback, and instructional dialogue).

[insert Figure 3 here]

At the same time, we see that scores are moderately correlated with one another (0.20-0.45), especially once measurement error is taken into consideration (Table B13). This means PSTs that receive higher scores on Instructional Support also tend to receive higher scores on Classroom Organization. This does not necessarily contradict the previous findings. However, it raises the possibility that there may be differences in individual PSTs' relative strengths and areas for improvement that CLASS scores are not able to detect. This would be the case, for example, if raters tend to perceive Instructional Support as more challenging—resulting in lower average scores—but also form a general impression of each PSTs overall skills (i.e. halo effects), rather than considering each domain individually (Cohen & Goldhaber, 2016). Under these conditions, PSTs with relatively higher Instructional Support scores could potentially be given high Classroom Organization scores, even if their underlying management skills were weaker. Unfortunately, we cannot determine whether this is the case in our context. It is equally possible that PSTs at Lambeth simply have similar relative strengths and areas of improvement.

Discussion

In many ways, Lambeth University is at the forefront of measuring pre-service teachers' instructional skills. Instead of relying on problematic supervisor ratings (Bartanen & Kwok, 2021), the university employs trained and certified raters who do not have personal relationships with the PSTs they rate, resulting in high inter-rater reliability. Furthermore, raters are randomly assigned to videos to avoid systematic bias in scores from differences in rater standards. The university also draws on a validated protocol for in-service teacher observations (CLASS) and innovative researcher-developed simulation-based measures that allow the university to standardize the lesson context in which a PST is observed. The university's data collection procedures are well-aligned with established best practices for in-service teacher observations, where there has been substantial attention to validity and reliability (Ho & Kane, 2013). In these ways, the university's procedures represents substantial improvements over typical approaches to measuring pre-service teacher practice (Bartanen & Kwok, 2021; Mancenido, 2022).

Our results suggest that Lambeth's attention to these measurement issues has been fruitful and highlight several strategies that TPPs may consider employing. First, our results suggest the benefit of supplementing global assessments of instructional quality with finer-grained measures. With a measure as broad as CLASS, differences between PSTs in their use of feedback loops to scaffold student thinking, for example, are likely to be drowned out by broad similarities in the instructional practices employed. By zooming on finer-grained aspects of instruction, simulation scores can detect larger differences between individual PSTs than CLASS scores can. Understanding these finer-grained details are also potentially more helpful for guiding TPP curricula and providing more nuanced feedback to PSTs (Hill & Grossman, 2013; Wylie, 2020). For example, it is far easier to support a PST with providing timely and specific redirections than supporting them with the far more amorphous and multifaceted CLASS dimension of "Positive Climate."

Second, TPPs can increase the extent to which the conditions of observation are standardized or randomly assigned to reduce the likelihood that these contextual conditions influence PSTs' scores. The higher reliability of Lambeth's simulation scores demonstrates that when randomly assigned to observations, raters have little systematic influence on PST scores. Similar to the use of randomized control trials (RCTs) to isolate the casual effect of a treatment, randomizing raters to videos ensures that specific PSTs are not systematically assigned to raters with stricter or more lenient standards. This greatly reduces the likelihood that observed differences in scores or growth between PSTs stem from differences in rater standards, a serious issue when PSTs are rated by a single supervisor or mentor (Bartanen & Kwok, 2021). This approach is costly, however. Lambeth invested in having PSTs video-record their observations and paying additional personnel to score each observation, instead of leveraging supervisor or mentor ratings. Alternatively, TPPs could ask supervisors to collect videos and then randomly assign which supervisors rate which videos. Despite the up-front investments, this approach could also pay dividends in terms of the accuracy of information gleaned from these observations.

Standardizing the conditions of observations offers a valuable alternative to randomization and ensures that all PSTs are observed under the same conditions. In the same way that we administer standardized academic assessments to all students under the same conditions, this standardization reduces the likelihood that differences between scores reflect differences between circumstantial conditions. Our results comparing scores from standardized simulations with CLASS scores reinforces the potential value of this approach. Though simulations are, by design, artificial and do not reflect the full complexity of the classroom, our results suggest that they can be helpful for identifying differences in how candidates enact their knowledge and skills when they are required to face common problems of practice, without the aid of a mentor or teacher educator. Observations during clinical placements cannot feasibly be standardized to the same extent as simulations, but they are more realistic and may benefit from at least some additional standardization. For example, TPPs may be able to provide all PSTs with a set of standardized instructional activities to complete at set times during clinical placements (e.g. facilitating a discussion about a word problem or orchestrating an analysis of a historical text).

At the same time, our results demonstrate that pre-service observational measures suffer from the same challenges that in-service observational measure do. At most, only 20% of the variation in CLASS or simulation scores represents consistent differences between PSTs. Drawing conclusions about individual PSTs' skills based on these scores is risky. When high scores stem not from strong instructional skills, but instead from contextual characteristics, we are missing the opportunity to provide requisite support during the pre-service period. This low reliability also masks differences between individual PSTs' baseline scores and growth over time. Even when we use multi-level models to account for measurement error, our results suggest it is difficult to use CLASS or simulation scores to identify which PSTs may benefit from additional support.

In practice, neither randomization nor standardization is likely to safeguard against a central finding here: mentors' influence on assessments of PST skills. Our results indicate that these effects are not limited to issues of management, but also influence estimates of PSTs' instruction. Standardizing would require ensuring that all PSTs are observed within the same mentor's classroom, an approach that would be logistically infeasible and disruptive for both PSTs and mentors. Randomization may be logistically more feasible but would limit TPPs' ability to intentionally match PSTs with specific geographic areas, grade levels, content areas, school contexts, or mentor characteristics.[2] Clinical experiences and mentors have an important role in scaffolding PST learning and supporting their development (Goldhaber et al., 2022; Ronfeldt, 2012). This means that the influence of mentors on assessments of PSTs is likely to remain a persistent issue for teacher education.

While we recognize the importance of observing how PSTs enact their knowledge and skills in real classrooms, we argue that it's equally important for the measures we use to capture what PSTs can do on their own, instead of what they can do only when assisted by mentors. More work is needed to understand how mentors influence PST scores and develop strategies for minimizing this influence. Indeed, we are not aware of any other work that discusses this issue or provides potential solutions. At a minimum, our work suggests the value of exploring new or modified measures whose indicators focus squarely on PST practices to minimize the potential

[2] In theory, TPPs could exert some control over these issues by first dividing PSTs into groups based on preferences for geographic area, grade level, content area, school context, and/or mentor characteristics. However, this requires a sufficiently large number of PSTs and mentor classrooms within each grouping to allow for random assignment. The more characteristics a TPP wants to influence, the more groups would be required and the smaller the size of each group. While it may be feasible, therefore, for a TPP to randomly assign PSTs to mentors within geographic areas, it may not be feasible for them to randomly assign PSTs to mentors within content area, grade levels, and geographic areas.

for mentor effects. Measures focused on *classroom* quality, like the CLASS, might be less useful in the pre-service context.

TPPs face complex decisions and trade-offs in managing their PST data systems. Systematic collection of PST data requires substantial resources and time investment on the part of TPPs and teacher educators. Altering these systems to address the measurement challenges highlighted in this paper requires even more, especially when new technologies (e.g. simulation equipment) or more intensive data collection efforts (e.g. doubling the number of observations) are required. In addition to financial and logistical constraints, TPPs must navigate potential tensions between improving the reliability, sensitivity, and validity of PST data and ensuring that PSTs' preparation experiences continue to support PST learning and well-being. Standardizing lesson objectives, for example, may improve reliability, but is also challenging when PSTs work in a wide range of grade levels and school contexts.

TPPs must also navigate tensions between the many questions PST data can be used to answer. Measures that detect differences between groups of PSTs, for example, may look different than measures that can detect differences between individual PSTs. Obtaining reliable estimates of PST growth over time requires a different observation schedule than obtaining reliable estimates of PSTs' skills at a specific moment in time. Prior work on in-service teacher observations suggests that expecting a single measure to serve several different purposes is unwise (Hill & Grossman, 2013; Papay, 2012). Instead, distinct measures should be used to draw distinct conclusions about PSTs. This means that TPPs must be crystal clear about what conclusions they wish to draw when making decisions about what measure(s) to use, especially if logistical and financial constraints prevent the use of multiple observational measures.

TPPs and teacher educators stand to reap large benefits when they develop systems that enables data-driven programmatic decision-making. TPPs can make better decisions about how best to allocate limited time and personnel to support the individual PSTs and areas of instructional practice with the greatest need. PST data can also help TPPs and researchers better evaluate the effects of specific preparation experiences on PST learning and skill development. However, the details of the data matter, if we want them to improve the quality of teacher preparation and not just serve as a compliance exercise for program accreditation and evaluation. This is especially true in the pre-service period where PSTs likely exhibit smaller differences in skill, as compared with in-service teachers with a range of experience. There are also more contextual factors, such as mentors, that may influence assessments of PSTs' skills. Finally, accounting for these and other contextual factors is especially important for PSTs. Drawing conclusions about PSTs' skills in the context of clinical placements is of limited interest if those conclusions do not extend to the classroom context(s) in which PSTs ultimately teach. We need PST data systems that are 1) sensitive enough to detect differences between PSTs and facets of instruction, 2) reliable enough to identify consistent differences between individual PSTs, and 3) allow TPPs to generate valid conclusions about PST learning and development.

References

Allen, J., Gregory, A., Mikami, A., Lun, J., Hamre, B., & Pianta, R. (2013). Observations of effective teacher–student interactions in secondary school classrooms: Predicting student achievement with the classroom assessment scoring system—secondary. *School Psychology Review*, 42(1), 76–98.

Allen, J., Hafen, C. A., Gregory, A. C., Mikami, A. Y., & Pianta, R. (2015). Enhancing secondary school instruction and student achievement: Replication and extension of the My Teaching Partner-secondary intervention. *Journal of Research on Educational Effectiveness*, 8(4), 475–489.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science*, *333*(6045), 1034-1037.

Allen, M., & Coble, C. (2018). Creating a data culture in educator preparation: The role of the states. In E. Mandinach and E. Gummer (Eds.), *Data for continuous programmatic improvement* (pp. 35–67). Routledge.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *The Quarterly Journal of Economics*, 131(3), 1415–1453.

Bartanen, B., & Kwok, A. (2020). *Pre-service teacher quality and workforce entry.* EdWorkingPaper No. 20-223. Annenberg Institute at Brown University.

Bartanen, B., & Kwok, A. (2021). Examining clinical teaching observation scores as a measure of preservice teacher quality. *American Educational Research Journal*, 58(5), 887–920.

Bassok, D., Magouirk, P., & Markowitz, A. J. (2021). Systemwide Quality Improvement in Early Childhood Education: Evidence From Louisiana. *AERA Open*, 7(1), 1-19.

Bastian, K. C., Henry, G. T., Pan, Y., & Lys, D. (2016). Teacher candidate performance assessments: Local scoring and implications for teacher preparation program improvement. *Teaching and Teacher Education*, 59, 1–12.

Bastian, K. C., Lys, D., & Pan, Y. (2018). A framework for improvement: analyzing performance-assessment scores for evidence-based teacher preparation program reforms. *Journal of Teacher Education*, 69(5), 448–462.

Bell, C. A., Dobbelaer, M. J., Klette, K., & Visscher, A. (2019). Qualities of classroom observation systems. *School Effectiveness and School Improvement*, 30(1), 3–29.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, 17(2–3), 62–87.

Bell, C. A., Qi, Y., Croft, A. J., Leusner, D., McCaffrey, D. F., Gitomer, D. H., & Pianta, R. C. (2015). Improving observational score quality: Challenges in observer thinking. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 50–97). John Wiley & Sons.

Boyd, D., Lankford, H., Loeb, S., Rockoff, J., & Wyckoff, J. (2008). The narrowing gap in New York City teacher qualifications and its implications for student achievement in high-poverty schools. *Journal of Policy Analysis and Management*, 27(4), 793–818.

Briggs, D. C., & Alzen, J. L. (2019). Making inferences about teacher observation scores over time. *Educational and Psychological Measurement*, 79(4), 636–664.

Casabianca, J. M., Lockwood, J. R., & McCaffrey, D. F. (2015). Trends in classroom observation scores. *Educational and Psychological Measurement*, 75(2), 311–337.

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, 19(1), 5–51.

Charney, R. S. (1993). *Teaching children to care: Management in the responsive classroom*. Northeast Foundation for Children.

Cohen, J. (2015). *"Data Driven" Teacher Education at a Research University: Promising Practices and Potential Pitfalls*. Unpublished manuscript.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, 45(6), 378-387.

Cohen, J., Wiseman, E., & Anglin, K. (In review). Teacher supports for text-based instruction: Experimental evidence from simulations in teacher preparation.

Cohen, J., Wong, V., Krishnamachari, A., & Berlin, R. (2020). Teacher coaching in a simulated environment. *Educational Evaluation and Policy Analysis*, *42*(2), 208-231.

Davis, S. C., & Peck, C. A. (2020). Using data for program improvement in teacher education: A study of promising practices. *Teachers College Record*, 122(3), 1–48.

Dieker, L. A., Rodriguez, J. A., Lignugaris/Kraft, B., Hynes, M. C., & Hughes, C. E. (2014). The potential of simulated environments in teacher education: Current and future possibilities. *Teacher Education and Special Education*, 37(1), 21–33.

Diez, M. E. (2010). It is complicated: Unpacking the flow of teacher education's impact on student learning. *Journal of Teacher Education*, 61(5), 441–450.

Downer, J. T., Booren, L. M., Lima, O. K., Luckner, A. E., & Pianta, R. C. (2010). The Individualized Classroom Assessment Scoring System (inCLASS): Preliminary reliability and validity of a system for observing preschoolers' competence in classroom interactions. *Early Childhood Research Quarterly*, 25(1), 1–16.

Feuer, M., Floden, R., Chudowsky, N., & Ahn, J. (2013). *Evaluation of teacher preparation programs: Purposes, methods, and policy options*. National Academy of Education.

Forzani, F. M. (2014). Understanding "core practices" and "practice-based" teacher education: Learning from the past. *Journal of Teacher Education*, 65(4), 357–368.

Gitomer, D. (Ed.). (2009). *Measurement issues and assessment for teaching quality*. Sage Publications.

Gitomer, D., Bell, C., Qi, Y., McCaffrey, D., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record*, 116(6), 1–32.

Gitomer, D., Martínez, J. F., Battey, D., & Hyland, N. E. (2021). Assessing the assessment: Evidence of reliability and validity in the edTPA. *American Educational Research Journal*, 58(1), 3–31.

Goldhaber, D. (2019). Evidence-based teacher preparation: Policy context and what we know. *Journal of Teacher Education*, 70(2), 90–101.

Goldhaber, D., Ronfeldt, M., Cowan, J., Gratz, T., Bardelli, E., & Truwit, M. (2022). Room for improvement? Mentor teachers and the evolution of teacher preservice clinical evaluations. *American Educational Research Journal*, 59(5), 1011-1048.

Grossman, P., Compton, C., Igra, D., Ronfeldt, M., Shahan, E., & Williamson, P. (2009). Teaching practice: A cross-professional perspective. *Teachers College Record*, 111(9), 2055–2100.

Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., & Pianta, R. C. (2015). Teaching through interactions in secondary school classrooms: Revisiting the factor structure and practical application of the classroom assessment scoring system–secondary. *The Journal of Early Adolescence*, 35(5–6), 651–680.

Henry, G. T., Campbell, S. L., Thompson, C. L., Patriarca, L. A., Luterbach, K. J., Lys, D. B., & Covington, V. M. (2013). The predictive validity of measures of teacher candidate programs and performance: Toward an evidence-based approach to teacher preparation. *Journal of Teacher Education*, 64(5), 439–453.

Hill, H., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, 41(2), 56–64.

Hill, H., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review*, 83(2), 371–384.

Hill, H., Mancenido, Z., & Loeb, S. (2020). *New research for teacher education.* EdWorkingPaper No. 20-252. Annenberg Institute at Brown University.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel.* Bill & Melinda Gates Foundation.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112(3), 527.

Kane, T., & Staiger, D. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Bill & Melinda Gates Foundation.

Kraft, M. A., & Gilmour, A. F. (2017). Revisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, 46(5), 234–249.

La Paro, K. M., Pianta, R. C., & Stuhlman, M. (2004). The Classroom Assessment Scoring System: Findings from the prekindergarten year. *The Elementary School Journal*, 104(5), 409–426.

Lampert, M., & Graziani, F. (2009). Instructional activities as a tool for teachers' and teacher educators' learning. *The Elementary School Journal*, 109(5), 491–509.

Malmberg, L.-E., Hagger, H., Burn, K., Mutton, T., & Colls, H. (2010). Observed classroom quality during teacher education and two years of professional practice. *Journal of Educational Psychology*, 102(4), 916–932.

Mancenido, Z. (2022). *Impact evaluations of teacher preparation practices: Challenges and opportunities for more rigorous research.* EdWorkingPaper No. 22–534. Annenberg Institute at Brown University.

McDonald, M., Kazemi, E., & Kavanagh, S. S. (2013). Core practices and pedagogies of teacher education: A call for a common language and collective activity. *Journal of Teacher Education*, 64(5), 378–386.

Pacheo, A. (2009). Mapping the terrain of teacher quality. In D. Gitomer (Ed.), *measurement issues and assessment for teaching quality* (pp. 168–178). Sage Publications.

Papay, J. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. Harvard Educational Review, 82(1), 123–141.

Park, Y. S., Chen, J., & Holtzman, S. L. (2015). Evaluating efforts to minimize rater bias in scoring classroom observations. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing* *teacher evaluation systems: New guidance from the Measures of Effective Teaching Project* (pp. 381-414). John Wiley & Sons.

Peck, C. A., & McDonald, M. (2013). Creating "cultures of evidence" in teacher education: Context, policy, and practice in three high-data-use programs. *The New Educator*, 9(1), 12–28.

Peck, C. A., Singer-Gabella, M., Sloan, T., & Lin, S. (2014). Driving blind: Why we need standardized performance assessment in teacher education. *Journal of Curriculum and Instruction*, 8(1), 8–30.

Pianta, R. C., & Hamre, B. K. (2009). Classroom processes and positive youth development: Conceptualizing, measuring, and improving the capacity of interactions between teachers and students. *New Directions for Youth Development*, 2009(121), 33–46.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Second Edition). Sage Publications.

Responsive Classroom. (2014). *The responsive classroom approach: Good teaching changes the future*.

Reznitskaya, A., Kuo, L.-J., Clark, A.-M., Miller, B., Jadallah, M., Anderson, R. C., & Nguyen-Jahiel, K. (2009). Collaborative reasoning: A dialogic approach to group discussions. *Cambridge Journal of Education*, 39(1), 29–48.

Ronfeldt, M. (2012). Where should student teachers learn to teach? Effects of field placement school characteristics on teacher retention and effectiveness. *Educational Evaluation and Policy Analysis*, 34(1), 3–26.

St. John, E., Goldhaber, D., Krieg, J., & Theobald, R. (2021). How the match gets made: Exploring student teacher placements across teacher education programs, districts, and schools. *Journal of Education Human Resources*, 39(3), 261–288.

van der Lans, R. M., van de Grift, W. J., van Veen, K., & Fokkens-Bruinsma, M. (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation*, 50, 88–95.

von Hippel, P. T., Bellows, L., Osborne, C., Lincove, J. A., & Mills, N. (2016). Teacher quality differences between teacher preparation programs: How big? How reliable? Which programs are different? *Economics of Education Review*, 53, 31–45.

Wallace, T. L., Parr, A. K., & Correnti, R. J. (2020). Assessing teachers' classroom management competency: A case study of the classroom assessment scoring system–secondary. *Journal of Psychoeducational Assessment*, 38(4), 475–492.

Weisberg, D., Sexton, S., Mulhern, J., Keeling, D., Schunck, J., Palcisco, A., & Morgan, K. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness*. New Teacher Project.

Wylie, E. C. (2020). Observing formative assessment practice: Learning lessons through validation. *Educational Assessment*, 25(4), 251–258.

Figures

Figure 1. Conceptual framework articulating how data on PST knowledge, skills, and dispositions can contribute to improvements in teacher preparation.

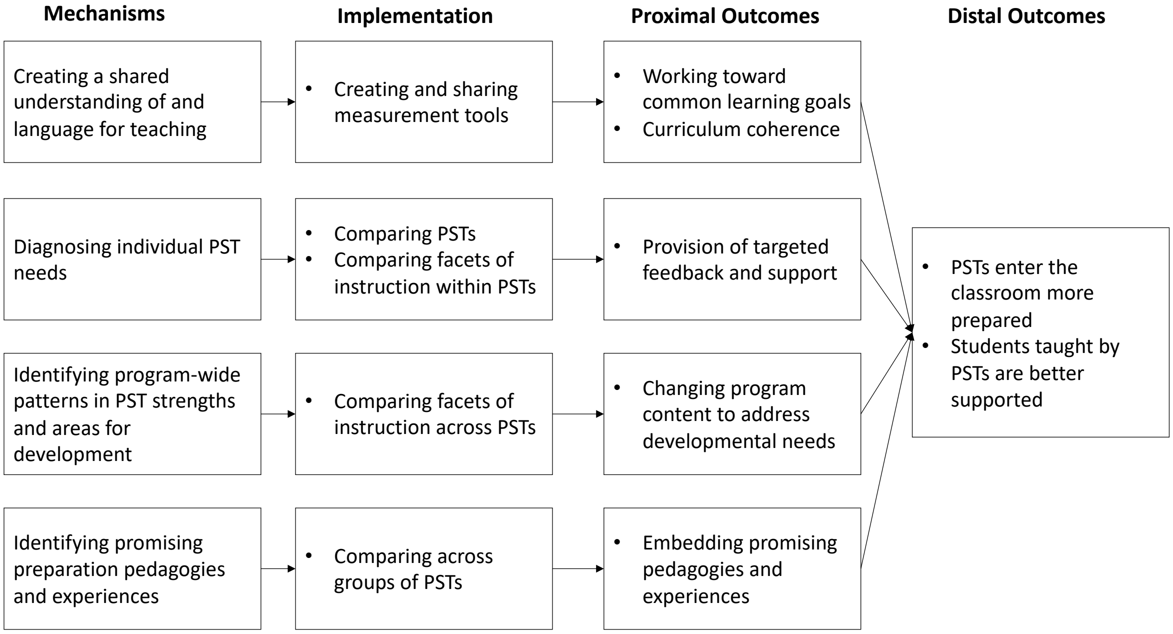
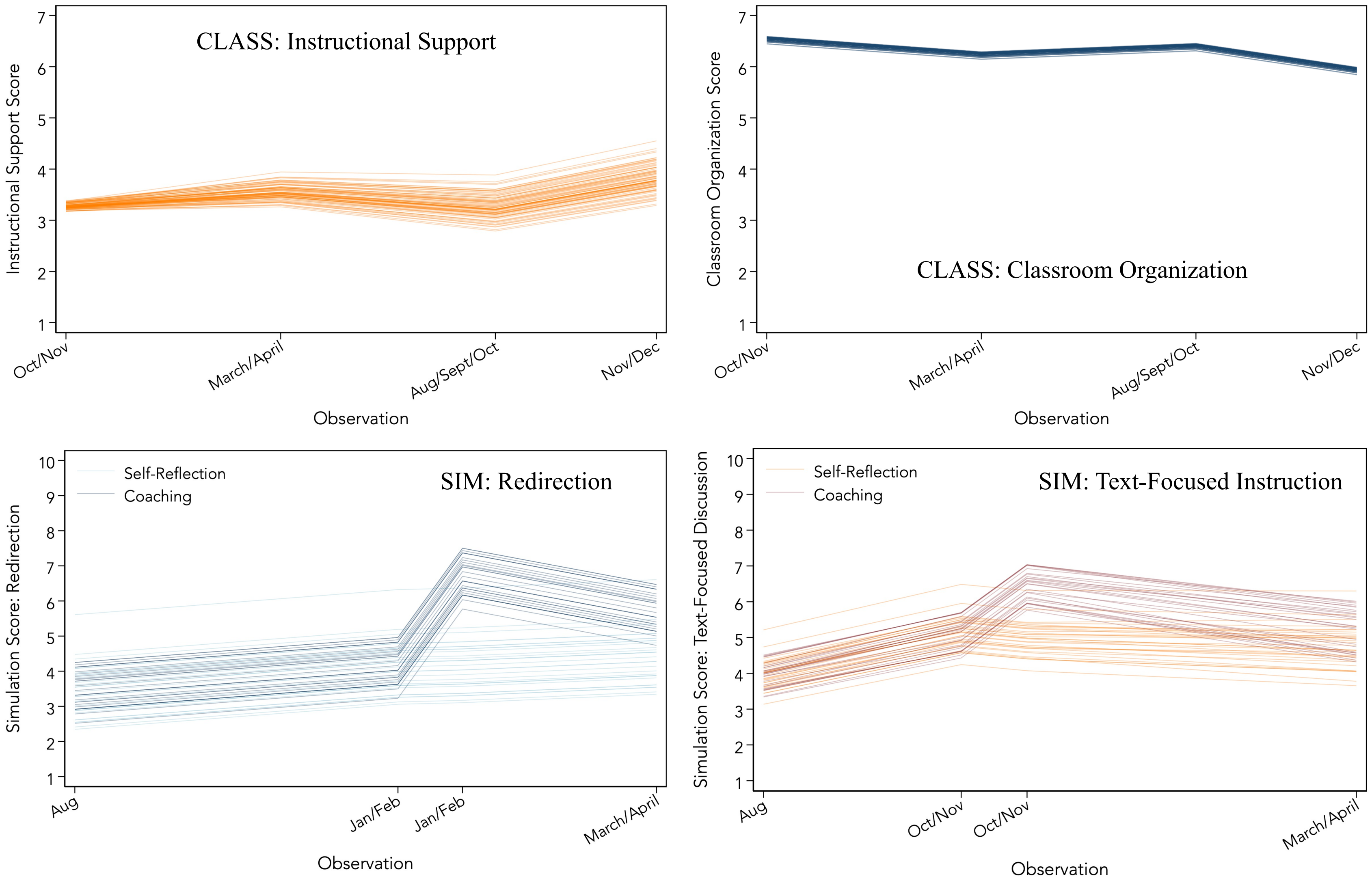

Mechanisms
Implementation
Proximal Outcomes
Distal Outcomes
Creating a shared understanding of and language for teaching
Creating and sharing measurement tools
Working toward common learning goals
Curriculum coherence
Diagnosing individual PST needs
Comparing PSTs
Comparing facets of instruction within PSTs
Provision of targeted feedback and support
PSTs enter the classroom more prepared
Students taught by PSTs are better supported
Identifying program-wide patterns in PST strengths and areas for development
Comparing facets of instruction across PSTs
Changing program content to address developmental needs
Identifying promising preparation pedagogies and experiences
Comparing across groups of PSTs
Embedding promising pedagogies and experiences

Figure 2. Predicted PST scores after accounting for measurement error to isolate between-PST variation in scores.



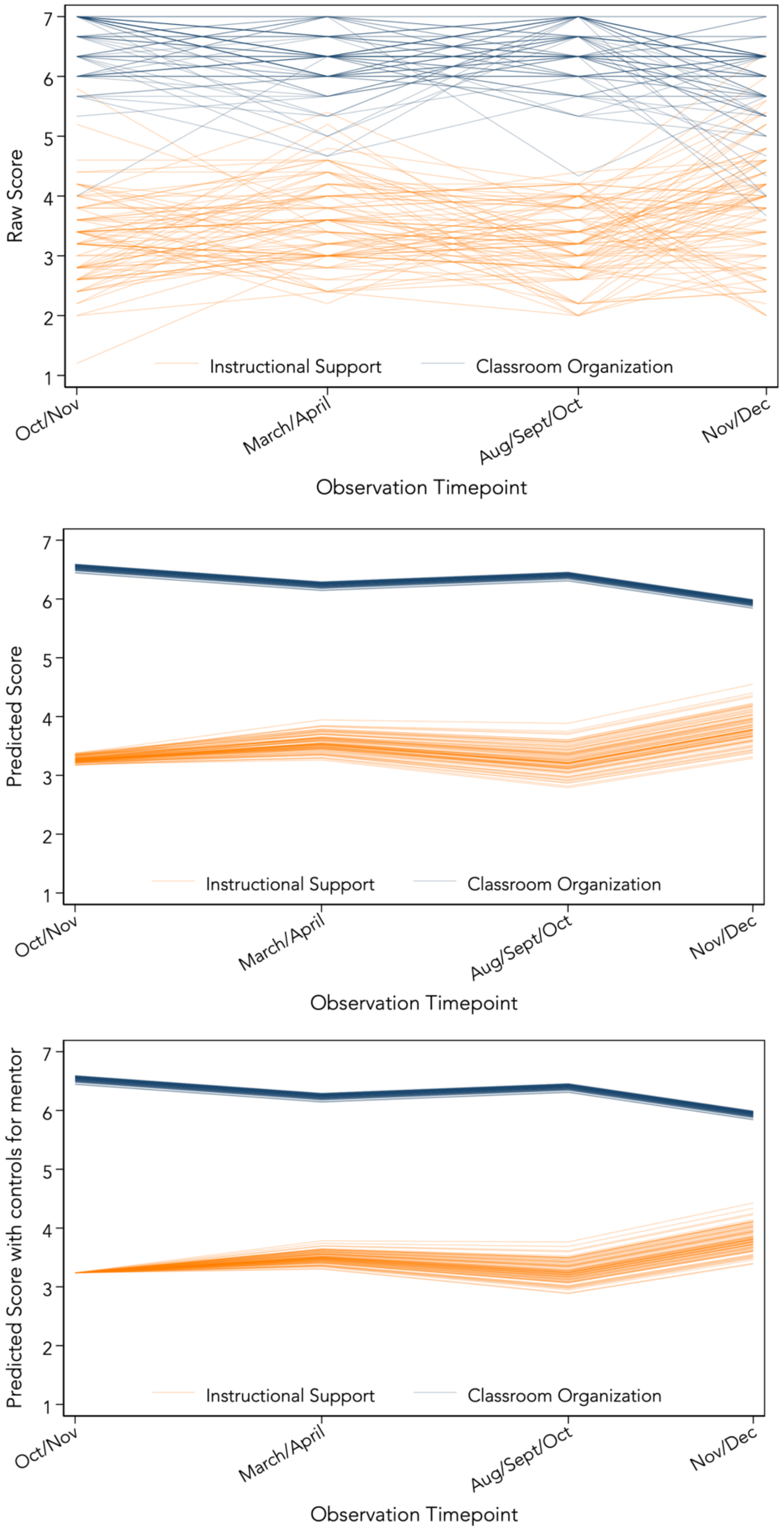
CLASS: Instructional Support

CLASS: Classroom Organization

SIM: Redirection

SIM: Text-Focused Instruction

Figure 3. Comparison between Instructional Support and Classroom Organization scores using raw scores and predicted scores.



Appendices

Appendix A: Methodological Appendix

RQ1: To what extent does each measure provide the intended information about differences in PSTs' instructional practice?

To answer this question, we fit a multi-level model to first partition the variance in CLASS and simulation scores. Our primary specification is the two-level model shown in Equations 1 and 2.

$$(1) \qquad y_{jt} = \pi_{0j} + \pi_{1j} + \pi_{2j} + \pi_{3j} + \varepsilon_{jt}$$

$$(2) \qquad \pi_{0j} = \beta_{00} + u_{0j}$$

$$\pi_{\mathrm{tj}} = \beta_{tj} + u_{1j}$$

In the Level 1 model (Equation 1), y_{jt} represents PST j's CLASS or simulation score at time *t*, and is modelled as a function of a PST's baseline score at t=0 (π_{0j}), three time splines to flexibly allow for non-linear changes in scores at each subsequent observation timepoint (π_{1j}, π_{2j}, π_{3j}) and a residual error term, (e_{jt}), where $e_{jt} \sim N(0, \sigma^2)$. Level 2 models between-PST variation in the Level 1 parameters as shown in Equation 2. Specifically, we estimate each individual PSTs' baseline score (intercept) as a function of the mean baseline score (β_{00}) and a random effect representing each PSTs' unique deviation from the mean (u_{0j}). We also estimate each PSTs' trajectory over time (slope) as a function of the mean slope at time *t* (β_{tj}) and a random effect representing each PSTs' unique time-invariant deviation from the mean trajectory (u_{1j}), where empirically supported.[3] In estimating the slope random effect, we center time as a

[3] We do not include a time random effect for analyses of simulation scores for the redirection scenario or CLASS Classroom Organization scores because empirical results and model fit statistics suggest that there is not sufficient variation to warrant the inclusion of a time random effect.

continuous variable representing the number of months since the first period of observation (t=0). Additionally, for models with simulation outcomes we include interactions between an indicator for whether the PST received coaching and any post-coaching timepoints to allow PSTs that received coaching to exhibit different trajectories in line with prior work demonstrating coaching treatment effects (Cohen et al., 2020; Cohen & Wiseman, 2021).

We then use the estimated variance of each random effect to calculate the proportion of the variation that reflects consistent differences between PSTs, using two methods. First, we calculate reliability as the proportion of the total variance explained by PST random effects and time random effects as shown in Equation 3, following traditional definitions of reliability (Raudenbush & Bryk, 2002). Second, we follow Briggs & Alzen's (2019) recently proposed approach for calculating the reliability of PST growth, which requires estimating a separate proportion for initial differences between PSTs (intercept), shown in Equation 4, and differences in PST growth over time (slope), shown in Equation 5.

$$(3)\ \rho = 100 * \frac{\sigma^2_{u0j} + \sigma^2_{uj}}{\sigma^2_{u0j} + \sigma^2_{uj} + \sigma^2_{\varepsilon}}$$

$$(4)\ \rho(\pi_{0j}) = \frac{\sigma^2_{u0j}}{\sigma^2_{u0j} + \sigma^2_{\varepsilon}}$$

$$(5)\ \rho(\pi_{\mathrm{tj}}) = \frac{\sigma^2_{uj}}{\sigma^2_{uj} + \ ^{\sigma^2_{\varepsilon}}/_{SST}}$$

We report these as proportions rather than percentages to reflect the fact that they cannot be added together to total 100% of the variation since these equations incorporate different denominators. In Equation 5, SST is calculated as $\sum_{t=0}^{T}(Time_{tj} - \overline{Time_j})^2$ and serves to adjust the proportion to account for the number and spacing of observations. Holding the time-period over which growth is estimated constant, conducting more observations and/or ensuring that observations are more widely spread out will increase the SST and therefore increase the estimated proportion. The intuition here is that scores from more or more widely spread-out observations will be more representative of each PST's growth over time than scores from fewer or more narrowly spaced observations. Since the exact spacing between observations varies across PSTs, we calculate the SST using the modal number of months between observations, resulting in an SST of 101 for Instructional Support and 26.75 for the Text-Focused Instruction simulation. We do not estimate Equation 5 for Classroom Organization and the Redirection simulation scores because we cannot estimate time random effects for these models.

RQ2: To what extent do raters and mentors influence PST scores?

To answer this question, we modify Equations 1 and 2 to further decompose the measurement error identified into distinct sources. Specifically, we include a rater random effect for simulation scores to capture systematic patterns in raters' influence on scores and a mentor teacher random effect for CLASS scores to capture systematic patterns in how the placement context influences scores. Because PSTs are not perfectly nested within raters or mentors, we add these random effects as crossed effects, where possible. However, for Classroom Organization and Redirection scores, the crossed model does not converge. As an alternative, we therefore estimate a simplified version of the model with raters or mentors as fixed rather than random effects. We then recalculate Equations 3 and 4, adding the variance of the rater or mentor random effect to the denominator where necessary. We also modify these equations to calculate the proportion of the variation explained by rater or mentor effects.

RQ3: How well can each measure differentiate between individual PSTs, groups of PSTs, and facets of instruction?

To explore the sensitivity of CLASS and simulation scores to individual differences, we review results from RQ1with the aim of understanding the extent to which we can detect between-PST variation in scores and trajectories. Specifically, we use log likelihood tests to compare models with and without PST intercept and slope random effects to determine whether there is significant between-PST variation in baseline scores and growth trajectories. We also generate graphs of PST trajectories, using both raw scores and the predicted Best Linear Unbiased Predictors (BLUPs) from RQ1 models to allow visual inspection of how much scores and trajectories vary across PSTs. We use BLUP estimates from RQ1 models without controls for rater or mentor effects as an upper bound of sensitivity.

To explore the sensitivity of scores to group differences, we review results from RQ1 for simulation scores, where a control for participation in coaching vs. self-reflection allows us to evaluate whether simulation scores are sufficiently sensitive to detect differences between PSTs exposed to these different preparation experiences. Unfortunately, we are not aware of any systematic differences in preparation experiences that can be used to compare groups for the CLASS scores.

To explore the sensitivity of scores to differences between different facets of instructional skill, we visually compare the magnitude of Classroom Organization and Instructional Support scores, using both raw scores and estimated BLUPs. We also estimate the correlation between Classroom Organization and Instructional Support scores at each timepoint. We do not investigate the sensitivity of simulation scores because the self-reflection and coaching treatments make results difficult to interpret. Furthermore, we expect that programs are more likely interested in comparing PST skills across broad domains of instruction, such as those represented by CLASS, which might inform courses and curricula in future years, rather than the narrow facets of instruction capture by the two simulation measures.

Appendix B: Additional Results

Table B1: Proportion of the variation that reflects consistent differences between PSTs relative to the overall variation in scores.

	Instructional Support	Classroom Organization	Redirection	Text-Focused Instruction
Overall variation, assuming no growth in scores over time	0.07	0.03	0.17	0.11
Overall variation, allowing for growth in scores over time	0.04	0.03	0.22	0.21
Variation in baseline scores	0.04			0.21
Variation in growth over time	0.15			0.15

Table B2. Coefficients from multi-level models for Instructional Support.

	(1)		(2)		(3)	
			Instructional Support			
Fixed Effects						
Time 0	3.27*	(0.08)	3.23*	(0.09)	3.24*	(0.08)
Time 1	0.28*	(0.11)	0.27*	(0.11)	0.27*	(0.11)
Time 2	-0.30*	(0.11)	-0.24*	(0.12)	-0.25*	(0.12)
Time 3	0.57*	(0.11)	0.57*	(0.11)	0.57*	(0.11)
Random Effects (S.D.)						
PST intercept	0.15				0.02	
PST slope	0.03				0.03	
Residual	0.71		0.71		0.69	
Mentor Teacher			0.30		0.22	
Variance Components						
Overall Reliability	0.04		n/a		0.0002	
Intercept Reliability	0.04		n/a		0.0004	
Slope Reliability	0.15		n/a		0.14	
Mentor Teacher	n/a		0.15		0.09	
Residual	0.96		0.85		0.91	
Observations (PSTs)	332	(83)	332	(83)	332	(83)

Table B3. Coefficients from multi-level models for Classroom Organization.

	(1)		(2)		(3)	
			Classroom Organization			
Fixed Effects						
Time 0	6.54*	(0.06)	6.54*	(0.07)	6.48*	(0.10)
Time 1	-0.30*	(0.09)	-0.28*	(0.08)	-0.25*	(0.09)
Time 2	0.16	(0.09)	0.15	(0.09)	0.04	(0.18)
Time 3	-0.47*	(0.09)	-0.47*	(0.08)	-0.47*	(0.08)
Mentor fixed effects					x	
Random Effects (S.D.)						
PST intercept	0.10					
Residual	0.55		0.52		0.00	
Mentor Teacher			0.23		0.51	
Variance Components						
Intercept Reliability	0.03		n/a		0.00	
Mentor Teacher	n/a		0.17		n/a	
Residual	0.97		0.83		1.00	
Observations (PSTs)	332	(83)	332	(83)	332	(83)

Table B4. Coefficients from multi-level models for Text-Focused Instruction simulation.

	(1)		(2)	
		Text-Focused Instruction		
Fixed Effects				
Time 0	3.97*	(0.16)	3.95*	(0.17)
Time 1	1.17*	(0.20)	1.16*	(0.20)
Time 2	-0.17	(0.24)	-0.15	(0.24)
Time 3	-0.25	(0.28)	-0.24	(0.29)
Time 2 * Coaching	1.51*	(0.30)	1.49*	(0.30)
Time 3 * Coaching	-0.98*	(0.41)	-0.96*	(0.41)
Random Effects (S.D.)				
PST intercept	0.56		0.56	
PST slope	0.09		0.09	
Residual	1.08		1.07	
Rater			0.14	
Variance Components				
Overall Reliability	0.21		0.21	
Intercept Reliability	0.21		0.21	
Slope Reliability	0.15		0.15	
Rater	n/a		0.01	
Residual	0.79		0.78	
Observations (PSTs)	240	(60)	240	(60)

Table B5. Coefficients from multi-level models for Redirection simulation.

	(1)		(2)	
	Redirection Simulation			
Fixed Effects				
Time 1	3.44*	(0.24)	3.47*	(0.28)
Time 2	0.71*	(0.29)	0.72*	(0.29)
Time 3	0.04	(0.37)	0.11	(0.36)
Constant	0.24	(0.41)	0.07	(0.43)
Time 2 * Coaching	2.50*	(0.45)	2.45*	(0.44)
Time 3 * Coaching	-1.28*	(0.59)	-1.23*	(0.58)
Random Effects (S.D.)				
PST intercept	0.86		0.89	
Residual	1.61		1.57	
Rater			0.33	
Variance Components				
Intercept Reliability	0.22		0.23	
Rater	n/a		0.03	
Residual	0.78		0.73	
Observations (PSTs)	240	(60)	240	(60)

Table B6. Proportion of the variation in CLASS scores that reflects consistent differences between PSTs vs. measurement error, separating out the influence of mentors on scores.

	Instructional Support	Classroom Organization
Variation in scores between mentors, assuming no additional between-PST variation	0.15	0.17
Variation in scores between mentors, allowing for additional between-PST variation	0.09	n/a
Variation in scores between PSTs, after accounting for between-mentor variation	0.0002 (baseline scores) 0.0004 (growth over time)	0.00

Table B7. Proportion of the variation in simulation scores that reflects consistent differences between PSTs vs. measurement error, separating out the influence of raters on scores.

	Redirection	Text-Focused Instruction
Variation in scores between raters, allowing for additional between-PST variation	0.03	0.01
Variation in scores between PSTs, after accounting for between-rater variation	0.23	0.21 (baseline scores) 0.15 (growth over time)

Table B8. Between-PST variation in baseline Instructional Support scores and growth trajectories.

	16th percentile	50th percentile	84th percentile
	CLASS: Instructional Support		
Baseline	3.12	3.27	3.42
+ 6 months	0.25	0.28	0.31
+ 5 months	-0.33	-0.30	-0.27
+ 2 months	0.54	0.57	0.60

Table B9. Between-PST variation in baseline Text-Focused Instruction scores and growth trajectories.

	16th percentile	50th percentile	84th percentile
	SIM: Text-Focused Instruction		
Baseline	3.41	3.97	4.52
+ 2 months	1.09	1.17	1.26
+ 5 min	-0.26 SR 1.25 C	-0.17 SR 1.34 C	-0.08 SR 1.42 C
+ 5 months	-0.34 SR -1.32 C	-0.25 SR -1.23 C	-0.16 SR -1.15 C

Note: *Estimates labelled "SR" are for candidates who participated in self-reflection, while estimates labelled "C" are for candidates who participated in coaching.*

Table B10. Between-PST variation in baseline Classroom Organization scores and growth trajectories.

	16th percentile	50th percentile	84th percentile
	CLASS: Classroom Organization		
Baseline	6.44	6.54	6.64
+ 6 months		-0.30	
+ 5 months		0.16	
+ 2 months		-0.47	

Table B11. Between-PST variation in baseline Redirection scores and growth trajectories.

	16th percentile	50th percentile	84th percentile
		SIM: Redirection	
Baseline	2.59	3.44	4.30
+ 5 months		0.71	
+ 5 min		0.04 SR 2.54 C	
+ 2 months		0.24 SR -1.03 C	

Note: *Estimates labelled "SR" are for candidates who participated in self-reflection, while estimates labelled "C" are for candidates who participated in coaching.*

Table B12. P-values from likelihood ratio tests testing the significance of between-PST variation in scores (i.e. PST random effects).

	Instructional Support	
	PST Random Intercept	Mentor Random Intercept
Fixed Effects Only	0.0324	0.0038
	PST Random Intercept	PST Random Intercept + Slope (unstructured covariance)
PST Random Intercept + Slope (covariance = 0)	0.0056	0.2084
	Classroom Organization	
	PST Random Intercept	Mentor Random Intercept
Fixed Effects Only	0.4861	0.0396
	Text-Focused Instruction	
	PST Random Intercept	Rater Random Intercept
Fixed Effects Only	0.0002	0.5755
	PST Random Intercept	PST Random Intercept + Slope (unstructured covariance)
PST Random Intercept + Slope (covariance = 0)	0.4728	0.3853
	Redirection	
	PST Random Intercept	Rater Random Intercept
Fixed Effects Only	0.0002	0.4876
	PST Random Intercept	PST Random Intercept + Slope (unstructured covariance)
PST Random Intercept + Slope (covariance = 0)	0.1084	

Table B13. Spearman's rank correlations between Instructional Support and Classroom Organization.

	Raw CLASS Scores	Predicted CLASS scores, without mentor effects	Predicted CLASS scores, with mentor effects
Observation 0	0.20*	0.45***	0.45***
Observation 1	0.12	0.44***	0.41***
Observation 2	0.39***	0.43***	0.41***
Observation 3	0.43***	0.42***	0.41***

*** p<0.01, ** p<0.05, * p<0.1

CHAPTER 2

Parsing Coaching Practice: A Systematic Framework for Describing Coaching Discourse

Arielle Boguslav

Abstract

Coaching is one of the most promising levers for educational equity. However, we don't yet understand how more impactful coaching interactions differ from less impactful interactions. Despite the common title of "coach," what coaches *do* to support teachers is highly variable. This leaves practitioners with a many choices and little evidence-based direction. Furthermore, the literature provides only rare glimpses into the concrete discourse strategies coaches can use. To address these gaps, this paper introduces a taxonomy of coaching discourse practices. In developing the taxonomy, I conduct a conceptual, qualitative review of the coaching literature to identify potential discourse moves. The taxonomy serves as a common language to describe coaches' interactions with teachers and how they may influence teacher development.

Over the last three decades, coaches, mentors, and consultants have become a regular fixture in schools around the world (Domina et al., 2015; Kraft et al., 2018; Lochmiller, 2021). During student teaching placements, pre-service teachers regularly meet with their mentor teachers and supervisors to discuss prior lessons and plan for future ones (Matsko et al., 2020). Early career teachers are often assigned mentors or instructional coaches (Kutsyuruba & Godden, 2019). As part of teacher evaluation systems, administrators provide feedback and support in debriefs following classroom observations (Donaldson & Woulfin, 2018; Hunter & Springer, 2022). Increasingly, schools are incorporating one-on-one instructional coaching as a central component of professional development for all teachers (Galey, 2016; K. Johnson, 2016; Neufeld & Roper, 2003). While important distinctions can be made between these programs, they are unified by a key component of their theory of action: that engaging in dialogue about the day-to-day details of a teacher's classroom and instruction with another education professional (teacher, administrator, coach, etc.) can spur improvements in teachers' instruction and student learning. For this reason, I use the term coaching here to refer collectively to coaching, mentoring, and consultation programs.

One reason coaching is so widespread is the growing evidence it can enhance teachers' instruction and improve student learning, unlike most other forms of professional development (Davis & Higdon, 2008; Hardt et al., 2020; Kraft et al., 2018; Ronfeldt et al., 2018; Ronfeldt & Reininger, 2012; R. Stanulis & Floden, 2009). This kind of personalized support is also highly valued by teachers (Clark & Byrnes, 2012; Gross, 2010; Ronfeldt et al., 2020). Together, this makes coaching one of the most promising levers for ensuring equitable access to educational opportunities (Alston et al., 2018; Hiebert & Morris, 2012; Kloser et al., 2019).

Yet realizing this promise is not straightforward. The administrators and coaches responsible for implementing coaching programs face a dizzying array of different coaching models and ideas about what coaches can say and do in their interactions with teachers to support their development. Knight's (2009) edited volume on coaching alone introduces seven different types of coaching, including instructional coaching, cognitive coaching, and content coaching. While studies exploring the effects of specific models abound, few studies make comparisons across different models of coaching to understand what coaching practices are the most helpful, for whom, and under what circumstances. Furthermore, synthesizing across studies is complicated by the lack of a common language for describing the practices coaches use. This makes it difficult to identify patterns across studies in how coaching practice supports teacher development. As a result, coaches and administrators are left with many options and little evidence-based direction for how to select among them (Galey, 2016; Gibbons & Cobb, 2016). These challenges are further exacerbated by the literature's focus on more abstract features of coaching practice, with limited attention to the concrete discourse strategies coaches can use to achieve these aims. For example, there is broad consensus about the need for coaches to build trusting relationships with their teachers, but little evidence-based guidance that highlights what coaches can say and do to build such relationships. Thus, coaches and administrators are left largely on their own to identify specific discourse strategies, such as validating a teachers' emotions, that can help them reach these goals.

Rather than relying on administrators to recruit coaches with strong "people skills" or placing the onus on practitioners to "figure it out" on their own, we need a systematic program of research designed to identify effective coaching discourse strategies across contexts and program models. To do so, we first need a coherent framework that can provide a common language for describing coaching discourse strategies and outline potentially promising strategies that warrant further investigation. This paper therefore introduces a taxonomy of concrete coaching discourse "moves," or questioning and feedback discourse strategies coaches may use in their interactions with teachers as the foundation for a framework of coaching practice (Boerst, et al., 2011). This work is informed by the literature on frameworks of teaching practice, which have helped researchers refine their understanding of high-quality teaching practice and are shaping teacher education (Boerst, et al., 2011; Cohen, 2015; Grossman et al., 2009; Grossman & McDonald, 2008; Lampert & Graziani, 2009). Contemporary frameworks of teaching practice use a nested structure, beginning with high-level practices and instructional purposes that are successively decomposed into ever more detailed and specific components, culminating in concrete discourse moves. These frameworks are built on decades of research identifying specific discourse strategies, like wait time and revoicing student contributions, that contribute to student learning (O'Connor & Michaels, 1993; Rowe, 1986; Tobin, 1987). Similar foundational work has yet to be conducted for coaching practice.

This paper, and the taxonomy it introduces, focus exclusively on coaching discourse moves because they are under explored in the literature and likely influence teacher learning. While the coaching literature is filled with discussions of high-level coaching practices and purposes, such as building trust and supporting teacher self-reflection, there is a dearth of analogous research on how concrete coaching discourse strategies support teacher development (Heineke, 2013; King et al., 2004; L'Allier et al., 2010; Obara, 2010; Robertson et al., 2020; Sisson & Sisson, 2017; Walpole et al., 2010). Yet, there is good reason to believe that coaching discourse strategies matter for teacher development. Given the substantial evidence of the role of teacher discourse in student learning, it seems unlikely that teacher learning would not also be influenced by coach discourse (Demszky & Hill, 2022; O'Connor & Michaels, 1993; Rowe, 1986; Tobin, 1987). Additionally, the limited available literature provides suggestive evidence of the importance of coach discourse strategies (Heineke, 2013; Hunt, 2016; Robertson et al., 2020; Sims & Fletcher-Wood, 2021). Finally, though the existing literature does not focus on discourse moves as the unit of analysis, examples and descriptions of coaching discourse are frequently used in the academic and practitioner literature to illustrate how coaches can implement these practices (Aguilar, 2013; Ippolito, 2010; Knight, 2019). This suggests a widespread belief in the importance of coaching discourse for teacher development and for differentiating between different approaches to coaching.

In developing the taxonomy, I conduct a conceptual, qualitative review of the coaching literature to identify potential discourse moves. Instead of following a systematic review process, I strategically sample several kinds of resources, including academic research and practitioner resources. to understand the nature and breadth of observed coaching discourse. In this way, the methods that I use are more akin to those used to develop qualitative codebooks (Miles et al., 2014). In identifying relevant literature, I use a broad definition of a coaching conversation as a dialogue between two or more education professionals, where:

- at least one participant is a classroom teacher, and the primary focus of the dialogue is on this teacher's classroom, this teacher's current teaching practice, and/or opportunities for the teacher to improve or change their teaching practice; and
- a different participant – the coach – serves as the facilitator to maintain focus on these topics.

This definition is purposely broad to be applicable to a variety of coaching models and contexts.

In introducing a taxonomy of coaching discourse moves, this paper makes three primary contributions. First, the taxonomy can serve as a conceptual framework for future empirical research, providing a common language for describing coaching discourse and articulating aspects of coaches' interactions with teachers that warrant further investigation. Second, the taxonomy can provide a practical toolkit and technical vocabulary for coaching practitioners, synthesizing our existing knowledge of coaching discourse into a flexible repertoire of discourse strategies that can be used to reflect on and plan for coaching conversations. Third, in serving as a common language for coaching research and practice, the taxonomy can foster greater integration between coaching research and practice. A shared framework and language for describing coaching discourse will facilitate the systematic accumulation, synthesis, and application of new knowledge about coaching (Boerst, et al., 2011; Charalambous & Praetorius, 2020; Hiebert & Morris, 2012; Kloser et al., 2019; McDonald et al., 2013). Currently, studies of coaching include a wide variety of programs, defined and operationalized in different ways, and described in varying levels of detail with inconsistent terminology. This requires authors of reviews and meta-analyses to dedicate substantial energy to making sense of these differences and developing a common coding scheme or conceptual framework to enable comparison across studies (Kraft et al., 2018). However, when individual studies use a common language for describing coaching, identifying patterns across studies, and conducting meta-analyses will be considerably easier. Furthermore, when coaching practitioners use the same language as one another to discuss their work, they will be better able learn from and support each other. Finally, when researchers and practitioners use the same language, it will be easier for researchers to communicate their insights to practitioners and for practitioners to act on these insights in their daily practice.

In the sections that follow I review the literature on coaching, describe my methods for developing the taxonomy, describe the taxonomy's structure and content, and illustrate how the taxonomy may be used by researchers and practitioners.

Background

What is coaching?

Like many popular educational interventions and innovations, how coaching is operationalized and implemented is highly variable. Indeed, a non-trivial portion of the literature focuses on defining and categorizing specific approaches to coaching. In their foundational work, Joyce and Showers (1981) describe coaching as ongoing cycles of "observation and feedback" (p. 170) where a coach aims to help improve a teacher's implementation of new instructional strategies introduced as part of professional development workshops or other programming. Later work introduces additional conversational structures, distinguishes between different kinds of coaching, and differentiates mentoring and consultation from coaching.

In attempt to identify potentially productive coaching activities, Gibbons and Cobb (2017) describe 19 structures coaches can use in their interactions with teachers. In addition to observation and feedback cycles, they include structures like co-teaching (where the coach and teacher together plan and teach a lesson), modeling instruction (where the teacher observes the coach's or another teacher's instruction and then debriefs the observation with the coach), lesson planning (where the coach and teacher plan a future lesson together), examining student work, and facilitating opportunities for a teacher to rehearse new instructional practices and receive feedback.

Other work focuses on more nuanced features of coaches' interactions with teachers to distinguish among different coaching approaches. Several scholars, for example, reference the distinction between responsive coaching, where the coach allows the teacher's self-reflections and goals to guide content of coaching, and directive coaching, where the coach draws on their own expertise to provide directive suggestions about what the teacher should change (Deussen et al., 2007; Dozier, 2006; McGatha, 2017). Several specific models of coaching are also defined by expectations about how coaches should interact with teachers. Knight's (2007) Instructional Coaching approach, for example, highlights seven partnership principles, including collaborating with teachers as equal partners and promoting teacher choice and decision-making. Whereas Instructional Coaching emphasizes the coach's role as a partner that works together with the teacher, Cognitive Coaching emphasizes the role of the coach as a facilitator whose goal is to help teachers exercise self-direction without offering their own judgment or advice (Costa & Garmston, 2002).

Other types of coaching are differentiated by the goals of the support provided. Literacy coaches, for example, are expected to work with students and teachers to promote student literacy, while mathematics coaches are expected to support teachers with developing students' mathematical skills (Obara, 2010; Toll, 2009; West, 2009). Some coaching programs focus on specific professional skills such as classroom management or data analysis (Marsh et al., 2010; Means et al., 2010; Reinke et al., 2009). Still other coaching programs look beyond individual teachers' practice to foster school or district-wide instructional reform (Woulfin, 2018; Woulfin & Rigby, 2017).

Approaches to coaching are also sometimes differentiated by *who* coaches and teachers are rather than *what* they do in their interactions. Ackland (1991), for example, distinguishes between expert peer coaching, where a more accomplished teacher supports the development of a less accomplished teacher, and reciprocal peer coaching, where teachers of similar skill levels work together. Other models rely on accomplished teachers who leave the classroom to focus primarily on supporting other teachers (Coggins et al., 2003; Kane & Rosenquist, 2019; Knight, 2007). Additionally, while some authors include conversations between a facilitator and multiple teachers as a form of coaching, others define coaching as one-on-one meetings between a single teacher and coach (Gibbons & Cobb, 2017; Kraft et al., 2018).

It is notable that few of these definitions are mutually exclusive. A coaching program in which teachers of similar skill levels coach each other on instructional strategies for mathematics reflects elements of both peer-to-peer coaching and mathematics coaching. Similarly, one can imagine a variety of different content-specific coaching programs that might rely on different structures and coaching practices. Existing work suggests that coaches may draw on both directive and responsive strategies even within a single coaching conversation (Ippolito, 2010).

This complexity is even more evident in the literature on mentoring and consultation. Not only is there no consensus about what distinguishes mentoring from consultation from coaching, but there is also considerable overlap in many of the definitions provided (Downer et al., 2018; Lancer et al., 2016). For example, both coaching and mentoring are sometimes described as relationship-oriented and may involve providing emotional support for teachers (Downer et al., 2018; Mena et al., 2016). Similarly, both coaching and consultation are sometimes described as individualized supports provided to help teachers with implementing specific instructional practices (Joyce & Showers, 1981; Kurz et al., 2017; Reinke et al., 2008).

With all this overlap, how to synthesize findings across studies is not at all clear. Can findings about effective strategies for building relationships in a mentoring program generalize to other kinds of programs? Can a district combine evidence on peer coaching with evidence on literacy coaching to create peer literacy coaching? Or does the evidence on literacy coaching only apply when coaches are literacy experts, as literacy coaching typically requires? Without a common language for differentiating between coaching programs and describing how coaches' interactions vary, we cannot develop a systematic understanding of coaching. In the meantime, coaches and administrators are left to muddle through this complexity on their own.

How do coaching conversations support teacher development?

The literature offers a range of ideas about how coaches' interactions with teachers can support teachers' professional development and instructional practice. Common theories are summarized in Figure 1. Many scholars, for example, highlight the *job-embedded* nature of coaching interactions as a key mechanism that makes coaching conversations valuable for teachers. Because coaching conversations are grounded in the details of teachers' day-to-day instruction, content, and students, they are responsive to teachers' needs and provide authentic opportunities for teachers to make connections between theory and the practical details and challenges of instruction in the context of their daily work (Collet, 2012; Croft et al., 2010; Koh & Neuman, 2006; Putnam & Borko, 2000; Terehoff, 2002).

Coaching conversations can also be conceptualized as *active learning* opportunities where, instead of serving as only passive recipients of information, teachers actively participate in tasks such as self-reflection, problem-solving, data-analysis, lesson-planning, and practicing instructional strategies (Desimone & Pak, 2017; Shernoff et al., 2015). Active participation necessitates a deeper level of mental engagement than that required by more passive activities and provides opportunities for teachers to construct new knowledge in collaboration with the coach (Lieberman, 1995; Niemi et al., 2016). Coaches can also use their time with teachers to help them make sense of the many competing pressures and challenges they face (e.g. district priorities, content standards, and principal priorities, student needs) and determine how to navigate and respond to them (Desimone & Pak, 2017). In this way, coaches can help create *coherence and alignment* between what teachers are working on with their coach, what teachers have previously worked on, other expectations placed upon them outside of coaching, and teachers' own viewpoints and beliefs (Coburn & Russell, 2008; Coggins et al., 2003; Lowenhaupt et al., 2014; Swinnerton, 2007).

Whether teachers implement and maintain the practices discussed in coaching conversations depends in large part on teachers' *motivation* to participate in coaching and develop a particular area of their practice (Gorozidis & Papaioannou, 2014; Hill et al., 2021; Kennedy, 2016a; Power & Goodnough, 2019). One way that coaches do this is by helping teachers recognize the potential benefits and purposes of a particular instructional goal or strategy (e.g. through explaining the research base, describing benefits for students, or modelling the strategy and asking the teacher to observe the impacts) (Gibbons & Cobb, 2016; Sims & Fletcher-Wood, 2021). Self-Determination Theory (Korthagen, 2017; Ryan & Deci, 2000) highlights several other routes through which coaches may support teacher motivation. First, coaches can help support teachers' feelings of competence by orchestrating mastery experiences, providing encouragement, and drawing teachers' attention to their own professional growth, strengths, and positive impacts on students (Collet, 2012; Knight, 2009; Kurz et al., 2017). Second, coaches can help support teachers' feelings of relatedness by building a strong relationship of mutual respect, trust, and support (Lowenhaupt et al., 2014; Power & Goodnough, 2019; Shernoff et al., 2015). Third, coaches can help support teachers' autonomy by creating opportunities for teachers to express their views and exercise choice and influence over what happens in the coaching session and its implications for their classroom instruction (Kennedy, 2016a; Knight, 2009; Power & Goodnough, 2019).

In addition to supporting teacher motivation and commitment to developmental goals, asking authentic questions that provide opportunities for teachers to express their own views, interpretations, and ideas also serves as a scaffold to support teachers' reflection and problem-solving skills (Collet, 2012; Heineke, 2013; Koh & Neuman, 2006). Coaches can help teachers develop professional expertise and judgment through strategic questioning that requires teachers to *analyze and reflect* on classroom events, their students' responses and needs, their own instruction, and their goals for future lessons, (Barnhart & van Es, 2015; Denton & Hasbrouck, 2009; Hiebert & Morris, 2012; Santagata & Angelici, 2010; Teemant, 2014; Winch et al., 2015).

Some scholars also highlight the coach's role as a source of instructional expertise to scaffold teachers' instructional practice and decision-making. As an expert "other," coaches can provide teachers with valuable feedback, instructional ideas, and support with implementing new ideas based on the coach's assessment of the teacher's and students' needs (Bean et al., 2010; Coburn & Woulfin, 2012; Cohen et al., 2020; Collet, 2012; Heineke, 2013). In providing *feedback* on prior instruction, coaches can help teachers understand their strengths and weaknesses, and identify problems, challenges, and manageable goals for improvement (Blazar & Kraft, 2015; Cassidy et al., 2008; Hunter & Springer, 2022; Kurz et al., 2017; Tung et al., 2004).

Coaches also bring additional *knowledge of content and pedagogy,* which they can use to identify and suggest ideas and strategies that teachers may not have been able to identify on their own (Joyce & Showers, 1981; Knight, 2009; Reddy et al., 2019). Coaches can also demonstrate how they use this knowledge in action by *modelling* their thinking processes and how a particular instructional strategy can be implemented in practice (Joyce & Showers, 1981; Knight, 2009; Kurz et al., 2017; Reddy et al., 2019). Finally, coaches can help teachers successfully enact new practices (Kennedy, 1999) by facilitating opportunities for rehearsals and other kinds of *deliberate practice* (Cohen et al., 2020; Ippolito, 2010; Reddy et al., 2019).

Of course, what coaches do and say during coaching conversations is not the only thing that influences teacher development. Coaches' activities outside of these conversations, including how coaches plan for their interactions with teachers, collaborate with administrators, or analyze teacher instruction and student learning during classroom observations also matter (Bean & DeFord, 2012; Gibbons & Cobb, 2016; Walpole & McKenna, 2012; Woulfin, 2018). Coaches' relationships with the teachers they coach are also influenced by prior experiences, coach reputation, interactions outside of coaching conversations, and perhaps even coach characteristics such as race and gender (Anderson et al., 2014; Blazar et al., 2021; S. Johnson et al., 2018; Tschannen-Moran, 2004). Other literature highlights the role of additional coach characteristics and features of the broader context, including coach expertise, the nature of coaches' job responsibilities, school leadership and culture, and local policy context (Blazar & Kraft, 2015; Booker & Russell, 2022; Deussen et al., 2007; Gallucci et al., 2010; Lowenhaupt et al., 2014; Marsh et al., 2012).

Coaching Discourse Strategies

Despite the substantial attention to how coaching can support teacher development, literature exploring the nature and effects of coaches' discourse strategies is comparatively limited (Gibbons & Cobb, 2016; Kurz et al., 2017; Robertson et al., 2020). While only a handful of studies specifically focus on identifying coaching discourse strategies, descriptions and examples of coach discourse are often included in other coaching literature. Combining these examples into a coherent framework of coaching discourse is not straightforward, however, because of the different approaches used to describe and distinguish between different strategies. Discussions of coaching discourse vary especially along three dimensions: grain size, framing, and normativity. Below, I define each of these features and provide examples of how they play out in the literature.

Grain Size

Grain size refers to the level of specificity and concreteness used to describe the components of coaches' interactions with teachers (Boerst, et al., 2011; Kennedy, 2016b). Existing studies of coaching tend to identify components of a relatively large grain size, offering only glimpses into the concrete details of coaching discourse. Many studies highlight broad conversational activities, (e.g. providing feedback based on an observation, setting goals, modelling, or planning for future instruction), topics (e.g. content-specific instructional strategies, teacher emotions), and goals (e.g. developing a trusting and equal partnership) that can support teachers (Gibbons & Cobb, 2017; Hunt, 2016; Marsh et al., 2015; Matsumura et al., 2013; Teemant, 2014).

What is largely unspecified, however, are the concrete details of what coaches can say and do to achieve these aims. A handful of qualitative studies have begun to investigate features of coach dialogue, such as coach versus teacher talk time, the use of open-ended questions, and patterns of interaction (Collet, 2012; Heineke, 2013; Shernoff et al., 2015). Robertson et al. (2020), for example, highlight three patterns of coach-teacher interaction observed in coaching interactions and associated with teachers' uptake of the instructional ideas discussed. Though some of these studies include some attention to specific kinds of utterances or coaching "moves," such as "affirms an action or statement made by the teacher," (Robertson et al., 2020, p. 412) such granular-level detail is not the primary focus in these studies.

Resources designed by practitioners for practitioner audiences, however, often describe specific coaching moves and provide exemplar coach dialogue to illustrate how coaches can achieve the broader coaching styles these resources aim to inspire. For example, in addition to describing the purpose and idea behind active listening during coaching conversations, Aguilar (2013) explains that one strategy coaches can use to achieve this is to “repeat back or paraphrase what the other person says” (p. 153). Researcher-created handbooks designed to guide coaches in implementing researcher-designed coaching models provide similar concrete suggestions (Knight, 2007; L’hospital et al., 2016). Unfortunately, because of the lack of attention to this small grain size in the research literature, there is little empirical evidence to support the suggestions made in these resources (Cornett & Knight, 2009; Gibbons & Cobb, 2017).

Framing

Framing refers to the extent to which the literature describes observable practitioner actions (i.e. what teachers and coaches do) versus the intended outcomes or purposes behind the actions a practitioner may take (i.e. why coaches and teachers do it) (Forzani, 2014; Kennedy, 2016b). This is often correlated with grain size. For example, the coaching strategy of building a strong coach-teacher relationship both describes an intended outcome and is at a large grain size. At the same time, providing positive praise both describes an observable action and is at a smaller grain size.

This is also evident in the coaching literature, where discussion of more granular coaching discourse strategies tends to highlight observable coaching actions (e.g. suggesting an instructional strategy) rather than goals and purposes (e.g. draw teacher’s attention to a specific instructional strategy) (Heineke, 2013; Hunt, 2016; Robertson et al., 2020). However, many of the discussions of coaching interactions at larger grain sizes also describe *what* coaches do rather than *why* they do it as seen in the discussion of conversational activities and topics highlighted above. Practitioner-facing resources, on the other hand, tend to describe both observable actions and purposes (Aguilar, 2013; L'hospital et al., 2016).

Normativity

Normativity refers to whether the identified components of coaching discourse are intended to reflect *high-quality* coaching practice that coaches *should* use or simply describe observed coaching practice (Goe et al., 2008; Kennedy, 2016b). In general, the coaching literature, practitioner-created resources, and coaching handbooks tend to provide normative guidance for what coaches should do based on an underlying theory of how coaching can influence teacher development (Aguilar, 2013; Desimone & Pak, 2017; Gibbons & Cobb, 2017; Sisson & Sisson, 2017). However, several descriptive studies illustrate how withholding normative judgment enables the development of new, empirically grounded ideas about what may constitute high quality coaching practice. For example, drawing on a descriptive analysis of how coaching practice varied among 20 Reading First coaches, Bean (2010) identifies qualities of coaches' practice that are more and less valued by the teachers with whom they work. Similarly, Robertson et al. (2020) draws on a descriptive analysis of patterns in coach and teacher discourse to identify patterns of interaction that are associated with teacher learning.

Methods

Phase 1

I began by conducting a conceptual review of the literature on coaching using broad search of the Google Scholar (scholar.google.com) database to identify an initial set of empirical and conceptual studies that could provide insight on the nuances of coach-teacher interactions. Search terms included references to coaching (e.g. coaching, instructional coaching, teacher coaching), references to coaching discourse (e.g. discourse, dialogue, coach-teacher interactions), and references to components of coaching (e.g. moves, activities, strategies). In reviewing these initial results, I discarded studies that did not include any discussion of coaching discourse, only keeping studies that included at least one example of coach dialogue or description of coaching conversation content. I also identified additional studies of interest from the citations included in the initial search results. I supplemented my review of the academic literature with an exploration of the limited body of researcher- and practitioner-developed literature that focuses on the nuances of coach dialogue. In this way, I aimed to ensure that the resulting coaching moves taxonomy would be grounded in our existing understanding of coaching practice.

Phase 2

Though the literature provides little attention to concrete coaching moves, many studies I found in my initial search included examples of coach dialogue and other indirect clues about what coaches should say or do. In the second phase, I used this as a starting point for identifying specific discourse moves that coaches may use to achieve the goals discussed in these studies. First, I read and coded the retained studies and practitioner resources to develop an initial list of potential coaching moves. In doing so, I employed an inductive coding approach, developing in vivo codes to describe the moves illustrated in the literature, while preserving each author's language and approach to describing a given move (Miles et al., 2014). This process resulted in a list of hundreds of potential coaching moves. I continued reviewing additional literature and adding to my list of coaching move codes until I ceased identifying distinct moves. This approach is akin to the method of saturation in coding qualitative data (Glaser & Strauss, 1967; Hennink et al., 2017).

Phase 3

In the third phase, I focused on shaping the long list of potential coaching moves into a coherent and well-organized framework. First, I grouped the codes by grain size, framing, and normativity. I then compared the moves in each group to identify where there was conceptual overlap in the moves. Where gaps were present, I created additional moves to ensure that each group included the full breadth of concepts I identified in the literature. In this way, I was able to compare the different approaches to defining coaching moves and think through their affordances and constraints. Ultimately, I selected the approach that would maximize the level of detail, clarity, and practical relevance of the moves, while also ensuring that the moves could be flexibly grouped in many ways to reflect the different coaching models, approaches, and purposes. I discuss the details of and rationale for my chosen approach in the Results section below.

Phase 4

I also engaged in two additional steps to ensure that the framework could describe a wide range of coaching practice and that the definitions and distinctions between moves were clear. First, I returned to the coaching literature, comparing the moves in my taxonomy with literature I had not yet read and revising the moves as needed to accommodate gaps. I continued this process until I stopped identifying additional revisions. Second, I hired four undergraduate students to pilot the framework by applying it as a coding scheme to a random sample of coaching transcripts from a previous study of coaching (see Cohen et al., 2021). In each round of piloting, several coders coded the same set of transcripts, then met to discuss their codes and identify codebook adjustments that would improve the clarity, reliability, and face validity.

After several rounds of piloting and revision, I shared the framework with five well-respected coaching researchers to ensure content validity. These experts offered feedback on ways to improve the clarity of the taxonomy, better distinguish between closely related moves, organize the moves into groups based on conceptual themes, and additional elements of coaching discourse to consider incorporating into the taxonomy. I then engaged in several additional rounds of revision and piloting with undergraduate coders to implement the feedback I received.

Results

Here, I present the final taxonomy of coaching discourse moves. First, I describe the grain size, framing, and normativity of the taxonomy. Then I introduce the 45 moves that make up the taxonomy. Finally, I use stylized vignettes to illustrate how the taxonomy can enhance our understanding of coaching discourse.

Grain Size, Framing, and Normativity

The final taxonomy articulates what coaches do at the highly granular level of "moves," (Boerst, et al., 2011; Heineke, 2013). As I discovered in Phase 3 of my analysis, discourse moves can be defined in many ways. While some descriptions focused on individual coach utterances or turns of talk, others focused on broader sequences of dialogue. There was also variety in the level of detail used to describe their structure, with strategies as general as questions (Aguilar, 2013) and as specific as affirming a teacher's prior instructional decision (Collet, 2012). Finally, while some strategies were defined in terms of their structure and function (e.g. asking for clarification [Robertson et al., 2020] or asking the teacher to justify a claim [Gibbons et al., 2018]), others were defined by the object or subject to which a coach was referring (e.g. asking questions about student thinking [Gibbons et al., 2018]).

In the final taxonomy, coaching moves are defined as individual coach utterances characterized primarily by their structure and function, with limited reference to the objects or subjects included in an utterance. Thus, the move labelled Cause & Effect is defined as “questioning that explicitly asks the teacher to reflect on the effect(s) that stemmed from a particular cause and/or the cause(s) that led to a particular effect” without specifying whether the cause or effect mentioned relates to teacher actions, student actions, or something else. This approach is purposely modelled after prior work on teaching moves, especially moves used to lead classroom discussions (Chapin et al., 2003; O’Connor & Michaels, 2019).

The moves included in the final taxonomy also describe the observable content of coaches’ discourse rather than the purposes such discourse may serve. This approach to framing aims to address the limited attention to *what* coaches can say and do to support teacher development in the existing coaching literature. Defining the Cause & Effect move by purpose, which might sound like “drawing the teacher’s attention to cause and effect relationships,” creates considerable ambiguity as to what a coach should say to enact such a move. Rather than prioritizing coaching purposes and leaving coaches to determine the specific discourse strategies they can use to achieve them, the coaching moves taxonomy instead prioritizes core discourse techniques, leaving coaches to determine the purposes they may serve and how they may be combined to achieve broader goals that facilitate teacher development (Hiebert et al., 2002; Reisman et al., 2019; Winch et al., 2015).

One potential criticism of this approach to framing is that it may reduce coaching practice to a disconnected set of rote discourse techniques that belies its complex, integrated, and context-specific nature and may encourage coaches to apply these techniques in mechanical and potentially inappropriate ways (Forzani, 2014; Grossman & McDonald, 2008; Kennedy, 2016b; O’Connor & Michaels, 2019). I agree that this is a potential danger. For example, borrowing from Aguilar (2013) and the MTP + 4Rs Coaching Handbook (Morningside Center for Teaching Social Responsibility, 2012), I define the mirroring move as “repeating or rephrasing what a teacher has just said”. Repeated, rote use of this move every time a teacher speaks is unlikely to be helpful. Instead, mirroring is one tool, among many, that coaches can draw on to help teachers feel heard, build a strong coach-teacher relationship, and support teacher *motivation* (Aguilar, 2013; Hunt, 2016).

Mirroring may also serve other purposes. Hearing one’s own ideas repeated back may also support teachers with *analyzing and reflecting* on their own beliefs or interpretations of a particular situation (Ippolito, 2010; O’Connor & Michaels, 1993). Of course, not all teachers will benefit from the mirroring move or be spurred to question their own beliefs because of it. This reinforces the contextual nature of coaching where the same moves may be used for different purposes and different moves may be necessary in different contexts to achieve the same purpose (Russell et al., 2020). This does not diminish the value of mirroring as a discourse strategy coaches may use. However, it does make it challenging to define or group moves together by purpose. For this reason, I organize the taxonomy based on the content and structure of the moves. However, I also incorporate attention to purpose by highlighting potential mechanisms (see Figure 1) each move may serve. In this way, the taxonomy recognizes the dynamic nature of coaching discourse, while still providing a common language for describing and operationalizing these details across studies and contexts.

Finally, while I connect each move included in the taxonomy to existing literature on coaching and the mechanisms they may serve, the taxonomy is not designed to provide a normative vision of what high-quality coaching practice should look like. Instead, the taxonomy is purposely designed to provide a descriptive view into the breadth of coaching practice given the lack of prior research focused on the granular details of coaching practice and its effects on teachers.

The Coaching Moves

Moves are grouped into six larger categories (Figure 2) to reflect structural distinctions between them, with 5-10 moves per group. In Tables 1-7, I list the corresponding moves, their definitions, exemplar coach dialogue to illustrate each move in action, the potential purposes they may serve, and the supporting literature for each group.

I first distinguish between *asking* moves, in which coaches pose open-ended questions that may prompt teacher reflection, analysis, and sense-making (Collet, 2012; Desimone & Pak, 2017; Heineke, 2013; Koh & Neuman, 2006; Shernoff et al., 2015), and *telling* moves where the coach provides the teacher with information and more directive feedback (Bean et al., 2010; Coburn & Woulfin, 2012; Ippolito, 2010; Kurz et al., 2017; Reddy et al., 2019). I also distinguish between *backward-facing* moves, which focus on processing and providing feedback on what has previously occurred, and *forward-facing moves*, which focus on planning for future lessons and changes to instruction (Hattie & Timperley, 2007; Sisson & Sisson, 2017).

The first four groups consist of all possible combinations of these labels. Group 1 (Table 1) consists of moves that are *asking* and *backward-facing* (a.k.a. AskBack moves). By virtue of their structure as questions, I hypothesize that these moves may support teacher analysis and reflection and embody active learning principles. In focusing on prior instruction, these moves help ensure that coaching conversations are also job-embedded. Specific moves may also serve additional purposes, as noted in Table 1. Group 2 (Table 2) consists of moves that are *telling* and *backward-facing* (a.k.a. TellBack moves). By virtue of their structure as more directive statements, I hypothesize that these moves may serve as important sources of feedback and may also serve to scaffold teachers' analysis and reflection, among other move-specific purposes. As with the first group, these moves also contribute to ensuring that coaching conversations are job-embedded. Group 3 (Table 3) consists of moves that are *asking* and *forward-facing* (a.k.a AskForward moves) and may provide a job-embedded, active learning opportunity while also supporting teacher analysis and reflection. Finally, Group 4 (Table 4) consists of moves that are *telling* and *backward-facing* (TellBack moves), which may allow coaches to communicate feedback and share their knowledge of content and pedagogy.

The remaining two groups focus on moves that fall outside of the four groups above, but may nonetheless be used by coaches for important purposes. The fifth group (Table 5) consists of moves in which the coach facilitates a structured *activity*, such as analyzing study data or reviewing curricular materials with the teacher. In addition to providing an opportunity for active learning, each activity may address other coaching mechanisms as illustrated in Table 5. The final group (Table 6) consists of moves that may not directly support teachers' instruction or professional knowledge but may promote a stronger coach-teacher relationship and teacher motivation more broadly through building *rapport* (Knight, 2009; Lowenhaupt et al., 2014; Power & Goodnough, 2019; Shernoff et al., 2015).

Applying the Taxonomy

There are two main ways scholars and practitioners might use the taxonomy in practice. First, it might be used retrospectively to analyze coaching discourse. Second, it might serve as a prospective tool for planning coaching conversations or articulating the components of a particular approach to coaching. Below, I illustrate both uses through stylized vignettes inspired by the coaching literature.

Coaching Moves in Research

Rebecca is an education researcher whose work focuses on coaching. Recently, she's become especially interested how coaches discuss their observations of teachers' lessons during coaching conversations. In partnership with a local school district, Rebecca identifies several experienced coaches who are regarded as highly effective and obtains consent to record several coaching sessions from each coach. Rebecca also identifies several less experienced coaches whose coaching skills are still developing. After transcribing these recordings, Rebecca uses the backward facing moves (Tables 1 and 2) as a coding scheme to code the transcriptions.

In analyzing the coded data, Rebecca notices an interesting pattern in the moves used by the less experienced coaches, as compared with the more experienced coaches. First, Rebecca notices that more experienced coaches often described their observations of the connection between a particular cause and effect (TellBack: connection) and/or asked teachers to reflect on the link between a particular cause and effect in the lesson (AskBack: cause and effect). Less experienced coaches, on the other hand, rarely used these moves drawing teachers' attention to the causal link between events. Less experienced coaches also tended to use a series of asking and backward facing moves to open the conversation about the teacher's previous lesson and then shift to using a series of telling and backward facing moves. For example, the coach might begin by asking the teacher to reflect on the success of their lesson (AskBack: self-assessment), then ask the teacher to justify their reflections (AskBack: justification) or ask the teacher to recall a particular moment in the lesson (AskBack: noticing). Then, the coach might transition to explaining their understanding of the lesson by describing what they observed (TellBack: observation), providing positive praise (TellBack: positive evaluation), and identifying a moment in the lesson or element of the teacher's instruction that was less successful (TellBack: observation, negative evaluation). More experienced coaches, on the other hand, tended to intersperse both asking and telling moves throughout the conversation, asking a question about the previous lesson, and responding with their own observations and interpretations before moving onto a second question.

In follow-up interviews with the coaches, Rebecca learns that all coaches were cognizant of following the school's provided coaching protocol, but that more experienced coaches had tweaked their use of the protocol over time, as they observed how teachers reacted to different moves. More experienced coaches typically described the desire to ensure that coaching conversations felt like a lively dialogue with the teacher as the key reason for interspersing asking and telling moves.

Interested in understanding how these different discourse patterns affect teacher development, Rebecca designs two follow-up experiments. In the first, half of the coaches are told to intersperse asking and telling moves and the other half are told to first use a series of asking moves and then shift to a series of telling moves. In the second, half of the coaches are told to make sure to use the cause & effect and connection moves, and the other half are told to avoid those moves. For each experiment, Rebecca compares teacher observation scores across the two coach groups. Rebecca also uses the coaching moves taxonomy to code transcripts from a subset of the conversations conducted as part of the experiment to confirm that coaches complied with their assigned protocol.

Coaching Moves in Practice

Lacy is a full-time middle school literacy coach. Recently, she has noticed that one teacher she works with, Sarah, has seemed resistant during coaching conversations (Jacobs et al., 2018). When Lacy identifies an instructional challenge or suggests something she can change in the next lesson, Sarah tends to push back, offering alternative interpretations of the instructional challenge and offering reasons why Lacy's suggestions won't work (Ippolito, 2010). The conversations always seem to end with Lacy imploring Sarah to at least “try out” what she suggested and Sarah reluctantly agreeing. Lacy knows that little progress will be made if Sarah and Lacy can’t establish agreed-upon goals for instructional improvement (Kochmanski, 2020), but she’s not sure how to establish those goals with Sarah. At her next meeting with the principal, Lacy describes this challenge and asks for advice. The principal introduces Lacy to a toolkit of coaching moves and asks Lacy to spend some time reflecting on what kinds of moves she uses with Sarah and then pick a few new moves to try out.

As Lacy looks through the different moves and thinks about her previous conversations with Sarah, she realizes that she primarily uses TellBack and TellForward moves, providing few opportunities for Sarah to express and process her own ideas. Lacy often begins the conversation by praising something about Sarah’s lesson (TellBack: positive evaluation, observation) and then describing a moment in the lesson where Sarah or her students encountered a challenge (TellBack: observation, interpretation, cause and effect, negative evaluation). Then, Lacy usually shifts to suggesting a change that Sarah can make to prevent this challenge in future lessons (TellForward: instructional strategy) and explaining why and how it will work (TellForward: student goal, demonstration, challenge). Lacy wonders if it would help to start by soliciting Sarah’s views about her instructional challenges and how they might be addressed before offering her own ideas and suggestions. Looking at the AskForward (Table 3) and AskBack moves (Table 1) from the toolkit, Lacy decides to start the conversation using the AskBack: self-assessment move to ask Sarah to provide her own views about the key instructional challenges she’s facing before Lacy provides any feedback of her own. Lacy also decides to try eliciting Sarah’s views about her goals for improvement (AskForward: goal-setting) and offering Sarah an opportunity to share her own ideas about what specific strategies will help her reach these goals (AskForward: generation).

Discussion

Though coaching programs have demonstrated effects on teachers' instruction and student learning (Kraft et al., 2018), they require a cadre of highly skilled coaches who can meet regularly with teachers. This makes coaching logistically complex and resource intensive, especially compared to more traditional forms of professional development (D. Knight, 2012). We need to provide coaches with a concrete understanding of effective coaching strategies to ensure that this commitment of resources will make a difference for students. This paper provides a key tool for addressing this challenge. To my knowledge, this is the first framework of coaching discourse that is applicable across coaching models and approaches, provides concrete and clear explanations of how coaches can use questions and feedback to support teacher development, and is grounded in the available empirical research.

In serving as a coding scheme for analyzing coaching dialogue, the taxonomy can support researchers in answering important qualitative and quantitative questions about coaching. For example, work is currently underway in collaboration with a methodologist to code coaching conversations, quantify variation in coaching discourse, and identify moves that predict improvements in teacher practice. We also plan to develop an automated approach for identifying moves in transcripts that will reduce coding costs and increase efficiency. Because the moves are of a small grain size and defined by low inference structural features, they are likely easier to automate than high-inference frameworks like teacher observation rubrics.

In providing a quantitative method of describing coaching discourse at scale, the coaching moves taxonomy will allow researchers to systematically investigate both the causes and effects of coaches' discourse strategies to answer questions such as: what discourse strategies help facilitate improvements in teacher instruction? What supports help coaches learn to skillfully use evidence-based discourse practices? And what hiring processes help administrators select skilled coaches? Furthermore, as more studies using the coaching moves taxonomy are conducted, researchers will be able to aggregate findings through conceptual reviews and quantitative meta-analyses with relative ease. Finally, researchers may also qualitatively explore how and why coaches use specific moves and how teachers perceive them.

The taxonomy also provides a key tool for supporting coaches in their daily work. For coaches, the taxonomy may serve as a valuable framework for guiding professional practice. It can serve as a technical language for reflecting on patterns in their current practice, identifying ways to improve their practice (e.g. by trying out new moves or altering the order in which moves are used), and planning for future coaching conversations (Lofthouse & Hall, 2014). Those who support and develop coaches may also use the taxonomy to develop their own curricula for supporting coaching practice.

Finally, the taxonomy can help support coaches with incorporating existing and future research insights about the features of high-quality coaching into their daily practice. It is only when practitioners can understand the concrete implications of research for their daily practice that research can even begin to have an impact. Creating this understanding is infinitely easier when researchers and practitioners use the same language to describe what coaches do and say in their interactions with teachers. The coaching moves taxonomy can provide this shared language. In future work, I plan to share the taxonomy with coaching practitioners and empirically explore its affordances and constraints.

Of course, the taxonomy does not capture every variation or characteristic of coaching practice that may influence teacher learning and development. Future work can move beyond the frequency and patterns with which moves are used to understand how the quality of these moves may vary across contexts. Additional frameworks can also be created to capture additional elements of coaching practice, including tone-of-voice or coach planning. Finally, as our understanding of coaching practice and its effects on teachers develops, we may ultimately be able to create complex multi-layered frameworks that provide a vision of high-quality coaching practice and articulate the purposes behind different techniques (Boerst, et al., 2011; Grossman & Dean, 2019; Kennedy, 2016b).

References

Ackland, R. (1991). A review of the peer coaching literature. *Journal of Staff Development*, *12*(1), 22–27.

Aguilar, E. (2013). *The art of coaching: Effective strategies for school transformation*. Jossey-Bass, A Wiley Brand.

Alston, C. L., Danielson, K. A., Dutro, E., & Cartun, A. (2018). Does a Discussion by Any Other Name Sound the Same? Teaching Discussion in Three ELA Methods Courses. *Journal of Teacher Education*, *69*(3), 225–238.

Anderson, R., Feldman, S., & Minstrell, J. (2014). Understanding relationship: Maximizing the effects of science coaching. *Education Policy Analysis Archives*, *22*, 54–54.

Barnhart, T., & van Es, E. (2015). Studying teacher noticing: Examining the relationship among pre-service science teachers' ability to attend, analyze and respond to student thinking. *Teaching and Teacher Education*, *45*, 83–93.

Bean, R., & DeFord, D. (2012). Do's and Don'ts for Literacy Coaches: Advice from the Field. *Literacy Coaching Clearinghouse*.

Bean, R., Draper, J. A., Hall, V., Vandermolen, J., & Zigmond, N. (2010). Coaches and Coaching in Reading First Schools: A Reality Check. *The Elementary School Journal*, *111*(1), 87–114.

Blazar, D., & Kraft, M. A. (2015). Exploring Mechanisms of Effective Teacher Coaching: A Tale of Two Cohorts From a Randomized Experiment. *Educational Evaluation and Policy Analysis*, *37*(4), 542–566.

Blazar, D., McNamara, D., & Blue, G. (2021). *Instructional Coaching Personnel and Program Scalability*. EdWorkingPaper No. 21-499. Annenberg Institute at Brown University.

Boerst, T., Sleep, L., Ball, D., & Bass, H. (2011). Preparing teachers to lead mathematics discussions. *Teachers College Record*, *113*(12), 2844–2877.

Booker, L. N., & Russell, J. L. (2022). *Improving teaching practice with instructional coaching.* Design Principles Series, EdResearch for Recovery.

Cassidy, T., Mallett, C., & Tinning, R. (2008). Considering conceptual orientations of coach education research: A tentative mapping. *International Journal of Coaching Science*, *2*(2), 43–58.

Chapin, S. H., O'Connor, M. C., & Anderson, N. C. (2003). *Classroom discussions: Using math talk to help students learn, grades 1-6*. Math Solutions Publications.

Charalambous, C. Y., & Praetorius, A.-K. (2020). Creating a forum for researching teaching and its quality more synergistically. *Studies in Educational Evaluation*, *67*, 100894.

Clark, S. K., & Byrnes, D. (2012). Through the eyes of the novice teacher: Perceptions of mentoring support. *Teacher Development*, *16*(1), 43–54.

Coburn, C. E., & Russell, J. L. (2008). District Policy and Teachers' Social Networks. *Educational Evaluation and Policy Analysis*, *30*(3), 203–235.

Coburn, C. E., & Woulfin, S. L. (2012). Reading Coaches and the Relationship Between Policy and Practice. *Reading Research Quarterly*, *47*(1), 5–30.

Coggins, C. T., Stoddard, P., & Cutler, E. (April 21-25, 2003). *Improving Instructional Capacity through School-Based Reform Coaches.* Paper presented at the Annual Meeting of the American Educational Research Association. Chicago, IL.

Cohen, J. (2015). Challenges in identifying high-leverage practices. *Teachers College Record*, *117*(7).

Cohen, J., Krishnamachari, A., & Wong, V. C. (2021). *Experimental Evidence on the Robustness of Coaching Supports in Teacher Education*.

Cohen, J., Wong, V., Krishnamachari, A., & Berlin, R. (2020). Teacher coaching in a simulated environment. *Educational Evaluation and Policy Analysis*, *42*(2), 208–231.

Collet, V. S. (2012). The Gradual Increase of Responsibility Model: Coaching for Teacher Change. *Literacy Research and Instruction*, *51*(1), 27–47.

Cornett, J., & Knight, J. (2009). Research on coaching. *Coaching: Approaches and Perspectives*, 192–216.

Costa, A. L., & Garmston, R. J. (2002). *Cognitive coaching: A foundation for renaissance schools*. ERIC.

Croft, A., Coggshall, J. G., Dolan, M., & Powers, E. (2010). *Job-embedded professional development: What it is, who is responsible, and how to get it done well*. Issue Brief. National Comprehensive Center for Teacher Quality.

Davis, B., & Higdon, K. (2008). The Effects of Mentoring/Induction Support on Beginning Teachers' Practices in Early Elementary Classrooms (K-3). *Journal of Research in Childhood Education*, *22*(3), 261–274.

Demszky, D., & Hill, H. (2022). The NCTE Transcripts: A Dataset of Elementary Math Classroom Transcripts. ArXiv Preprint, *2211.11772*.

Denton, C. A., & Hasbrouck, J. (2009). A description of instructional coaching and its relationship to consultation. *Journal of Educational and Psychological Consultation*, *19*(2), 150–175.

Desimone, L. M., & Pak, K. (2017). Instructional Coaching as High-Quality Professional Development. *Theory Into Practice*, *56*(1), 3–12.

Deussen, T., Coskie, T., Robinson, L., & Autio, E. (2007). *"Coach" Can Mean Many Things: Five Categories of Literacy Coaches in Reading First.* Issues & Answers. REL 2007-No. 005. Regional Educational Laboratory Northwest.

Domina, T., Lewis, R., Agarwal, P., & Hanselman, P. (2015). Professional Sense-Makers: Instructional Specialists in Contemporary Schooling. *Educational Researcher*, *44*(6), 359–364.

Donaldson, M. L., & Woulfin, S. (2018). From tinkering to going "rogue": How principals use agency when enacting new teacher evaluation systems. *Educational Evaluation and Policy Analysis*, *40*(4), 531–556.

Donegan, M. M., Ostrosky, M. M., & Fowler, S. A. (2000). Peer coaching: Teachers supporting teachers. *Young Exceptional Children*, *3*(3), 9–16.

Downer, J. T., Williford, A. P., Bulotsky-Shearer, R. J., Vitiello, V. E., Bouza, J., Reilly, S., & Lhospital, A. (2018). Using data-driven, video-based early childhood consultation with teachers to reduce children's challenging behaviors and improve engagement in preschool classrooms. *School Mental Health*, *10*(3), 226–242.

Dozier, C. (2006). *Responsive literacy coaching: Tools for creating and sustaining purposeful change*. Stenhouse Publishers.

Forzani, F. M. (2014). Understanding "Core Practices" and "Practice-Based" Teacher Education: Learning From the Past. *Journal of Teacher Education*, *65*(4), 357–368.

Galey, S. (2016). The evolving role of instructional coaches in US policy contexts. *The William & Mary Educational Review*, *4*(2), 11.

Gallucci, C., Van Lare, M. D., Yoon, I. H., & Boatright, B. (2010). Instructional Coaching: Building Theory About the Role and Organizational Support for Professional Learning. *American Educational Research Journal*, *47*(4), 919–963.

Gibbons, L. K., & Cobb, P. (2016). Content-Focused Coaching: Five Key Practices. *The Elementary School Journal*, *117*(2), 237–260.

Gibbons, L. K., & Cobb, P. (2017). Focusing on Teacher Learning Opportunities to Identify Potentially Productive Coaching Activities. *Journal of Teacher Education*, *68*(4), 411–425.

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Aldine Publishing.

Goe, L., Bell, C., & Little, O. (2008). Approaches to Evaluating Teacher Effectiveness: A Research Synthesis. In *National Comprehensive Center for Teacher Quality*. National Comprehensive Center for Teacher Quality. https://eric.ed.gov/?id=ED521228

Gorozidis, G., & Papaioannou, A. G. (2014). Teachers' motivation to participate in training and to implement innovations. *Teaching and Teacher Education*, *39*, 1–11.

Gregory, A., Ruzek, E., Hafen, C. , Mikami, A., Allen, J., & Pianta, R. (2017). My Teaching Partner-Secondary: A video-based coaching model. *Theory into Practice*, *56*(1), 38–45.

Gross, P. A. (2010). Not Another Trend: Secondary-Level Literacy Coaching. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *83*(4), 133–137.

Grossman, P., & Dean, C. G. (2019). Negotiating a common language and shared understanding about core practices: The case of discussion. *Teaching and Teacher Education*, *80*, 157–166.

Grossman, P., Hammerness, K., & McDonald, M. (2009). Redefining teaching, re-imagining teacher education. *Teachers and Teaching*, *15*(2), 273–289.

Grossman, P., & McDonald, M. (2008). Back to the Future: Directions for Research in Teaching and Teacher Education. *American Educational Research Journal*, *45*(1), 184–205.

Guba, E. G., & Lincoln, Y. S. (1994). Competing paradigms in qualitative research. *Handbook of Qualitative Research*, *2*(163–194), 105.

Guskey, T. R. (1986). Staff development and the process of teacher change. *Educational Researcher*, *15*(5), 5–12.

Hardt, D., Nagler, M., & Rincke, J. (2020). *Can peer mentoring improve online teaching effectiveness? An rct during the covid-19 pandemic*.

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

Heineke, S. F. (2013). Coaching Discourse: Supporting Teachers' Professional Learning. *The Elementary School Journal*, *113*(3), 409–433.

Hennink, M. M., Kaiser, B. N., & Marconi, V. C. (2017). Code saturation versus meaning saturation: How many interviews are enough? *Qualitative Health Research*, *27*(4), 591–608.

Hiebert, J., Gallimore, R., & Stigler, J. W. (2002). A knowledge base for the teaching profession: What would it look like and how can we get one? *Educational Researcher*, *31*(5), 3–15.

Hiebert, J., & Morris, A. K. (2012). Teaching, Rather Than Teachers, As a Path Toward Improving Classroom Instruction. *Journal of Teacher Education*, *63*(2), 92–102.

Hiebert, J., Morris, A. K., Berk, D., & Jansen, A. (2007). Preparing teachers to learn from teaching. *Journal of Teacher Education*, *58*(1), 47–61.

Hill, H. C., Papay, J. P., Schwartz, N., Johnson, S., Freitag, E., Donohue, K., Berry III, R. Q., Loeb, S., Anderson, M., & Baker, M. (2021). *Improving Teacher Professional Learning at Scale*.

Hoffman, J. V., Wetzel, M. M., Maloch, B., Greeter, E., Taylor, L., DeJulio, S., & Vlach, S. K. (2015). What can we learn from studying the coaching interactions between cooperating teachers and preservice teachers? A literature review. *Teaching and Teacher Education*, *52*, 99–112.

Hunt, C. S. (2016). Getting to the heart of the matter: Discursive negotiations of emotions within literacy coaching interactions. *Teaching and Teacher Education*, *60*, 331–343.

Hunter, S. B., & Springer, M. G. (2022). Critical Feedback Characteristics, Teacher Human Capital, and Early-Career Teacher Performance: A Mixed-Methods Analysis. *Educational Evaluation and Policy Analysis*, 01623737211062913.

Ippolito, J. (2010). Three Ways That Literacy Coaches Balance Responsive and Directive Relationships with Teachers. *The Elementary School Journal*, *111*(1), 164–190.

Jacobs, J., Boardman, A., Potvin, A., & Wang, C. (2018). Understanding teacher resistance to instructional coaching. *Professional Development in Education*, *44*(5), 690–703.

Jewett, P., & MacPhee, D. (2012). Adding collaborative peer coaching to our teaching identities. *The Reading Teacher*, *66*(2), 105–110.

Johnson, K. (2016). Instructional Coaching Implementation: Considerations for K-12 Administrators. *Journal of School Administration Research and Development*, *1*(2), 37–40.

Johnson, S., Pas, E. T., Bradshaw, C. P., & Ialongo, N. S. (2018). Promoting Teachers' Implementation of Classroom-Based Prevention Programming Through Coaching: The

Mediating Role of the Coach–Teacher Relationship. *Administration and Policy in Mental Health and Mental Health Services Research*, *45*(3), 404–416.

Joyce, B. R., & Showers, B. (1981). Transfer of training: The contribution of "coaching." *Journal of Education*, *163*(2), 163–172.

Kane, B. D., & Rosenquist, B. (2019). Relationships between instructional coaches' time use and district-and school-level policies and expectations. *American Educational Research Journal*, *56*(5), 1718–1768.

Kennedy, M. (1999). The role of preservice teacher education. *Teaching as the Learning Profession: Handbook of Policy and Practice*, 54–85.

Kennedy, M. (2016a). How does professional development improve teaching? *Review of Educational Research*, *86*(4), 945–980.

Kennedy, M. (2016b). Parsing the Practice of Teaching. *Journal of Teacher Education*, *67*(1), 6–17.

King, D., Neuman, M., Pelchat, J., Potochnik, T., Rao, S., & Thompson, J. (2004). *Instructional Coaching: Professional Development Strategies that Improve Instruction*. Annenberg Institute at Brown University.

Kloser, M., Wilsey, M., Madkins, T. C., & Windschitl, M. (2019). Connecting the dots: Secondary science teacher candidates' uptake of the core practice of facilitating sensemaking discussions from teacher education experiences. *Teaching and Teacher Education*, *80*, 115–127.

Knight, D. (2012). Assessing the cost of instructional coaching. *Journal of Education Finance*, 52–80.

Knight, J. (2007). *Instructional coaching: A partnership approach to improving instruction*. NSDC : Corwin Press.

Knight, J. (2009). *Coaching: Approaches and perspectives*. Corwin Press.

Knight, J. (2019). Instructional Coaching for Implementing Visible Learning: A Model for Translating Research into Practice. *Education Sciences*, *9*(2), 101.

Knight, J., & van Nieuwerburgh, C. (2012). Instructional coaching: A focus on practice. *Coaching: An International Journal of Theory, Research and Practice*, *5*(2), 100–112.

Kochmanski, N. (2020). *Aspects of High-Quality Mathematics Coaching: What Coaches Need to Know and Be Able to Do to Support Individual Teachers' Learning* [PhD Thesis].

Koh, S., & Neuman, S. B. (2006). Exemplary elements of coaching. *Unpublished Manuscript, University of Michigan, Ann Arbor*.

Korthagen, F. (2017). Inconvenient truths about teacher learning: Towards professional development 3.0. *Teachers and Teaching*, *23*(4), 387–405.

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.

Kurz, A., Reddy, L. A., & Glover, T. A. (2017). A multidisciplinary framework of instructional coaching. *Theory into Practice*, *56*(1), 66–77.

Kutsyuruba, B., & Godden, L. (2019). The role of mentoring and coaching as a means of supporting the well-being of educators and students. *International Journal of Mentoring and Coaching in Education*, *8*(4), 229–234.

L'Allier, S., Elish-Piper, L., & Bean, R. M. (2010). What Matters for Elementary Literacy Coaching? Guiding Principles for Instructional Improvement and Student Achievement. *The Reading Teacher*, *63*(7), 544–554.

Lampert, M., & Graziani, F. (2009). Instructional Activities as a Tool for Teachers' and Teacher Educators' Learning. *The Elementary School Journal*, *109*(5), 491–509.

Lancer, N., Clutterbuck, D., & Megginson, D. (2016). *Techniques for coaching and mentoring*. Routledge.

L'hospital, A., Matthew, K., Downer, J. T., Williford, A., Weaver, W., Villanueva, D. R., & Ampudia, S. (2016). *Learning to Objectively Observe Kids Consultancy Manual*.

Lieberman, A. (1995). Practices that support teacher development: Transforming conceptions of professional learning. In F. Stevens (Ed.), *Innovating and Evaluating Science Education: NSF Evaluation Forums, 1992-94.* (pp. 67–78). Westat, Inc.

Lochmiller, C. R. (2021). Guest editorial: Coaching for improvement in education: new insights and enduring questions. *International Journal of Mentoring and Coaching in Education*.

Lofthouse, R., & Hall, E. (2014). Developing practices in teachers' professional dialogue in England: Using Coaching Dimensions as an epistemic tool. *Professional Development in Education*, *40*(5), 758–778.

Lowenhaupt, R., McKinney, S., & Reeves, T. (2014). Coaching in context: The role of relationships in the work of three literacy coaches. *Professional Development in Education*, *40*(5), 740–757.

Marsh, J. A., Bertrand, M., & Huguet, A. (2015). Using Data to Alter Instructional Practice: The Mediating Role of Coaches and Professional Learning Communities. *Teachers College Record*, 40.

Marsh, J. A., McCombs, J. S., & Martorell, F. (2012). Reading coach quality: Findings from Florida middle schools. *Literacy Research and Instruction*, *51*(1), 1–26.

Marsh, J. A., Sloan McCombs, J., & Martorell, F. (2010). How Instructional Coaches Support Data-Driven Decision Making: Policy Implementation and Effects in Florida Middle Schools. *Educational Policy*, *24*(6), 872–907.

Matsko, K. K., Ronfeldt, M., Nolan, H. G., Klugman, J., Reininger, M., & Brockman, S. L. (2020). Cooperating teacher as model and coach: What leads to student teachers' perceptions of preparedness? *Journal of Teacher Education*, *71*(1), 41–62.

Matsumura, L. C., Garnier, H. E., & Spybrook, J. (2013). Literacy coaching to improve student reading achievement: A multi-level mediation model. *Learning and Instruction*, *25*, 35–48.

McDonald, M., Kazemi, E., & Kavanagh, S. S. (2013). Core practices and pedagogies of teacher education: A call for a common language and collective activity. *Journal of Teacher Education*, *64*(5), 378–386.

McGatha, M. B. (2017). Elementary Mathematics Specialists: Ensuring the Intersection of Research and Practice. *North American Chapter of the International Group for the Psychology of Mathematics Education*.

Means, B., Padilla, C., & Gallagher, L. (2010). Use of education data at the local level: From accountability to instructional improvement. *US Department of Education*.

Mena, J., García, M., Clarke, A., & Barkatsas, A. (2016). An analysis of three different approaches to student teacher mentoring and their impact on knowledge generation in practicum settings. *European Journal of Teacher Education*, *39*(1), 53–76.

Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative data analysis: A method sourcebook*. Sage Publications.

Morningside Center for Teaching Social Responsibility. (2012). *4Rs + MTP Coaching Handbook for Staff Developers*. University of Virginia and Fordham University.

Neufeld, B., & Roper, D. (2003). *Coaching: A strategy for developing instructional capacity*.

Niemi, H., Nevgi, A., & Aksit, F. (2016). Active learning promoting student teachers' professional competences in Finland and Turkey. *European Journal of Teacher Education*, *39*(4), 471–490.

Obara, S. (2010). Mathematics coaching: A new kind of professional development. *Teacher Development*, *14*(2), 241–251.

O'Connor, C., & Michaels, S. (2019). Supporting teachers in taking up productive talk moves: The long road to professional learning at scale. *International Journal of Educational Research*, *97*, 166–175.

O'Connor, M. C., & Michaels, S. (1993). Aligning Academic Task and Participation Status through Revoicing: Analysis of a Classroom Discourse Strategy. *Anthropology & Education Quarterly*, *24*(4), 318–335.

Perkins, S. J. (1998). On becoming a peer coach: Practices, identities, and beliefs of inexperienced coaches. *Journal of Curriculum and Supervision*, *13*(3), 235–254.

Power, K., & Goodnough, K. (2019). Fostering teachers' autonomous motivation during professional learning: A self-determination theory perspective. *Teaching Education*, *30*(3), 278–298.

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher*, *29*(1), 4–15.

Reddy, L. A., Glover, T., Kurz, A., & Elliott, S. N. (2019). Assessing the effectiveness and interactions of instructional coaches: Initial psychometric evidence for the instructional coaching assessments–teacher forms. *Assessment for Effective Intervention*, *44*(2), 104–119.

Reinke, W., Lewis-Palmer, T., & Merrell, K. (2008). The classroom check-up: A classwide teacher consultation model for increasing praise and decreasing disruptive behavior. *School Psychology Review*, *37*(3), 315–332.

Reinke, W., Sprick, R., & Knight, J. (2009). Coaching classroom management. In Knight (Ed.), *Coaching Approaches & Perspectives* (pp. 91–112). Corwin Press.

Reisman, A., Cipparone, P., Jay, L., Monte-Sano, C., Schneider Kavanagh, S., McGrew, S., & Fogo, B. (2019). Evidence of emergent practice: Teacher candidates facilitating historical discussions in their field placements. *Teaching and Teacher Education*, *80*, 145–156.

Reiss, K. (2009). Leadership coaching. In Knight (Ed.), *Coaching Approaches & Perspectives* (pp. 166–191). Corwin Press.

Robertson, D. A., Ford-Connors, E., Frahm, T., Bock, K., & Paratore, J. R. (2020). Unpacking productive coaching interactions: Identifying coaching approaches that support instructional uptake. *Professional Development in Education*, *46*(3), 405–423.

Ronfeldt, M., Bardelli, E., Truwit, M., Mullman, H., Schaaf, K., & Baker, J. C. (2020). Improving Preservice Teachers' Feelings of Preparedness to Teach Through Recruitment of Instructionally Effective and Experienced Cooperating Teachers: A Randomized Experiment. *Educational Evaluation and Policy Analysis*, *42*(4), 551–575.

Ronfeldt, M., Brockman, S. L., & Campbell, S. L. (2018). Does cooperating teachers' instructional effectiveness improve preservice teachers' future performance? *Educational Researcher*, *47*(7), 405–418.

Ronfeldt, M., & Reininger, M. (2012). More or better student teaching? *Teaching and Teacher Education*, *28*(8), 1091–1106.

Rowe, M. B. (1986). Wait time: Slowing down may be a way of speeding up! *Journal of Teacher Education*, *37*(1), 43–50.

Russell, J. L., Correnti, R., Stein, M. K., Bill, V., Hannan, M., Schwartz, N., Booker, L. N., Pratt, N. R., & Matthis, C. (2020). Learning From Adaptation to Support Instructional Improvement at Scale: Understanding Coach Adaptation in the TN Mathematics Coaching Project. *American Educational Research Journal*, *57*(1), 148–187.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68.

Santagata, R., & Angelici, G. (2010). Studying the impact of the lesson analysis framework on preservice teachers' abilities to reflect on videos of classroom teaching. *Journal of Teacher Education*, *61*(4), 339–349.

Shernoff, E. S., Lakind, D., Frazier, S. L., & Jakobsons, L. (2015). Coaching Early Career Teachers in Urban Elementary Schools: A Mixed-Method Study. *School Mental Health*, *7*(1), 6–20.

Sims, S., & Fletcher-Wood, H. (2021). Identifying the characteristics of effective teacher professional development: A critical review. *School Effectiveness and School Improvement*, *32*(1), 47–63.

Sisson, D., & Sisson, B. (2017). *The literacy coaching handbook: Working with teachers to increase student achievement*. Routledge, Taylor & Francis Group.

Stanulis, R., & Floden, R. E. (2009). Intensive Mentoring as a Way to Help Beginning Teachers Develop Balanced Instruction. *Journal of Teacher Education*, *60*(2), 112–122.

Stanulis, R. N. (1994). Fading to a whisper: One mentor's story of sharing her wisdom without telling answers. *Journal of Teacher Education*, *45*(1), 31–38.

Swinnerton, J. (2007). Brokers and boundary crossers in an urban school district: Understanding central-office coaches as instructional leaders. *Journal of School Leadership*, *17*(2), 195–221.

Teemant, A. (2014). A Mixed-Methods Investigation of Instructional Coaching for Teachers of Diverse Learners. *Urban Education*, *49*(5), 574–604.

Terehoff, I. I. (2002). Elements of adult learning in teacher professional development. *NASSP Bulletin*, *86*(632), 65–77.

Tobin, K. (1987). The role of wait time in higher cognitive level learning. *Review of Educational Research*, *57*(1), 69–95.

Toll, C. (2009). Literacy Coaching. In Knight (Ed.), *Coaching Approaches & Perspectives* (pp. 56–69). Corwin Press.

Tschannen-Moran, M. (2004). *Trust Matters: Leadership for successful schools*. Jossey Bass.

Tung, R., Ouimette, M., & Feldman, J. (2004). The Challenge of Coaching: Providing Cohesion among Multiple Reform Agendas. *Center for Collaborative Education*.

van der Linden, S., van der Meij, J., & McKenney, S. (2021). Teacher Video Coaching, From Design Features to Student Impacts: A Systematic Literature Review. *Review of Educational Research*, 92(1), 114-165.

Vanderburg, M., & Stephens, D. (2009). What Teachers Say They Changed Because of Their Coach and How They Think Their Coach Helped Them. *Literacy Coaching Clearinghouse*.

Vanderburg, M., & Stephens, D. (2010). The impact of literacy coaches: What teachers value and how teachers change. *The Elementary School Journal*, *111*(1), 141–163.

Walpole, S., & McKenna, M. C. (2012). *The literacy coach's handbook: A guide to research-based practice*. Guilford Press.

Walpole, S., McKenna, M. C., Uribe-Zarain, X., & Lamitina, D. (2010). The relationships between coaching and instruction in the primary grades: Evidence from high-poverty schools. *The Elementary School Journal*, *111*(1), 115–140.

West, L. (2009). Content coaching: Transforming the teaching profession. In Knight (Ed.), *Coaching Approaches & Perspectives* (pp. 113–144). Corwin Press.

Winch, C., Oancea, A., & Orchard, J. (2015). The contribution of educational research to teachers' professional learning: Philosophical understandings. *Oxford Review of Education*, *41*(2), 202–216.

Woulfin, S. L. (2018). Mediating Instructional Reform: An Examination of the Relationship Between District Policy and Instructional Coaching. *AERA Open*, *4*(3), 1-16.

Woulfin, S. L., & Rigby, J. G. (2017). Coaching for Coherence: How Instructional Coaches Lead Change in the Evaluation Era. *Educational Researcher*, *46*(6), 323–328.

Tables

Table 1. Asking & Backward-Facing (AskBack) Moves (Group 1)

Move	Definition	Examples	Mechanisms	Relevant Literature
Noticing	questioning that only asks the teacher to recall information about themselves, a lesson, or their students based on prior experiences or their general familiarity with themselves or their students	• What did you notice about student x's behavior? • How did student x respond to the prompt? • What did you do when…?	Analyze and reflect Job-embedded Active learning	Barnhart & van Es, 2015 Collet, 2012 Gibbons & Cobb, 2016 Gregory et al., 2017 Perkins, 1998 Robertson et al., 2020
Cause & Effect	questioning that explicitly asks the teacher to reflect on the effect(s) that stemmed from a particular cause and/or the cause(s) that led to a particular effect.	• How do you think giving wait time influenced students? • What did you notice about how that technique influenced students' responses? • How did implementing the strategy we talked about last time help students?	Analyze and reflect Job-embedded Active learning Motivation: competence	Barnhart & van Es, 2015 Collet, 2012 Gibbons & Cobb, 2016 Gibbons & Knapp, 2018 Gregory et al., 2017 Hiebert et al., 2007 Hoffman et al., 2015 Robertson et al., 2020
Justification	questioning that explicitly prompts the teacher to provide evidence, rationale, and/or purpose for a claim, decision, or action they have made previously.	• Why did you choose to do x when student y was talking? • What were you hoping x move would accomplish	Analyze and reflect Job-embedded Active learning Motivation: autonomy	Barnhart & van Es, 2015 Collet, 2012 Gibbons & Cobb, 2016 Gibbons & Knapp, 2018 Hoffman et al., 2015 Perkins, 1998
Interpretation	questioning that explicitly asks the teacher to develop a hypothesis, draw a conclusion, or make an inference about their students (e.g. a student's motivations, rationale, understanding, or skill level), their instruction (e.g., or themselves (other than identifying a cause/effect or providing justification).	• What do you think y shows about this student's understanding? • What do you think this lesson shows about your strengths as a teacher? • How well do you think that student understood the text?	Analyze and reflect Job-embedded Active learning Motivation: competence	Barnhart & van Es, 2015 Collet, 2012 Gibbons & Cobb, 2016 Gibbons & Knapp, 2018 Hoffman et al., 2015 Perkins, 1998 Robertson et al., 2020

Move	Definition	Examples	Mechanisms	Relevant Literature
Vision	questions that explicitly prompt the teacher to articulate their goals or vision for a previous lesson or activity. This can include goals for students and for the teacher's own instruction	• What did you hope would happen in today's lesson? • What objectives did you hope students would learn? • What did you want students to understand about the text?	Analyze and reflect Active learning Job-embedded Motivation: autonomy Coherence & Alignment	Barnhart & van Es, 2015 Collet, 2012 Desimone & Pak, 2017 Gibbons & Cobb, 2016 Robertson et al., 2020
Lessons Learned	questioning that explicitly asks the teacher to identify something that they have learned from a prior experience or were working on implementing in the prior lesson	• What lessons have you learned about addressing kids' challenging behavior so far? • What did the resource I asked you to read teach you about giving feedback?	Analyze and reflect Active learning Motivation: competence Coherence & Alignment	Desimone & Pak, 2017 Perkins, 1998
Self-Assessment	questioning that asks the teacher to make a judgement about the success and quality of their own instructional practice	• How successful were you at x? • What do you think you did well in terms of feedback?	Analyze and reflect Active learning Motivation: competence Motivation: autonomy	Cohen et al., 2020
Grading	questioning that ask the teacher to locate themselves within a particular performance framework	• Thinking about the Essential Practices/CLASS/other, how would you rate yourself for the domain of questioning? • What language from the rubric do you think best describes your classroom management in today's lesson?	Analyze and reflect Active learning Motivation: competence Motivation: autonomy Coherence & Alignment	Gregory et al., 2017

Table 2. Telling & Backward-Facing (TellBack) Moves (Group 2)

Move	Definition	Examples	Mechanisms	Relevant Literature
Observation	feedback that describes specific factual information about students, a lesson, or the teacher based on the coach's observation of prior instruction or general familiarity with the students or teacher's instructional practice.	• I noticed that when student did x, you did y, and then z happened • I saw that you said "xyz" in response to the student's question	Feedback Analyze & Reflect Job-embedded	Heineke, 2013 Hoffman, 2015 Robertson et al., 2020 Russell et al., 2020
Connection	feedback that explicitly discusses the connection between a particular cause and its effect	• Giving wait time allowed that student to process and generate a more complete answer • I think the students were distracted and had trouble paying attention today because Ethan was making a lot of noise	Feedback Analyze & Reflect Knowledge of content and pedagogy Job-embedded	Heineke, 2013 Hiebert et al., 2007
Justification	feedback where the coach makes an inference about the teacher's rationale for a particular decision, claim, or action. The coach must explicitly use language to indicate that they are making an inference rather than making a simple statement of fact.	• I'm guessing that you asked Ethan to share a norm because you hoped it would refocus him to be on-task. • I think when you did that, you were trying to like bring it back to the rules and expectations a couple of times.	Feedback Analyze & Reflect Motivation: competence, relatedness Job-embedded	Robertson et al., 2020
Interpretation	feedback in which the coach communicates a hypothesis, draws a conclusion, or makes an inference about something that does not meet the criteria for Connection or Justification.	• When Ethan gave the answer that Lisa was excited, this suggested that Ethan didn't fully understand the text • You seemed a little frustrated	Feedback Analyze & Reflect Motivation: competence, relatedness Knowledge of content and pedagogy Job-embedded	Heineke, 2013 Robertson et al., 2020

Move	Definition	Examples	Mechanisms	Relevant Literature
Positive Evaluation	feedback that communicates a positive judgment about a teacher's general skill as a teacher, specific elements of the teacher's practice, or provides a general affirmation of the teacher's instruction in a specific lesson or time-period.	• You did a really great job managing student behavior	Feedback Motivation: competence Job-embedded	Perkins, 1998 Collet, 2012 L'Allier et al., 2010 Robertson et al., 2020 Sims et al., 2022
Negative Evaluation	feedback that communicates a negative judgment about specific elements of the teacher's practice, about the teacher's instruction in a specific lesson or time-period, or about a teacher's general weaknesses/problems.	• You struggled to give descriptive feedback	Feedback Analyze & Reflect Job-embedded	L'Allier et al., 2010 Perkins, 1998
Grading	feedback that explicitly makes a connection between the teacher's instructional practice and a specific framework of instructional practice or performance/evaluation rubric	• On the district's evaluation framework, I think you would score…	Feedback Analyze & Reflect Coherence & Alignment Job-embedded	Gregory et al., 2017
Check-in	dialogue that references a topic of discussion from a previous coaching conversation or professional development activity.	• So last week we talked about implementing a new behavior management strategy	Coherence & Alignment	Desimone & Pak, 2017 Sims et al., 2022

Table 3. Asking & Forward-Facing (AskForward) Moves (Group 3)

Move	Definition	Examples	Mechanisms	Relevant Literature
Generation	questioning that prompts the teacher to generate or identify new ideas, action steps, or strategies that the teacher can use in future lessons, including to meet a pre-specified goal.	• What do you want to do differently next time? • What strategy could you use to better support student engagement next lesson?	Job-embedded Active-learning Motivation: autonomy Analyze & reflect	Barnhart & van Es, 2015 Desimone & Pak, 2017 Gibbons & Cobb, 2016 Gibbons et al., 2018
Goal-setting	questioning that prompts the teacher to identify a goal or outcome for their classroom or students for the teacher to work towards.	• What do you want students to learn in the next lesson? • What reading strategies do you want students to use when they read poetry?	Job-embedded Active-learning Motivation: autonomy Analyze & reflect	Desimone & Pak, 2017 Perkins, 1998 Teemant, 2014 Russell et al., 2020
Anticipation	questioning that explicitly prompts the teacher to elaborate on the consequences of an instructional strategy, action, or goal, including the importance or purpose, potential negative consequences, or challenges the teacher may face in using the strategy	• Why is asking for text evidence important? • Why is it important for students to use context clues when they read? • What do you think would happen if you never redirected misbehaviors?	Job-embedded Active-learning Analyze & reflect	Barnhart & van Es, 2015 Gibbons et al., 2017
Application	questioning that prompts the teacher to decide when and/or how to apply a specific instructional strategy.	• How will you apply this strategy to your lesson tomorrow? • How could you give that redirection in a more specific way next time?	Job-embedded Active-learning Analyze & reflect	Cohen et al., 2020
Check-for-Understanding	questioning that checks for a teacher's understanding of a pedagogical strategy or other professional concepts that the	• So, what would a non-example of a succinct redirection be?	Analyze & reflect	Cohen et al., 2020

Move	Definition	Examples	Mechanisms	Relevant Literature
	coach or teacher have been discussing. These questions tend to require the teacher to synthesize or apply previously discussed content in order to answer them.	• What is the difference between the strategy I just suggested and the one that you used originally? • How would you summarize what wait time is?	Motivation: competency	
Content Understanding	questioning which supports the teacher in understanding the details of specific subject-matter content	• What paragraph in the text allows you to make that conclusion? • What is the correct answer to that math problem?	Job-embedded Active-learning Knowledge of content and pedagogy	Gibbons et al., 2017

Table 4. Telling & Forward-Facing (TellForward) Moves (Group 4)

Move	Definition	Examples	Mechanisms	Relevant Literature
Reinforcement	feedback in which the coach explicitly reinforces that the teacher should, in future lessons, continue using a strategy that the teacher has already using	• I noticed that you did x in your last lesson, and I want you to keep doing that	Feedback Motivation: Autonomy	Collet, 2012 Robertson et al., 2020 Sims et al., 2022 Teemant, 2014
Challenge	feedback that articulates a challenge or problem of teaching	• So sometimes students struggle to comprehend the text that they're reading, they can make claims that sometimes can't be supported with the text or may even be refuted with the text. • Sometimes we as teachers aren't aware of students' emotions because they don't know to communicate them	Knowledge of content and pedagogy Feedback Analyze & reflect	Cohen et al., 2020
Student Goal	feedback that <u>articulates</u> an instructional goal(s) <u>for students</u> for future lessons.	• We would like to increase positive task engagement. • I think it's important that we focus on helping students with writing topic sentences.	Feedback Knowledge of content and pedagogy	Teemant, 2014 Russell et al., 2020
Instructional Strategy	feedback that explicitly proposes a new strategy that a teacher can or should use.	• Next time, I want you to work on being more specific with your redirections • One thing you can do is try to avoid using negative tone of voice and instead…	Feedback Knowledge of content and pedagogy	Collet, 2012 Heineke, 2013 Robertson et al., 2020
Demonstration	dialogue where the coach illustrates *how* a specific instructional strategy can be	• A calm tone would sound like, "Ethan, please be quiet".	Modelling	Collet, 2012 Heineke, 2013

Move	Definition	Examples	Mechanisms	Relevant Literature
	used or implemented. This includes defining what a particular strategy means.	- Succinct redirections use as few words as possible	Knowledge of content and pedagogy	Robertson et al., 2020 Coburn & Woulfin, 2012 Matsumura et al., 2013
Implementation	dialogue where the coach provides a specific direction or suggestion for how the teacher should handle a specific future situation or how they could have improved a specific prior situation.	- To remind yourself to use this new strategy, you should add a note to your lesson plan - I would like you try what we talked about in your lesson tomorrow	Feedback Deliberate Practice Knowledge of content and pedagogy	Sims et al., 2022
Content Understanding	dialogue which supports the teacher in understanding the details of specific subject-matter content	- So, the text doesn't explicitly give an answer, instead the reader has to make an inference - I actually had a different answer for that math problem…	Feedback Knowledge of content and pedagogy Modelling	Gibbons et al., 2017 Matsumura et al., 2013

Table 5. Activities (Group 5)

Move	Definition	Mechanisms	Relevant Literature
Practice	dialogue where the coach initiates and facilitates a role-play activity or other approximation of practice.	Active-learning Deliberate practice Motivation: competence	Cohen et al., 2020 Sims et al., 2022
Data-Analysis	reviewing student-created materials or summary data on student learning (e.g. test score data) to analyze student understanding, learning, strengths, weaknesses, needs, etc.	Active learning Deliberate practice Motivation: autonomy Motivation: relatedness Analyze and reflect Modelling	Aguilar, 2013 Donegan et al., 2000 Downer et al., 2018 Jewett & MacPhee, 2012 L'Allier et al, 2010 Marsh et al., 2015
Co-planning	reviewing curricular materials, state-standards, student-facing material (e.g. a book or problem-set) or other documents that teachers might reference	Active learning Deliberate practice Motivation: autonomy Motivation: relatedness Analyze and reflect Modelling	Gibbons & Cobb, 2016 Gibbons et al., 2017 Jewett & McPhee, 2012 Matsumura et al., 2013
Instructional Artifact	reviewing a lesson plan, student-facing handout, video, or other artifact of the teacher's or another person's prior instruction (not including student-created materials)	Active learning Motivation: autonomy Motivation: relatedness Analyze and reflect Modelling	Gregory et al., 2017 Jewett & McPhee, 2012 Stanulis, 1994 van der Linden et al., 2021
Professional Resource	reviewing professional resources that provides general information about content or pedagogy, e.g. an article or video about strategies for teaching fractions.	Motivation: autonomy Motivation: relatedness Analyze and reflect Active learning	Aguilar, 2013 Downer et al., 2018 Jewett & McPhee, 2012 Vanderburg & Stephens, 2009, 2010
Instructional Rubric	reviewing or explaining a specific rubric or framework of high-quality instruction	Coherence & alignment Analyze and reflect	Allen et al., 2015

Note: these activities are often accompanied by moves from the first four categories that serve to facilitate teacher analysis and reflection about the activity or the materials used in the activity or provide feedback on the teachers' engagement in the activity.

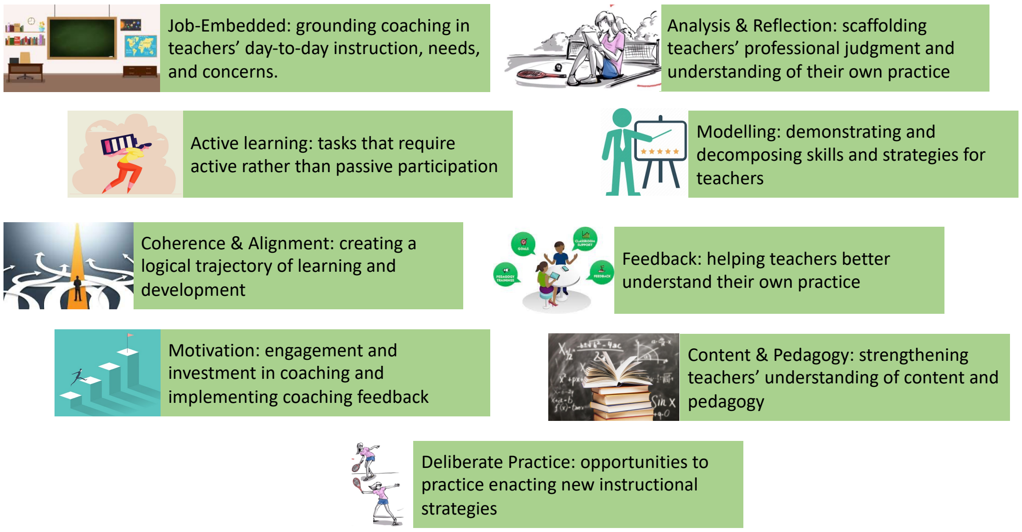
Table 6. Rapport Moves (Group 6)

Move	Definition	Examples	Mechanisms	Relevant Literature
Sharing	sharing personal information about the coach, asking about teacher personal information, or demonstrating a personal understanding of the teacher	• I love hiking too! • When I taught 8th grade, I really struggled with classroom management	Motivation: relatedness Motivation: competence	Collet, 2012 Knight, 2007
Assistance	dialogue where the coach offers to provide specific assistance or provides an opportunity for the teacher to request specific assistance	• All right. Um, we have plenty of time. We can use it to think through or talk through anything on your mind. • I can do some research and get back to you about that • Is there any additional support you would like from me?	Motivation: relatedness	Collet, 2012 L'Allier et al., 2010 Lowenhaupt et al., 2014 Perkins, 1998
Encouragement	dialogue where the coach expresses positive expectations for the teacher's future work	• you've got this! • you're going to be great!	Motivation: relatedness Motivation: competence	Perkins, 1998 Shernoff et al., 2015 Teemant, 2014
Normalizing struggle	dialogue where the coach communicates that facing challenges and struggles in teaching is normal	• Most teachers struggle with wait time	Motivation: relatedness Motivation: competence	Shernoff et al., 2015
Permission	dialogue where the coach asks the teacher for permission to do or say something.	• Is it okay if I give you some advice? • Is it okay if I ask you about that topic?	Motivation: relatedness Motivation: autonomy	L'Allier et al., 2010
Empathy	dialogue in which the coach asks about, anticipates, or expresses an understanding of the teacher's emotions or perspective.	• that must have been hard • I know it can be very hard to keep track of everything that's going on all at once	Motivation: relatedness Job-embedded	Hunt, 2016 Shernoff et al., 2015 Teemant, 2014

Move	Definition	Examples	Mechanisms	Relevant Literature
Coaching Feedback	dialogue in which the coach invites the teacher to provide feedback on how the coaching session is going or how well the coach is meeting the teacher's needs	• How was this coaching session for you? • Is there anything you'd like me to do differently in our next conversation?	Motivation: autonomy Motivation: relatedness Job-embedded	Perkins, 1998
Agenda-setting	dialogue where the coach previews things that will happen in future as part of the teacher's participation in coaching or the purpose of coaching, a specific conversation, or a specific part of the conversation	• My goal as a coach is to be as helpful to you as possible • First, we'll review the video from your lesson, then we will…	Motivation: relatedness Coherence & Alignment Job-embedded	Sims et al., 2022
Mirroring	repeating or rephrasing what a teacher just said in the previous turn of talk.	• **Teacher**: So with that one student I could just bring him right back to the… • **Coach**: Bring him back to the text, so…	Motivation: relatedness	Aguilar, 2013 Perkins, 1998 Shernoff et al, 2015
Revoicing	dialogue in which the coach rephrases what the teacher has said in a recent turn-of-talk. This must move beyond mirroring to introduce different language or build on the teacher's ideas.	• **Teacher**: So with that one student I could just bring him right back to the… • **Coach**: Yes, you can focus his attention on the text by asking him to find text evidence to support his response.	Motivation: relatedness Analyze & reflect	Aguilar, 2013 Perkins, 1998 Shernoff et al, 2015

Figures

Figure 1. Summary of the prevailing theories about the mechanisms that explain how coaching conversations support teacher development.



Job-Embedded: grounding coaching in teachers' day-to-day instruction, needs, and concerns.

Analysis & Reflection: scaffolding teachers' professional judgment and understanding of their own practice

Active learning: tasks that require active rather than passive participation

Modelling: demonstrating and decomposing skills and strategies for teachers

Coherence & Alignment: creating a logical trajectory of learning and development

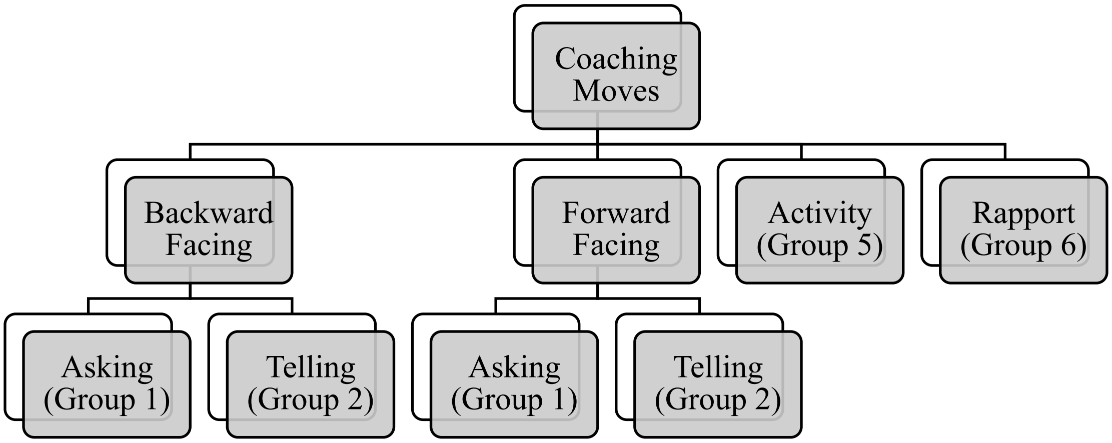
Feedback: helping teachers better understand their own practice

Motivation: engagement and investment in coaching and implementing coaching feedback

Content & Pedagogy: strengthening teachers' understanding of content and pedagogy

Deliberate Practice: opportunities to practice enacting new instructional strategies

Figure 2. Organization of the coaching moves taxonomy



CHAPTER 3

Identifying Promising Coaching Moves to Support Teacher Development

Arielle Boguslav and Kylie Anglin

Abstract

There is growing evidence that coaching programs can enhance teachers' instruction and improve student learning. This makes coaching one of the most promising levers for ensuring students have access to high-quality instructional experiences. However, we don't yet understand how the details of coaches' interactions with teachers contribute to teacher development. The coaching literature describes general features of high-quality coaching but provides only rare glimpses into the concrete discourse strategies coaches can use to achieve these aims. Developing a better understanding of effective coaching requires robust quantitative tools for measuring what coaches say and do in their interactions with teachers. In this chapter, we use a comprehensive taxonomy of coaching "moves" presented in Chapter 2 to code 186 transcripts of coaching conversations from prior studies. We then calculate the frequency of each move for each transcript to provide a quantitative measure of how coaches' discourse strategies vary. Finally, we use this novel measure to identify coaching moves that are associated with teachers' instructional practice. This serves as a proof-of-concept for the taxonomy and as a model for future work using the taxonomy to understand and improve coaches' interactions with teachers.

Introduction

Across the country, tens of thousands of coaches work every day to help teachers better support student learning and development (Domina et al., 2015). During student teaching placements, pre-service teachers regularly meet with supervisors to discuss prior lessons and plan for future ones (Matsko et al., 2020). Early career teachers are often assigned instructional coaches (Kutsyuruba & Godden, 2019). Increasingly, schools and districts are also incorporating one-on-one instructional coaching as a central component of their approach to teacher professional development for teachers at all career stages (Galey, 2016; Johnson, 2016; Neufeld & Roper, 2003). In the context of the growing evidence that coaching opportunities are highly valued by teachers (Boguslav et al., 2022; Clark & Byrnes, 2012; Gross, 2010; Ronfeldt et al., 2020) and lead to improvements in teaching practice and student learning (Davis & Higdon, 2008; Hardt et al., 2020; Kraft et al., 2018; Ronfeldt et al., 2018; Ronfeldt & Reininger, 2012; Stanulis & Floden, 2009), the expansion of coaching programs holds great promise.

At the same time, realizing this promise is challenging given the wide variability in programs and our limited understanding of the features that characterize effective and high-quality coaching programs. To date, most of what we know about the implementation of coaching at scale relates to broad structural elements of implementation such as who coaches are, how coaches are assigned to teachers, coaching dosage, the guidance and support provided to coaches, and how coaches split their time across broad areas of responsibility (e.g. Bean et al., 2010; Deussen et al., 2007; Marsh et al., 2010). While important, these studies provide little guidance as to what coaches should do and say in their interactions with teachers. The large qualitative literature on coaching provides greater insight into coach-teacher interactions, highlighting their role in supporting teacher development and identifying the features of these interactions that better support that development (e.g. Collet, 2012; Heineke, 2013; Hunt, 2016; Ippolito, 2010; Power & Goodnough, 2019; Robertson et al., 2020; Teemant, 2014). However, these studies are typically very limited in sample size. Additionally, they often focus on non-representative programs operating under ideal conditions and/or recruit individual coaches judged to be particularly successful, raising questions about the generalizability of their findings. Furthermore, these studies typically focus on more abstract features of coach-teacher interaction, rarely identifying concrete discourse strategies that coaches can use to support teacher development. For example, there is broad consensus about the need for coaches to build trusting relationships with their teachers, but little evidence-based guidance that highlights what coaches can say and do to build such relationships (Heineke, 2013; King et al., 2004; L'Allier et al., 2010; Obara, 2010; Robertson et al., 2020; Sisson & Sisson, 2017; Walpole & McKenna, 2012). Thus, coaches, program directors, and policymakers are left to navigate a dizzying array of coaching models and approaches with little concrete evidence-based guidance.

Quantitative measures that detail variation in coaches' discourse and interactions with teachers offer an important route for understanding how coaching interactions influence teacher improvement. Unfortunately, few such tools exist and those that do typically rely on teacher surveys and/or coach self-reports, which may not accurately reflect what happens in coaching conversations (Reddy et al., 2019; Richardson et al., 2020). These tools are often also designed for the specific needs of a single study context to evaluate, for example, the extent to which coaching interactions conform to a particular coaching model (Huff et al., 2013; Powell & Diamond, 2013; Wayne & Coggshall, 2022). This limits the extent to which study findings may be applicable to other models and contexts. Furthermore, such tools have limited utility for conducting research on other models, requiring researchers interested in conducting such studies to invest substantial time and effort to develop their own measurement tool(s) (Anglin et al., 2021).

To advance our understanding of the kinds of coaching interactions that best support teacher development, we need measures that leverage direct observation of coaching interactions and are applicable across a wide range of program models and contexts. Such measures will enable a systematic program of research to identify effective coaching discourse strategies that can ultimately provide coaching practitioners with clearer and more concrete guidance when navigating decisions about how coaches should interact with the teachers they support. When individual studies use common measures for describing coaching discourse, identifying patterns across studies and conducting meta-analyses will be considerably easier, leading to more efficient knowledge aggregation. These benefits can be seen, for example, in the widespread use of teacher observation rubrics, which have helped researchers refine their understanding of high-quality instruction (Cohen, 2015; T. Kane & Staiger, 2012).

In this paper, we introduce a new measurement tool for quantitatively describing how the nuances of coaches' interactions with teachers vary at scale and is applicable across a broad range of coaching models and contexts. We then apply this tool to 186 transcripts of coaching conversations from prior studies of coaching to provide a proof-of-concept for how this tool can be used to advance our understanding of the features of effective coaching. The measurement tool we introduce is structured around the comprehensive taxonomy of concrete coaching discourse "moves" introduced in Chapter 2. This taxonomy highlights 45 questioning and feedback strategies that coaches may use in their conversations with teachers. For example, the *Cause & Effect* move refers to questions that ask the teacher to identify or reflect on a causal relationship between two classroom events. We then use this taxonomy to code the coaching moves used in our sample of coaching conversations. Finally, we transform the resulting coded data into quantitative measures of coaching discourse by calculating the frequency of each move for each transcript, thereby allowing us to answer the following research questions:

4. What coaching moves do coaches tend to use?
5. How do the moves that coaches use vary across transcripts, teachers, coaches, and contexts?
6. What is the relationship between variation in the coaching moves used and teachers' observed instructional practice after participating in coaching?

In answering the first two research questions, we explore coaches' interactions at a larger scale than prior qualitative work, while preserving a level of detail that is typically lost in larger-scale quantitative studies of coaching. Additionally, the concrete nature of the discourse moves included in the taxonomy helps ensure that our findings can be clearly interpreted by researchers and coaching practitioners alike.

In answering the third research question, we identify coaching moves that are associated with stronger post-coaching instructional practice. In an ideal world, we would want to evaluate the causal relationship between coaches' use of coaching moves and teachers' subsequent instructional practice. Unfortunately, our data do not allow us to make such causal claims because neither coaches nor coaching moves were randomly assigned to teachers. This means that any estimates of the naïve relationship between coaching moves and instructional practice after coaching may be biased by omitted variables that are correlated with coaches' use of coaching moves and teachers' observed instructional practice. For example, coaches may employ certain coaching moves for teachers with greater initial skill. In this case, a positive relationship between those moves and teachers' subsequent instructional practice may be due to this initial difference in teacher skill instead of any benefit those specific coaching moves have on teacher skill development. Because of this concern, we include controls for characteristics like teachers' initial skill level that we believe are the most likely sources of bias. However, we cannot rule out the possibility that other sources of bias remain, preventing us from making causal claims.

We make both substantive and methodological contributions. Substantively, we provide a new glimpse into the black box of coaching interactions. We highlight overall trends in the discourse moves coaches use and identify how the patterns of moves coaches use vary across the transcripts. In doing so, we gain insight into the ways that coaches implemented and adapted the coaching protocol they were asked to follow. We find, for example, that coaches tended to primarily employ discourse moves that were emphasized in the standardized coaching protocol provided to them. At the same time, we find variation in the emphasis coaches placed on different components of the coaching protocol. Rather than reflecting differences between coaches, we find that individual coaches emphasized different components of the protocol in different transcripts, raising questions about how these different approaches to implementing the protocol might have affected teacher development. In exploring the relationship between this variation and teachers' subsequent instructional practice, we also identify several coaching moves that may be particularly effective for supporting teacher development, though more work is needed to confirm that this relationship is causal.

Methodologically, this study introduces a measurement tool and analytic method that other researchers can apply to a variety of other contexts and research questions. Because the coaching moves taxonomy is applicable across many different coaching models, other researchers may apply it to understand coaching practice in other coaching programs. Researchers may also focus their analysis on a specific subset of moves to investigate coach fidelity to specific coaching protocols in the context of randomized controlled trials (RCTs), thereby avoiding the substantial cost and challenge of creating study-specific measures of fidelity (Anglin et al., 2021). When used on an ongoing basis throughout a study, researchers may also use the coaching moves tool to guide the feedback and support provided to coaches throughout a study. Additionally, this study serves as the first step in a broader program of work in which we intend to develop an automated Natural Language Processing-based tool for coding coaching transcripts. Once developed, this kind of automatic tool will dramatically reduce the resources required to apply the coaching moves taxonomy to transcriptions of coaching conversations. As more studies using the coaching moves tool are conducted, researchers will be able to aggregate findings through conceptual reviews and quantitative meta-analyses with relative ease, ultimately allowing us to provide clearer, evidence-based guidance to coaching practitioners.

Though coaching programs hold great promise for supporting teacher development and student learning, realizing this promise rests on understanding the specific concrete coaching practices that can support this development. This is the gap that we address in this paper. Without this understanding, we risk establishing coaching programs that are expensive and logistically complex but ineffective. Furthermore, we risk that the promise of coaching will fade to a passing fad as more schools and districts fail to see the benefits of their efforts. As coaching continues to grow and new programs are developed, the need to understand coaching practice and its effects will become even more pressing.

Background

What is teacher coaching?

The literature on teacher coaching offers no single, agreed-upon definition of what constitutes a coach, a coaching intervention, or even a coaching conversation. Instead, the literature offers a range of definitions and categorizations of different kinds of coaching that reflect variation in how researchers conceptualize coaching and distinguish it from other professional development activities. In their foundational work, Joyce and Showers (1981) describe coaching as ongoing cycles of “observation and feedback” (p. 170) where a coach aims to help improve a teacher’s implementation of new instructional strategies introduced as part of professional development workshops or other programming.

Later work introduces additional conversational structures and distinguishes between different kinds of coaching. Gibbons and Cobb (2017), for example, describe 19 structures coaches can use in their interactions with teachers, including modeling instruction (where the teacher observes the coach’s or another teacher’s instruction and then debriefs the observation with the coach) and lesson planning (where the coach and teacher plan a future lesson together). Other work focuses on more nuanced features of coaches’ interactions with teachers to distinguish among different coaching approaches. Several scholars, for example, reference the distinction between responsive coaching, where the coach allows the teacher’s self-reflections and goals to guide content of coaching, and directive coaching, where the coach draws on their own expertise to provide directive suggestions about what the teacher should change (Deussen et al., 2007; Dozier, 2006; McGatha, 2017). Other models of coaching are also defined by expectations about how coaches should interact with teachers. Knight’s (2007) Instructional Coaching approach, for example, highlights seven partnership principles, including collaborating with teachers as equal partners and promoting teacher choice and decision-making. Whereas Instructional Coaching emphasizes the coach’s role as a partner that works together with the teacher, Cognitive Coaching emphasizes the role of the coach as a facilitator whose goal is to help teachers exercise self-direction without offering their own judgment or advice (Costa & Garmston, 2002).

Other types of coaching are differentiated by the goals of the support provided or who the coaches and teachers are. Some coaching programs focus on specific content areas (e.g. literacy or mathematics coaching), while others focus on general teaching skills such as classroom management or data analysis (Kutsyuruba & Godden, 2019; Marsh et al., 2010; Obara, 2010; Reinke et al., 2009; Toll, 2009; West, 2009). Still other coaching programs look beyond individual teachers' practice to foster school or district-wide instructional reform (Woulfin, 2018; Woulfin & Rigby, 2017). Some models rely on accomplished teachers who leave the classroom to coach other teachers, while others match classroom teachers with one another (Ackland, 1991; Coggins et al., 2003; B. D. Kane & Rosenquist, 2019; J. Knight, 2007). Coaching programs may also serve different populations of teachers, including pre-service teachers, early career teachers, teachers that are considered less effective, or schools' entire faculty (Cohen & Wiseman, 2023; Kutsyuruba & Godden, 2019; Matsko et al., 2020). Additionally, while some authors include conversations between a facilitator and multiple teachers as a form of coaching, others define coaching as one-on-one meetings between a single teacher and coach (Gibbons & Cobb, 2017; Kraft et al., 2018).

In reviewing the existing literature on coaching, we employ the broad definition provided in Chapter 2 when developing the coaching moves taxonomy. Specifically, we define coaching conversations as a dialogue between two or more education professionals, where:

- at least one participant is a classroom teacher, and the primary focus of the dialogue is on this teacher's classroom, this teacher's current teaching practice, and/or opportunities for the teacher to improve or change their teaching practice; and
- a different participant – the coach – serves as the facilitator to maintain focus on these topics.

In this way, we aim to encompass all the different models described above.

How do coaching interactions vary?

In addition to variation in how coaching is defined, the literature suggests there is substantial variation in what coaches do and say in their interactions with teachers across conversations, coaches, programs, and contexts. For example, in their investigation of Reading First Implementation across five states, Deussen et al. (2007), identify five different types of coaches who employed different foci in their conversations with teachers. Coaches classified as data-oriented coaches, for example, typically reported using coaching meetings to discuss student assessment data with teachers. Student-oriented coaches, who dedicated more time toward working directly with students in the classroom, reported using coaching conversations to debrief the lessons where a coach was present.

Prior work also highlights potential sources for the observed variation in coaches' interactions with teachers. Deussen et al. (2007), for example, find that coaches' classification tended to vary based on the state in which a coach worked. Alongside evidence that state coaching policies differed along similar lines, these findings suggest the key role of local policy in shaping coaching practice. Other studies highlight the influence of district policy, school leadership, allocation of resources, coach training, and coach support on coach-teacher interactions (Coburn & Russell, 2008; Marsh et al., 2010, 2015; Woulfin, 2018). Several studies also highlight evidence of relationships between coaches' professional experience, knowledge, interpersonal skills, and personality and coaching practice (Gibbons & Cobb, 2016; Marsh et al., 2010, 2015).

Individual coaches also often employ a range of strategies and foci within their work. Given the conceptualization of coaching as individualized support (Desimone & Pak, 2017; Killion, 2012; Kraft et al., 2018), coaches unsurprisingly often report purposefully differentiating their coaching practice to meet teachers' individual needs, both in terms of what they discuss with teachers during coaching conversations and how they facilitate these conversations (Collet, 2012; Dusenbury et al., 2010; Gibbons & Cobb, 2016; Grierson & Woloshyn, 2013; Stover et al., 2011; Tomlinson & McTighe, 2006). Other studies highlight the ways in which coaches vary their practice over time for a single teacher or even within a single coaching conversation. Ippolito (2010), for example, highlights how coaches shift from more directive feedback strategies to more responsive strategies that promote teacher reflection even within a single conversation. Collet et al., (2012), on the other hand, highlights how coaches' practice shifts over the course of 11 weeks, as teachers develop greater skill and need less support. Whereas coaches tended to make recommendations and model instructional strategies during earlier weeks, they tended instead to rely on probing questions in later weeks to allow teachers to take greater responsibility for their instructional decision-making.

Together, this literature suggests that we might also expect to see variation in the discourse strategies coaches use across contexts, coaches, teachers, and conversations.

What are the features of effective coaching interactions?

There is no shortage of literature that aims to identify features of coaching interactions that support teacher development. Distilling this work into concrete guidance for coaches and program administrators is challenging for two key reasons. First, evidence about the impacts of specific features is often mixed, resulting in little consensus about how to interpret it. For example, some scholars highlight challenges with coaching interactions that take place as part of teacher evaluation systems and recommend providing coaching solely for the purposes of professional development and/or by coaches who are not involved with making evaluative decisions about the teachers they coach (Booker & Russell, 2022). Other studies, however, suggest that coaching in the context of teacher evaluation systems can be impactful (Marsh et al., 2017; Phipps & Wiseman, 2021; Woulfin & Rigby, 2017). Another area of debate concerns the extent to which coaches should provide more directive feedback, where coaches offer their own interpretations of what they observed in the teachers' classroom and suggestions for improvement, or more reflective questioning, scaffolding teachers' own self-reflection and problem-solving (Deussen et al., 2007; Dozier, 2006; McGatha, 2017). In addition to studies highlighting the benefits of one approach over the other, additional work suggests that both reflective and directive coaching may be valuable in different contexts or even at different moments within a single conversation (Cohen et al., 2020; Ippolito, 2010).

Even when the evidence is more consistent, it is still limited by the lack of concrete details about what coaches can say and do to enact such features in practice. One area of consensus, for example, is the importance of strong coach-teacher relationships for fostering teacher development (Lowenhaupt et al., 2014; Power & Goodnough, 2019; Shernoff et al., 2015). What is less clear, however, is what coaches can say and do to develop such relationships. We also know, for example, that coaches can provide teachers with valuable feedback on teachers' prior instruction to help teachers understand their strengths and weaknesses as well as identify problems, challenges, and manageable goals for improvement (Bean et al., 2010; Blazar & Kraft, 2015; Cassidy et al., 2008; Coburn & Woulfin, 2012; Cohen et al., 2020; Collet, 2012; Heineke, 2013; Hunter & Springer, 2022; Kurz et al., 2017; Tung et al., 2004). However, what constitutes high quality feedback or the specific discourse strategies coaches can use to explore and negotiate improvement goals with teachers, is unclear (Kochmanski, 2020). Similarly, coaches can use questioning strategies to provide opportunities for teachers to express their own views, interpretations, and ideas as well as scaffold teachers' reflection and problem-solving skills (Barnhart & van Es, 2015; Collet, 2012; Denton & Hasbrouck, 2009; Heineke, 2013;

Hiebert & Morris, 2012; Kennedy, 2016; J. Knight, 2009; Koh & Neuman, 2006; Power & Goodnough, 2019; Santagata & Angelici, 2010; Teemant, 2014; Winch et al., 2015). However, we do not know what kinds of questions are most effective for achieving these different aims, for different kinds of teachers, and in what contexts.

A handful of qualitative studies have begun to investigate features of coach dialogue, such as coach versus teacher talk time, the use of open-ended questions, and patterns of interaction (Collet, 2012; Heineke, 2013; Shernoff et al., 2015). Robertson et al. (2020), for example, highlight three patterns of coach-teacher interactions that are associated with teachers' uptake of the instructional ideas discussed. Though some of these studies include some attention to specific kinds of utterances or coaching "moves," such as "affirms an action or statement made by the teacher," (Robertson et al., 2020, p. 412) such granular-level detail is not the primary focus in these studies. In this paper, we build on this existing work by employing a much larger sample size and zooming in on the nuances of coaches' discourse moves as the primary unit of analysis.

Data and Methods

Study Context

This study draws on transcripts of audio and video recorded coaching conversations from a series of multi-site, multi-year replication studies of the effects of a coaching intervention for pre-service teachers, known as TeachSim (Cohen et al., 2020, 2021, In review). As part of the TeachSim model, teachers randomized to receive the coaching treatment participate in simulation-based practice opportunities and receive directive feedback from a coach during a five-minute coaching conversation. The simulations employ "mixed-reality" technology, where a trained actor provides voices and controls the movement of multiple virtual student avatars. As part of the study, participating teachers first complete a five-minute simulation in which they practice leading five virtual students in a discussion of classroom norms[4] and redirecting off-task student behaviors while a coach observes the simulation. Immediately after completing the simulation, the coach facilitates a five-minute coaching conversation to provide feedback on the teachers' practice. Finally, immediately after the coaching conversation, teachers repeat the simulation scenario, providing an opportunity for them to incorporate the coach's feedback. As part of the broader program of TeachSim research, many teachers participated in additional simulation-based practice opportunies focused on other instructional skills and some participated in additional coaching opportunities. However, for simplicity, we focus exclusively on the classroom norms intervention, which included only two five-minute simulations and one five-minute coaching conversation.

Participating teachers were recruited from teacher preparation programs at three university sites in the United States. Site 1 is a public university in the southeast. Participating teachers were full-time graduate and undergraduate students enrolled in a traditional teacher preparation program for elementary education. Site 2 is a private university in a large southern city. Participating teachers were enrolled in an alternative preparation program in which they worked full-time as teachers of record while simultaneously completing a university-based Master of Teaching. Finally, Site 3 is a public university in a rural area near the US-Mexico border. Participating teachers were undergraduate students enrolled in an interdisciplinary program for bilingual and second-language teaching. At all three sites, participants were

[4] In leading this discussion, teachers were prompted to ask students to generate their own ideas about what norms, or behavioral expectations, the class should adopt and explain their rationale for their ideas. In addition to being a common activity conducted by teachers of all content areas at the beginning of the school year, this kind of discussion is an important foundation for building a positive classroom culture and is an important activity for pre-service teachers to practice (Cohen et al., 2020).

primarily female, over 21 years old, and attended high school in a suburban area. While participants from Site 1 were primarily white, participants from the other sites were more likely to be persons of color. Finally, participants from Site 3 were more likely to be first-generation college students.

Logistical constraints meant that teachers at different sites participated in the TeachSim activities at different times and at different points in their program. Site 1 teachers completed the activities after completing two full semesters of coursework and field placements. Participants at this site also include teachers from three different cohorts, who participated in the TeachSim activities in the spring semester of three different academic years. Site 2 teachers completed the activities in the summer before entering the classroom as teachers of record. Finally, Site 3 teachers completed the TeachSim activities in their first semester of the program, before participating in any formal field placement experiences.

Participating coaches included doctoral students, full-time teachers, and full-time instructional coaches. All coaches were required to have at least three years of teaching experience and complete 5-10 hours of TeachSim-specific training. Coaches were trained to use a protocol to observe pre-service teachers during the simulation and identify one of four targeted instructional skills that best reflected the candidate's Zone of Proximal Development (Collet, 2012; Shabani et al., 2010; Vygotsky & Cole, 1978; Warford, 2011). These skills were selected in accordance with the Responsive Classroom framework and included: 1) responding to off-task behaviors immediately, 2) providing a specific redirection that unambiguously communicates the teacher's expectations for how the student's actions should change, 3) providing a succinct redirection that takes up minimal class time, and 4) using positive or neutral tone of voice and body language in delivering the redirection (Charney, 1993; Responsive Classroom, 2014).

Coaches were also trained to follow a highly structured coaching protocol that focused on providing teachers with specific, targeted, and actionable feedback based on their observations of the teacher's prior simulation practice. The coaching protocol consisted of five stages: (1) ask the candidate to assess their own performance; (2) affirm an observed teaching practice, explaining why the practice was effective; (3) identify and explain one targeted skill for the candidate to target in the next session; (4) engage the candidate in roleplay so that the candidate can practice the target skill; and (5) close the coaching session with positive reinforcement. In addition to following this structured format for the coaching conversation, coaches were also expected to limit their discussion to the four targeted instructional skills included in the observation protocol. While this coaching protocol is more standardized than many of the protocols currently in practice, we believe it is a valuable context for providing a proof-of-concept for the coaching moves taxonomy. If we can detect nuanced differences in coaches' interactions with teachers in the context of such a highly standardized protocol where the variation is likely to be relatively limited, then we can feel confident that the taxonomy will be capable of detecting the much greater variation we would expect to see in less standardized coaching protocols.

For more detailed descriptions of coach selection and training, the coaching model, and the simulation scenario see Cohen et al. (2020, 2021) and Anglin et al. (2021).

Analytic Sample

Our primary analytic sample consists of 186 transcripts of coaching conversations between 186 teachers and 13 coaches.[5] This sample includes three different cohorts from Site 1 and one cohort each from Sites 2 and 3. Table 1 shows the number of transcripts and coaches included in our sample for each cohort by site combination, which we hereafter refer to as

[5] Of the 186 transcripts, there are 6 for which we cannot identify the coach. These transcripts are still included in our sample, however.

site/cohort. Table 2 provides a sense of the characteristics of these transcripts. On average, transcripts contained 5.22 minutes of conversation, with 80% of the words in the transcript being spoken by the coach.

Analyses that include data on teacher characteristics and observation scores include slightly smaller samples because of missing data.

Measures

Coaching Moves

We draw on the coaching moves framework created in Chapter 2 as a coding scheme to qualitatively document variation in coaching practice in each of the 186 transcripts included in the analytic sample. The coaching moves taxonomy decomposes coaching practice into the granular level of "moves" (Boerst, et al., 2011; Heineke, 2013). This approach prioritizes the concrete details of coaches' interactions with teachers as the core techniques coaches can draw on to achieve more complex goals that facilitate teacher development (Hiebert et al., 2002; Lortie, 2002; Reisman et al., 2019; Winch et al., 2015). As described in detail in Chapter 2, the specific moves included in the taxonomy are drawn from theoretical and empirical literature on coaching, documentation of practitioner- and researcher-developed coaching protocols, and transcripts of coaching conversations. The taxonomy also connects each move to the potential mechanisms of teacher development it may serve, according to existing literature on coaching. For example, borrowing from Aguilar (2013) and the MTP + 4Rs Coaching Handbook (Morningside Center for Teaching Social Responsibility, 2012), Chapter 2 defines the *mirroring* move as repeating or rephrasing what a teacher has just said and suggests that this move may serve as one tool that coaches can draw on to help teachers feel heard, build a strong coach-teacher relationship, and support teacher motivation (Aguilar, 2013; Hunt, 2016). Hearing one's own ideas repeated back may also be an important tool for supporting teachers with analyzing and reflecting on their own beliefs or interpretations of a particular situation (Ippolito, 2010; O'Connor & Michaels, 1993).

Because individual moves may be used for multiple purposes, the coaching moves taxonomy organizes moves into six groups, each with 5-10 moves, according to their structure and content rather than their underlying purpose or goal. This approach also allows for efficient navigation of the taxonomy so that coders can easily and reliably narrow down their choices to an individual group before selecting the most appropriate code within the group. Attention to structural rather than purpose-oriented distinctions between moves also helps ensure consistent use of these moves when coding. Below we summarize the six groups of moves included in the taxonomy. We include an overview of these groups in Figure A1.

The first four groups consist of combinations of two structural properties: 1) whether the moves are *asking* moves which pose open-ended questions or are *telling* moves that provide the teacher with information and feedback; and 2) whether the moves *are backward-facing* moves, which focus on processing and providing feedback on what has previously occurred or *forward-facing moves*, which focus on planning for future lessons and changes to instruction (Hattie & Timperley, 2007; Sisson & Sisson, 2017). This creates the following four groups: AskBack (asking and backward-facing), TellBack (telling and backward facing), AskForward (asking and forward-facing), and TellForward (telling and forward-facing). The remaining two groups focus on moves that cannot be easily categorized in terms of their structural properties but may nonetheless be used by coaches for important purposes. The fifth group, Activities, consists of moves in which the coach facilitates a structured activity, such as analyzing study data or reviewing curricular materials with the teacher. Finally, the sixth group, Rapport, consists of moves that facilitate connection and trust.

Together, these moves reflect a variety of potential mechanisms that may underpin the influence of coaching on teacher learning and development. For example, backwards-facing moves can help ensure that coaching conversations are *job-embedded* by grounding teacher learning in the details of teachers' day-to-day instruction, content, and students. This can help ensure that coaching conversations are responsive to teachers' needs and concerns while providing authentic opportunities for teachers to make connections between theory and the practical details and challenges of instruction in the context of their daily work (Collet, 2012; Croft et al., 2010; Denton & Hasbrouck, 2009; Gibbons & Cobb, 2017; Joyce & Showers, 1981; Koh & Neuman, 2006; Putnam & Borko, 2000; Terehoff, 2002). Asking and activity moves can help coaches provide *active learning* opportunities where teachers are not passive recipients of information, but instead actively construct new knowledge through problem-solving, data-analysis, lesson-planning, and other professional activities (Desimone & Pak, 2017; Lieberman, 1995; Niemi et al., 2016; Shernoff et al., 2015). Additionally, activity moves can help teachers successfully enact new practices (Kennedy, 1999) by providing opportunities for *deliberate practice* during coaching conversations and in the classroom (Cohen et al., 2020; Ericsson & Pool, 2016; Ippolito, 2010; Reddy et al., 2019). For a comprehensive discussion of the mechanisms that underpin the moves in the coaching taxonomy see Chapter 2.

Coaching Moves Coding. In applying the moves in the coaching moves framework to code transcripts, we used coach turns-of-talk as the unit-of-analysis. Specifically, coders were instructed to label all applicable moves within a coach turn-of-talk in the order that they appeared. If a coach used the same move to convey different information multiple times within a single turn-of-talk then the same code would be listed multiple times. However, coders were not asked to identify the specific sentences or sentence fragments that represent an individual move within a coach turn-of-talk. In this way, we retained our focus on the move-level details of coaches' discourse while also taking advantage of unambiguous and naturally occurring segments of the conversation.[6] Table 2 includes summary statistics to provide a sense of what these turns-of-talk looked like. On average, coach turns-of-talk include 48.12 words, although the large standard deviation (67.98 SD) suggests that this was highly variable. There was also substantial variation in the total number of coach turns-of-talk per transcript, with an average of 15.12 and SD of 8.34.

Because the recorded conversations could theoretically include dialogue that does not fit into any of the 45 move categories in the taxonomy, coders could also label relevant turns of the talk with four additional codes, which we refer to collectively as "Other Moves": 1) Unclear for dialogue that is difficult to interpret (this includes parts of a transcript where poor audio quality prevented complete transcription), 2) Non-Coaching for dialogue that is not part of the coaching conversation (e.g. discussion of logistical or technical issues), 3) Other Coaching for dialogue that is part of the coaching conversation but does not fit into any of the move categories in the taxonomy, and 4) No Moves for turns-of-talk that consisted of short interjections that did not fit within any of these categories, such as "yeah" and "okay" were coded as having no coaching moves. We include the full codebook given to coders in Appendix A.

[6] Ensuring that several coders can reliably code individual utterances would require coders to reliably identify the beginning and end of every individual move. This is costly and challenging but would provide little additional benefit for quantitative analysis purposes. Alternatively, if coders could only select a primary code for each turn-of-talk we would be obscuring much of the detail and variation in coaches' practice. This approach also poses challenges for inter-rater reliability if raters disagree about which code should be considered primary in a given turn-of-talk, especially when coach turns-of-talk are long.

Within each of the five site/cohort combinations, transcripts were randomly assigned to coders to avoid systematic rater effects. Coders included one of the authors as lead coder and one undergraduate research assistant with prior experience coding coaching transcripts. The undergraduate coder completed coder training and was required to reach at least 80% agreement with the lead coder for a randomly selected set of coach turns-of-talk before being certified to code (Miles et al., 2013). Coders also participated in weekly norming meetings to reduce rater drift and support inter-rater reliability over time. To enable the calculation of inter-rater reliability, 15% of transcripts were randomly selected for double coding.

Quantifying Coaching Moves Codes. Once all transcripts were coded, we constructed a transcript by turn-of-talk dataset consisting of all applicable moves identified for that turn-of-talk in the order they were selected. For each turn-of-talk, we then transformed the list of moves into move count variables, identifying the number of times each move was present within each turn-of-talk. Using all turns-of-talk for all double-coded transcripts, we then calculated inter-rater reliability for each move using percent agreement and Krippendorf's alpha. Coders were considered "in agreement" if the total count for a given move identified by one coder exactly matched the total count identified by the second coder. Inter-rater reliability statistics for each individual move and move-group are included in Table 3. For moves that were rarely present in our sample of transcripts, Krippendorf's alpha statistics are mechanically very low and difficult to interpret. For example, the Rapport: Revoicing move was used less than one time per transcript on average and has a Krippendorf's alpha of 0.24. However, the percent agreement for this move is a robust 95%. Indeed, the lowest percent agreement for any move is 86% with most moves falling in the 90-99% range, which is encouraging. To be conservative, we focus our analysis of individual moves (in RQ3) on those that meet minimum frequency standards, which eliminates moves with the lowest Krippendorf's alpha statistics.[7] We also maintain caution when drawing conclusions about the relationships between teacher observation outcomes and moves where reliability is particularly low.

While the coding of moves occurred at the turn-of-talk level, we ultimately use these codes to characterize the coaching at the transcript level. We capture this with two descriptive statistics at the transcript level: presence and count. To capture the presence of a given move, we first construct 45 indicator variables that are coded to 1 if a transcript does include a move and 0 if not (e.g., move 1 presence, move 2 presence…move 45 presence). To capture the count of moves, for each transcript, we create a set of 45 variables that count the total number of times that a given move is used (e.g., move 1 count, move 2 count…move 45 count). In theory, the values for these variables can be equal to 0 (a given move never occurred in the transcript) or any positive integer (representing the number of times a given move occurred in a transcript). In practice, the largest value we observe is 15 for the Other: Non-Coaching move (Table 3). For some analyses, we seek to move beyond the more granular description of the 45 individual moves and instead capture 1) whether a transcript contains any moves of each of the 7 move-groups (e.g., TellBack presence, Other Moves, etc.) and 2) the count of moves in each move-group (e.g., TellBack count, TellForward count). For each transcript we connect these coaching move variables to the specific coach and teacher who participated in the coaching conversation as well as additional data on teacher characteristics, which we describe further below.

Teacher Observation Scores

[7] Specifically, we exclude all coaching moves where fewer than 25% of the 186 transcripts in our sample have a non-zero value. This means that all coaching moves from Other: Other Coaching through AskForward: Generation in Table 3 are excluded because the proportion of transcripts in which they are present is less than 25%.

We also obtain teacher observation scores that capture teachers' instructional practice during each of the two simulations (pre-coaching and post-coaching). The raters responsible for generating these scores were undergraduate research assistants who had no pre-existing relationships with the teachers they observed. These raters were not the same as those used to code the coaching transcripts for coaching moves. To generate the observations scores, raters evaluated each teachers' instructional skill in redirecting off-task student behavior during each simulation using a rubric based on the Responsive Classroom framework, in alignment with the coaching protocol described above (Charney, 1993; Responsive Classroom, 2014). We use the post-coaching observation scores as the outcome of interest for RQ3 analyses and the pre-coaching observation score in RQ2 and RQ3 analyses.

Based on their observations while watching a videorecording of a simulation and the indicators included in the rubric, raters selected a score from 1-10 with a score of 1-3 representing less skillful redirections, a score of 4-6 representing somewhat skillful redirections, and a score of 7-10 representing more skillful redirections. The mean pre-coaching score is 3.70 with a SD of 1.63. The mean post-coaching score is 6.12 with a SD of 1.65. For ease of interpretation, we standardize both pre-coaching and post-coaching scores for all analyses.

Rating procedures followed established best practices, including providing rater training, requiring rater certification, conducting weekly norming meetings, and randomly assigning videos to raters. Additionally, raters were blinded to whether each videorecording captured a teacher's pre-coaching simulation or post-coaching simulation. Inter-rater reliability was calculated using Krippendorf's alpha and varied from 0.75-0.88 across sites/cohorts. For additional details about the observational measure, see Cohen et al., 2020.

Additional Teacher Characteristics

In addition to pre-coaching observation scores, we draw on a several additional variables reflecting baseline teacher characteristics collected as part of the TeachSim work. Specifically, we explore the relationship between the three variables described in Table 4 and the moves coaches use to explore how coaching moves vary across teachers in RQ2. The coach rating variable reflects a coach's perception of a teachers' skill in redirecting off-task student behavior during the pre-coaching simulation (as opposed to a trained rater's perception captured by the pre-coaching observation score). The self-efficacy variable captures teachers' baseline sense of self-efficacy for teaching. Finally, the nervousness variable captures the extent to which teachers experienced nervousness or anxiety when engaging with the simulation technology at baseline. We focus on these three variables because they suffer the least from missing data challenges and because prior literature suggests that coaches might employ different strategies in their conversations with teachers depending on these characteristics (Deussen et al., 2007; Hunt, 2016; Killion, 2008; Kowal & Steiner, 2007). Because of the different scales used across these three variables, we standardize them for ease of interpretation.

Additional data on teacher characteristics, including the NEO Five-Factor Inventory capture personality traits (Costa Jr & McCrae, 1992), and high school Grade Point Average (GPA) are also included in supplementary analyses using the subset of teachers for whom such additional data are available.

Analytic Approach

RQ1: What coaching moves do coaches tend to use?

Using the transcript-level move count variables, we calculate descriptive statistics for each move variable and move-group variable. We then use these results to identify which of the 45 individual moves are used most frequently. To characterize how the moves used fall into the seven less granular move-groups, we also generate a graph illustrating how the count for each move-group varies across the 186 transcripts in our sample.

RQ2a: How do the moves that coaches use vary across transcripts?

With 45 different coaching move variables, it is difficult to obtain a holistic understanding of how coaching moves vary across transcripts from looking at the variation in individual moves separately. For this reason, we instead use hierarchical agglomerative clustering methods to identify distinct conversational profiles based on differences in the frequencies of each move between transcripts (Everitt, 2011). The hierarchical agglomerative clustering method allows us to assign each transcript to a coaching profile (or cluster). Using the move frequency variables, this method first compares the moves identified in each transcript with the moves identified in every other transcript to generate a matrix of dissimilarity statistics. Then, using this dissimilarity matrix, every transcript is matched to the transcript with which it has the lowest dissimilarity (or greatest similarity). Then these pairs of transcripts are successively matched with other pairs to generate successively larger groupings until all transcripts are connected within a single group. The result of this process is a cluster tree illustrating a hierarchy of nested clusters of transcripts based on similarity. Conducting this analysis at the transcript level allows us to empirically explore potential sources of variation in coaching moves, instead of assuming, for example, that this variation might stem from differences in coach style.

In executing this clustering method, we have several choices for how to calculate dissimilarity between transcripts (dissimilarity method) and how to calculate dissimilarity between groups of transcripts (linkage method). Following common practice, we first generate cluster trees using a variety of different dissimilarity metrics. We then select the approach that maximizes the distance between clusters while also resulting in a manageable number of clusters (i.e. 2-6), with a reasonable number of transcripts in each cluster (e.g. 25-50) and reasonable balance in sample size across clusters to facilitate analysis of relationship between clusters, covariates, and outcomes of interest. Figure C1 includes all the cluster trees we generated. Ultimately, we felt that the cluster tree using correlation as the dissimilarity method (to compare individual transcripts) and complete linkage as the linkage method (to compare groups of transcripts) best met our criteria. The correlation dissimilarity method calculates a correlation coefficient between each pair of transcripts using a modified version of the standard Pearson correlation coefficient equation and takes a value between -1 and 1. Instead of comparing the relationship between two variables across multiple observations, this approach compares the relationship between two observations (i.e. two transcripts) across multiple variables (i.e. the 45 move count variables). The complete linkage method defines the dissimilarity between two groups of transcripts as the largest correlation dissimilarity exhibited between a transcript in Group 1 and a transcript in Group 2. For example, if Group 1 consisted of Transcript A and B and Group 2 consisted of Transcript C and D, we would compare the correlation dissimilarities across the following four pairs: A-C, A-D, B-C, and B-D and then select the maximum correlation dissimilarity among these pairs.

While the hierarchical agglomerative clustering method allows us to assign each transcript to a profile (or cluster), it does not directly provide any information about how the coaching moves used differ across clusters. To understand this, we use a three-step process. First, we calculate descriptive statistics for each move and move-group separately by profile to identify which moves and move-groups differ the most across profiles. Second, for the moves and move-groups identified, we create color-coded scatterplots to identify visual patterns in how the profiles differ on these variables. Finally, we consider these patterns altogether to create descriptions for each profile that characterize how they differ.

RQ2b: What predicts variation in the moves included in coaching transcripts?

Once we have identified how the frequency of coaching moves vary across transcripts, we turn our attention to exploring potential explanations for this variation. In line with prior literature, we explore the extent to which coaching interactions vary across contexts (captured by the five site/cohort combinations), coaches, and teachers (Collet, 2012; Deussen et al., 2007; Dusenbury et al., 2010; Gibbons & Cobb, 2016; Grierson & Woloshyn, 2013; Marsh et al., 2010, 2015; Stover et al., 2011; Tomlinson & McTighe, 2006). Unfortunately, we cannot investigate how coaching interactions vary within teachers because we only have one transcript per teacher.

To understand the extent to which coaching profiles vary by site/cohort and coach, we generate color-coded bar graphs illustrating the prevalence of each profile for each site/cohort combination and coach. To understand how coaching profiles vary across teachers, we use Ordinary Least Squares (OLS) regression analyses to investigate the relationship between coaching profiles and four teacher covariates measured before teachers participated in coaching: coach rating, teacher self-efficacy, pre-coaching observation score, and teacher nervousness. In doing so, our goal is to provide a descriptive picture of the kinds of teachers that are exposed to the different coaching profiles rather than estimating the effect of specific teacher characteristics on the probability a teacher is exposed to a particular profile. Specifically, we estimate the following model separately for each of our four baseline teacher characteristics:

$$(1)\ Teacher\ Characteristic_i = \alpha_i + \sum_{p=1}^{4} ProfileP_i + \delta_c + \pi_r + \varepsilon_i$$

Here, teacher i's self-efficacy, pre-coaching score, coach rating, or nervousness is modeled as a function of the profile to which they are exposed ($\sum_{p=1}^{4} ProfileP_i$). We operationalize the profile to which a teacher was exposed by including an indicator variable for each of the four profiles that takes on the value of 1 if teacher *i* was exposed to that profile or the value of 0 if teacher *i* was not exposed to that profile. Since all coaches use nearly all profiles, we also include coach fixed effects (δ_c) to focus our analysis on within-coach variation in the use of coaching profiles.[8] We also include coder fixed effects (π_r) to ensure that differences in teacher characteristics between profiles do not result from differences in how different coders applied the coaching moves taxonomy to the transcripts during coding. The coefficients of interest are the coefficients on the profile indicator variables, which reflect the average value of a given teacher characteristic for teachers exposed to each profile and assigned to the same coach.

RQ3: What is the relationship between variation in the coaching moves used and teachers' observed instructional practice after participating in coaching?

As previewed in the introduction, in an ideal world we would want to evaluate the causal relationship between coaches' use of coaching moves and teachers' subsequent instructional practice. The ideal experiment to do this would require the random assignment of the count of coaching moves to teachers and random assignment of coaches to teachers. Without random assignment of coaching moves to teachers, we worry about selection bias in what teachers are exposed to what coaching moves. For example, coaches may rely more heavily on certain moves for teachers that are less skilled to begin with, resulting in a negative relationship between those coaching moves and post-coaching observation scores not because of the benefit of those moves on teacher development, but because of this initial difference in skill. Without random

[8] Because coaching was conducted via Zoom, individual coaches were often assigned to coach teachers at multiple sites. Some coaches also provided coaching for multiple cohorts of teachers at Site 1.

assignment of coaches to teachers, we worry about selection bias in what coaches are assigned to what teachers. For example, certain coaches may be assigned to less skilled teachers. If there are also differences in the moves different coaches use, then any estimated relationship between the coaching moves used and post-coaching teacher observation scores may reflect the underlying differences in teachers' initial skill levels.

Unfortunately, in our case, neither coaches nor coaching moves were randomly assigned to teachers. Since all coaches were trained to use the same coaching protocol, use of coaching moves is decidedly idiosyncratic. At the same time, coach assignment was primarily dictated by coach and teacher schedules for logistical reasons and when we estimate the relationship between coach fixed effects and teacher covariates, we find some evidence of non-random selection (Table B1). As we discuss further below, in lieu of random assignment of coaching moves and coaches to teachers, we include several controls in our analytic models to account for some of the most common sources of selection bias, namely teacher characteristics and the coach to which a teacher was assigned.

For RQ3, the key predictors of interest reflect the counts of different coaching moves that a teacher was exposed to in the transcript of their coaching conversation. We explore this at three levels of granularity in three different models: First, at the least granular level, we include each transcript's coaching profile assignment based on the clustering technique described above. Second, we include seven predictors that capture the raw number (or count) of moves in each transcript for each of the seven move-groups (TellForward, AskBackward, etc.). Finally, at the most fine-grained level, for each transcript, we include 21 predictors that capture the raw number (or count) of moves in each transcript.[9] Our primary specifications are shown in Equations 2-4.

[9] Move frequency variables are not highly correlated with one another so we are not concerned about issues of collinearity here. However, as Table B2 shows, there are some move frequency variables with very limited

Here, teacher i's standardized post-coaching observation score is modeled as a function of a set of coaching moves variables – profiles, move-groups, or moves – that reflect the coaching moves teacher *i* was exposed to in the coaching conversation captured by one of the 186 transcripts in our sample.

$$(2)\ Score_i = \ \alpha_i + \sum_{p=1}^{4} ProfileP_i \ + \ PreCoachingScore_i + CoachRating_i + \ \theta_s + \delta_c + \ \pi_r + \ \varepsilon_i$$

$$(3)\ Score_i = \ \alpha_i + \sum_{g=1}^{7} MoveGroupGFreq_i + \ PreCoachingScore_i + \ \theta_s + \delta_c + \ \pi_r + \ \ \varepsilon_i$$

$$(4)\ Score_i = \ \alpha_i + \sum_{m\ =1}^{21} MoveMFreq_i + \ PreCoachingScore_i + \ \theta_s + \ \delta_c + \ \pi_r + \ \ \varepsilon_i$$

All three equations include controls for the teacher i's standardized pre-coaching observation score, site/cohort fixed effects θ_s, coach fixed effects δ_c, and coder fixed effects π_r. In this way, we ensure that any resulting differences in observation scores captured by coefficients on the key predictors of interest— moves, move-groups, or profiles—do not result from differences in teachers' external learning context (site/cohort), pre-existing differences in instructional skill, differences in the kinds of teachers assigned to different coaches, or differences in how different coders applied the coaching moves taxonomy to the transcripts during coding. For Equation 2, we also include controls for teacher i's standardized coach rating given our finding in RQ2 that coach ratings differ significantly across the different profiles. By including this control, we

variation, which raises the concern that estimated coefficients for these moves might be determined by a very small number of observations. Rare moves also suffer from lower Krippendorf's alpha inter-rater reliability statistics. As noted above, we therefore estimate Equation 4 using a limited set of move frequency variables for which at least 25% of the 186 transcripts in our sample have a non-zero value. This means that all coaching moves from Other: Other Coaching through AskForward: Generation in Table 3 are excluded because the proportion of transcripts in which they are present is less than 25%.

ensure that any resulting differences in observation scores by coaching profile do not result from coaches selecting different profiles for teachers they perceive to be less skillful.

In Equation 2, the coefficients of interest are the profile indicator variables ($\sum_{p=1}^{4} ProfileP_i$), representing covariate-adjusted group mean differences in post-coaching observation scores across the four profiles of coaching to which a teacher could be exposed. The coefficient on the Rapport & Report profile, for example, reflects the difference in mean post-coaching observation scores (in SDs) between a teacher exposed to the Rapport & Report profile and a teacher exposed to the Modeling profile (the reference category), among teachers in the same site assigned to the same coach and coaching moves coder, and adjusted for pre-coaching score and coach rating. In Equation 3, the coefficients of interest are those on move-group variables ($\sum_{g=1}^{7} MoveGroupGFreq_i$), representing how observation scores vary depending on the frequency of each move-group. The coefficient on the TellForward variable, for example, reflects the change in post-coaching observation scores (in SDs) associated with exposure to one additional TellForward move, among teachers in the same site assigned to the same coach and coaching moves coder, and adjusted for pre-coaching score. Finally, in Equation 4, the coefficients of interest are the move frequency variables ($\sum_{m=1}^{21} MoveMFreq_i$), representing how observation scores vary depending on the frequency of each move. The coefficient on the TellForward: Instructional Strategy variable, for example, reflects the change in post-coaching observation scores (in SDs) associated with exposure to one additional TellForward: Instructional Strategy move, among teachers in the same site assigned to the same coach and coaching moves coder, and adjusted for pre-coaching score.

Limitations

While we believe that our analysis makes important progress toward understanding the features of effective coaching, we also recognize its limitations. Specifically, we highlight the limitations of our sample and the limitations of our coding scheme. We discuss the limitations of our analytic models and their implications for interpreting our results in the discussion section.

While the coaching conversations included in our sample reflect authentic coaching conversations embedded within authentic educational contexts, they are not necessarily representative of the day-to-day coaching conversations that occur in schools and districts that are not actively participating in a research study. Additionally, the coaches that facilitated these conversations may not be representative of the broader population of coaches currently working in schools and districts across the country. This limits the generalizability of our findings about coaching practice but does not negate the value of this study as a proof-of-concept and model for future work where new data can be collected from schools and districts implementing their own coaching programs.

As with any coding scheme, the coaching moves coding scheme prioritizes describing certain features of coaching practice to the exclusion of others that may also play a role in teacher learning and development. In coding transcripts, rather than audio or video recordings, we focus exclusively on coach's spoken dialogue and effectively ignore non-verbal features of coaching practice, such as coach tone of voice and body language. This is a fruitful area for further work, especially given the subtle and nuanced nature of non-verbal communication that makes it difficult to code reliably. Additionally, in coding only coach dialogue, we cannot consider teacher's interpretations of and responses to coach dialogue, which we recognize are a key part of how coaching conversations may influence teacher practice (Heineke, 2013;

Robertson et al., 2020). This reflects the trade-off between breadth and depth in conducting qualitative analysis.

Results

RQ1: What coaching moves do coaches tend to use?

As shown in Table 2, coaches on average employed 33.78 unique moves per transcript. When we account for repetition of the same move multiple times within a transcript, coaches on average employed 37.65 moves per transcript.

Given the standardized and detailed nature of the TeachSim protocol, we can pinpoint a specific set of moves we would expect coaches to use. Descriptive statistics for each move and move-group (included in Table 3) allow us to explore the extent to which coaches employed the specific moves included in the coaching protocol and made use of moves not explicitly included in the protocol. For the most part, coaches tended to rely most heavily on moves and move-groups that align with the standardized TeachSim protocol. For example, Figure 2, which plots the frequency of each move-group for each transcript, illustrates that coaches tended to provide directive feedback about the observed simulation (TellBack moves) and directive suggestions for how teachers could improve (TellForward moves). In particular, coaches tended to use the moves included in Table 5a, the majority of which were explicitly included and emphasized in the coaching protocol. However, we also see that coaches employed some moves that were not specifically highlighted in the TeachSim protocol. Figure 2, for example, illustrates that coaches often emphasized relationship-building moves (Rapport moves), while Table 5a illustrates the specific Rapport moves used. At the same time, several of the moves explicitly emphasized in the coaching protocol were employed less frequently by coaches, as shown in Table 5b. Detailed definitions and examples for all move-groups and moves are included in Appendix A.

RQ2: How do the moves that coaches use vary across transcripts and what predicts this variation?

Using hierarchical agglomerative clustering methods to identify distinct profiles of coaching conversations allows us to move beyond considering individual moves in isolation to better understand how coaching conversations differed in terms of all the moves used. Our preferred clustering approach assigns each transcript to one of four profiles. Table 6 summarizes the key differences between these profiles in terms of the coaching moves used and includes descriptive statistics illustrating how move-group frequency varies across profiles. The first profile, which we refer to as "Modeling," consists of transcripts where coaches more frequently modeled how teachers might implement a particular redirection strategy (TellForward: Demonstration). The second profile, which we refer to as "Modeling Plus," consists of transcripts where coaches supplemented the TellForward: Demonstration move with providing feedback about what they noticed during the teacher's pre-coaching simulation (TellBack: Observation). The third profile, "Glow & Grow," consists of transcripts where coaches provided a "glow," or positive praise related to the teacher's pre-coaching simulation (TellBack: Positive Evaluation), and a "grow," a suggestion for a specific instructional strategy to implement in the next simulation (TellForward: Instructional Strategy). Finally, the "Rapport & Report" profile consists of transcripts where coaches made more frequent use relationship-building moves (Rapport moves) alongside positive praise and feedback on the teacher's pre-coaching simulation. While the first three profiles are distinguished by the relative emphasis on different components included in the TeachSim coaching protocol, the Rapport & Report profile, is distinguished by greater emphasis on relationship-building moves that were not explicitly included in the coaching protocol. Figure C2 includes the scatterplots used to identify these differences.

We then turn our attention to the predictors of profile membership to better understand the extent to which coaching moves vary across context, coaches, and teachers, as prior literature suggests we might expect (Collet, 2012; Deussen et al., 2007; Dusenbury et al., 2010; Gibbons & Cobb, 2016; Grierson & Woloshyn, 2013; Marsh et al., 2010, 2015; Stover et al., 2011; Tomlinson & McTighe, 2006). When we graph profile membership by site/cohort and coach, we do not see clear evidence that the transcripts from certain sites tend to be categorized into the same profiles (Figure 2), nor do we find that transcripts from certain coaches tend to be categorized into the same profiles (Figure 3). As Figures 2 and 3 illustrate, all four profiles are present across all sites and most coaches, albeit in different proportions. Given the randomization of transcripts to coders, we are also reassured to find that the coder is not a significant predictor of coaching profile, except for the Glow & Grow and Rapport & Report profiles (Table D3). We discuss this issue in more detail in Appendix D. Because we cannot rule out that differences in coder standards may contribute to whether a transcript is assigned to a particular profile, our preferred specifications for all regressions include coder fixed effects.

When we investigate the relationship between teacher characteristics and coaching profile (Table 7), we find some evidence that coaches may, consciously or unconsciously, emphasize different coaching moves for different kinds of teachers. Specifically, we find that teachers exposed to the Modeling profile report being approximately 1 SD more nervous at baseline than teachers exposed to any other profile (Table 7, columns 7-8). We also see some evidence that teachers exposed to the Rapport & Report profile tend to be perceived as more skillful by coaches (Coach Rating), though this relationship is not statistically significant in all models. We do not observe any other statistically significant differences in the characteristics of teachers that were exposed to different coaching profiles. Results are similar with and without coder fixed effects.

RQ3: What is the relationship between variation in the coaching moves used and teachers' observed instructional practice after participating in coaching?

Across all models, we find little evidence of significant differences in teachers' post-coaching observed instructional practice based on the coaching profiles, groups, and moves to which they were exposed. However, because of limited precision for some estimates, we cannot definitively interpret these results to mean that there is no relationship between the moves coaches use and teachers' instructional practice. Instead, we turn to the magnitude of the estimates to identify profiles, groups, and moves that may warrant further investigation. Tables 8-10 include our estimates of the relationship between coaching profile (Table 8), move-groups (Table 9), and individual moves (Tables 10), respectively.

We first begin by exploring regression-adjusted group-mean differences in outcomes, by profile (Table 8). The largest differences we see in teacher observation scores by coaching profile are the differences between the Modeling profile (the reference category) and the Rapport & Report profile and the Modeling profile and the Modeling Plus profile. In our preferred model (Table 8, column 6), we find that teachers exposed to the Rapport & Report and Modeling Plus profiles score approximately 0.35 SDs lower than teachers exposed to the Modeling profile, among teachers in the same site assigned to the same coach and coaching moves coder, and adjusted for pre-coaching score and coach rating. Though this difference is not statistically significant, it is non-trivial in magnitude. When we consider all four profiles together, we see that average teacher observation scores for the Modeling and Glow & Grow are similar, while scores for Modeling Plus and Rapport & Report are notably lower.

When we explore the relationship between move-groups and teacher observation outcomes (Table 9), the magnitude of these relationships is relatively small, corresponding to a difference of 9% of a standard deviation, at most. For instance, we find that for every additional AskBack move that appears in a transcript, teachers' post-coaching observation scores tend to be 0.093 SDs higher on average, among teachers in the same site assigned to the same coach and coaching moves coder, and adjusted for pre-coaching score (Table 9, column 5). These coefficients reflect the difference between a teacher exposed to one additional move compared with a teacher exposed to one less move. In practice, however, we observe much greater variation from one coaching session to the next than just one move. In fact, the SD of move-count ranges from a low of 1.09 moves per transcript (for the AskBack group of moves) to up to 4.92 moves per transcript (for the TellForward group of moves). This means that the coefficients estimated in Table 9 (columns 1-5) may understate the observed magnitude of the relationship between move frequency and teacher observation scores. To gain a more realistic understanding of this relationship, therefore, we multiply each coefficient from our preferred model (Table 8, column 5) by the SD of the move-group variable. As shown in Table 8, column 6, the largest association between a coaching move-group and post-coaching teacher observation scores is the relationship between the number of AskForward moves and post-coaching teacher observation scores. Specifically, a 1 SD increase in the number of AskForward moves (2.16) is associated with a 0.12 SD increase in post-coaching teacher observation scores.

When we explore the relationship between individual moves and post-coaching teacher observation scores (Table 9), we identify several moves that are positively related to post-coaching teacher observation scores and several moves that are negatively related to post-coaching scores.[10] First, we find that one additional Rapport: Empathy move is associated with a decrease of 0.14 SD in post-coaching teacher observation scores among teachers in the same site assigned to the same coach and coaching moves coder, and adjusted for pre-coaching score (Table 9, column 5). This is consistent with the fact that teachers exposed to the Modeling profile tend to receive lower post-coaching observation scores and this profile is the only one for which we see relatively less emphasis on the Rapport: Empathy move. However, we caution that inter-rater reliability for Rapport: Empathy is relatively low (0.89 for rater agreement and 0.42 for Krippendorf's alpha).

We also observe negative relationships between post-coaching teacher observation scores and the following moves: Rapport: Sharing, TellForward: Student Goal, and TellForward: Reinforcement, with coefficients ranging from 0.105-0.158 SD. Sharing moves reflect dialogue where a coach shares personal information about themselves, elicits personal information about the teacher, or demonstrates personal knowledge of the teacher, such as "The simulations are weird for me too!" or "When I taught 8th grade, I really struggled with redirecting off-task behavior." Reinforcement moves include feedback where a coach suggests that a teacher continue using an instructional strategy, such as "I really want you to continue maintaining that positive tone of voice in your next simulation." Finally, Student Goal moves include feedback where a coach articulates a goal for student learning or behavior for the teacher to work towards, such as "so we really want students to be engaged in the discussion throughout the lesson and

[10] We acknowledge that including more than 20 move variables together in one regression raises the concern of multiple hypothesis testing. When we estimate p-values using a Bonferroni correction, we find that no move coefficients are even marginally statistically significant. Post-estimation testing also indicates that we cannot reject the null hypothesis that all move coefficients are equal to zero. However, given the exploratory nature of these analyses, we privilege attention to the magnitude of the estimated relationships over the statistical significance.

minimize time off-task." Notably, none of these moves were highlighted as distinguishing features between the different coaching profiles, nor were they among the most commonly used moves.

Finally, we identify three moves that are positively associated with post-coaching observation scores with at least marginal statistical significance. First, we find that one additional Rapport: Encouragement move is associated with a 0.151 SD increase in post-coaching teacher observation scores. Encouragement moves consist of dialogue where a coach expresses positive expectations for the teacher's future instruction, such as "You've got this! You'll be great!" Second, we find that one additional TellForward: Implementation move is associated with a 0.119 SD increase in teacher observation scores. Implementation moves consist of specific suggestions for when a teacher should use a specific instructional strategy, such as "I would like you to try using this strategy next time you notice that student is off-task." Finally, we find that one additional TellBack: Interpretation move is associated with a 0.126 SD increase in teacher observation scores. Interpretation moves include feedback where coaches share a hypothesis or interpretive inference about what occurred during the teacher's first simulation, such as "when that student was tapping on the table, it suggested that he was distracted and not engaged in the lesson" or "you seemed a little frustrated at the end." As with the moves discussed above, none of these moves were highlighted as distinguishing features between the different coaching profiles, nor were they among the most commonly used moves.

Discussion

In providing a quantitative measure of coach discourse, coding coaching transcripts using the coaching moves taxonomy allows us to gain a detailed picture of coach-teacher interactions as part of the TeachSIM coaching model. We find, for example, that coaches tend to use coaching moves that were aligned to the standardized protocol they were expected to follow. We also identify additional moves, namely relationship-building Rapport moves, that were not a key part of the coaching protocol or theory of change yet were frequently used by coaches. Beyond understanding general patterns in coach-teacher interactions, we also gain insight into how these interactions vary. We find little evidence that different coaches have different coaching styles, instead finding that most of the variation stems from individual coaches using different patterns of moves (a.k.a coaching profiles) with different teachers. We also find relationships between the pattern of moves coaches use and teacher characteristics, suggesting that coaches may consciously or unconsciously adapt their discourse strategies for different kinds of teachers. Finally, we find that beyond variation in the use of Rapport moves, much of the remaining variation in coaching moves reflected differences in the relative emphasis on moves that were strongly aligned with the TeachSIM protocol.

For short, standardized and directive coaching programs similar to TeachSIM and implemented in similar contexts, these findings provide valuable insight into coaches' adherence to a specific coaching protocol and the kinds of adaptations and implementation challenges that might occur. This kind of insight can be helpful for proactively designing coach supports that can support implementation success. First, our findings suggest that coach supports like detailed coaching protocols can be helpful in shaping coaches' interactions with teachers. Second, our findings suggest that coaches attend to the relational aspects of their interactions, often employing relationship-building moves even when they are not emphasized in the coaching protocol. Rather than leaving coaches to figure out how to use these strategies on their own, this suggests that program implementers should consider the role relationship-building discourse has in their theory of change and provide coaches with explicit guidance in this regard. Finally, we note that coaches employed different patterns of moves for different teachers. This is notable given that the coaching protocol was designed to be highly standardized and encouraged differentiating coaching conversations by employing different focus areas rather than altering the discourse strategies used. This suggests the need to consider these different approaches to differentiation and provide coaches with explicit guidance on the kinds of differentiation program implementers would like to see. Furthermore, coaches would likely benefit from guidance as to the criteria they should consider in making decisions about how to differentiate their coaching for different teachers.

At the same time, we recognize that many of the coaching programs discussed in the research and implemented in practice diverge substantially from the TeachSIM model, offering longer, repeated coaching conversations with a variety of instructional foci and built around coaching protocols that have different theories of change and often less standardization (Cohen et al., 2020; J. Knight, 2009; Kraft et al., 2018). While the specific patterns we identify in our analyses may be less relevant to those programs, our analyses provide a valuable proof-of-concept for how the coaching moves taxonomy can be used to understand the content of coaching interactions. Researchers can use the tool to monitor implementation which can, in turn, inform ongoing implementation support and support the interpretation of program evaluation results. Coaching program managers can similarly use the tool to understand what their coaches are doing and provide tailored feedback. Because the taxonomy encompasses a broad definition of coaching and does not reflect a particular coaching model, it can be easily tailored to a variety of different programs. One need only identify the specific moves or groups of moves that are most aligned with the model or protocol of interest. While relevant groupings might align to the seven groups around which the taxonomy is structured (e.g. TellBack and

Rapport), it is also possible to create new groups of moves tailored to a particular coaching protocol's theory of change. Depending on the questions of interest, coaching transcripts can then be coded using the full taxonomy or just the specific moves that are aligned with the coaching model. Finally, researchers and program managers can investigate trends and variation in move frequency using similar methods to those we included in this paper.

In relying on human coders to read each transcript and identify each coaching move, we recognize that our approach to measuring coaching practice requires substantial time and resources to use at scale. In the context of well-funded, large-scale field trials we believe that this expenditure is likely to be feasible and worthwhile as it can both contribute to our understanding of how coaching is implemented and to the interpretation of the findings that result from the study. We also see this paper as the first step toward creating an automated Natural Language Processing-based tool for coding coaching transcripts. Once developed, this kind of automatic tool would be comparatively easy and inexpensive to apply.

As our RQ3 analyses show, we can also leverage variation in the coaching moves used to explore the relationship between this variation and teachers' instructional practice after participating in coaching. Given the limited time that coaches and teachers have for coaching conversations and the many choices coaching practitioners face for how that time can be spent, it's important that we develop a clearer understanding of the implications these different choices have for teachers' learning and development (Gibbons & Cobb, 2017; J. Knight, 2009; Kraft et al., 2018). By identifying moves that are associated with teachers' post-coaching observation scores we gain a partial sense of these implications and identify moves that may be especially fruitful for further investigation.

In this paper, we identify several moves that are positively associated with teachers' observation scores and several moves that are negatively associated with teachers' observation scores. Specifically, we find that teachers have stronger observation scores when coaches provide them with 1) more expressions of positive future expectations (Rapport: Encouragement), 2) more feedback on when to implement a particular instructional strategy (TellForward: Implementation), and 3) more discussions of the inferences coaches made about what happened during the teacher's first simulation (TellBack: Interpretation). At the same time, teachers have weaker observation scores when coaches 1) engage in more discussion of personal information about the coach or teacher (Rapport: Sharing), 2) provide more feedback on student goals the teacher should work towards (TellForward: Student Goal), and 3) provide more positive reinforcement for instructional strategies the teacher used in the pre-coaching simulation (TellForward: Reinforcement).

It is tempting to interpret these relationships as an indication of which coaching moves are more effective in supporting teacher development and which moves are less effective. Under this interpretation, we might be surprised that prior literature suggests that the moves exhibiting a negative relationship with teacher observation scores should, in fact, help support teacher development (Coburn & Woulfin, 2012; Collet, 2012; Heineke, 2013; Hoffman et al., 2015; J. Knight, 2007; L'Allier et al., 2010; Lowenhaupt et al., 2014; Matsumura et al., 2013; Perkins, 1998; Robertson et al., 2020; Russell et al., 2020; Sims et al., 2022; Teemant, 2014). However, we caution against interpreting our results this way for several reasons. First, our results lack statistical precision, meaning that the relationships we observe may result from sampling error instead of systematic patterns. Second, while we include a variety of controls in our preferred models, we cannot rule out the possibility that our results suffer from omitted variable bias. In particular, we worry the moves coaches use are influenced by unobserved teacher characteristics that are not accounted for by teachers' pre-coaching observation scores and are correlated with teachers' post-coaching observation scores, especially given existing evidence that coaches often adapt how they facilitate coaching conversations to teachers' unique needs (Collet, 2012; Dusenbury et al., 2010; Gibbons & Cobb, 2016; Grierson & Woloshyn, 2013; Stover et al., 2011; Tomlinson & McTighe, 2006). For example, coaches may have employed more Rapport: Sharing moves for teachers who were less engaged in the coaching conversation in attempt to generate engagement through personal connection. If this disengagement also makes teachers less likely to implement the coach's feedback, this could explain the negative relationship we observe between observation scores and Rapport: Sharing moves. Third, the relationship we observe may be specific to the TeachSIM context. For example, it's possible that dialogue focused on building a personal connection is simply less helpful for and less valued by teachers when they know that they will not have any future interactions with their coach. Unfortunately, our data do not allow us to evaluate these different hypotheses.

Finally, in a time-limited conversation, spending time on specific coaching moves necessarily reduces coaches' opportunity to use other coaching moves. A coach that uses more Rapport: Sharing, TellForward: Student Goal, or TellForward: Reinforcement moves may be forced to spend less time on other moves or run out of time for certain components of the coaching protocol altogether. However, when we estimate the pairwise correlations between each individual move, we find little evidence of systematic trade-offs. Additionally, the relatively low mean frequency for Rapport: Sharing and TellForward: Reinforcement, suggest that coaches may not have spent enough time on these moves to crowd-out other moves.

Our results highlight several avenues for future research. First, they reinforce the need for studies that directly compare coaching models that employ different discourse strategies with designs that allow us to confidently draw causal conclusions about the effects of these differences on teachers' development and instructional practice. The specific moves that we find to be related to teachers' observation scores in this study provide a helpful starting point for moves to investigate. Second, our results suggest the value of future work exploring how individual coaches tailor their interactions to different teachers' needs. Though this issue has been discussed in small-scale, qualitative studies, we have limited understanding of these processes at larger scales. If, as our results suggest, there is substantial within-coach variation in coaches' interactions with teachers, then understanding the sources of this variation is vital to developing a better understanding of the features of effective coaching and how they may vary for different teachers. Finally, and perhaps most importantly, our analysis illustrates how the coaching moves taxonomy can be used to make progress on these questions. Researchers can use the taxonomy to explore coaching discourse and how and why it varies in other contexts. The taxonomy can also be used to take advantage of existing variation in coaching practice to further investigate the implications this variation has for teacher development. Finally, researchers can use the taxonomy (and prior research generated from it) as a starting point for designing experimental comparisons of different coaching protocols and as a tool for monitoring implementation of those models. When individual studies use common terms and measures for describing coaching discourse, identifying patterns across studies and conducting meta-analyses will be considerably easier.

In the interim before these studies are conducted, our results have several implications for coaching practitioners. First, our results highlight the need for explicit attention to the time-limited nature of coaching conversations and the tradeoffs these limits create. Though many coaching conversations are longer than the 5-minute TeachSIM coaching conversations, few would argue that time is an abundant resource in schools (Kraft et al., 2018). Indeed, one persistent challenge for implementing coaching programs is finding time for coaches to meet with teachers at all (Bean et al., 2010). It is therefore vital that coaches are prepared to make explicit, informed choices about how they spend the limited time they have with teachers, considering not only what kinds of interactions might support a teacher's development, but also what kinds of interactions and discourse strategies may be most efficient or can be discarded to free up time for those that might be more effective at supporting teacher development.

Finally, coaching practitioners may want to experiment with the three coaching moves we found to be associated with stronger teacher observation scores. Given the limitations of our analyses, we cannot guarantee that using these moves will better support teacher development, but the observed positive relationship suggests they may be promising, especially given that these patterns are consistent with prior coaching research (Heineke, 2013; Hunt, 2016; Perkins, 1998; Robertson et al., 2020; Shernoff et al., 2015; Sims et al., 2022; Teemant, 2014). We do not, however, advocate that coaching practitioners avoid the coaching moves that we found to be associated with lower teacher observation scores, given the many alternative explanations for these findings and the robust literature base suggesting their positive effects on teacher learning.

Though coaching programs have demonstrated effects on teachers' instruction and student learning (Kraft et al., 2018), they require a cadre of highly skilled coaches who can meet regularly with teachers. This makes coaching logistically complex and resource intensive, especially compared to more traditional forms of professional development (D. S. Knight, 2012).

We need to provide coaches with a concrete understanding of effective coaching strategies to ensure that this commitment of resources will make a difference for students.

References

Ackland, R. (1991). A review of the peer coaching literature. *Journal of Staff Development*, *12*(1), 22–27.

Aguilar, E. (2013). *The art of coaching: Effective strategies for school transformation*. Jossey-Bass, A Wiley Brand.

Anglin, K. L., Wong, V. C., & Boguslav, A. (2021). A natural language processing approach to measuring treatment adherence and consistency using semantic similarity. *AERA Open*, *7*, 23328584211028616.

Barnhart, T., & van Es, E. (2015). Studying teacher noticing: Examining the relationship among pre-service science teachers' ability to attend, analyze and respond to student thinking. *Teaching and Teacher Education*, *45*, 83–93.

Bean, R. M., Draper, J. A., Hall, V., Vandermolen, J., & Zigmond, N. (2010). Coaches and Coaching in Reading First Schools: A Reality Check. *The Elementary School Journal*, *111*(1), 87–114. https://doi.org/10.1086/653471

Blazar, D., & Kraft, M. A. (2015). Exploring Mechanisms of Effective Teacher Coaching: A Tale of Two Cohorts From a Randomized Experiment. *Educational Evaluation and Policy Analysis*, *37*(4), 542–566. https://doi.org/10.3102/0162373715579487

Boerst, T., Sleep, L., Ball, D., & Bass, H. (2011). Preparing teachers to lead mathematics discussions. *Teachers College Record*, *113*(12), 2844–2877.

Boguslav, A., Cohen, J., Katz, V., Sadowski, K., Wiseman, E., & Wyckoff, J. (2022). *Implementing targeted professional development at scale in the District of Columbia Public Schools* [Manuscript submitted for publication].

Booker, L. N., & Russell, J. L. (2022). *Improving teaching practice with instructional coaching* (Design Principles Series). EdResearch for Recovery.

Cassidy, T., Mallett, C., & Tinning, R. (2008). Considering conceptual orientations of coach education research: A tentative mapping. *International Journal of Coaching Science*, *2*(2), 43–58.

Charney, R. S. (1993). *Teaching children to care: Management in the responsive classroom.* ERIC.

Clark, S. K., & Byrnes, D. (2012). Through the eyes of the novice teacher: Perceptions of mentoring support. *Teacher Development*, *16*(1), 43–54. https://doi.org/10.1080/13664530.2012.666935

Coburn, C. E., & Russell, J. L. (2008). District Policy and Teachers' Social Networks. *Educational Evaluation and Policy Analysis*, *30*(3), 203–235. https://doi.org/10.3102/0162373708321829

Coburn, C. E., & Woulfin, S. L. (2012). Reading Coaches and the Relationship Between Policy and Practice. *Reading Research Quarterly*, *47*(1), 5–30. https://doi.org/10.1002/RRQ.008

Coggins, C. T., Stoddard, P., & Cutler, E. (2003). *Improving Instructional Capacity through School-Based Reform Coaches.*

Cohen, J. (2015). Challenges in identifying high-leverage practices. *Teachers College Record*, *117*(7).

Cohen, J., Krishnamachari, A., & Wong, V. C. (2021). *Experimental Evidence on the Robustness of Coaching Supports in Teacher Education*. https://doi.org/10.26300/DGF9-CA95

Cohen, J., & Wiseman, E. (2023). Supporting Professional Learning at Scale: Evidence from the District of Columbia Public Schools. *Teachers College Record*, 01614681221147738.

Cohen, J., Wiseman, E., & Anglin, K. L. (In review). *Teacher supports for text-based instruction: Experimental evidence from simulations in teacher preparation.*

Cohen, J., Wong, V., Krishnamachari, A., & Berlin, R. (2020). Teacher coaching in a simulated environment. *Educational Evaluation and Policy Analysis*, *42*(2), 208–231.

Collet, V. S. (2012). The Gradual Increase of Responsibility Model: Coaching for Teacher Change. *Literacy Research and Instruction*, *51*(1), 27–47. https://doi.org/10.1080/19388071.2010.549548

Costa, A. L., & Garmston, R. J. (2002). *Cognitive coaching: A foundation for renaissance schools*. ERIC.

Costa Jr, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual.* Psychological Assessment Resources.

Croft, A., Coggshall, J. G., Dolan, M., & Powers, E. (2010). Job-Embedded Professional Development: What It Is, Who Is Responsible, and How to Get It Done Well. Issue Brief. *National Comprehensive Center for Teacher Quality*.

Davis, B., & Higdon, K. (2008). The Effects of Mentoring/Induction Support on Beginning Teachers' Practices in Early Elementary Classrooms (K-3). *Journal of Research in Childhood Education*, *22*(3), 261–274. https://doi.org/10.1080/02568540809594626

Denton, C. A., & Hasbrouck, J. (2009). A Description of Instructional Coaching and its Relationship to Consultation. *Journal of Educational and Psychological Consultation*, *19*(2), 150–175. https://doi.org/10.1080/10474410802463296

Desimone, L. M., & Pak, K. (2017). Instructional Coaching as High-Quality Professional Development. *Theory Into Practice*, *56*(1), 3–12. https://doi.org/10.1080/00405841.2016.1241947

Deussen, T., Coskie, T., Robinson, L., & Autio, E. (2007). " Coach" Can Mean Many Things: Five Categories of Literacy Coaches in Reading First. Issues & Answers. REL 2007-No. 005. *Regional Educational Laboratory Northwest*.

Domina, T., Lewis, R., Agarwal, P., & Hanselman, P. (2015). Professional Sense-Makers: Instructional Specialists in Contemporary Schooling. *Educational Researcher*, *44*(6), 359–364. https://doi.org/10.3102/0013189X15601644

Dozier, C. (2006). *Responsive literacy coaching: Tools for creating and sustaining purposeful change*. Stenhouse Publishers.

Dusenbury, L., Hansen, W. B., Jackson-Newsom, J., Pittman, D. S., Wilson, C. V., Nelson-Simley, K., Ringwalt, C., Pankratz, M., & Giles, S. M. (2010). Coaching to enhance quality of implementation in prevention. *Health Education*.

Ericsson, A., & Pool, R. (2016). *Peak: Secrets from the New Science of Expertise*. HarperCollins.

Galey, S. (2016). The evolving role of instructional coaches in US policy contexts. *The William & Mary Educational Review*, *4*(2), 11.

Gibbons, L. K., & Cobb, P. (2016). Content-Focused Coaching: Five Key Practices. *The Elementary School Journal*, *117*(2), 237–260. https://doi.org/10.1086/688906

Gibbons, L. K., & Cobb, P. (2017). Focusing on Teacher Learning Opportunities to Identify Potentially Productive Coaching Activities. *Journal of Teacher Education*, *68*(4), 411–425. https://doi.org/10.1177/0022487117702579

Grierson, A. L., & Woloshyn, V. E. (2013). Walking the talk: Supporting teachers' growth with differentiated professional learning. *Professional Development in Education*, *39*(3), 401–419.

Gross, P. A. (2010). Not Another Trend: Secondary-Level Literacy Coaching. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas*, *83*(4), 133–137. https://doi.org/10.1080/00098651003774844

Hardt, D., Nagler, M., & Rincke, J. (2020). *Can peer mentoring improve online teaching effectiveness? An rct during the covid-19 pandemic*.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Heineke, S. F. (2013). Coaching Discourse: Supporting Teachers' Professional Learning. *The Elementary School Journal*, *113*(3), 409–433. https://doi.org/10.1086/668767

Hiebert, J., Gallimore, R., & Stigler, J. W. (2002). A knowledge base for the teaching profession: What would it look like and how can we get one? *Educational Researcher*, *31*(5), 3–15.

Hiebert, J., & Morris, A. K. (2012). Teaching, Rather Than Teachers, As a Path Toward Improving Classroom Instruction. *Journal of Teacher Education*, *63*(2), 92–102. https://doi.org/10.1177/0022487111428328

Hoffman, J. V., Wetzel, M. M., Maloch, B., Greeter, E., Taylor, L., DeJulio, S., & Vlach, S. K. (2015). What can we learn from studying the coaching interactions between cooperating teachers and preservice teachers? A literature review. *Teaching and Teacher Education*, *52*, 99–112.

Huff, J., Preston, C., & Goldring, E. (2013). Implementation of a coaching program for school principals: Evaluating coaches' strategies and the results. *Educational Management Administration & Leadership*, *41*(4), 504–526.

Hunt, C. S. (2016). Getting to the heart of the matter: Discursive negotiations of emotions within literacy coaching interactions. *Teaching and Teacher Education*, *60*, 331–343. https://doi.org/10.1016/j.tate.2016.09.004

Hunter, S. B., & Springer, M. G. (2022). Critical Feedback Characteristics, Teacher Human Capital, and Early-Career Teacher Performance: A Mixed-Methods Analysis. *Educational Evaluation and Policy Analysis*, 01623737211062913.

Ippolito, J. (2010). Three Ways That Literacy Coaches Balance Responsive and Directive Relationships with Teachers. *The Elementary School Journal*, *111*(1), 164–190. https://doi.org/10.1086/653474

Johnson, K. G. (2016). Instructional Coaching Implementation: Considerations for K-12 Administrators. *Journal of School Administration Research and Development*, *1*(2), 37–40.

Joyce, B. R., & Showers, B. (1981). Transfer of training: The contribution of "coaching." *Journal of Education*, *163*(2), 163–172.

Kane, B. D., & Rosenquist, B. (2019). Relationships between instructional coaches' time use and district-and school-level policies and expectations. *American Educational Research Journal*, *56*(5), 1718–1768.

Kane, T., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (Measures of Effective Teaching Project). Bill & Melinda Gates Foundation.

Kennedy, M. (1999). The role of preservice teacher education. *Teaching as the Learning Profession: Handbook of Policy and Practice*, 54–85.

Kennedy, M. (2016). How does professional development improve teaching? *Review of Educational Research*, *86*(4), 945–980.

Killion, J. (2008). Are you coaching heavy or light. *Teachers Teaching Teachers*, *3*(8), 1–4.

Killion, J. (2012). Coaching in the K-12 context. *The SAGE Handbook of Mentoring and Coaching in Education*, 273–295.

King, D., Neuman, M., Pelchat, J., Potochnik, T., Rao, S., & Thompson, J. (2004). *Instructional Coaching: Professional Development Strategies that Improve Instruction*. Annenberg Institute at Brown University.

Knight, D. S. (2012). Assessing the cost of instructional coaching. *Journal of Education Finance*, 52–80.

Knight, J. (2007). *Instructional coaching: A partnership approach to improving instruction*. NSDC : Corwin Press.

Knight, J. (2009). *Coaching: Approaches and perspectives*. Corwin Press.

Kochmanski, N. (2020). *Aspects of High-Quality Mathematics Coaching: What Coaches Need to Know and Be Able to Do to Support Individual Teachers' Learning* [PhD Thesis].

Koh, S., & Neuman, S. B. (2006). Exemplary elements of coaching. *Unpublished Manuscript, University of Michigan, Ann Arbor*.

Kowal, J., & Steiner, L. (2007). Instructional Coaching. Issue Brief. *Center for Comprehensive School Reform and Improvement*.

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. https://doi.org/10.3102/0034654318759268

Kurz, A., Reddy, L. A., & Glover, T. A. (2017). A multidisciplinary framework of instructional coaching. *Theory into Practice*, *56*(1), 66–77.

Kutsyuruba, B., & Godden, L. (2019). The role of mentoring and coaching as a means of supporting the well-being of educators and students. *International Journal of Mentoring and Coaching in Education*, *8*(4), 229–234. https://doi.org/10.1108/IJMCE-12-2019-081

L'Allier, S., Elish-Piper, L., & Bean, R. M. (2010). What Matters for Elementary Literacy Coaching? Guiding Principles for Instructional Improvement and Student Achievement. *The Reading Teacher*, *63*(7), 544–554. https://doi.org/10.1598/RT.63.7.2

Lieberman, A. (1995). Practices that support teacher development: Transforming conceptions of professional learning. In F. Stevens (Ed.), *Innovating and Evaluating Science Education: NSF Evaluation Forums, 1992-94.* (pp. 67–78). Westat, Inc.

Lortie, D. C. (2002). *Schoolteacher: A sociological study*. University of Chicago Press.

Lowenhaupt, R., McKinney, S., & Reeves, T. (2014). Coaching in context: The role of relationships in the work of three literacy coaches. *Professional Development in Education*, *40*(5), 740–757.

Marsh, J. A., Bertrand, M., & Huguet, A. (2015). Using Data to Alter Instructional Practice: The Mediating Role of Coaches and Professional Learning Communities. *Teachers College Record*, 40.

Marsh, J. A., Bush-Mecenas, S., Strunk, K. O., Lincove, J. A., & Huguet, A. (2017). Evaluating Teachers in the Big Easy: How Organizational Context Shapes Policy Responses in New

Orleans. *Educational Evaluation and Policy Analysis*, *39*(4), 539–570. https://doi.org/10.3102/0162373717698221

Marsh, J. A., Sloan McCombs, J., & Martorell, F. (2010). How Instructional Coaches Support Data-Driven Decision Making: Policy Implementation and Effects in Florida Middle Schools. *Educational Policy*, *24*(6), 872–907. https://doi.org/10.1177/0895904809341467

Matsko, K. K., Ronfeldt, M., Nolan, H. G., Klugman, J., Reininger, M., & Brockman, S. L. (2020). Cooperating teacher as model and coach: What leads to student teachers' perceptions of preparedness? *Journal of Teacher Education*, *71*(1), 41–62. https://doi.org/10.1177/0022487118791992

Matsumura, L. C., Garnier, H. E., & Spybrook, J. (2013). Literacy coaching to improve student reading achievement: A multi-level mediation model. *Learning and Instruction*, *25*, 35–48. https://doi.org/10.1016/j.learninstruc.2012.11.001

McGatha, M. B. (2017). Elementary Mathematics Specialists: Ensuring the Intersection of Research and Practice. *North American Chapter of the International Group for the Psychology of Mathematics Education*.

Miles, M., Huberman, M., & Saldana, J. (2013). *SAGE: Qualitative Data Analysis: A Methods Sourcebook*. California: SAGE.

Morningside Center for Teaching Social Responsibility. (2012). *4Rs + MTP Coaching Handbook for Staff Developers*. University of Virginia and Fordham University.

Neufeld, B., & Roper, D. (2003). *Coaching: A strategy for developing instructional capacity*.

Niemi, H., Nevgi, A., & Aksit, F. (2016). Active learning promoting student teachers' professional competences in Finland and Turkey. *European Journal of Teacher Education*, *39*(4), 471–490.

Obara, S. (2010). Mathematics coaching: A new kind of professional development. *Teacher Development*, *14*(2), 241–251.

O'Connor, M. C., & Michaels, S. (1993). Aligning Academic Task and Participation Status through Revoicing: Analysis of a Classroom Discourse Strategy. *Anthropology <html_ent Glyph="@amp;" Ascii="&amp;"/> Education Quarterly*, *24*(4), 318–335. https://doi.org/10.1525/aeq.1993.24.4.04x0063k

Perkins, S. J. (1998). On becoming a peer coach: Practices, identities, and beliefs of inexperienced coaches. *Journal of Curriculum and Supervision*, *13*(3), 235–254.

Phipps, A. R., & Wiseman, E. A. (2021). Enacting the rubric: Teacher improvements in windows of high-stakes observation. *Education Finance and Policy*, *16*(2), 283–312.

Powell, D. R., & Diamond, K. E. (2013). Implementation fidelity of a coaching-based professional development program for improving Head Start teachers' literacy and language instruction. *Journal of Early Intervention*, *35*(2), 102–128.

Power, K., & Goodnough, K. (2019). Fostering teachers' autonomous motivation during professional learning: A self-determination theory perspective. *Teaching Education*, *30*(3), 278–298.

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researcher*, *29*(1), 4–15.

Reddy, L. A., Glover, T., Kurz, A., & Elliott, S. N. (2019). Assessing the effectiveness and interactions of instructional coaches: Initial psychometric evidence for the instructional coaching assessments–teacher forms. *Assessment for Effective Intervention*, *44*(2), 104–119.

Reinke, W., Sprick, R., & Knight, J. (2009). Coaching classroom management. In J. Knight (Ed.), *Coaching Approaches & Perspectives* (pp. 91–112). Corwin Press.

Reisman, A., Cipparone, P., Jay, L., Monte-Sano, C., Schneider Kavanagh, S., McGrew, S., & Fogo, B. (2019). Evidence of emergent practice: Teacher candidates facilitating historical discussions in their field placements. *Teaching and Teacher Education*, *80*, 145–156. https://doi.org/10.1016/j.tate.2018.12.014

Responsive Classroom. (2014). *The responsive classroom approach: Good teaching changes the future*. https://www.responsiveclassroom .org/sites/default/files/pdf_files/RC_approach_ White_paper.pdf

Richardson, G., Yost, D., Conway, T., Magagnosc, A., & Mellor, A. (2020). Using instructional coaching to support student teacher-cooperating teacher relationships. *Action in Teacher Education*, *42*(3), 271–289.

Robertson, D. A., Ford-Connors, E., Frahm, T., Bock, K., & Paratore, J. R. (2020). Unpacking productive coaching interactions: Identifying coaching approaches that support instructional uptake. *Professional Development in Education*, *46*(3), 405–423. https://doi.org/10.1080/19415257.2019.1634628

Ronfeldt, M., Bardelli, E., Truwit, M., Mullman, H., Schaaf, K., & Baker, J. C. (2020). Improving Preservice Teachers' Feelings of Preparedness to Teach Through Recruitment of Instructionally Effective and Experienced Cooperating Teachers: A Randomized Experiment. *Educational Evaluation and Policy Analysis*, *42*(4), 551–575. https://doi.org/10.3102/0162373720954183

Ronfeldt, M., Brockman, S. L., & Campbell, S. L. (2018). Does cooperating teachers' instructional effectiveness improve preservice teachers' future performance? *Educational Researcher*, *47*(7), 405–418.

Ronfeldt, M., & Reininger, M. (2012). More or better student teaching? *Teaching and Teacher Education*, *28*(8), 1091–1106. https://doi.org/10.1016/j.tate.2012.06.003

Russell, J. L., Correnti, R., Stein, M. K., Bill, V., Hannan, M., Schwartz, N., Booker, L. N., Pratt, N. R., & Matthis, C. (2020). Learning From Adaptation to Support Instructional Improvement at Scale: Understanding Coach Adaptation in the TN Mathematics Coaching Project. *American Educational Research Journal*, *57*(1), 148–187. https://doi.org/10.3102/0002831219854050

Santagata, R., & Angelici, G. (2010). Studying the impact of the lesson analysis framework on preservice teachers' abilities to reflect on videos of classroom teaching. *Journal of Teacher Education*, *61*(4), 339–349.

Shabani, K., Khatib, M., & Ebadi, S. (2010). Vygotsky's zone of proximal development: Instructional implications and teachers' professional development. *English Language Teaching*, *3*(4), 237–248.

Shernoff, E. S., Lakind, D., Frazier, S. L., & Jakobsons, L. (2015). Coaching Early Career Teachers in Urban Elementary Schools: A Mixed-Method Study. *School Mental Health*, *7*(1), 6–20. https://doi.org/10.1007/s12310-014-9136-6

Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S., Stansfield, C., Goodrich, J., Van Herwegen, J., & Anders, J. (2022). *Effective teacher professional development: New theory and a meta-analytic test*. UCL Centre for Education Policy and Equalising Opportunities.

Sisson, D., & Sisson, B. (2017). *The literacy coaching handbook: Working with teachers to increase student achievement*. Routledge, Taylor & Francis Group.

Stanulis, R., & Floden, R. E. (2009). Intensive Mentoring as a Way to Help Beginning Teachers Develop Balanced Instruction. *Journal of Teacher Education*, *60*(2), 112–122. https://doi.org/10.1177/0022487108330553

Stover, K., Kissel, B., Haag, K., & Shoniker, R. (2011). Differentiated coaching: Fostering reflection with teachers. *The Reading Teacher*, *64*(7), 498–509.

Teemant, A. (2014). A Mixed-Methods Investigation of Instructional Coaching for Teachers of Diverse Learners. *Urban Education*, *49*(5), 574–604. https://doi.org/10.1177/0042085913481362

Terehoff, I. I. (2002). Elements of adult learning in teacher professional development. *NASSP Bulletin*, *86*(632), 65–77.

Toll, C. (2009). Literacy Coaching. In J. Knight (Ed.), *Coaching Approaches & Perspectives* (pp. 56–69). Corwin Press.

Tomlinson, C. A., & McTighe, J. (2006). *Integrating differentiated instruction & understanding by design: Connecting content and kids*. ASCD.

Tung, R., Ouimette, M., & Feldman, J. (2004). The Challenge of Coaching: Providing Cohesion among Multiple Reform Agendas. *Center for Collaborative Education*.

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard university press.

Walpole, S., & McKenna, M. C. (2012). *The literacy coach's handbook: A guide to research-based practice*. Guilford Press.

Warford, M. K. (2011). The zone of proximal teacher development. *Teaching and Teacher Education*, *27*(2), 252–258.

Wayne, A. J., & Coggshall, J. G. (2022). How to ensure high-quality instructional coaching at scale. *Phi Delta Kappan*, *103*(5), 42–46.

West, L. (2009). Content coaching: Transforming the teaching profession. In J. Knight (Ed.), *Coaching Approaches & Perspectives* (pp. 113–144). Corwin Press.

Winch, C., Oancea, A., & Orchard, J. (2015). The contribution of educational research to teachers' professional learning: Philosophical understandings. *Oxford Review of Education*, *41*(2), 202–216.

Woulfin, S. L. (2018). Mediating Instructional Reform: An Examination of the Relationship Between District Policy and Instructional Coaching. *AERA Open*, *4*(3), 233285841879227. https://doi.org/10.1177/2332858418792278

Woulfin, S. L., & Rigby, J. G. (2017). Coaching for Coherence: How Instructional Coaches Lead Change in the Evaluation Era. *Educational Researcher*, *46*(6), 323–328. https://doi.org/10.3102/0013189X17725525

Tables

Table 1. Analytic sample by site and cohort.

Site and Cohort	Number of Transcripts & Teachers	Number of Coaches
Site 1, Cohort 1	45	6
Site 1, Cohort 2	53	8
Site 1, Cohort 3	31	4
Site 2	37	4
Site 3	20	3
Total Unique	186	13

Note: Many coaches provided coaching for multiple sites/cohorts, which is why the total number of unique coaches is less than the sum of the number of coaches in each site/cohort combination.

Table 2. Descriptive statistics reflecting transcript duration and number of words spoken by coaches (N=186).

	Mean	SD	Min	Max
Total number of words per transcript	915.96	160.71	483	1409
Total number of coach words per transcript	727.51	129.09	429	1230
Proportion of total words spoken by coach per transcript	0.80	0.78	0.56	0.97
Transcript duration (in minutes)	5.22	0.85	3.2	9.78
Words per coach turn-of-talk	48.12	67.98	1	669
Words per teacher turn-of-talk	12.46	23.09	1	463
Total number of coach turns-of-talk	15.12	8.34	4	53
Number of moves per transcript	37.65	9.28	21	76
Number of unique moves per transcript	33.78	9.02	20	72

Table 3. Descriptive statistics each move and move-group, ordered from most frequently used to least frequently used.

		Prevalence of Move	Count of the Number of Times Move Occurs per Transcript				Inter-rater Reliability of Move Coding	
Group	Move	Proportion of Transcripts	Mean	SD	Min	Max	Percent Agreement	Krippendorf's Alpha
TellBack	observation	0.98	3.75	1.86	0	11	0.91	0.78
TellForward	demonstration	0.98	3.44	2.29	0	12	0.89	0.78
TellForward	instructional strategy	0.99	3.18	1.99	0	11	0.86	0.56
TellBack	positive evaluation	0.99	3.12	1.67	0	10	0.95	0.85
Activity	practice	0.94	2.86	1.62	0	9	0.96	0.86
Rapport	agenda setting	0.93	2.78	1.69	0	9	0.92	0.73
TellForward	student goal	0.95	2.32	1.45	0	8	0.91	0.68
TellForward	challenge	0.84	1.92	1.57	0	7	0.92	0.61
Rapport	empathy	0.83	1.76	1.42	0	10	0.89	0.42
Other	non-coaching	0.77	1.71	1.78	0	15	0.95	0.72
TellForward	implementation	0.73	1.28	1.08	0	6	0.91	0.50
Other	no moves	0.53	0.99	1.26	0	6	0.97	0.83
Rapport	encouragement	0.53	0.83	1.00	0	4	0.96	0.55
AskForward	check-for-understanding	0.45	0.78	1.28	0	10	0.97	0.70
Rapport	assistance	0.66	0.78	0.66	0	3	0.99	0.89
Rapport	normalizing struggle	0.45	0.63	0.85	0	4	0.97	0.55
TellBack	connection	0.39	0.63	0.95	0	5	0.95	0.18
TellBack	interpretation	0.45	0.63	0.91	0	6	0.97	0.51
Rapport	sharing	0.33	0.62	1.05	0	5	0.98	0.71
AskForward	application	0.33	0.59	1.16	0	6	0.97	0.67
Rapport	revoicing	0.40	0.54	0.79	0	4	0.95	0.24
Other	unclear	0.34	0.53	0.94	0	6	0.96	0.41
AskBack	self-assessment	0.45	0.52	0.63	0	2	0.97	0.44
AskBack	noticing	0.38	0.48	0.71	0	4	0.96	0.35
TellForward	reinforcement	0.35	0.48	0.75	0	4	0.97	0.58
AskForward	anticipation	0.41	0.45	0.56	0	2	0.99	0.79

		Prevalence of Move	Count of the Number of Times Move Occurs per Transcript				Inter-rater Reliability of Move Coding	
Group	Move	Proportion of Transcripts	Mean	SD	Min	Max	Percent Agreement	Krippendorf's Alpha
Other	other coaching	0.22	0.27	0.55	0	3	0.97	0.24
TellBack	negative evaluation	0.22	0.25	0.52	0	2	0.98	0.00
Rapport	mirroring	0.18	0.22	0.50	0	3	0.98	0.22
AskBack	cause & effect	0.05	0.07	0.35	0	3	1.00	0.00
Rapport	permission	0.04	0.05	0.24	0	2	n/a	n/a
AskBack	interpretation	0.04	0.04	0.20	0	1	1.00	0.00
AskBack	vision	0.04	0.04	0.23	0	2	1.00	0.50
AskForward	goal setting	0.03	0.04	0.24	0	2	1.00	0.67
Rapport	coaching feedback	0.03	0.03	0.16	0	1	n/a	n/a
AskBack	justification	0.02	0.02	0.13	0	1	n/a	n/a
AskForward	generation	0.02	0.02	0.15	0	1	n/a	n/a
TellForward	all	1.00	12.62	4.92	4	28	0.77	0.81
TellBack	all	1.00	8.38	3.25	2	20	0.86	0.84
Rapport	all	1.00	8.24	3.67	2	25	0.78	0.69
Other	all	0.94	3.50	2.59	0	15	0.88	0.69
Activity	all	0.94	2.86	1.62	0	9	0.96	0.86
AskForward	all	0.88	1.88	2.16	0	16	0.96	0.82
AskBack	all	0.72	1.17	1.09	0	5	0.95	0.54
Overall							Moves: 0.96 Groups: 0.88	Moves: 0.53 Groups: 0.76

Table 4. Description of additional teacher characteristics included in analyses.

	Description	Timing of Data Collection	Score Range	Source
Coach Rating	Coach's rating of teachers' skill in redirecting off-task student behavior during the pre-coaching simulation. Coaches were not provided with a specific observation rubric to use when selecting a score. Instead, coaches were simply asked to rate how well they felt the teacher redirected off-task behavior during the pre-coaching simulation.	Immediately after the pre-coaching simulation	1-10	Researcher-created
Self-Efficacy	Teacher Sense of Efficacy Scale	Before the pre-coaching simulation	1-9	Tschannen-Moran & Hoy, 2001
Nervousness	Teacher's level of agreement with the statement "I felt nervous or anxious in the simulator."	Before the pre-coaching simulation, after engaging in an initial practice simulation	Likert: 1-5	Researcher-created

Note: The coach rating variable differs from the pre-coaching teacher observation score in that 1) the rating was completed by coaches instead of trained, external raters and 2) coaches were not provided with a standardized observation rubric to use when selecting a score.

Table 5a. Definitions and role in the TeachSim coaching protocol for moves and move-groups with the highest mean frequency.

Move or Move-Group	Definition	Emphasized in the coaching protocol
TellForward	Moves that are focused on the teacher's future instruction and provide the teacher with information, including the coach's opinions, analysis, and reflections.	Yes
TellBack	Moves that are focused on the teacher's prior instruction and provide the teacher with information, including the coach's opinions, analysis, and reflections.	Yes
Rapport	Moves that support the establishment of a positive coach-teacher relationship.	No
TellBack: Observation	Feedback that describes specific factual information about students, a lesson, or the teacher based on the coach's observation of prior instruction or general familiarity with the students or teacher's instructional practice.	Yes
TellForward: Demonstration	Dialogue where the coach explicitly illustrates *how* a specific instructional strategy can be used or implemented.	Yes
TellForward: Instructional Strategy	Feedback that explicitly proposes a new strategy that a teacher can or should use. This change can include using or tweaking a strategy that the teacher has previously used but needs to improve or has not yet applied to the specific situation.	Yes
TellBack: Positive Evaluation	Feedback that communicates a positive judgment about a teacher's general skill as a teacher, specific elements of the teacher's practice, or provides a general affirmation of the teacher's instruction in a specific lesson or time-period.	Yes
Activity: Practice	Dialogue where the coach initiates and facilitates a role-play activity.	Yes

Move or Move-Group	Definition	Emphasized in the coaching protocol
Rapport: Agenda-Setting	Providing an explanation of the purpose or agenda for coaching, including coaching as an ongoing program, an individual coaching conversation, or a specific part of an individual coaching conversation.	No
TellForward: Student Goal	Feedback that articulates a goal or outcome for students for the teacher to work towards.	Yes
TellForward: Challenge	Feedback that articulates a challenge or problem of teaching.	Yes
Rapport: Empathy	Dialogue in which the coach asks about, anticipates, or expresses an understanding of the teacher's emotions or perspective.	No
Other: Non-Coaching	Dialogue that is clearly not part of a coaching conversation, including logistical issues, technology issues, timing issues, or background dialogue from speakers not involved in the coaching conversation.	No

Table 5b. Definitions for the moves explicitly emphasized in the TeachSIM coaching protocol, but not included in Table 4a because they were rarely used on average.

Move or Group	Definition
AskBack: Noticing	Questioning that asks the teacher to recall information about themselves, a lesson, or their students based on prior experiences or their general familiarity with themselves or their students.
AskBack: Self-Assessment	Questioning that asks the teacher to make a judgement about the success and quality of their own instructional practice.
AskForward: Check-for-Understanding	Questioning that checks for a teacher's understanding of a pedagogical strategy or other professional concepts that the coach or teacher have been discussing. These questions tend to require the teacher to synthesize or apply previously discussed content in order to answer them.
AskForward: Anticipation	Questioning that explicitly prompts the teacher to elaborate on the consequences of an instructional strategy, classroom situation, or goal, including the importance or purpose, potential negative consequences, or challenges the teacher may face in using a strategy.
Askforward: Application	Questioning that prompts the teacher to decide when and/or how to apply a specific instructional strategy.
TellForward: Implementation	Dialogue where the coach provides a direction or suggestion for *when* a teacher should use a strategy in a future situation or how to prepare for using a strategy in a future lesson.
Rapport: Encouragement	Dialogue where the coach expresses positive expectations for the teacher's future work.

Table 6. Characterizing the four coaching profiles suggested by the hierarchical agglomerative clustering analysis.

	Modeling	Modeling Plus	Glow & Grow	Rapport & Report
More frequent moves (relative to other profiles)	TellForward: demonstration	TellForward: demonstration TellBack: observation	TellBack: positive evaluation TellForward: instructional strategy	Rapport: agenda-setting and empathy TellBack: observation and positive evaluation
Less frequent moves (relative to other profiles)	Rapport: agenda-setting, empathy, and encouragement TellBack: positive evaluation	Rapport: agenda-setting and encouragement	Rapport: agenda-setting and mirroring	TellForward: demonstration and instructional strategy
AskBack Count	1.49	1.12	1.13	1.04
(mean and sd)	(1.31)	(1.05)	(1.04)	(0.98)
TellBack Count	7.63	8.30	8.90	8.22
(mean and sd)	(3.12)	(3.46)	(3.41)	(2.93)
AskForward Count	2.89	1.36	1.88	1.51
(mean and sd)	(3.96)	(0.99)	(1.61)	(1.23)
TellForward Count	17.00	12.61	12.86	9.16
(mean and sd)	(5.18)	(3.65)	(4.83)	(2.43)
Rapport Count	7.60	6.36	8.75	9.24
(mean and sd)	(3.71)	(3.19)	(3.86)	(3.18)
Activity Count	2.97	2.30	3.12	2.80
(mean and sd)	(1.69)	(1.36)	(1.64)	(1.67)
Other Count	1.23	0.58	1.03	1.06
(mean and sd)	(1.55)	(0.79)	(1.11)	(1.43)
No Moves Count	2.34	3.27	2.51	2.10
(mean and sd)	(2.01)	(3.25)	(1.61)	(1.58)
N	35	33	69	49

Table 7. Regression estimates of the relationship between coaching profile and teacher baseline characteristics. Reference category is the Modeling coaching profile.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Coaching Rating (Standardized)		Pre-Coaching Score (Standardized)		Self-Efficacy (Standardized)		Nervousness (Standardized)	
Modeling Plus	-0.147	-0.124	0.146	0.172	-0.187	-0.218	1.072**	1.081**
	(0.260)	(0.265)	(0.203)	(0.205)	(0.450)	(0.422)	(0.389)	(0.403)
Rapport & Report	0.433	0.516+	0.0716	0.129	-0.466	-0.373	0.905*	0.928*
	(0.289)	(0.304)	(0.204)	(0.218)	(0.365)	(0.388)	(0.353)	(0.357)
Glow & Grow	0.173	0.152	0.329	0.328	0.0339	0.0471	1.004**	0.994**
	(0.254)	(0.253)	(0.207)	(0.208)	(0.277)	(0.291)	(0.320)	(0.323)
Coach Fixed Effects	X	X	X	X	X	X	X	X
Coder Fixed Effects		X		X		X		X
Constant	-0.655*	-0.619*	0.903	0.923	0.0927	0.0790	-0.740*	-0.720*
	(0.261)	(0.266)	(1.056)	(1.120)	(0.315)	(0.358)	(0.314)	(0.329)
Observations	129	129	181	181	95	95	95	95
R-squared	0.288	0.295	0.189	0.193	0.184	0.242	0.162	0.163

Robust standard errors in parentheses

** p<0.01, * p<0.05, + p<0.10

Table 8. Regression-adjusted differences in post-coaching teacher observation scores, by coaching profile. Reference category is the Modeling coaching profile.

	(1)	(2)	(3)	(4)	(5)	(6)
	Standardized Post-Coaching Observation Score					
Modeling Plus	-0.237	-0.262	-0.196	-0.197	-0.189	-0.362
	(0.255)	(0.239)	(0.209)	(0.217)	(0.220)	(0.250)
Rapport & Report	-0.130	-0.170	-0.284	-0.237	-0.214	-0.360
	(0.203)	(0.200)	(0.189)	(0.194)	(0.209)	(0.230)
Glow & Grow	-0.0897	-0.164	-0.163	-0.0257	-0.0266	-0.0722
	(0.202)	(0.201)	(0.197)	(0.201)	(0.202)	(0.229)
Pre-Coaching Score		X	X	X	X	X
Cohort/site Fixed Effects			X	X	X	X
Coach Fixed Effects				X	X	X
Coder Fixed Effects					X	X
Coach Rating						X
Constant	0.109	0.164	0.525*	1.004**	1.015**	-0.288
	(0.152)	(0.146)	(0.231)	(0.311)	(0.295)	(0.299)
Observations	176	174	174	174	174	122
R-squared	0.006	0.046	0.188	0.328	0.328	0.359

Robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.10

Table 9. Regression estimates of the relationship between move-group count and post-coaching observation score.

	(1)	(2)	(3)	(4)	(5)	(6)
	Standardized Post-Coaching Observation Score					
AskBack Count	0.065	0.080	0.084	0.088	0.093	0.101
	(0.076)	(0.074)	(0.067)	(0.069)	(0.071)	
TellBack Count	0.021	0.008	0.002	0.004	-0.003	-0.010
	(0.022)	(0.021)	(0.019)	(0.020)	(0.022)	
AskForward Count	0.083**	0.078*	0.060+	0.052	0.057	0.123
	(0.031)	(0.032)	(0.032)	(0.052)	(0.052)	
TellForward Count	-0.010	-0.006	0.002	0.006	0.002	0.010
	(0.015)	(0.015)	(0.016)	(0.016)	(0.017)	
Rapport Count	0.010	0.010	-0.019	-0.026	-0.028	-0.103
	(0.020)	(0.020)	(0.021)	(0.022)	(0.023)	
Activity Count	0.016	0.017	0.043	0.040	0.041	0.066
	(0.048)	(0.047)	(0.045)	(0.050)	(0.051)	
Other	0.059*	0.048	0.020	0.015	0.015	0.039
	(0.029)	(0.031)	(0.030)	(0.028)	(0.028)	
Pre-Coaching Score		X	X	X	X	X
Cohort/site Fixed Effects			X	X	X	X
Coach Fixed Effects				X	X	X
Coder Fixed Effects					X	X
Coefficients are multiplied by SD of predictor variable						X
Constant	-0.615*	-0.522+	-0.055	0.638+	0.792+	
	(0.297)	(0.287)	(0.331)	(0.383)	(0.406)	
Observations	176	174	174	174	174	174
R-squared	0.093	0.116	0.229	0.342	0.344	

Robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.10

Note: In column 6, the coefficients of interest from column 5 have been multiplied by the SD of the corresponding move-group variable to provide a more realistic picture of the magnitude of these relationships within our sample of transcripts.

Table 10. Regression estimates of the relationship between individual move count and post-coaching observation score.

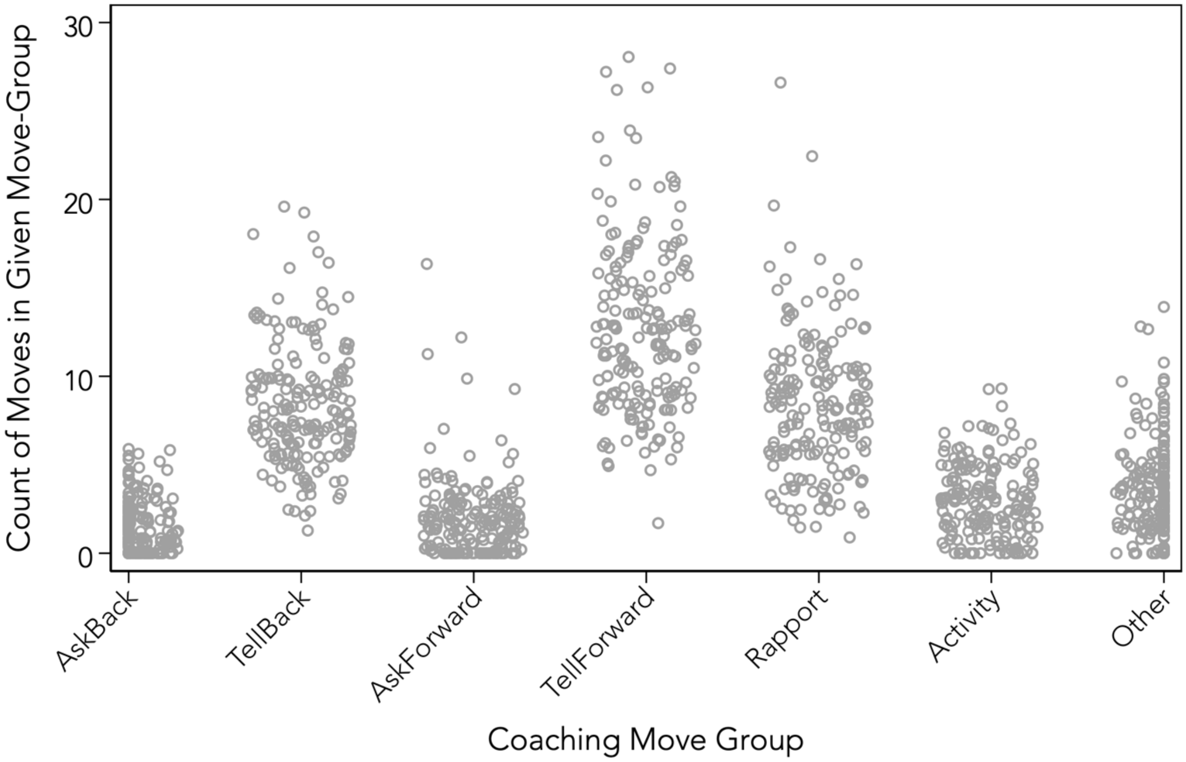
	(1)	(2)	(3)	(4)	(5)
			Standardized Post-Coaching Observation Score		
		Rapport			
Encouragement	0.029	0.047	0.006	0.141	0.154+
	(0.085)	(0.085)	(0.085)	(0.086)	(0.090)
Assistance	-0.102	-0.126	-0.119	-0.187	-0.180
	(0.133)	(0.134)	(0.121)	(0.125)	(0.129)
Empathy	-0.023	-0.037	-0.095	-0.120+	-0.136*
	(0.069)	(0.070)	(0.069)	(0.066)	(0.068)
Agenda-Setting	0.066	0.060	0.048	-0.011	-0.017
	(0.047)	(0.048)	(0.044)	(0.048)	(0.048)
Revoicing	-0.034	-0.036	0.015	0.012	0.022
	(0.12)	(0.12)	(0.11)	(0.098)	(0.100)
Normalizing Struggle	0.012	0.037	-0.019	-0.027	-0.012
	(0.103)	(0.106)	(0.099)	(0.093)	(0.095)
Sharing	-0.060	-0.040	-0.085	-0.153+	-0.158+
	(0.075)	(0.076)	(0.082)	(0.081)	(0.084)
		TellForward			
Implementation	0.018	0.036	0.080	0.134+	0.119
	(0.071)	(0.074)	(0.072)	(0.077)	(0.079)
Challenge	-0.012	-0.008	0.014	0.019	0.026
	(0.053)	(0.056)	(0.055)	(0.055)	(0.057)
Student Goal	-0.084	-0.055	-0.051	-0.096	-0.105+
	(0.052)	(0.054)	(0.053)	(0.060)	(0.062)
Reinforcement	-0.049	-0.088	-0.213+	-0.187	-0.196
	(0.117)	(0.121)	(0.116)	(0.129)	(0.132)
Demonstration	0.018	0.020	0.015	0.012	0.013
	(0.034)	(0.035)	(0.033)	(0.037)	(0.037)
Instructional Strategy	-0.001	-0.009	0.028	0.048	0.055
	(0.055)	(0.056)	(0.050)	(0.048)	(0.049)
		TellBack			
Observation	-0.017	-0.016	0.036	0.0077	0.0087
	(0.054)	(0.053)	(0.047)	(0.051)	(0.053)
Positive Evaluation	0.055	0.023	-0.024	-0.027	-0.016
	(0.048)	(0.048)	(0.043)	(0.056)	(0.056)
Interpretation	0.145+	0.160*	0.166*	0.115	0.126
	(0.078)	(0.078)	(0.074)	(0.080)	(0.085)
Connection	0.047	0.044	0.001	0.034	0.046
	(0.081)	(0.078)	(0.070)	(0.067)	(0.083)

		AskForward			
Application	0.059	0.055	0.013	0.034	0.033
	(0.076)	(0.080)	(0.081)	(0.113)	(0.113)
Anticipation	-0.018	0.017	-0.11	-0.039	-0.053
	(0.144)	(0.147)	(0.167)	(0.168)	(0.175)
Check for Understanding	0.127*	0.103	0.075	0.036	0.027
	(0.064)	(0.066)	(0.070)	(0.075)	(0.075)
		AskBack			
Noticing	-0.038	0.020	0.002	0.094	0.084
	(0.117)	(0.118)	(0.126)	(0.114)	(0.118)
Self-Assessment	0.081	0.092	0.110	0.055	0.064
	(0.115)	(0.115)	(0.116)	(0.116)	(0.119)
		Additional Moves			
Activity: Practice	-0.001	0.006	0.031	0.012	0.012
	(0.060)	(0.059)	(0.057)	(0.058)	(0.058)
Other: No Moves	-0.034	-0.042	-0.112+	-0.069	-0.070
	(0.069)	(0.071)	(0.066)	(0.072)	(0.072)
Other: Non-Coaching	0.041	0.036	0.031	-0.003	0.002
	(0.043)	(0.044)	(0.037)	(0.039)	(0.041)
Other: Unclear	0.116	0.120	0.066	0.104	0.096
	(0.105)	(0.109)	(0.094)	(0.094)	(0.097)
Pre-Coaching Score		X	X	X	X
Cohort/site Fixed Effects			X	X	X
Coach Fixed Effects				X	X
Coder Fixed Effects					X
Constant	-0.352	-0.731	0.069	1.149+	1.081
	(0.443)	(0.488)	(0.502)	(0.605)	(0.672)
Observations	176	174	174	174	174
R-squared	0.150	0.169	0.311	0.428	0.432

Robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.10

Figures

Figure 1. Distribution of the frequency of each move-group across transcripts.



Note: Within each move-group category (x-axis), each grey, circular marker represents one transcript. This means that each transcript appears seven times here, once for each move-group.

Figure 2. Coaching profiles by cohort/site.

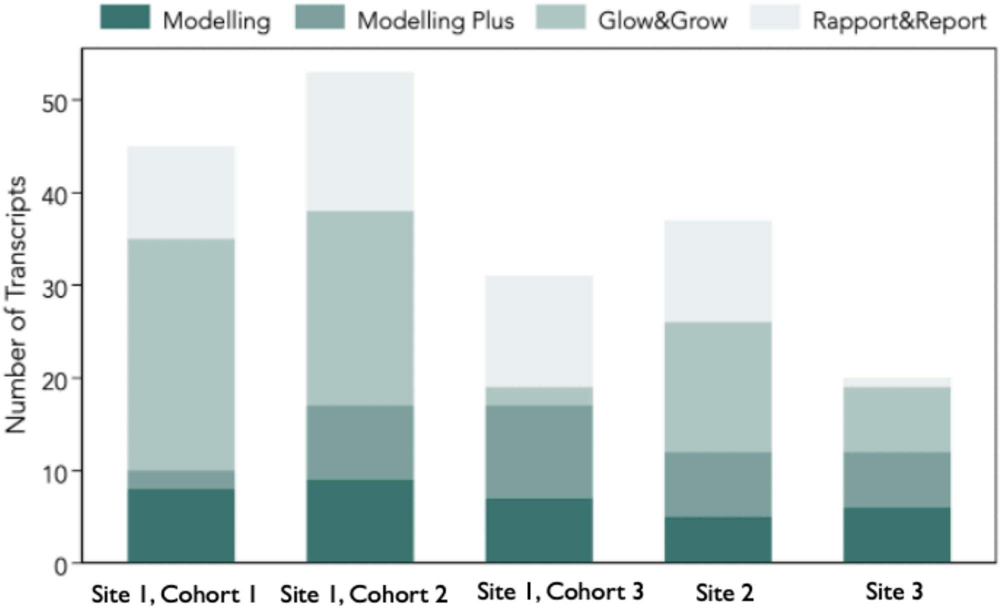
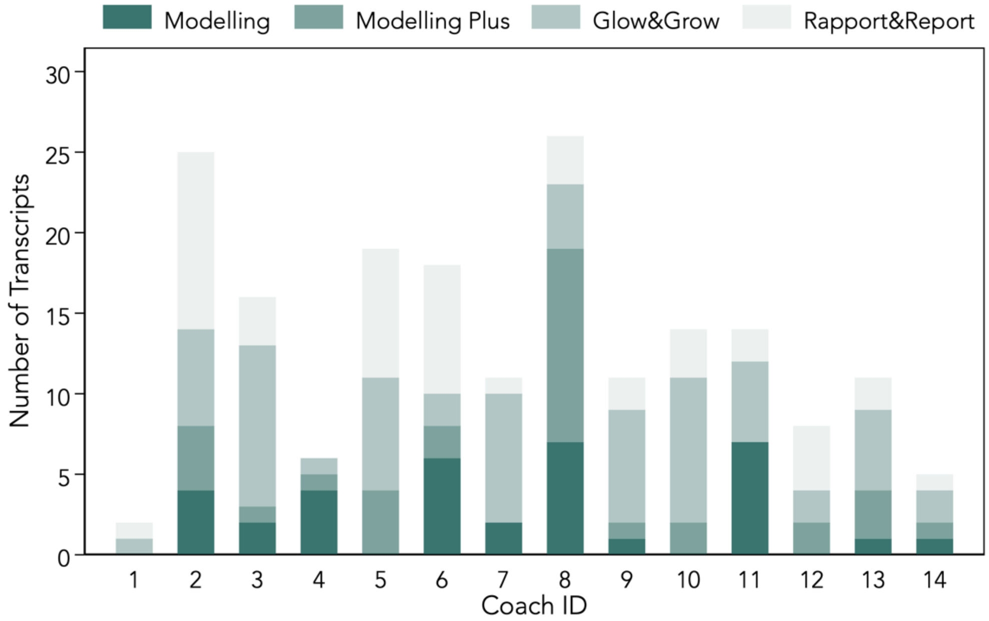


Figure 3. Coaching profiles by coach



Appendices

Appendix A: Coaching Moves Codebook

General Directions

This framework is focused on understanding the moves that coaches use to support teachers' instructional practice. This means that the focus is on coding coach dialogue that might meaningfully inform how a teacher teaches in a real or simulated classroom environment. Instructions are included below for how to code dialogue that covers logistics or other issues that do not directly relate to how a teacher can support student learning.

You should only code coach dialogue.

You should code one turn of talk at a time. A turn of talk is defined as all the coach dialogue that comes between two instances of teacher dialogue (see the example below).

> Coach: [00:00:00] So, UM how do you think that went? [00:00:02].
>
> Teacher Candidate: [00:00:04] Uh, I think it went better than the first time I did it. [00:00:06].
>
> Coach: [00:00:07] It was the first time that you did the management one. [00:00:08].
>
> Teacher Candidate: [00:00:09] Oh, yes. [00:00:09].

The dialogue highlighted in blue represents the first coach turn of talk. The dialogue highlighted in grey represents the second coach turn of talk. In some transcripts, a single turn of talk may have multiple "paragraphs."

Multiple "moves" may be present in a single turn of talk. In general, you should list the "moves" in the order that they appear in the turn of talk, including if there are repeated moves.

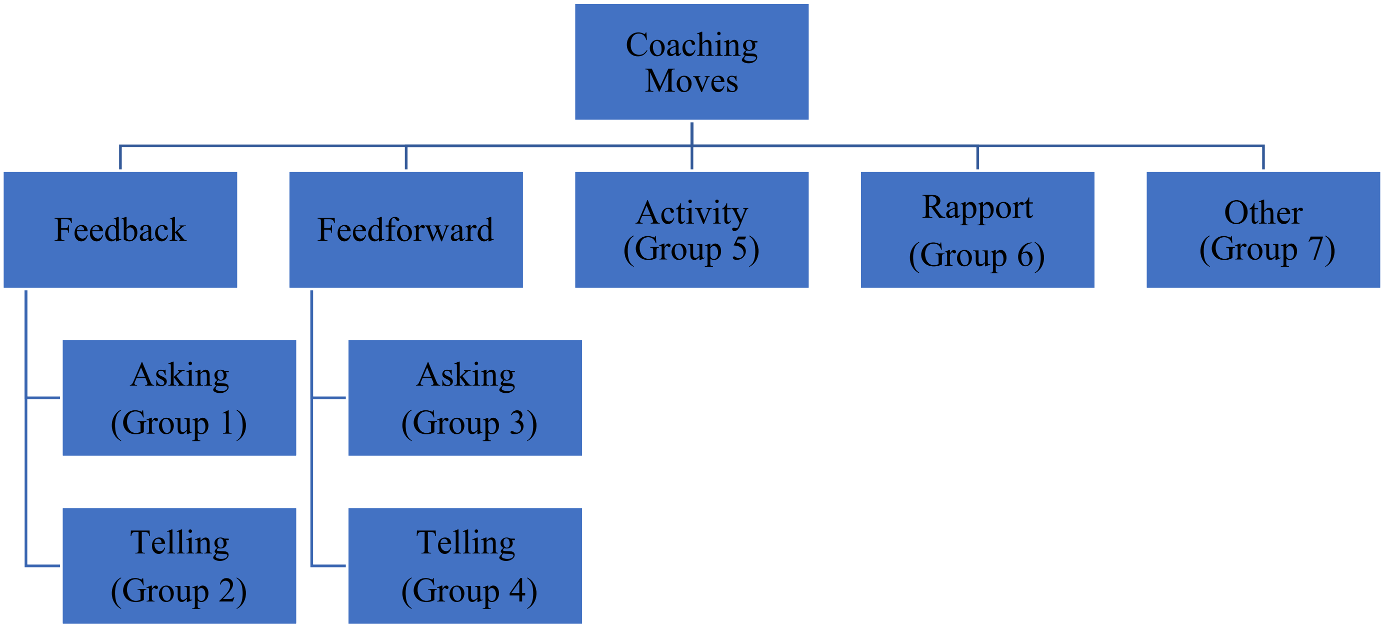
Coding decisions should be made primarily based on the function and structure of the coach's dialogue. You can and should use earlier dialogue as important context for interpreting coach dialogue, but you should be careful not to make coding decisions based primarily on how a teacher interprets the coach's dialogue. For example:

> Coach: what did you notice about that student?
>
> Teacher: telling the student to stop the behavior was really effective in refocusing the student's attention

In this case, the teacher makes a connection between a cause and an effect in their lesson. However, the coach's dialogue should be coded as AskBack Noticing not AskBack Cause&Effect because the coach did not explicitly ask the teacher to discuss a cause and effect, but rather asked a more general noticing question.

Coaching Moves & Definitions

Figure A1: Overview of coaching move-groups.



Step 1: Identifying the Group
The first thing you should ask yourself is whether the coach is discussing something that happened in the past or something that happens in the future. In other words, is it backwards facing or forward facing?

- Backward-facing = dialogue that focuses backward on the teacher's prior instruction, including
    - Information about a prior lesson (including information about the teacher or their students)
    - Information about a teacher's general strengths or weaknesses
    - Background information about students
    - General information about the teacher's student(s), including their strengths and weaknesses
- Forward-facing = dialogue that focuses forward on the teacher's future instruction and opportunities for improvement, including
    - planning instruction for a specific future lesson
    - making a general change to a teacher's instruction
    - talking about future student learning
    - providing background information about learning and education to inform teacher decision-making or instruction in the future
    - suggestions and questions about how a teacher could have done something differently in a previous lesson since this kind of feedback ultimately serves to inform what a teacher should do differently in the future

Then, you should ask yourself whether the coach is asking the teacher something or telling them something.

- Asking = dialogue that prompts the teacher to generate reflections or ideas about their instructional practice or student learning. This can include yes/no questions that support the teacher in expanding their ideas or providing more information. For example:
    - *Does he like the food?* (AskBack Noticing)
    - *And what about transitions? Are those also challenging?* (AskBack Noticing)
    - *And do you want to implement that with Ethan only, or…?* (AskForward Implementation)
- Telling = Dialogue that provides information about the teacher's instructional practice or student learning, with or without explicit evaluation or judgement. This dialogue may also include an invitation or request for the teacher to react to the feedback (e.g. what do you think about my idea? Do you agree? Does that make sense?), rhetorical questions, or simple yes/no questions that do not prompt the teacher to generate their own reflections or ideas, but rather simply offer an opportunity for the teacher to agree or disagree with the coach's assessment. For example
    - *None of those, so, allowing those to go on, you're not gonna get into that, the content engagement until you get the behavior to stop. Does that make sense?*
    - *So like, next time a student is doing something like I want you to either tell them exactly what you want him to do, or exactly what you want him to stop doing. And, feel free to use like non-verbals. So like, if he's like drumming, you could be like, Ethan our hands are folded together on desk. Do you know what I mean?*

    - *Did you notice when you said all of our devices are going to be turned off during the school day?*
    - *What about trying to wait for 5 seconds before calling on a student?*

Other coach dialogue that cannot be categorized in this way, should instead be identified as one or more of the following:

- Activity = Dialogue where the coach initiates, facilitates, or otherwise participates in a structured activity that includes dialogue that does not fall into any of the other groups. For example,
    - Engaging in a role-play
    - Planning a lesson to incorporate a change to instruction
    - grading student work
    - reviewing data from a student assessment
- Rapport = Coach dialogue that may contribute to maintaining, strengthening, or otherwise supporting the relationship between the coach and the teacher
- Other = any dialogue that does not fit into any of the above categories

Step 2: Identifying the Move

After identifying the correct group, you should select the appropriate move from the moves listed within that group on the following pages.

Group 1: Asking & Backward-Facing Moves

2. Noticing: questioning that only asks the teacher to recall information about themselves, a lesson, or their students based on prior experiences or their general familiarity with themselves or their students.
    - *What did you notice about student x's behavior?*
    - *How did student x respond to the prompt?*
    - *What did you do when…?*
    - *How do you usually respond when…?*
    - *And how did students respond after you did x?*
    - *How did you feel when Ethan was misbehaving?*
    - *What went through your mind when…?*
    - *Which students were confused?*
    - *Does he spend much time reading?*
    - *Tell me about when xyz happened… (*this question is too vague to be anything other than noticing)
    - *Tell me about the lesson…*
3. Cause & Effect: questioning that explicitly asks the teacher to reflect on the effect(s) that stemmed from a particular cause and/or the cause(s) that led to a particular effect. This does not include questions that ask the teacher to provide rationale for their own claim or action, which are instead coded as Justification. This also does not include questions that ask the teacher to identify another person's rationale or motivation, which are instead coded as Interpretation.
    - *How do you think giving wait time influenced students?*
    - *What did you notice about how that technique influenced students' responses?*
    - *How did implementing the strategy we talked about last time help students?*
    - *What do you think prevented you from providing wait time?*
    - *Why do you think you were able to successfully use x strategy during the last lesson?*
    - *How did using that strategy help to speed up the pace at the start of your lesson?*
    - *What would have happened if you had used a more specific redirection at that point?* (This is a hypothetical, but is still backward facing)
4. Justification: questions that explicitly prompt the teacher to provide evidence, rationale, and/or purpose for a claim, decision, or action they have made previously
    - *Why did you choose to do x when student y was talking?*
    - *What were you hoping x move would accomplish?*
    - *What about the student's behavior led you to give him a demerit?*
    - *How did you think that activity would promote students' understanding?*
    - *What makes you think that?*
5. Interpretation: questioning that explicitly asks the teacher to develop a hypothesis, draw a conclusion, or make an inference about their students (e.g. a student's motivations, rationale, understanding, or skill level), their instruction (e.g., or themselves (other than identifying a cause/effect or providing justification).
    - *What do you think y shows about this student's understanding?*
    - *What do you think this lesson shows about your strengths as a teacher?*
    - *How well do you think that student understood the text?*
    - *How successful do you think that student felt in class today?*

   - *Based on your observations, what do you think students had the most trouble understanding?*
   - *What do you think would be the most helpful thing to reteach based on how today's lesson went?*

6. Vision: questions that explicitly prompt the teacher to articulate their goals or vision for a previous lesson or activity, but DO NOT explicitly ask for the teacher to justify an instructional choice. This can include goals for students and for the teacher's own instruction
   - *What did you hope would happen in today's lesson?*
   - *What objectives did you hope students would learn?*
   - *What did you want students to understand about the text?*
   - *What action step were you focused on in your previous lesson to improve your teaching?*
   - *What goal for your own instruction were you focused on?*
7. Lessons Learned: questioning that explicitly asks the teacher to identify something that they have learned from a prior experience or were working on implementing in the prior lesson
   - *What lessons have you learned about addressing kids' challenging behavior so far?*
   - *What did the resource I asked you to read teach you about giving feedback?*
8. Self-Assessment: questioning that asks the teacher to make a judgement about the success and quality of their own instructional practice, including locating their performance within a particular performance framework
   - *How successful were you at x?*
   - *What do you think you did well in terms of feedback?*
   - *How did that simulation go?*
   - *How effective do you think you were at giving feedback today?*
   - *On a scale from 1-5, how effective do you think your redirections were?*
   - *Thinking about the Essential Practices/CLASS/other, how would you rate yourself for the domain of questioning?*
   - *What language from the rubric do you think best describes your classroom management in today's lesson?*
9. Grading: questioning that ask the teacher to locate themselves within a particular performance framework
   - *Thinking about the Essential Practices/CLASS/other, how would you rate yourself for the domain of questioning?*
   - *What language from the rubric do you think best describes your classroom management in today's lesson?*

Group 2: Telling & Backward-Facing Moves

10. Observation: feedback that describes specific factual information about students, a lesson, or the teacher based on the coach's observation of prior instruction or general familiarity with the students or teacher's instructional practice. This does not include references to a prior coaching conversation or PD session (which would be Check-in). Whether the information is factual should be judged based on how the coach positions the comment rather than based on your assessment of whether the information is an opinion or not.
    - *I noticed that when student did x, you did y, and then z happened*
    - *I saw that you said "xyz" in response to the student's question*
    - *Jonathan was really bored in class today* (though boredom is not directly observable and therefore requires some level of inference or opinion, the language the coach uses suggests that they see this as a fact rather than as an inference/interpretation/opinion).
    - *X student has ADHD*
    - *You tend to ask kids to raise their hands a lot*
11. Connection: feedback that explicitly discusses the connection between a particular cause and its effect.
    - *Giving wait time allowed that student to process and generate a more complete answer* (this would be coded as connection only, not observation)
    - *I think the students were distracted and had trouble paying attention today because Ethan was making a lot of noise* (this would be coded as connection only, not observation)
    - *I noticed that you waited a few seconds before calling on students. That really helped Susan participate more because it gave her more time to think about her answer.* (this would be coded as observation for the first sentence and connection for the second sentence)
12. Justification: feedback where the coach makes an inference about the teacher's rationale for a particular decision, claim, or action. The coach must explicitly use language to indicate that they are making an inference rather than making a simple statement of fact.
    - *I'm guessing that you asked Ethan to share a norm because you hoped it would refocus him to be on-task.*
    - *I think when you did that, you were trying to like bring it back to the rules and expectations a couple of times.*
13. Interpretation: feedback in which the coach communicates a hypothesis, draws a conclusion, or makes an inference about something that does not meet the criteria for Connection or Justification. You should use the coach's language to determine whether the statement is a hypothesis/conclusion/inference. Common indications would be language suggesting that the teacher isn't absolutely certain (e.g. "I think…", "it seemed…") and/or the presentation of a piece of evidence accompanied by a further interpretation of that evidence (e.g. "that showed me that…", "that indicates that…")
    - *When Ethan gave the answer that Lisa was excited, this suggested that Ethan didn't fully understand the text.* (Note, this would be Tellback observation and interpretation).
    - *So, he was trying to get off topic…it never felt like a disruption. However, it was a disruption from setting those class norms.*
    - *Your tone of voice suggested that you were a little frustrated*

- *You seemed a little frustrated* (the word "seemed" indicates that the coach is treating this as an inference rather than a statement of fact)
- *Wow, and he was looking for you, Teacher! That's how I know that your positive relationship with the children has improved. That's what a positive relationship looks like!* (Note, this would be Tellback Observation and Interpretation. The first statement provides the evidence, and the second part is the conclusion/inference that the coach is drawing from this evidence).

14. Positive Evaluation: feedback that communicates a positive judgment about a teacher's general skill as a teacher, specific elements of the teacher's practice, or provides a general affirmation of the teacher's instruction in a specific lesson or time-period. This may include locating a teacher's practice within a specific instructional rubric.
    - *You did a really great job managing student behavior*
    - *You did a great job!*
    - *So, he was trying to get off topic. And you are so kind and engaging with him.*
    - *You're clearly a strong teacher!*
    - *So, I can totally tell that you have a teacher mindset. You were really well poised.*
    - *I would rate you at the high end for Emotional Support*
15. Negative Evaluation: feedback that communicates a negative judgment about specific elements of the teacher's practice, about the teacher's instruction in a specific lesson or time-period, or about a teacher's general weaknesses/problems. This may include locating a teacher's practice within a specific instructional rubric.
    - *You struggled to give descriptive feedback*
    - *I don't think you met your objectives for that lesson.*
    - *That activity didn't really work…*
    - *Responding to student behavior is not a strength of yours*
    - *You don't quite meet the criteria for scoring Advanced on the Essential Practices rubric…*
16. Grading: feedback that explicitly makes a connection between the teacher's instructional practice and a specific framework of instructional practice or performance/evaluation rubric
    - *I would rate you at the high end for Emotional Support*
    - *I would give you a score of 5 for x domain on the CLASS*
    - *On the district's evaluation framework, I think you would score…*
    - *You were actively monitoring in line with the x dimension of the Danielson Framework* (note, this move also includes observation)
17. Check-in: dialogue that references a topic of discussion from a previous coaching conversation or professional development activity. Note, this will often be accompanied by other codes.
    - *So last week we talked about implementing a new behavior management strategy and when I observed you today, I thought this strategy worked really well!* (Note, this example would be both check-in and positive evaluation)
    - *In yesterday's professional development session, we talked about wait time. How did you do with wait time in your lesson today?* (Note, this example would be both check-in and ask-back self-evaluation)
    - *Last time we talked about effective commands…*(If this is the whole turn of talk, then this would just be coded as Check-in).

Group 3: Asking & Forward-Facing Moves

18. Goal-Setting: questioning that prompts the teacher to identify a goal or outcome for their classroom or students for the teacher to work towards.
    - *What do you want students to learn in the next lesson?*
    - *How do you want students' behavior to be different in the future?*
    - *How do you want the lesson to feel for students?*
    - *What reading strategies do you want students to use when they read poetry?*
19. Generation: questioning that prompts the teacher to generate or identify new ideas, action steps, or strategies that the teacher can use in future lessons
    - *What do you want to do differently next time?*
    - *What do you want to do better in the next lesson?*
    - *What strategy could you use to better support student engagement next lesson?*
    - *How can you adapt your lesson to better meet the needs of the students who struggle with reading comprehension?*
    - *What can you do when a student misunderstands the text to help them correct their misunderstanding?*
    - *What do you want to work on for next time?* (though a teacher could answer this question with a student-focused goal, the language of this question focuses on the teacher)
20. Application*:* questioning that prompts the teacher to decide when and/or how to apply a specific instructional strategy. Application questions are distinguished from generation questions by the fact that there is an implicit or explicit teacher action/strategy that the coach wants the teacher to work towards and is asking the teacher to identify more specifically how/when they could implement that strategy or action.
    - *So, for tomorrow's lesson, when will you use this new strategy? What cue will you look for?*
    - *When will you implement this in your classroom?*
    - *How will you apply this strategy to your lesson tomorrow?*
    - *How could you give that redirection in a more specific way next time?*
    - *How would your idea work in practice?*
    - *In what other cases would this strategy be useful? When would it not be useful?*
    - *So how can you use this strategy in your teaching?*
    - *How could you apply the strategy of text-based questioning to yesterday's lesson?*
    - *How can you shorten that redirection even more?*
    - Note: this is distinct from practice moves where the coach explains a hypothetical or real situation and asks the teacher to practice responding to that situation
21. Anticipation: questioning that explicitly prompts the teacher to elaborate on the consequences of an instructional strategy, classroom situation, or goal, including the importance or purpose, potential negative consequences, or challenges the teacher may face in using a strategy
    - *Why is asking for text evidence important?*
    - *Why is it important for students to use context clues when they read?*
    - *So, if we point them to specific text and then we ask a follow up question, how do you think that might lead them to a stronger answer than the author doesn't tell us?*
    - *What do you think would happen if you asked students to do xyz…?*

  - *What do you think would happen if you never redirected misbehaviors?*
  - *Which of these strategies do you think would have the biggest impact for students?*
  - *What could go wrong with this strategy?*
  - *Are there any reasons you think this strategy might not work for your specific students, classroom, teaching style, etc.?*
  - *When certain students lose focus, what does this do in terms of attention and focus for the rest of the class*

22. Check-for-Understanding: questioning that checks for a teacher's understanding of a pedagogical strategy or other professional concepts that the coach or teacher have been discussing. These questions tend to require the teacher to synthesize or apply previously discussed content in order to answer them.
    - *So, what would a non-example of a succinct redirection be?*
    - *What is the difference between the strategy I just suggested and the one that you used originally?*
    - *How would you summarize what wait time is?*
23. Content Understanding: questioning which supports the teacher in understanding the details of specific subject-matter content
    - *So, how would you answer that discussion question about who Lisa really is?*
    - *What paragraph in the text allows you to make that conclusion?*
    - *What's the answer to this math problem?*

Group 4: Telling & Forward-Facing Moves

24. Reinforcement: feedback in which the coach explicitly reinforces that the teacher should, in future lessons, continue using a strategy that they are already using
    - o I noticed that you did x in your last lesson, and I want you to keep doing that (note, this would be coded as tellback observation as well)
25. Challenge: feedback that articulates a challenge or problem of teaching
    - o *So sometimes students struggle to comprehend the text that they're reading, they can make claims that sometimes can't be supported with the text or may even be refuted with the text.*
    - o *Psychology research indicates that students have limited mental capacity* (Challenge) *so we as teachers must be careful not to overload them with information* (Student Goal)
    - o *Sometimes we as teachers aren't aware of students' emotions because they don't know to communicate them*
26. Student Goal: feedback that articulates a goal or outcome for students for the teacher to work towards
    - o *We would like to increase positive task engagement.*
    - o *I think it's important that we focus on helping students with writing topic sentences.*
    - o *We want students to really understand what they read and we want to help them become independent readers*
    - o *It's really important for students to use text evidence in their responses*
    - o *[Being more specific with your redirections] ensures that students understand your expectations and are able to comply with them*
    - o *it's really important that we can support all students in understanding the text*
27. Instructional Strategy: feedback that explicitly proposes a new strategy that a teacher can or should use. This change can include using or tweaking a strategy that the teacher has previously used but needs to improve or has not yet applied to the specific situation
    - o *Naming the student is really helpful*
    - o *Next time, I want you to work on being more specific with your redirections*
    - o *One thing you can do is try to avoid using negative tone of voice and instead*
    - o *What about trying to wait for 5 seconds before calling on a student?* (counts as telling rather than asking because it ultimately serves as a way for the coach to suggest an idea rather than ask the teacher to generate their own idea)
    - o *Next time, I want you to say, "please be quiet" (this is not Demonstration because the coach does not explicitly indicate that the quote is meant to be an example demonstration of a more generalizable strategy)*
    - o *At that point in the lesson you should have provided a specific redirection*
28. Demonstration: dialogue where the coach explicitly illustrates *how* a specific instructional strategy can be used or implemented. This includes defining what a particular strategy means.
    - o *For example, "please be quiet" is a specific redirection*
    - o *You could use a non-verbal like this.* (this implies that the coach demonstrates a non-verbal signal).
    - o *A calm tone would sound like, "Ethan, please be quiet".*
    - o *Succinct redirections use as few words as possible*

- *Wait time means waiting at least 5 seconds before you call on a student…*

29. Implementation: dialogue where the coach provides a direction or suggestion for *when* a teacher should use a strategy in a future situation or how to prepare for using a strategy in a future lesson
    - *To remind yourself to use this new strategy, you should add a note to your lesson plan*
    - *I would like you try what we talked about in your lesson tomorrow*
    - *Make sure that you ask the questions we talked about in your next simulation*
    - *I think it would be helpful for you to review my summary of your action plans before our next meeting*
    - *I want you to apply the strategy of asking follow-up questions to how you lead the discussion and make sure to ask a follow-up question after every student response*
    - *That would have been a great moment in the lesson to apply the strategy we've been talking about*
30. Content Understanding*:* dialogue which supports the teacher in understanding the details of specific subject-matter content
    - *So, the text doesn't explicitly give an answer, instead the reader has to make an inference*
    - *I actually had a different answer for that math problem…*
    - *So if we know that, we know that Lisa's heart was pounding, I think that comes up later in our discussion or in the text. And so when her heart was pounding, we know that that was a direct explanation of how she feels. We – it's not explicitly stating that she feels that way because of the lie detector test, but we think that those two things are connected because of her heart pounding and her being so nervous all those clues that the author is giving us*
    - *Okay, so we have agreed that you will try out using wait time*
    - *I will come observe you on Friday to see you do this*

Group 5: Activities

31. Practice: dialogue where the coach initiates and facilitates a role-play activity. To count as a role-play, the coach must make clear that the teacher should embody the role of the teacher as if they were in class. Any dialogue where the coach gives directions or acts as a student should be coded as practice, but it is not necessary for the coach to pretend to be a student for the directions to count as a role-play.
    - *So what if Ethan is tapping on the table, what would you say?*
    - *I'll pretend to be a misbehaving student and I want you to redirect me specifically and succinctly* (this should be coded as Practice only and not TellForward Suggestion because the coach frames it as the instructions for the practice rather than a suggestion for how the teacher should change their instruction in the future).
32. Data Analysis: reviewing student-created materials or summary data on student learning (e.g. test score data) to analyze student understanding, learning, strengths, weaknesses, needs, etc.
33. Co-planning: reviewing curricular materials, state-standards, student-facing material (e.g. a book or problem-set) or other documents that teachers might reference
34. Instructional Artifacts: reviewing a lesson plan, student-facing handout, video, or other artifact of a teacher's prior instruction (not including student-created materials)
35. Professional Resource: reviewing general professional resources that provides general information about content or pedagogy, e.g. an article or video about strategies for teaching fractions. Videos that only serve to provide a model of how to enact a specific instructional strategy should be coded as Demonstration. If the purpose of the video is unclear, then it should be coded as Professional Resource.
36. Instructional Rubric: reviewing or explaining a specific rubric or framework of high-quality instruction

Note, feedback about the candidate's practice/activity or follow-up questions asked by the coach should not be coded as an Activity, but should instead be coded with the appropriate backward-facing or forward-facing move. For example:

- A statement about the meaning of a state standard during co-planning should be coded as TellForward Content Understanding
- A question about what a teacher noticed after reading a professional learning resource would be AskBackward Noticing

Group 6: Rapport

37. Encouragement: dialogue where the coach expresses positive expectations for the teacher's future work
    - *you've got this!*
    - *you're going to be great!*
    - *You will get better with time*
38. Normalizing Struggle: dialogue where the coach communicates that facing challenges and struggles in teaching is normal
    - *Most teachers struggle with wait time*
    - *I've been teaching for 20 years and I still struggle with…*(this would also be coded as sharing)
    - *That's okay, you're not going to get it perfect the first time* (while the coach is likely responding to a teacher's negative emotions when making this comment, the coach does not explicitly mention or express what emotions the teacher might be feeling).
    - *These simulations are for practice so we don't expect you to get it perfect*
    - *This simulation is difficult (for most people)*
39. Empathy: dialogue in which the coach asks about, anticipates, or expresses an understanding of the teacher's emotions or perspective. This does not include short and generic affirmation of the teacher's previous comments like "uh-huh" or "I understand"
    - *What did you think? How do you feel about it?* (this question is too vague to be self-evaluation or noticing)
    - *So, how are you feeling about that first simulation?* (this question is too vague to be self-evaluation or noticing)
    - *that must have been hard*
    - *I know that this simulation can feel really challenging*
    - *I totally understand how you feel (*here, the use of the word "feel" is an explicit mention of feelings or emotions)
    - *I know it can be very hard to keep track of everything that's going on all at once*
    - *Yeah, that's tough…*
    - *Now that we've discussed…does that help you feel more prepared?*
    - *I'm so sorry, but for the sake of time, I'm going to interrupt you…*
    - *Non-example*: this should be coded as N/A
      Teacher: *That simulation was really hard. I feel so drained*
      Coach: *That's okay, I understand.* (here the coach does not engage with the content of the teacher's emotions or explicitly mention feelings/emotions)
40. Revoicing: dialogue in which the coach rephrases what the teacher has said in a recent turn-of-talk. To count the coach must use different language or extend the teacher's comments, it is not just the repetition of some of the words that the teacher previously used.
41. Assistance: offering assistance or provides an opportunity for a teacher to request specific assistance
    - *All right. Um, we have plenty of time. We can use it to think through or talk through anything on your mind.*
    - *Do you have any questions?*
    - *What do you want to talk about today?*
    - *Is there any additional support you would like from me?*

- *We can definitely talk about that, but first…*
- *I can copy those pages of the book for you if you'd like?*

42. Permission: asking for permission to say or do something or otherwise offering choice
    - *Is it okay if I give you some advice?*
    - *Is it okay if I ask you about that topic?*
    - *Are Mondays ok?*
43. Coaching Feedback: eliciting feedback from the teacher on the coaching session
    - *How was this coaching session for you*?
    - *Is there anything you'd like me to do differently in our next conversation*?
    - *How helpful was today's conversation for you*?
44. Sharing: sharing personal information about the coach, asking about teacher personal information, or demonstrating a personal understanding of the teacher
    - *I love hiking too!*
    - *When I taught 8th grade, I really struggled with classroom management* (this would also be coded as encouragement for normalizing struggle)
    - *I've been a coach for three years and I most of the teachers I work with struggle with the same thing* (this would also be coded as normalizing struggle)
    - *Have you worked with children or taught before?* (this is about the teacher's general teaching experience but is not specifically related to their instruction)
    - *So, one thing that I know and love about you is that you are a very kind and gentle person.*
45. Agenda-setting: providing an explanation of the purpose or agenda for coaching, including coaching as an ongoing program, an individual coaching conversation, or a specific part of an individual coaching conversation.
    - *My goal as a coach is to be as helpful to you as possible*
    - *We're going to use this time to help you feel more comfortable with leading a discussion*
    - *We're going to talk about your last lesson so that you can reflect on where you might want to make changes for tomorrow…*
    - *We'll check-in at our next meeting to see how things are going.*
    - *we're going to talk about how to make the next simulation better*
46. Mirroring: repeating a teacher's words or finishing a sentence for the teacher

Group 7: Other Codes

47. Other Coaching: questions and statements that coaches make that do not fall into any of the other categories but nonetheless support teachers with their instructional practice or otherwise meaningfully informs how a teacher might teach in a real or simulated classroom environment
    - General probing questions when focused on content related to instruction
    - *Tell me more…*
    - *It would help me understand if you gave an example of what you mean*
    - *What did you mean when you said "…"?*
48. Unclear: dialogue that is too incomplete, ambiguous or unclear to accurately code using the other categories.
    - "[indiscernable]"
49. Non-Coaching: This is dialogue that is clearly not part of a coaching conversation and should be coded anywhere it is present, even if there are other moves within the same turn of talk. Examples include:
    - Logistical issues, such as when the next coaching session or observation will be (if it cannot be considered follow-up)
    - Technology issues, such as being unable to hear what the teacher said
    - Timing issues, such as "your five minutes for coaching are up"
    - Dialogue from other speakers not included in the conversation, e.g. background noise from students

Appendix B: Descriptive Statistics

Table B1. Covariate balance across coaches for coaches assigned to teachers where covariate data are available.

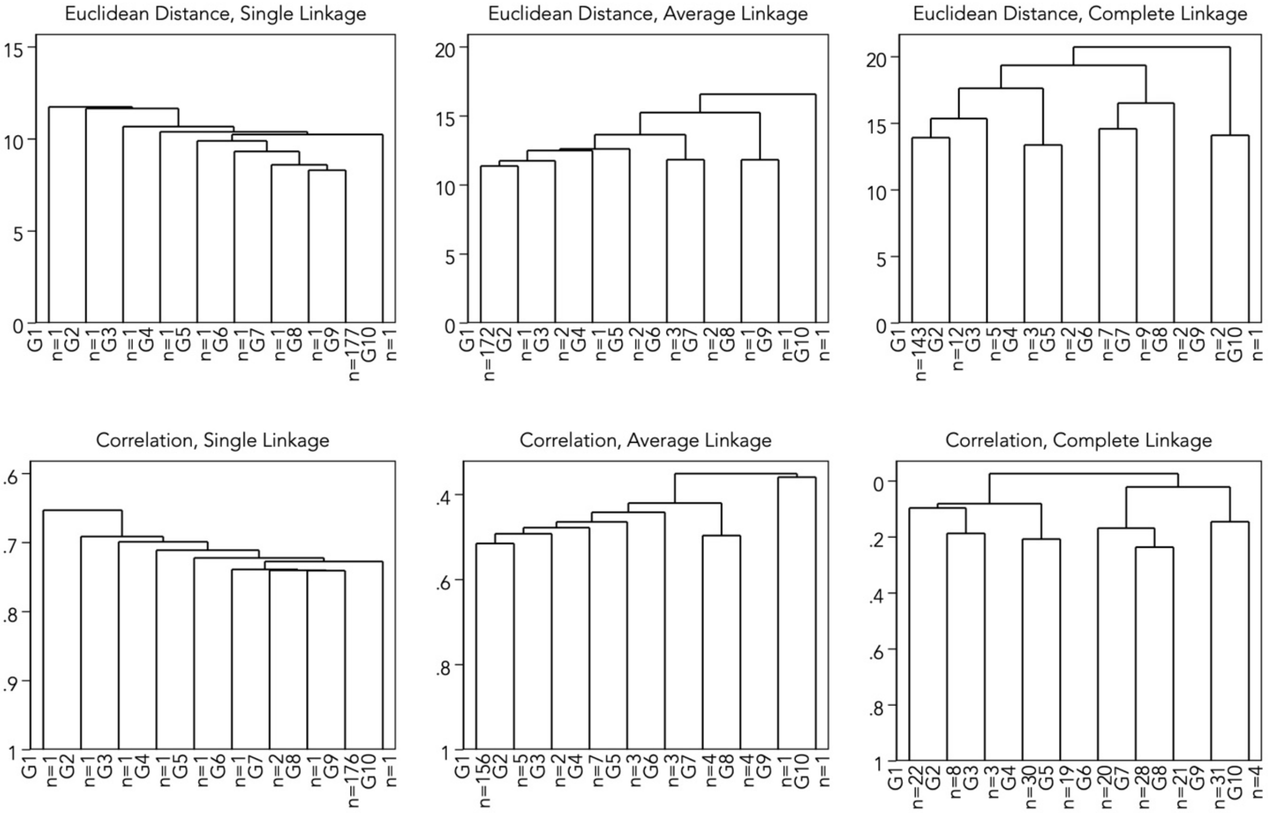
Covariates (Standardized)	N	Constant	1	2	3	4	5	6	7	8	9
Perceptions of simulation benefits	93	0.395	-0.000	-0.392	-1.026	-0.205	-0.840	0.264	-0.704	-0.154	-0.493
Teacher nervousness	93	0.215**	-0.444	0.161	-0.296	-0.296	-0.0807	-0.000	-0.380	-0.111	-0.621
GPA	89	0.733**	-0.791	-1.351**	-0.894	-0.674*	-0.530	-0.650	-0.848*	-0.573	-0.479
Culturally Responsive Teaching Self-Efficacy Scale	93	-0.650+	0.839	0.755	0.537	0.690+	0.865*	1.006*	0.737	0.120	0.656
Anxiety	93	-0.136	0.142	-0.271	-0.283	0.331	0.863*	0.385	0.364	-0.319	-0.340+
Depression	93	0.0209	-0.237	-0.517*	-0.426+	0.284	0.620	-0.081	0.345	-0.391	-0.455
Stress	93	-0.0103	-0.285	-0.333	-0.447	0.190	0.466	-0.070	0.523	-0.396	-0.244
Grit	93	-0.530	0.646	0.302	0.585	0.369	1.426*	0.105	0.554	0.646	0.240
Mindfulness in Teaching Scale	93	-0.519**	0.629*	0.248	0.349	0.559+	1.010**	0.629	0.629*	0.105	0.671*
NEO-5 Factor inventory composite: Agreeableness	91	-1.160**	1.259*	1.150+	0.630	1.217*	1.219*	1.214*	1.220*	1.349*	1.398*
NEO-5 Factor inventory composite: Conscientiousness	91	1.183+	-1.563*	-1.144	-1.371	-1.161	-1.219	-1.258	-1.095	-1.446+	-1.119
NEO-5 Factor inventory composite: Extraversion	91	0.285	0.333	-0.049	0.128	-0.734	-0.314	-0.516	-0.041	-0.384	-0.353
NEO-5 Factor inventory composite: Neuroticism	91	0.152	0.142	-0.463	-0.511	-0.087	-0.013	-0.303	0.011	-0.121	-0.269
NEO-5 Factor inventory composite: Openness to experience	91	0.147	0.0539	-0.537	-0.431	-0.090	-0.360	-0.543	0.255	-0.450	0.259
Adult Attachment Scale	90	-0.351**	0.667+	0.494	0.574	0.107	0.165	0.539+	0.126	0.406	0.671*
Teacher Multicultural Attitude Survey	93	-0.226	0.192	-0.105	-0.426	0.675+	0.128	0.256	0.621+	-0.0639	0.141
Factors Influencing Teaching Choice	93	-0.375	0.592	0.444	0.345	0.362	0.498	0.529	0.825	-0.111	-0.0592

Covariates (Standardized)	N	Constant	1	2	3	4	5	6	7	8	9
Perceptions of student avatars	93	-0.196	0.0734	0.407	0.147	-0.082	0.020	0.934*	0.168	0.073	0.617
Self-rating	91	-0.651**	1.080+	0.598+	0.914*	0.953**	0.340	0.107	0.712+	1.183*	-1.49e-08
Coach rating of first observation	129	-0.465*		0.361		0.714**		0.497	1.546**	-0.147	0.820**
Self-Efficacy Composite	93	-0.123	0.282	0.322	-0.005	0.0627	0.362	-0.009	0.410	-0.568	0.370
Likelihood of recommending punitive consequences for off-task avatars	93	0.547	-0.762	-0.430	-0.051	-0.474	-0.901	-0.544	-0.370	-0.724	-0.884
Pre-Coaching Observation Score	177	1.104	-1.444	-0.784	-0.510	-0.680	-1.224	-1.001	-0.721	-1.033	-0.544
Joint p-value by coach			0.00	0.22	0.11	0.13	0.11	0.03	n/a	n/a	n/a

Robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.10

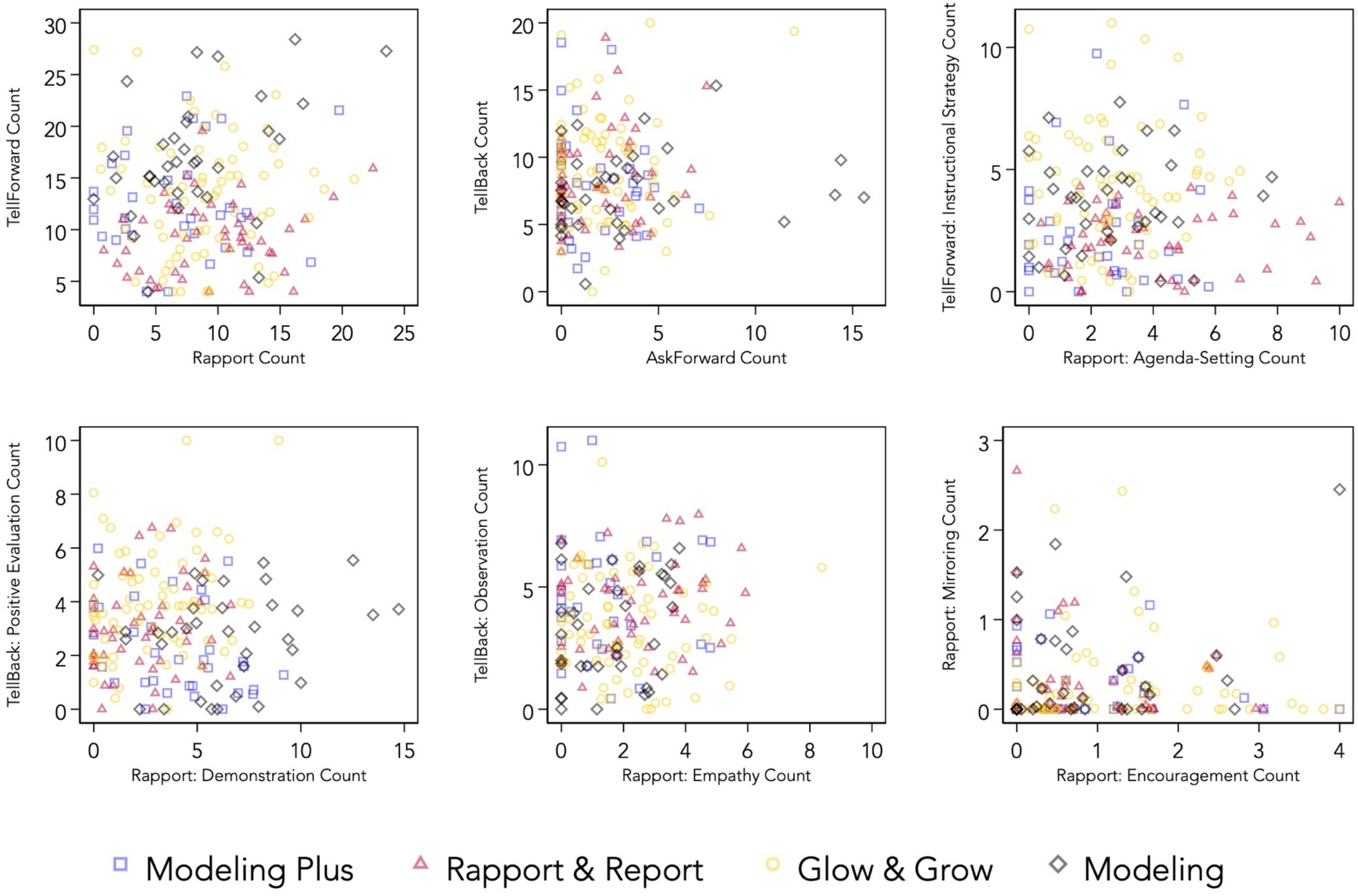
Appendix C: Identifying and Characterizing Coaching Profiles

Figure C1. Cluster trees resulting from hierarchical agglomerative clustering using six different dissimilarity methods.



Note: each cluster tree illustrates the how the specific dissimilarity method groups transcripts into clusters, showing only the first twenty splits for readability. The x-axis describes the number of transcripts in each terminal cluster. For example, the first terminal cluster (G1) in the cluster tree generated by the Euclidean Distance and Single Linkage approach has only one transcript in it (n=1). The y-axis reflects how similar the transcripts within a cluster are to another, as estimated by the specific dissimilarity metric employed. Numbers closer to the top of the graph reflect greater dissimilarity, while numbers closer to the bottom reflect greater similarity (or less dissimilarity). Across all six cluster trees, the main splits occur near the top of the tree, indicating that the transcripts within these clusters are still somewhat dissimilar, even though they are more similar to one another than the transcripts in another cluster at the same level on the y-axis.

Figure C2. Scatterplots comparing the frequency of moves and move-groups across the four coaching profiles. We focus here on the moves and move-groups that most differ across profiles.



Appendix D: Coaching Profiles by Coder

If our randomization of transcripts to coders was successful, we would expect to see balance in the characteristics of transcripts within each coder. Consistent with this, we find little evidence of systematic relationships between coders and teacher covariates (Table D1) or coders and coach assignment (Table D2). However, we do see significant differences in the proportion of transcripts assigned to the Rapport & Report and Glow & Grow coaching profiles (Table D3). This raises the concern that observed differences between these two profiles may reflect differences in coder standards instead of true differences in the coaching moves used. To better understand how concerned we should be, we compare move counts across coders using the cohort/site of double-coded transcripts, focusing on the key moves that differentiate the profiles from one another. Table D4 shows that only two out of eight moves are significantly different by coder when comparing double-coded transcripts. While this is encouraging, a joint F-test indicates that these significant results are not merely a consequence of testing multiple hypotheses. Additionally, though not the only moves that distinguish between the Glow & Grow and Rapport & Report profiles, the two that are significantly different (Rapport: Agenda-Setting and TellForward: Demonstration) are central to the differences between these profiles. For this reason, our preferred regression models include coder fixed effects as a control.

Table D1. Teacher covariates by coder.

	N	Coder 1 Mean	Difference between Coder 1 and Coder 2
Perceptions of simulation benefits	92	4.333	-0.291+
Teacher nervousness	92	3.689	0.120
GPA	88	3.503	-0.0152
Culturally Responsive Teaching Self-Efficacy Scale	92	63.63	5.726
Anxiety	92	0.397	0.0621
Depression	92	0.403	0.0406
Stress	92	0.832	0.0649
Grit	92	2.502	-0.0355
Mindfulness in Teaching Scale	92	3.924	0.0670
NEO-5 Factor inventory composite**:** Agreeableness	90	3.903	0.0513
NEO-5 Factor inventory composite**:** Conscientiousness	90	3.917	0.0507
NEO-5 Factor inventory composite**:** Extraversion	90	3.765	-0.249*
NEO-5 Factor inventory composite: Neuroticism	90	2.864	-0.0774
NEO-5 Factor inventory composite: Openness to experience	90	3.540	0.0128
Adult Attachment Scale	89	3.501	0.0916
Teacher Multicultural Attitude Survey	92	4.315	-0.0697
Factors Influencing Teaching Choice	92	5.893	0.128
Perceptions of student avatars	92	2.619	0.0390
Self-rating	90	3.818	0.0296
Coach rating of first observation	124	3.508	0.0740
Self-Efficacy Composite	92	6.319	0.278
Likelihood of recommending punitive **c**onsequences for off-task avatars	92	3.800	0.0936
Pre-Coaching Observation Score	171	3.613	0.140

Robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.10

Table D2. Coach assignment by coder.

	(1) Coder 2
Coach 1	0.0652
	(0.383)
Coach 2	-0.187
	(0.387)
Coach 3	0.100
	(0.433)
Coach 4	0.0263
	(0.387)
Coach 5	0.000
	(0.390)
Coach 6	0.0455
	(0.400)
Coach 7	0.0833
	(0.382)
Coach 8	0.000
	(0.403)
Coach 9	-0.0714
	(0.393)
Coach 10	0.000
	(0.393)
Coach 11	-0.0714
	(0.416)
Coach 12	0.000
	(0.403)
Constant	0.500
	(0.368)
Observations	171
R-squared	0.024
Joint significance test	p=0.98

Robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.10

Table D3. Relationship between coder and coaching profile assignment.

	(1) Modeling	(2) Modeling Plus	(3) Glow & Grow	(4) Rapport & Report
Coder 2	-0.092	-0.005	-0.215**	0.311**
	(0.058)	(0.056)	(0.070)	(0.060)
Constant	0.236**	0.180**	0.483**	0.101**
	(0.045)	(0.041)	(0.053)	(0.032)
Observations	186	186	186	186
R-squared	0.014	0.000	0.049	0.125

Robust standard errors in parentheses
** p<0.01, * p<0.05, + p<0.10

Table D4. Move count by coder for double-coded transcripts, focusing on the key moves that distinguish the Rapport & Report and Glow & Grow coaching profiles.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Rapport: Agenda-Setting	Rapport: Encouragement	Rapport: Empathy	Rapport: Mirroring	TellBack: Positive Evaluation	TellBack: Observation	TellForward: Instructional Strategy	TellForward: Demonstration
Coder 2	0.750+	-0.321	0.429	-0.0357	-0.357	-0.0714	-0.679	-1.036+
	(0.395)	(0.237)	(0.353)	(0.0998)	(0.394)	(0.474)	(0.504)	(0.603)
Constant	2.321**	0.929**	1.429**	0.179*	2.893**	4.036**	3.250**	4.071**
	(0.279)	(0.168)	(0.250)	(0.0706)	(0.279)	(0.335)	(0.356)	(0.426)
Observations	56	56	56	56	56	56	56	56
R-squared	0.063	0.033	0.027	0.002	0.015	0.000	0.032	0.052
Joint F-test	p = 0.00							

Standard errors in parentheses
** p<0.01, * p<0.05, + p<0.10

Note: For greater statistical precision, observations here represent individual turns of talk rather than individual transcripts.