ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 20-252

New Research for Teacher Education

Heather C. Hill
Harvard University

Zid Mancenido
Harvard University

Susanna Loeb
Brown University

Despite calls for more evaluative research in teacher education, formal assessments of the effectiveness of novel teacher education practices remain rare. One reason is that we lack designs and measurement approaches that appropriately meet the challenges of causal inference in the field. In this article, we seek to fill this gap. We first outline the difficulties of doing evaluative work in teacher education. We then describe a set of replicable practices for developing measures of key teaching outcomes, and propose evaluative research designs that can be adapted to suit the needs of the field. Finally, we identify community-wide initiatives that are necessary to advance useful evaluative research.

VERSION: July 2020

Suggested citation: Hill, Heather, Zid Mancenido, and Susanna Loeb. (2020). New Research for Teacher Education. (EdWorkingPaper: 20-252). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/zhhb-j781

New Research for Teacher Education

Heather C. Hill[1, 2], Zid Mancenido[1], and Susanna Loeb[2]

[1] Harvard Graduate School of Education

[2] Annenberg Institute at Brown University

Author Notes

Correspondence concerning this article should be addressed to Heather C. Hill, 445 Gutman Library, Harvard Graduate School of Education, 6 Appian Way, Cambridge, MA 02138. Phone: 617-495-1898. Email: heather_hill@harvard.edu. This research was generously supported by the National Science Foundation (ECR-1920616) and the Spencer Foundation (256629). We are grateful to participants in a Spencer-supported workshop on teacher education, who helped shape this paper. All errors and omissions are our own.

Abstract

Despite calls for more evaluative research in teacher education, formal assessments of the effectiveness of novel teacher education practices remain rare. One reason is that we lack designs and measurement approaches that appropriately meet the challenges of causal inference in the field. In this article, we seek to fill this gap. We first outline the difficulties of doing evaluative work in teacher education. We then describe a set of replicable practices for developing measures of key teaching outcomes, and propose evaluative research designs that can be adapted to suit the needs of the field. Finally, we identify community-wide initiatives that are necessary to advance useful evaluative research.

Over the past decade, scholars have developed promising practices for teacher education, where practices are the approaches, activities and processes supplied by teacher educators for the purpose of developing pre-service teachers' (PSTs) knowledge and skill. These promising practices include experiences that foster PSTs' ability to analyze classroom video (e.g., Santagata & Yeh, 2014; Sun & van Es, 2015); courses focused on preparing PSTs for addressing issues of race and racism in classrooms (e.g., Brown, 2014; Durden, Dooley, & Truscott, 2016; Haddix, 2017; Lee, 2018); instruction in how to meet the needs of diverse learners (e.g., Bravo et al., 2014; Hernandez & Shroyer, 2017; Kang & Zinger, 2019); field experiences that join coursework to community settings (e.g., Horn & Campbell, 2015; Wasburn-Moses, Noltemeyer & Schmitz, 2015); and, rehearsals and teaching simulations aimed at increasing PSTs' instructional skills (e.g., Kavanagh & Rainey, 2017; Windschitl et al., 2012). These new practices differ markedly from typical teacher education curricula and pedagogy, and often challenge the traditional separation between coursework and clinical experiences.

Promising as these practices are, researchers have seldom evaluated their efficacy in producing more skilled and thoughtful teaching. One reason for the lack of evaluation is that many of these practices are still in the development stage, necessitating careful exploration of design and initial implementation, and then adjustments based on these observations and participant feedback. Such work is necessary for refining emerging practices.

In this paper, we address a second reason for the paucity of evaluative studies: the lack of research designs and measures appropriate to understanding whether and how new practices in teacher education work to improve teaching. In particular, we lack implementable designs that compare otherwise similar individuals with and without the experience of the innovative teacher education practice. Such comparisons can help us identify the specific influence of the practice

on outcomes. We also lack high-quality, replicable measures of the PST outcomes we care about, including *in situ* teaching skills, as well as the knowledge and dispositions thought to mediate improved teaching outcomes. Thus despite calls for more evaluative research in teacher education over the years (e.g., Cochran-Smith & Zeichner, 2005; Diez, 2010; Fallon, 2006; Grossman, 2008), barriers to doing such work remain high.

To make progress, the field needs research designs and measures sensitive to the challenges of studying innovative practices in teacher education. In this paper, we review these challenges, then propose two pathways forward. First, we describe a set of replicable practices for building measures of key teaching outcomes. Second, we propose new research designs that take advantage of comparisons between groups or within individuals over time to evaluate the efficacy of new practices. Although such methods have been used widely in program evaluation for decades, we discuss adaptations to suit the needs of teacher education. Finally, we describe a plan of collective, community-wide work to bring together research designs, measurement, and the realities of teacher education programs.

Evaluative Research in Teacher Education

For several decades, most research on educating new teachers has fallen into one of three paradigms (Borko, Liston & Whitcomb, 2007): *interpretive research*, in which scholars trace PSTs' sense-making as they engage with teaching tasks; *practitioner research*, in which teacher education scholars describe the nuances of their and/or others' practice; and *design research*, in which scholars create blueprints for PST learning experiences, enact and critique those plans, then iteratively redesign those experiences. We argue that research in these traditions has improved the state of practice in the field. Teacher educators now have sources of expertise in designing learning experiences for PSTs, and the process of innovation, reflection, revision and

redesign has both improved local offerings and produced strong hypotheses about effective approaches to training PSTs (e.g., Cochran-Smith, Barnatt, Friedman & Pine, 2009; Draper, Broomhead, Jensen, & Nokes, 2012; Hyland & Noffke, 2005; Lustick, 2009).

Formally evaluating the efficacy of such practices on teaching and teaching-related outcomes, however, has been rare (for an exception see Boyd et al., 2009). A hypothetical example highlights how this situation might have come about. Imagine Gabrielle, a mathematics teacher educator and researcher at a mid-sized college of education. Gabrielle wants the 20 PSTs in her math methods course to be able to launch cognitively demanding mathematics tasks, but to do so with sensitivity to learners' prior knowledge, a topic of research and concern in STEM education (e.g., González & Eli, 2017; Jackson et al., 2012; Kang et al., 2016). In thinking about this goal, Gabrielle relies on theories and evidence developed as part of Universal Design for Learning (UDL) (Rose & Meyer, 2006), which urges teachers to think about and plan for variability in student responses to instruction. Gabrielle has also become interested in approximations of practice, such as rehearsals, an activity in which PSTs "publicly and deliberately practice with their peers how to teach rigorous content to particular students using particular instructional activities" (Lampert et al., 2013, p. 227).

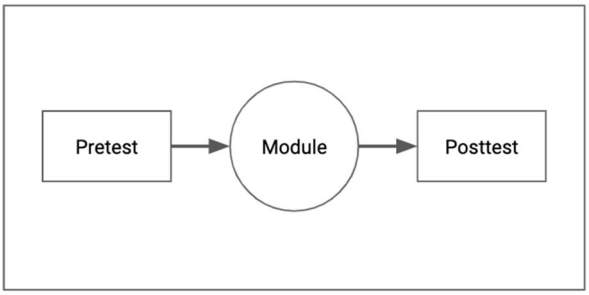
Gabrielle quickly builds a short module –about three hours of instruction – that takes PSTs through the process of task selection and task analysis, and then a micro-teaching simulation similar to a rehearsal of practice. Specifically, Gabrielle designs this module to help PSTs: (a) select cognitively demanding mathematics tasks that offer multiple entry points to students (Smith & Stein, 1998), so that students of all levels of prior knowledge and abilities can productively work on the task (ensuring equitable access for all students; e.g. Rose & Meyer, 2006); (b) identify the contextual features of mathematics problems (unknown vocabulary words,

unfamiliar settings; see Jackson, et al., 2013) that may affect the accessibility of the task to students; and (c) launch tasks with sensitivity to student prior knowledge and background, but without devolving the cognitive demand of the task (Jackson et al., 2012). Gabrielle then wonders whether and how to collect data about her module's effects on PSTs' skills and practice. Mindful that she has yet to achieve tenure, she hopes to publish a manuscript describing her results.

Some of Gabrielle's colleagues have promoted the idea that she can save time by doing research on her own class, designing a pretest and posttest on the key skills she hopes to improve, and then collecting and analyzing data to show that PSTs improved their performance on the test during the class (Figure 1).

Figure 1

Typical Teacher Education Evaluative Research Design



In fact, this is a typical evaluative research design in STEM teacher education. Following a review of science teacher education studies published between 2002-2015, Cochran-Smith et al. (2016) conclude:

> Many of the science-for-all studies used single-case, pre-post test mixed-methods designs, with science methods courses or programs… functioning as strategic research

> sites for the investigation of particular instructional interventions or teaching activities. Typically at least one member of a research team or the single researcher was the instructor of the course or the director of the program under investigation….Data were representations of teacher candidates' learning in response to the instructional intervention that was the object of study, including candidates performance on particular course tasks or in classrooms, their responses to surveys and questionnaires, and structured interviews that clarified survey responses. (p. 472)

However, Gabrielle has concerns about this design (see Table 1 for a summary). She knows that students will learn some content naturally, as they mature and gain life experience. She also knows that her PSTs have concurrent experiences in other courses and in their clinical setting, and that these may also improve PSTs' performance on selecting, analyzing, and launching tasks. Research designs that lack a comparison group cannot distinguish these effects – broadly known as *history* and *maturation* effects – from actual learning from her module.

Gabrielle also knows that pre-post studies of innovations in intact methods courses cannot disentangle the effect of a new practice, the effect of the specific instructor, and the effect of class composition (e.g., peer knowledge and skill; propensity to collaborate productively). She worries most about *instructor* effects; any estimate of the impact of her module will include some assessment of her basic effectiveness as a teacher as well. If she is particularly effective, the module will appear more successful; if she is particularly ineffective, her module will fail to show effects. Gabrielle also worries about simultaneously teaching and studying her PSTs and the conflict of interest and potential bias in her data that might result, particularly if she uses course assignments as an outcome measure (i.e., *instrumentation* effects).

Finally, she worries about whether students who opted into her class section might be different from those who enrolled in other instructors' sections, for she knows she has a reputation for paying attention to equity, a focus of her new module. Given her reputation, any gains she observes among her PSTs may result from having students well-positioned to learn this content, in terms of motivation and prior knowledge, rather than the effect of her module alone (i.e., *selection* effects).

Table 1

Common Threats to Validity in Evaluations of Teacher Preparation Practices

Threat to Validity	Description	Possible strategies to address threat
History and maturation effects	- Effects of events concurrent to treatment may be captured in study outcome measures (e.g., if Gabrielle's students also learn how to select, analyze, and launch tasks in other courses) - Effects due to natural growth (e.g., Gabrielle's students learn because of normal cognitive maturation)	- Use a comparison group that does not receive the intervention, or receives an alternative intervention aimed at a different outcome -Use multiple assessments to gauge growth on the outcome measure a) prior to and b) during the intervention

Instructor and instrumentation effects	- Effects of bias due to who rates the outcome measure (e.g., Gabrielle rates those who received treatment systematically higher) - Effects of bias if the outcome measure is a graded course assignment (e.g., Gabrielle's students game the outcome such that it does not represent their true ability) - Effects due to the designer of the treatment delivering it in a way that cannot be standardized/replicated (e.g., Gabrielle is much stronger than a typical instructor and carries that strength into the module)	- Use independent measures for course assignments and research outcomes - Use multiple raters to check the reliability and replicability of the scoring procedure - Double-blind the scoring procedure* - Develop standardized training and/or a manual for delivering the treatment - Collect implementation fidelity measures to check for standardization across groups that receive the intervention - Have research assistants deliver the intervention
Selection effects	- Effects due to selection into treatment (e.g., some of Gabrielle's students are highly motivated to take the module, or have pre-existing	- Randomize participants to the treatment vs comparison group

	knowledge that helps them learn more efficiently) - Effects due to selection into study (e.g., same as above, except into the study)	- Randomize offers to participate in the study from the broader population*

* While we do not explicitly discuss these strategies in this paper, we note them because they are good research practices. We refer readers to Shadish, Cook, and Campbell (2001) for further information.

Gabrielle's doctoral training included several classes in research design, including both positivist approaches focused on identifying causal impacts, and more interpretivist approaches focused on uncovering how research participants make meaning of social experiences and interactions. Gabrielle also considers herself to have reasonably strong quantitative skills. Gabrielle thus begins to consider some of the designs she learned about in her program, including randomized experiments and analytic approaches that attempt to mimic experiments using observational data (e.g., regression discontinuity designs). Yet when Gabrielle begins to think about using such designs in the context of teacher education, she identifies numerous challenges.

Because teacher education is widely dispersed, with almost 1,500 separate institutions educating teachers for initial licensure, most programs are small like hers (U.S. Department of Education, 2019), meaning they at most support one or (as in Gabrielle's case) two sections of each course. Further, similar to other universities, her students' schedules are often determined partly by preference and partly by external scheduling constraints, preventing Gabrielle from taking advantage of random assignment to classes. Even if she were able to randomly assign

students to math methods sections, however, students would still be taught by two different instructors, and Gabrielle would not be able to disentangle the effect of any new practice from her effect as an instructor. A further problem is that, as with most other teacher education programs (TEPs), all her students follow the same or very similar pathways through coursework, meaning approaches that take advantage of differences in these pathways are not possible.

For some perspective, Gabrielle consults her school's research methodologist, who first recommends launching an experimental trial of her new module in multiple universities, an idea with some interest in teacher education (e.g., Grossman & McDonald, 2008). Yet Gabrielle knows that working across multiple institutions would be logistically difficult and that a large sample of teacher education institutions (potentially in the dozens) would be necessary to achieve statistical power for detecting effects. She also knows the costs of recruiting and training teacher education faculty would be beyond what Gabrielle could bear.

Hearing this, the methodologist recommends Gabrielle offer a mini-course featuring her module, then track students who took this new course into their first years of teaching, comparing them to students who did not take the course. But Gabrielle knows that results from PSTs' performance as teachers of record are typically not available for at least a year, and possibly several years, after they graduate. This kind of performance data – either from classroom observations, or from student test scores – can also be difficult to obtain and to interpret, given the non-random sorting of PSTs into schools and districts, as well as the non-random attrition of PSTs between course enrollment, program graduation, and actual teaching assignments.

The methodologist's comment also makes Gabrielle think about how to measure her PSTs' outcomes. Even if teacher performance data from classrooms were easily available, it

seems unlikely that a distal measure such as her PSTs' eventual value-added or classroom observation scores would capture outcomes from her module, both because these measures can be insensitive to teacher learning (Sussman & Wilson, 2018) and because her one-week intervention on task selection and launch is unlikely to move the needle on either (Diez, 2010). She conducts a search of major academic databases and uses *EdInstruments* (EdInstruments.com) to locate scholarship that has assessed PST selection, analyses, and implementation of tasks, but finds that most such research is qualitative in nature, and the instruments that do exist don't quite capture the skills she cares about developing in her PSTs.

Given Gabrielle's concerns and challenges, what kinds of measures and designs are feasible? We explore possible answers to this question below. We first discuss how she could approach the challenge of measuring her outcomes of interest, noting that Gabrielle will need to conserve her limited time and research funds at the same time she chooses or builds instruments that return accurate and valid scores. We then turn to discussing Gabrielle's potential research designs. While not every design satisfies all her concerns, we explain how each enables her to progress toward making stronger inferences about PST learning.

Measurement

Key to improving researchers' ability to evaluate new practices in teacher education is work locating or developing outcome measures that: (1) assess the focal teaching knowledge or skill; (2) provide reliable and replicable scores; (3) can be feasibly implemented in the context of teacher education; and, ideally, (4) predict future valued outcomes such as student learning. Distal measures, such as generic classroom observation instruments or teacher value-added scores, are unlikely to capture the effects of new practices in teacher education, for these distal measures capture many different aspects of teaching rather than the specific practice under study,

making them less sensitive to that practice. As well, some of these distal measures may work at cross-purposes to the goals of teacher educators, necessitating measures development within the field. Ideally, these new measures should not double as course assignments. Course assignments are problematic as outcome measures because incentives in grading can distort PST performance (e.g., when PSTs game an assignment), and, if the researchers are teaching the course they plan to evaluate, because the researcher may unknowingly bias grades in favor of finding an effect of the new practice.

Recall that Gabrielle hopes to develop PSTs' skills in selecting cognitively demanding tasks that allow multiple entry points for students; identifying features of tasks that may present barriers to student work; and launching tasks with sensitivity to learners' knowledge and background while maintaining high cognitive demand. Gabrielle begins by breaking down these skills into five distinct constructs (see Table 2). She observes that she can potentially group these constructs together into two measures: (1) a written assessment, in which PSTs select high-cognitive demand tasks with multiple entry points and identify features that affect the accessibility of tasks to students; and (2) a simulation of PSTs actually launching a task in front of peers or graduate assistants playing the role of students (e.g., Shaughnessy & Boerst, 2018).

Gabrielle begins by examining existing instruments, hoping to locate ones that require minimal adaptation for her purposes, thus controlling cost, including the cost of her own time. Locating existing instruments would also make her findings comparable to those from other research teams, thus facilitating later research syntheses. Gabrielle finds two relevant instruments: the Mathematics Scan (M-Scan) (Berry et al., 2012) and the Instructional Quality Assessment (IQA) (Boston, 2017). Both instruments contain an item that captures the potential of a task to result in cognitively demanding work (1a from Table 2 below), and an item that

captures the enacted cognitive demand of tasks (2b). Both instruments define each construct carefully and provide raters guidance on how to assign score points associated with different levels of the construct, something that Gabrielle knows is important given her aim of having reliable and replicable scores.

After a close analysis of the items in M-Scan and IQA, Gabrielle chooses to use the items from M-Scan, as they are more aligned with how she conceptualizes task launch, and because she knows that other faculty in her department use M-Scan, allowing for some synergy across courses. To ensure she uses the instrument as intended, Gabrielle makes plans to attend an M-Scan training.

While Gabrielle has been lucky to find measures for two of her constructs, she cannot find an instrument that assesses whether selected tasks have multiple entry points (1b), whether PSTs can identify contextual features that affect task accessibility to students (1c), nor whether PSTs are sensitive to students' prior knowledge when launching tasks (2a). As such, she decides to develop these from scratch.

Table 2

Measurement Plan

Construct	Potential Assessment	Existing instruments that measure the construct	Plan of action
1a. Selecting a task with potential for high cognitive demand	Task selection and analysis written assessment	IQA (Boston, 2017) M-Scan (Berry et al., 2013)	Use "Cognitive Demand: Task

			Selection" item from M-Scan
1b. Selecting a task with multiple entry points	Task selection and analysis written assessment	Not found	Develop from scratch
1c. Identifying contextual features that affect task accessibility to students	Task selection and analysis written assessment	Not found	Develop from scratch
2a. Launching a task with sensitivity to students' prior knowledge	Simulation of task launch	Not found	Develop from scratch
2b. Launching a task in a way that maintains high cognitive demand	Simulation of task launch	IQA (Boston, 2017) M-Scan (Berry et al., 2013)	Use "Cognitive Demand: Teacher Enactment" item from M-Scan

Because Gabrielle has read several papers describing measures development, she knows that this process follows a relatively straightforward set of steps. First, she writes prompts that would enable PSTs to demonstrate their knowledge, reasoning and skills. She remembers from

her reading of the literature and her courses with Dr. Schilling, her measurement professor, that it is important these prompts gauge each of the constructs independently. For instance, a student's score on 1c, "identifying contextual features that affect task accessibility to students," cannot be contingent on them already getting a high score on 1a, "selecting a task with potential for high cognitive demand." In light of this, Gabrielle decides to standardize parts of her assessment. For task selection (rows 1a and 1b), she will have PSTs select tasks themselves, so she can measure whether these tasks have the potential for high cognitive demand and multiple entry points, but then she will have PSTs all analyze the same task for contextual barriers that affect accessibility. Gabrielle decides to similarly standardize the simulation, asking all PSTs to launch the same task rather than tasks they find, which may vary in potential quality. Her end result is a three-part assessment: (1) a prompt that asks PSTs to select a task that has both high cognitive demand and multiple entry points; (2) a prompt that asks PSTs to analyze a previously unseen task to identify contextual features that affect task accessibility to students; and (3) a prompt that asks PSTs to adapt and then launch a second previously unseen task to two research assistants posing as elementary students.

She pilots these prompts with several PSTs similar to those in her population of interest, because doing so improves the chance that the prompts will yield the data she needs to evaluate her module. For example, she wants to make sure that she sees some variation in responses to the measures across PSTs. If all PSTs respond the same way or if they all perform very well on the tasks, it is unlikely that she will be able to find an effect of her intervention. As well, Gabrielle must see to whether any features of the prompt lead respondents to interpret the prompt the wrong way, or to provide off-base or idiosyncratic answers. Based on her analysis of data returned from these pilots, Gabrielle iteratively revises and pilots the prompts.

Next, Gabrielle must develop a procedure for scoring the data she collects. The goal of this scoring system is to distill responses to the task into easy-to-use units, like scores on a rubric. For the multiple entry point measure, her research team sits down and provisionally scores the pilot data, developing definitions for score points iteratively by successively scoring and discussing PST responses. When the research team feels they have arrived at a set of items and score point definitions, they test for interrater agreement using a fresh set of responses.

If project resources permit, Gabrielle can determine score reliability through a generalizability study (see, e.g., Hill, Kraft & Charalambous, 2014). The benefit of a generalizability study is that it can help determine the optimal number of raters required to produce desired score reliabilities, something that is important if Gabrielle seeks to continue to develop the measure and use it on a larger scale. If project resources permit, Gabrielle can also determine whether scores have predictive validity – for instance, whether scores on the task launch part of the assessment predict the quality of task launches during PSTs' clinical placements. If project resources are constrained, Gabrielle can simply estimate Cohen's kappa to determine rater agreement.

Gabrielle concludes this phase of measure development by documenting the prompts and scoring guidelines in a codebook. Noting that some of her raters are more accurate in either the performance or simulation measures, she plans to have them specialize in rating one or the other to improve the reliability of the scoring procedure. These two elements – carefully designed prompts that elicit targeted PST responses, and scoring systems that compress information from responses into easily manipulated units – help ensure that Gabrielle's measures produce high-quality data for the evaluation of her module, and that that this data can be scored in a reliable and replicable way.

Finally, after this whole process of development, Gabrielle considers how her outcome measures can be not just tools for the evaluation of her module, but also authentic learning experiences for her students. For instance, Gabrielle plans to give the posttest task selection assessment as part of an in-class exercise, after which she will lead a reflective discussion. Gabrielle also plans to create a graded assignment where PSTs will be required to watch and reflect on their task launch. In this way, Gabrielle plans to integrate the measurement process within her course, extending her students' potential learning.

Gabrielle's work toward measuring valued outcomes is both practical and forward-looking. Where possible, she has built on measures developed by other scholars, reducing her cost and allowing for future comparisons between her studies and others. Investing in careful measurement of new outcomes, on the other hand, helps build a base for future researchers by creating measures that reflect the specific knowledge and skills targeted by teacher educators.

Research Designs

Having developed a plan for measuring outcomes, Gabrielle now turns her attention to research design. Through her reading, Gabrielle becomes interested in research designs that use three strategies to make stronger causal inferences. These strategies are: (1) making comparisons between 'untreated' PSTs and 'treated' PSTs (Shadish, Cook, & Campbell, 2001, p. 134); (2) using randomization to the 'treated' and 'untreated' groups in order to increase the likelihood that PSTs in those groups are similar before the treatment begins; and (3) increasing the number of measurement occasions to facilitate both within- and between-person comparisons, and to increase the precision of her estimates. Given Gabrielle's aim is to isolate the specific effects of her module on her PSTs' skills, using each of these three strategies together or in combination can help rule out alternative explanations for any changes in their performance.

Extra-treatment Designs

A design that may work particularly well in her situation, given that she isn't yet sure about incorporating the new module into her regular class, is an extra-treatment design. In this design, Gabrielle would deliver the module to selected participants from her class, but do so outside of regular class hours, as an extra class. Gabrielle finds this design appealing because the pulled-out students would attend class as usual with the control group, ensuring that they don't miss important material.

Figure 2

Extra-Treatment Research Design

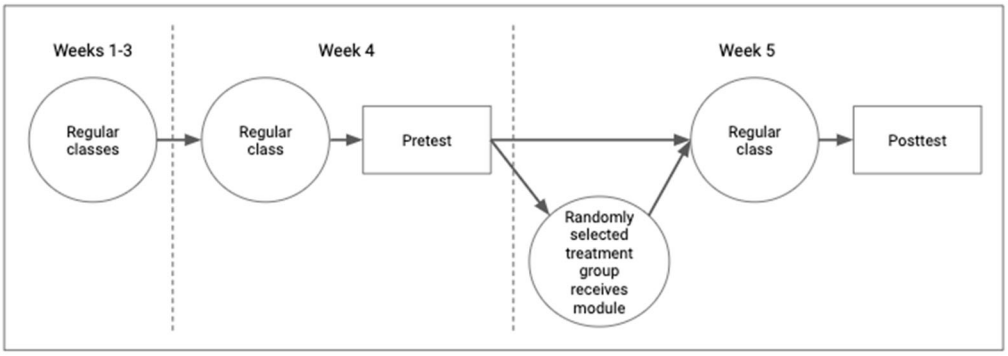


Figure 2 shows a hypothetical extra-treatment design and data collection plan. In this example, all PSTs in both groups (A and B) take the first four weeks of a course as it is typically taught, followed by a pretest. Following the pretest, a group is randomly assigned to treatment. Randomization of PSTs to the extra-treatment group addresses selection effects by making it more likely that the treated and untreated group will be equivalent at the outset of her study, something the pretest can help verify. But randomization can be difficult in her teacher education program – students take other classes, have jobs, or commute to school, and thus are not

available at all times. Gabrielle also knows she will have to "count" in her treatment group the students who are assigned to but cannot attend the extra Week 5 session; once random assignment occurs, students must stay, for analysis, in their randomly assigned group. Although this option seems unappealing, the alternative – contending with selection effects by allowing students to opt into the module – seems much worse. Therefore, Gabrielle decides to find times in her school's master schedule when it seems likely students will be on campus and free. She asks students at the beginning of the semester to hold this date.

After the 'treated' group takes the new module in addition to their regular class during Week 5, both groups complete the posttest. Because of randomization, the performance of groups on the posttest can be directly compared, controlling for pretest scores, to identify whether the module had an effect. The underlying assumption is that because those in the 'treated' group and those in the 'non-treated' group were randomly assigned, we would expect their average outcomes to have been the same if there was no treatment (or if everyone received the treatment). To undertake the comparison between groups, Gabrielle would likely use a simple regression or related statistical model.

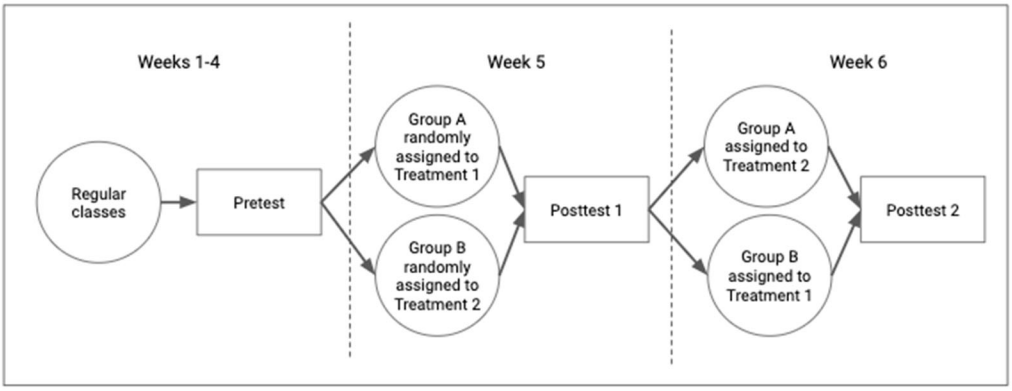
One major challenge of implementing the extra-treatment design in teacher education settings is equity. Gabrielle knows that while she can offer the module to the 'non-treated' group later in the semester, students may not be able to access it because of scheduling conflicts or the end-of-semester time crunch. She recognizes this is a particular concern for postgraduate and/or alternative teacher preparation programs, which are generally shorter and therefore have less flexibility in curriculum and scheduling.

Crossover Designs

Given this equity concern, Gabrielle might consider a crossover design, sometimes called a counterbalanced design. This design is most often used in early-stage research, typically with intact classrooms of instructors and students, making it particularly appealing to teacher educators (e.g. Baylor & Kitsantas, 2005; Bulunuz & Jarrett, 2009). In this design, Gabrielle would compare two treatments that are both implemented to all students but in a different order. As the name implies, the defining characteristic of this design is that participants "cross over" from one treatment to another in the middle of the study. As shown in Figure 3, each group of PSTs is randomly assigned to receive each treatment, yet in a different sequence.

Figure 3

Crossover Design



Recall that Gabrielle's main goals focus on task selection, analysis, and launch. While Gabrielle was developing her module, she considered two approaches to achieving the last goal (2b in Table 1): one that she chose, in which students examine a task and then conduct a simulation of it with peers (Treatment 1); and another option, in which students examine several written tasks and then analyze videos of those tasks launched in classrooms (Treatment 2). Both possibilities are rooted in the work of teaching, and both feature an examination and critique of a

number of tasks. However, during development of her module Gabrielle was not sure which would lead to her PSTs being better able to launch tasks. On one hand, she worried that PSTs may not take the simulation seriously enough, and that doing a simulation with peers absent a close video analysis of expert teachers would deprive her PSTs of the opportunity to study strong practice. On the other hand, she worried that a video-based decomposition would not be a close enough approximation of practice to improve PSTs' skills. Because Gabrielle still remains somewhat undecided about the benefit of one approach over another, she considers the crossover design.

As Figure 3 shows, her intact class would experience the first weeks of the course as a group. During the first phase of the study in Week 5, Gabrielle splits the class randomly, with half (Group A) examining a task then rehearsing, it and the other half (Group B) examining tasks and then watching video. At the end of this first phase of the study, Gabrielle collects a wave of outcome data (Posttest 1) so she can identify the initial conditions' effects on her PSTs. The groups then "cross over" to the other treatment condition. A second round of outcome data collection after this second treatment administration (Posttest 2) enables Gabrielle to generate within-person comparisons. While not common, Gabrielle's approach is similar to how other researchers have adapted crossover designs in the field of teacher education. For example, Baylor & Kitsantas (2005) compare the effects of two different instructional planning scaffolds by having two sections "cross over" in an educational technology course; similarly, Bulunuz & Jarrett (2009) compare the effects of readings, hands-on learning stations, and concept mapping using two sections of a science methods course.

This approach uses two types of comparisons. First, the initial round of experiences provides a comparison of the two treatments. Specifically, the posttest after Week 5 allows for a

direct comparison of the two approaches. The randomization of students to each group helps assure that the posttest would have been similar between the two groups without the specific difference in experience. Second, the posttest after the second treatment in Week 6 allows Gabrielle to see whether the order of Treatment 1 and 2 affects the outcomes (e.g., at the conclusion of the study, PSTs in Group A on average grew 3 units more than PSTs in Group B, suggesting a benefit in videos first then simulations). Again, these comparisons are possible because of randomization. For this design to work best, the outcome of interest needs to be something that PSTs will get better at, but not yet fully master, during the first treatment, and the outcome measure needs to capture performance equally well at all levels of skill such that growth can be reliably identified.

One major challenge of implementing the crossover design in teacher education settings is that intact classes may not contain a sufficient number of students (i.e., power) to detect the effects of treatment with confidence. Class sizes of twenty, for instance, may be too small for differences between treatment groups to be statistically significant, unless these differences are quite large. One solution may be for Gabrielle to join forces with another mathematics teacher educator to conduct her study across multiple classes. Doing so, however, would open up a whole range of other challenges such as ensuring treatment fidelity, guarding against instructor effects, and standardizing curriculum across courses to ensure a fair comparison.

Lab Experiments

If her class size is too small for the extra-treatment or crossover design, and the challenge of partnering with other teacher educators is too great, another option Gabrielle might consider is a lab experiment. While this design has not been used much in teacher education (see Fives & Barnes (2017) for an exception), it is common in other fields like social psychology and

behavioral economics (e.g., Jacoby-Senghor, Sinclair & Shelton, 2016). Here, Gabrielle would recruit students independently of her class and randomly assign them to participate in the module or to serve as a control, with a pretest and posttest for all.

Similar to the extra-treatment and crossover designs, random assignment makes it more likely that the treatment and control group will be equivalent at pretest in terms of key variables such as knowledge, skill, and motivation to learn. This allows for the control group's performance in the posttest to stand in for treatment group performance in the absence of treatment. In other educational settings, such as schools, random assignment studies can be quite large, involving dozens of schools, hundreds of teachers, thousands of students and, often, millions of dollars. In Gabrielle's case, however, a much smaller randomized trial is possible, though it still comes with some costs.

A drawback of this approach centers around the fact that instead of using students in her class as the sole participants in the study, Gabrielle would need to recruit, treat, collect data from, and compensate some number of additional PSTs. To determine how many PSTs, Gabrielle might conduct a power analysis (perhaps by using freeware catalogued at http://powerandsamplesize.com) to determine an adequate study sample size given the magnitude of impact on performance that she anticipates and the level of statistical significance she hopes to reach. Given an outcome measure that is well-predicted by a pretest, a measure sensitive to treatment, and a treatment with a moderate expected effect, as few as 40 PSTs – 20 in the treatment group and 20 in the control – may be possible.

Gabrielle would then need to think about where she could recruit those PSTs. One option is to use all PSTs in Gabrielle's and her colleague's class, as together they may reach the number of PSTs indicated by the power analysis. But Gabrielle knows that it's unlikely that everyone

enrolled in these classes would take part in the study. Instead, Gabrielle may think about recruiting from the intending teacher population generally. In her school, for instance, PSTs take two years of general education and discipline-based coursework before enrolling in math methods, and this population would thus be good candidates for her study. Engaging PSTs who are not enrolled in Gabrielle's class would also alleviate her concern about studying her own students.

Once Gabrielle decides whom to enroll in the study, she would then need to finalize other parts of her design. One decision she faces involves how the module gets administered to those in the study (i.e., the unit of assignment to treatment). One possibility is that the treatment could be delivered individually to each PST; in Gabrielle's case, that means that her module would be delivered as a tutorial. Gabrielle finds this idea appealing, because it would enhance the argument that she has a well-defined treatment that can be replicated across settings and in her absence. However, Gabrielle also wonders whether conducting her task launch module in groups, simulating a regular class, would enhance the external validity of her experiment. Delivery in groups is important given the module is supposed to be completed in a class setting, and because she expects students to learn from one another. Further, delivering the module to small groups instead of one big group is appealing, because it would allow Gabrielle to understand whether there are group-level effects in her data (e.g., the extent to which one group performed better than another because of group composition, day of the week, or some other variable). Whatever she decides, data analysis of randomized trials requires relatively simple methods, typically regression models comparing treatment impact while controlling for pretest performance.

Lab experiments can be implemented quite flexibly by teacher educators. For instance, a study on feedback to PSTs by clinical supervisors could randomize the feedback to be self-reflective or more directive in nature. "Clinical supervisor" impacts could be controlled with a series of binary variables indicating each supervisor and thus representing his or her "effect" (i.e., fixed effects), and treatment fidelity could be ensured through logs and video recordings. Or, a program piloting a new community-based teacher education experience may choose to randomly assign half of a cohort to that experience, half to the more traditional course it replaces. Because most lab experiments need to occur outside class settings, this design is also flexible with regard to what is taught and how intensively; interventions outside the typical curriculum and that last for as long as investigators desire (and that PSTs will participate) are possible. For example, to assess the effects of a scaffold on novice teachers' skills in assessment development, Fives & Barnes (2016) recruit students from their education psychology class to participate in an outside-class research session in exchange for extra credit. Upon entry to the study room, participants were randomly distributed study materials, half of which included the scaffold.

Other Designs

Gabrielle briefly considers two other designs.

Designs that use non-equivalent dependent variables

In these designs, researchers assess a range of similar outcomes before and after participants receive treatment. Some outcomes are specifically targeted by the treatment, other outcomes (i.e., the non-equivalent dependent variables) are not (summary in Figure 4). The effect of the treatment is identified by comparing gains in targeted outcomes (A-C) vs non-targeted outcomes (D). The underlying assumption is that without the treatment, gains in the target outcomes would have been the same as the non-targeted outcomes. Some examples of

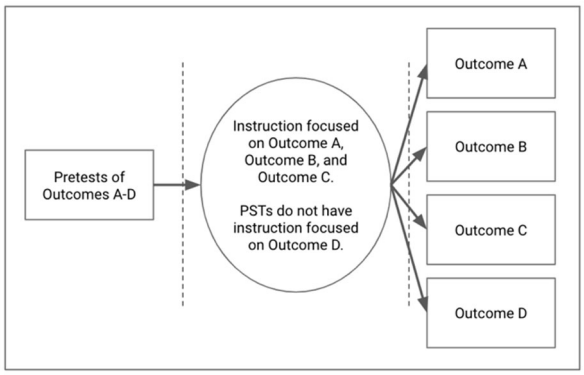
non-equivalent dependent variables that Gabrielle may choose include: PSTs' skills in using redirections to manage student behavior; or, their skills in using student productions to move the lesson forward.

The primary challenge of this design is choosing non-equivalent variables that satisfy all necessary assumptions (Shadish, Cook, & Campbell, 2001). The non-equivalent variable cannot be so similar to the outcome of interest such that any learning is transferred. For instance, Gabrielle cannot use PSTs' knowledge of what makes a complex task, or clarity of instruction-giving as non-equivalent dependent variables, because they are too close to the target outcomes listed in Table 2. Nor can the variable be learned during the study period elsewhere, such as in their practicum or concurrent coursework. Either of these situations would understate the magnitude of any effects, as the non-equivalent variable would be inflated. Alternatively, the non-equivalent variable cannot be so different (e.g., in Gabrielle's case, a list of words to translate into Russian or PSTs' knowledge of professional ethics) such that it does not serve as an adequate baseline and/or capture the effects of other possible influences on the outcomes. This would overstate the magnitude of any effects.

Despite this challenge, this design may be useful to teacher educators who have a large number of potential course topics, but must choose a narrow set for focus during a semester-long course. For instance, Morris and Hiebert (2017) and Kavanagh and Rainey (2017) both compared PST performance on knowledge and skills taught and not taught in their teacher education courses, finding that students performed better on the content taught. While the not-taught content could have been valuable to include in the course, time constraints prevented their teaching it.

Figure 4

Non-equivalent Dependent Variable Design



Non-equivalent comparison groups

Given her students' or institution's scheduling constraints or equity concerns, Gabrielle might think about using one of the above designs but without random assignment, or she may choose to run the module in her class and compare her students' outcomes with those from other classes (i.e., she may use a non-equivalent comparison group design). These designs make many analysts' hearts sink: even with a strong pretest on the outcome of interest, there are few ways to ensure that the two groups are similar enough – in prior skills and knowledge, motivation to improve, or other key characteristics – to serve as fair comparisons to one another.

That said, when a non-equivalent comparison group design is necessary, investigators can try to identify, investigate and eliminate as many threats to validity (see Table 1) as possible. This approach is well-detailed in Shadish, Cook, & Campbell (2001, p. 105). Specific features of teacher education programs may be useful in mitigating such threats. One opportunity stems from the fact that students typically enroll in teacher education programs over multiple years. PSTs drawn from the cohort prior to (or ahead of) those treated may serve as adequate comparisons. For example, Santagata & Yeh (2014) compared the Performance Assessment for

California Teachers (PACT) scores and videos from two cohorts of students: one who participated in their video-analysis course, and one which preceded the development of the course. Investigators also may take advantage of the course enrollment process, which in some places does not allow PSTs much choice in the selection of classes and sections. Investigators can gather information about the enrollment process and, if possible, make the case that assignment to treatment is not contingent on PSTs' preferences for experiences.

Another opportunity stems from the fact teacher education researchers typically have access to students before they become research participants. This means that prior to the introduction of the new practice, researchers can either directly assess the outcome targeted by the study or use a proxy measure (e.g., grade in a foundation course). Pretests of this sort have several advantages. They can help establish baseline equivalence between comparison groups. In the case attrition from either the treated or untreated group, pretests also allow investigators to examine the nature of that attrition, determining whether higher or lower performers dropped out and qualifying conclusions appropriately. Finally, because pretests and posttests are typically correlated, including pretests in analytic models helps explain posttest variability, and can make the identification of treatment effects more efficient, meaning a smaller sample size. Teacher education researchers can also collect multiple pretests from study participants, which allows the researcher to compare pre-treatment learning trajectories and during-treatment learning trajectories.

Other

Finally, investigators can also create new research designs that combine the basic elements discussed above. For inspiration, we refer readers to the section in Shadish, Cook, and Campbell (2001) titled: "Untreated matched controls with multiple pretests and posttests,

nonequivalent dependent variables, and removed and repeated treatments." (p. 153). Clearly, design possibilities are endless. However, in keeping with the ideas we have presented here, teacher educators should gravitate towards designs that apply key principles of rigorous evaluative research design: establishing comparison groups to guard against history and maturation effects, using randomization to increase the likelihood that the comparison and treatment groups are similar, and using pretests to ensure baseline equivalence and to adjust for pre-intervention statuses or trends.

External and Internal Validity

Regardless of the design she selects, Gabrielle knows that she still needs to consider a number of external threats to validity – concerns that the effects she identifies have limited generalizability to other treatments and other settings. She must also think about internal threats to validity – concerns that the effects she identifies are not just due to her module alone.

One issue involves the extent to which results from the studies described above would generalize to other settings. For instance, Gabrielle knows that evaluating a discrete practice like the simulations in her module may provide direct information about its efficacy in this particular context; however, it offers little information about the use of simulations more broadly. This is because simulations differ in their design across teacher education settings (e.g., Kavanagh et al., 2020; Stroupe & Gotwals, 2018) and because, even if replicated, Gabrielle's simulations would be offered in different contexts by different instructors with a different set of students, variations that may affect its efficacy. Many fields address this issue by repeatedly studying new practices and programs across a wide range of settings, building not only general evidence of efficacy but also, ideally, a sense for what adaptations are generative for producing learning.

A second issue is instructor-by-treatment effects, which occur when a single teacher educator provides the treatment under consideration, and which are not solved by our research designs. In these cases, threats to the external validity of study results arise when Gabrielle's enthusiasm for her module means that she unwittingly provides better instruction for those in the treatment condition than those in the control condition, and/or when Gabrielle's expertise means that other instructors in other contexts are unlikely to deliver the intervention with equivalent skill. We see several ways to mitigate the bias associated with instructor-by-treatment effects. First, instructors can document the quality of instruction, describing both the delivery of the new practice and any teaching done in a comparison condition. This documentation may involve a close description of practice or coding for a set of practice-specific and generic indicators (e.g., teacher educator questioning and modeling, PST engagement) and, when carefully used, can serve as evidence for the equivalence of instructional quality across different conditions. A second (and better) strategy to mitigate this issue involves having other instructors, such as graduate students, deliver the new practice to PSTs; doing so would show that the intervention can be replicated by individuals who are not the designers of the program, albeit with close training and supervision. Finally, similar to how both IES and NSF fund progressively larger and more complex studies, teacher educators can replicate their designs with different samples of PSTs, different instructors, and even in different settings.

A third issue is spillover effects – essentially, when treated students share the knowledge, practices, and skills with untreated students – which can affect the internal validity of study results. Several recent studies have identified effects on teacher A's practice when teacher A's peers attend professional development (Gonzalez, 2020; Sun, Penuel, Frank, Gallagher & Youngs, 2013). In an extra-treatment design, for instance, students may share knowledge gained

in the extra class time with students in the control condition. Such sharing would have the result of decreasing any treatment impact. Investigators who choose designs with comparison groups may want to take steps to limit spillovers, and to understand the extent to which they occur. For instance, if Gabrielle were to use an extra-treatment design, she may ask the extra-treatment group to not discuss the treatment with the rest of the class until the conclusion of data collection; she may also survey her students to determine how much social interaction between treatment conditions occurred during the study period, and which study-related topics students discussed during those interactions.

A fourth issue relates to random assignment, in particular, equity concerns regarding PSTs who may not be selected for treatment but – given their background or interests – may need or want to participate. In these cases, Gabrielle may consider pulling these PSTs out of the official study prior to random assignment, allowing them to experience the treatment but not using their data to evaluate the efficacy of the new practice. The key is making this determination prior to the random assignment process. Similarly, to preserve the integrity of random assignment, once students are assigned to one group, they cannot be switched to another, even if the reasons are relatively neutral (e.g., scheduling difficulties).

Conclusion

This paper has introduced approaches to measurement and research design that meet the needs of teacher education researchers interested in assessing the effects of promising new practices. To conclude, we offer some thoughts on how these approaches might fit into the larger field of teacher education research.

First, the evaluative studies we describe above can only answer a narrow set of questions in teacher education, questions focused on identifying which new practices work to improve

certain teacher-level outcomes, and therefore, hopefully, related student-level outcomes. There is clearly a larger universe of questions – for instance, those investigating PSTs' and teacher educators' beliefs, thinking, experiences, and the interplay between them. These questions call for other methodologies, including critical race theory, ethnography, surveys or case studies. We see the research designs proposed above as complementary to others already established in the field, expanding options for researchers interested in questions that involve formal assessments of the efficacy of new teacher education practices.

Second, we see three additional shifts that must take place within the field before we can improve the research designs in evaluative studies. The first shift relates to how research is conceptualized and designed. Similar to harder sciences (Becher, 1989), teacher education must establish relatively more linear and durable lines of research, work that addresses common questions, uses common measures, and thus accumulates knowledge (Grossman & McDonald, 2008). For instance, teacher education researchers may propose a central challenge for PSTs – for instance, responding to students' ideas (Kavanagh et al., 2020) – and then investigate different methods by which program experiences can prepare PSTs to engage in this skill. Or, researchers may choose to focus on a promising practice, like rehearsals, and provide evidence about its efficacy across many contexts, teacher educators, and potential designs. Whatever the case, establishing durable lines of research means that over time, teacher education researchers may establish robust and nuanced evidence regarding the efficacy of practice, something that is not possible in a situation in which teacher education research tends toward entropy in its foci and aims.

The second shift relates to the publishing process. When authors seek to make inferences about PST learning from teacher education experiences, reviewers and editors must become

more demanding of their measures and research designs, preferring those with fewer threats to the validity of the conclusion. Reviewers and editors should ensure that each intervention is defined in enough detail that it is replicable by others wishing to take up the line of research. Further, authors must describe what happens in the "business as usual" condition, so the effects of treatment can be better interpreted. Finally, a common template for reporting methods and results would be useful, for we have noticed that studies inconsistently report several key study characteristics, including how investigators select PSTs to join the study, PST attrition, and the reliability and validity of PST scores on the study's outcomes of interest.

The final shift relates to expectations and support for evaluative teacher education research. Instead of assuming that teacher education researchers can conduct their research "on the side" – often alongside busy teaching schedules – colleges of education must support them with the time and resources necessary to carry off these more labor-intensive designs. Newly minted faculty need start-up packages that fund masters and doctoral students through at least one or two projects, and the field needs funders willing to invest in both better research designs and better measures. Doctoral training institutions must encourage their graduate students to apprentice with faculty carrying out rigorous evaluative research, similar to how STEM graduate students often apprentice in the labs of established scholars. Doctoral training institutions must also examine their curriculum to ensure that coursework focuses on producing scholars who can be active in different epistemological traditions, from interpretivism to positivism, and who can use both quantitative and qualitative data to achieve their goals (see also Wilson, 2006).

References

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. https://doi.org/10.1016/j.tate.2014.09.005

Baylor, A. L., & Kitsantas, A. (2005). A comparative analysis and validation of instructivist and constructivist self-reflective tools (IPSRT and CPSRT) for novice instructional planners. *Journal of Technology and Teacher Education*, *13*(3), 433.

Berry, III, R. Q., Rimm-Kaufman, S. E., Ottmar, E. M., Walkowiak, T. A., & Merritt, E. (2010). *The Mathematics Scan (M-Scan): A Measure of Mathematics Instructional Quality*. Unpublished measure, University of Virginia.

Becher, T. (1989). *Academic tribes and territories: Intellectual enquiry and the cultures of disciplines*. The Society for Research into Higher Education & Open University Press.

Borko, H., Liston, D., & Whitcomb, J. A. (2007). Genres of empirical research in teacher education. *Journal of Teacher Education, 58*(1), 3-11. https://doi.org/10.1177/0022487106296220

Boston, M. (2017). *Instructional Quality Assessment in Mathematics Classroom Observation Toolkit*. Unpublished measure. Duquesne University.

Boyd, D. J., Grossman, P. L., Lankford, H., Loeb, S., & Wyckoff, J. (2009). Teacher preparation and student achievement. *Educational Evaluation and Policy Analysis*, *31*(4), 416–440. https://doi.org/10.3102/0162373709353129

Bravo, M. A., Mosqueda, E., Solís, J. L., & Stoddart, T. (2014). Possibilities and Limits of Integrating Science and Diversity Education in Preservice Elementary Teacher

Preparation. *Journal of Science Teacher Education*, *25*(5), 601–619. https://doi.org/10.1007/s10972-013-9374-8

Brown, K. D. (2014). Teaching in color: A critical race theory in education analysis of the literature on preservice teachers of color and teacher education in the US. *Race Ethnicity and Education*, *17*(3), 326-345. https://doi.org/10.1080/13613324.2013.832921

Bulunuz, N., & Jarrett, O. S. (2009). Understanding of earth and space science concepts: Strategies for concept-building in elementary teacher preparation. *School Science and Mathematics*, *109*(5), 276–289. https://doi.org/10.1111/j.1949-8594.2009.tb18092.x

Cochran-Smith, M., & Zeichner, K. M. (Eds.). (2005). *Studying teacher education: The report of the AERA panel on research and teacher education*. Routledge.

Cochran-Smith, M., Barnatt, J., Friedman, A., & Pine, G. (2009). Inquiry on inquiry: Practitioner research and student learning. *Action in Teacher Education*, 31(2), 17-32. https://doi.org/10.1080/01626620.2009.10463515

Cochran-Smith, M., Villegas, A. M., Abrams, L. W., Chavez-Moreno, L. C., Mills, T., & Stern, R. (2016). Research on teacher preparation: Charting the landscape of a sprawling field. In D. H. Gitomer & C. Bell (Eds.), *Handbook of research on teaching* (5th ed., pp. 439–547). American Educational Research Association. https://doi.org/10.3102/978-0-935302-48-6_7

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. https://doi.org/10.1177/0022487110372927

Draper, R. J., Broomhead, P., Jensen, A. P., & Nokes, J. D. (2012). (Re)imagining literacy and teacher preparation through collaboration. *Reading Psychology*, *33*(4), 367–398. https://doi.org/10.1080/02702711.2010.515858

Durden, T., Dooley, C. M., & Truscott, D. (2016). Race still matters: Preparing culturally relevant teachers. *Race Ethnicity and Education*, *19*(5), 1003-1024. https://doi.org/10.1080/13613324.2014.969226

Fallon, D. (2006). The buffalo upon the chimneypiece: The value of evidence. *Journal of Teacher Education*, *57*(2), 139-154. https://doi.org/10.1177/0022487105285675

Fives, H., & Barnes, N. (2017). Informed and Uninformed Naïve Assessment Constructors' Strategies for Item Selection. *Journal of Teacher Education*, *68*(1), 85–101. https://doi.org/10.1177/0022487116668019

Grossman, P. (2008). Responding to our critics: From crisis to opportunity in research on teacher education. *Journal of Teacher Education*, *59*(1), 10–23. https://doi.org/10.1177/0022487107310748

Grossman, P., & McDonald, M. (2008). Back to the future: Directions for research in teaching and teacher education. *American Educational Research Journal* 48(1), 184-205. https://doi.org/10.3102/0002831207312906

Haddix, M. M. (2017). Diversifying teaching and teacher education: Beyond rhetoric and toward real change. *Journal of Literacy Research*, *49*(1), 141-149. https://doi.org/10.1177/1086296x16683422

Hernandez, C., & Shroyer, M. G. (2017). The use of culturally responsive teaching strategies among Latina/o student teaching interns during science and mathematics instruction of

CLD students. *Journal of Science Teacher Education*, *28*(4), 367–387. https://doi.org/10.1080/1046560x.2017.1343605

Hill, H. C., 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. https://doi.org/10.3102/0013189X12437203

Horn, I. S., & Campbell, S. S. (2015). Developing pedagogical judgment in novice teachers: Mediated field experience as a pedagogy for teacher education. *Pedagogies: An International Journal*, *10*(2), 149-176. https://doi.org/10.1080/1554480x.2015.1021350

Hyland, N. E., & Noffke, S. E. (2005). Understanding diversity through social and community inquiry: An action-research study. *Journal of Teacher Education*, *56*(4), 367–381. https://doi.org/10.1177/0022487105279568

Jackson, K. J., Shahan, E. C., Gibbons, L. K., & Cobb, P. A. (2012). Launching complex tasks. *Mathematics Teaching in the Middle School*, *18*(1), 24-29. https://doi.org/10.5951/mathteacmiddscho.18.1.0024

Jackson, K., Garrison, A., Wilson, J., Gibbons, L., & Shahan, E. (2013). Exploring relationships between setting up complex tasks and opportunities to learn in concluding whole-class discussions in middle-grades mathematics instruction. *Journal for Research in Mathematics Education*, *44*(4), 646-682. https://doi.org/10.5951/jresematheduc.44.4.0646

Jacoby-Senghor, D. S., Sinclair, S., & Shelton, J. N. (2016). A lesson in bias: The relationship between implicit racial bias and performance in pedagogical contexts. *Journal of Experimental Social Psychology*, *63*, 50-55. https://doi.org/10.1016/j.jesp.2015.10.010

Kang, H., & Zinger, D. (2019) What do core practices offer in preparing novice science teachers for equitable instruction? *Science Education*, *103*(4), 823– 853.

https://doi.org/10.1002/sce.21507

Kang, H., Windschitl, M., Stroupe, D., & Thompson, J. (2016). Designing, launching, and implementing high quality learning opportunities for students that advance scientific thinking. *Journal of Research in Science Teaching*, *53*(9), 1316-1340. https://doi.org/10.1002/tea.21329

Kavanagh, S. S., & Rainey, E. C. (2017). Learning to support adolescent literacy: Teacher educator pedagogy and novice teacher take up in secondary English Language Arts teacher preparation. *American Educational Research Journal*, *54*(5), 904–937. https://doi.org/10.3102/0002831217710423

Kavanagh, S. S., Metz, M., Hauser, M., Fogo, B., Taylor, M. W., & Carlson, J. (2020). Practicing responsiveness: Using approximations of teaching to develop teachers' responsiveness to students' ideas. *Journal of Teacher Education*, *71*(1), 94–107. https://doi.org/10.1177/0022487119841884

Lampert, M., Franke, M. L., Kazemi, E., Ghousseini, H., Turrou, A. C., Beasley, H., Cunard, A., & Crowe, K. (2013). Keeping it complex: Using rehearsals to support novice teacher learning of ambitious teaching. *Journal of Teacher Education*, *64*(3), 226–243. https://doi.org/10.1177/0022487112473837

Lee, R. E. (2018). Breaking down barriers and building bridges: Transformative practices in community- and school-based urban teacher preparation. *Journal of Teacher Education*, *69*(2), 118-126. https://doi.org/10.1177/0022487117751127

Lustick, D. (2009). The failure of inquiry: Preparing science teachers with an authentic investigation. *Journal of Science Teacher Education*, *20*(6), 583–604. https://doi.org/10.1007/s10972-009-9149-4

Morris, A. K., & Hiebert, J. (2017). Effects of teacher preparation courses: Do graduates use what they learned to plan mathematics lessons? *American Educational Research Journal*, *54*(3), 524-567. https://doi.org/10.3102/0002831217695217

Rose, D. H., & Meyer, A. (Eds.) (2006). *A Practical Reader in Universal Design for Learning*. Harvard Education Press.

Santagata, R., & Yeh, C. (2014). Learning to teach mathematics and to analyze teaching effectiveness: Evidence from a video- and practice-based approach. *Journal of Mathematics Teacher Education*, *17*(6), 491–514. https://doi.org/10.1007/s10857-013-9263-2

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). Cengage Learning.

Shaughnessy, M., & Boerst, T. A. (2018). Uncovering the skills that preservice teachers bring to teacher education: The practice of eliciting a student's thinking. *Journal of Teacher Education*, *69*(1), 40-55. https://doi.org/10.1177/0022487117702574

Smith, M. S., & Stein, M. K. (1998). Selecting and creating mathematical tasks: From research to practice. *Mathematics teaching in the middle school*, *3*(5), 344-50.

Stroupe, D., & Gotwals, A. W. (2017). "It's 1000 degrees in here when I teach": Providing preservice teachers with an extended opportunity to approximate ambitious instruction. *Journal of Teacher Education*, 69(3), 294–306. https://doi.org/10.1177/0022487117709742

Sun, J., & van Es, E. A. (2015). An exploratory study of the influence that analyzing teaching has on preservice teachers' classroom practice. *Journal of Teacher Education*, *66*(3), 201-214. https://doi.org/10.1177/0022487115574103

Sun, M., Penuel, W. R., Frank, K. A., Gallagher, H. A., & Youngs, P. (2013). Shaping professional development to promote the diffusion of instructional expertise among teachers. *Educational Evaluation and Policy Analysis*, *35*(3), 344–369. https://doi.org/10.3102/0162373713482763

Sussman, J., & Wilson, M. R. (2019). The use and validity of standardized achievement tests for evaluating new curricular interventions in mathematics and science. *American Journal of Evaluation*, *40*(2), 190–213. https://doi.org/10.1177/1098214018767313

U.S. Department of Education (2019). *Integrated Postsecondary Education Data System (IPEDS)*. National Center for Education Statistics. https://nces.ed.gov/ipeds

Wasburn-Moses, L., Noltemeyer, A. L., & Schmitz, K. J. (2015). Initial results of a new clinical practice model: Impact on learners at risk. *The Teacher Educator*, *50*(3), 203–214. https://doi.org/10.1080/08878730.2015.1041313

Wayne, D. B., Didwania, A., Feinglass, J., Fudala, M. J., Barsuk, J. H., & McGaghie, W. C. (2008). Simulation-based education improves quality of care during cardiac arrest team responses at an academic teaching hospital: a case-control study. *Chest*, *133*(1), 56-61. https://doi.org/10.1378/chest.07-0131

Wilson, S. M. (2006). Finding a canon and core: Meditations on the preparation of teacher educator-researchers. *Journal of Teacher Education*, *57*(3), 315-325. https://doi.org/10.1177/0022487105285895

Windschitl, M., Thompson, J., Braaten, M., & Stroupe, D. (2012). Proposing a core set of instructional practices and tools for teachers of science. *Science Education*, *96*(5), 878–903. https://doi.org/10.1002/sce.21027