ANNENBERG
BROWN UNIVERSITY

EdWorkingPaper No. 22-534

Impact Evaluations of Teacher Preparation Practices: Challenges and Opportunities for More Rigorous Research

Zid Mancenido
Harvard University

Many teacher education researchers have expressed concerns with the lack of rigorous impact evaluations of teacher preparation practices. I summarize these various concerns as they relate to issues of internal validity, external validity, and measurement. I then assess the prevalence of these issues by reviewing 166 impact evaluations of teacher preparation practices published in peer-reviewed journals between 2002-2019. Although I find that very few studies address issues of internal validity, external validity and measurement, I highlight some innovative approaches and present a checklist of considerations to assist future researchers in designing more rigorous impact evaluations.

VERSION: February 2022

Suggested citation: Mancenido, Zid. (2022). Impact Evaluations of Teacher Preparation Practices: Challenges and Opportunities for More Rigorous Research. (EdWorkingPaper: 22-534). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/ywnd-4c31

Impact Evaluations of Teacher Preparation Practices: Challenges and Opportunities for More Rigorous Research

Zid Mancenido*
Harvard Graduate School of Education

Abstract

Many teacher education researchers have expressed concerns with the lack of rigorous impact evaluations of teacher preparation practices. I summarize these various concerns as they relate to issues of internal validity, external validity, and measurement. I then assess the prevalence of these issues by reviewing 166 impact evaluations of teacher preparation practices published in peer-reviewed journals between 2002-2019. Although I find that very few studies address issues of internal validity, external validity and measurement, I highlight some innovative approaches and present a checklist of considerations to assist future researchers in designing more rigorous impact evaluations.

Keywords: Impact evaluation, teacher education, literature review

* mancenido@g.harvard.edu; (646)-467-1027; 13 Appian Way, Cambridge MA, 02138.
I thank Heather Hill, Susan Moore Johnson, Eric Taylor. I thank Laura Frustaci for research support. This material is based upon work supported by the National Science Foundation under Grant No.1920616.

Introduction

Researchers and policymakers have long sought to identify the most effective teacher preparation practices, i.e., the "approaches, activities and processes that can be undertaken by teacher educators to develop pre-service teachers' knowledge, skill, and dispositions" (Hill, Mancenido, & Loeb, 2021). A wide variety of teacher preparation practices have been studied, including general education, content area, and methods courses, as well as modules or workshops run within or in parallel to coursework (e.g., Floden & Meniketti, 2005); teacher educator pedagogies such as practice-based approaches (e.g., Forzani, 2014); and, ways of structuring elements of teacher education programs like paired placements during teaching practicum (e.g., Nokes et al., 2008).

Much has been learned from studying these teacher preparation practices through the most prevalent research approaches in the field, including interpretive research, practitioner research, and design research (Borko, Liston & Whitcomb, 2007). Some teacher educator researchers, however, have noted the lack of rigorously designed impact evaluations, i.e., studies able to identify the effects of specific teacher preparation practices on pre-service teachers and their (future) students (Cochran-Smith & Villegas, 2015; Cochran-Smith & Zeichner, 2005; Grossman, 2008; Grossman & McDonald, 2008; Sleeter, 2014; Wilson, Floden, & Ferrini-Mundy, 2002; Zientek et al., 2008). This lack of impact evaluations is a problem because it means researchers are limited in the guidance they can provide teacher educators and policymakers about what works for whom and in what contexts.

The aim of this article is to support teacher education researchers in designing more rigorous impact evaluations of teacher preparation practices. I do this by first identifying the primary issues of internal validity, measurement, and external validity that have hampered the

ability of researchers to make causal inferences about teacher preparation practices over the past two decades. Then I provide examples and guidance for researchers on how they may be able to overcome these issues in future studies.

I begin by defining impact evaluations and justifying the need for improving their quantity and quality in the field of teacher education. Then I review the research designs and measurement approaches of 166 impact evaluations published in peer-reviewed journals between 2002-2019. Finally, I develop a checklist of considerations for designing and reporting on impact evaluations of teacher preparation practices. This checklist aims to support teacher education researchers to move closer towards developing a shared evaluation paradigm (Grossman & McDonald, 2008) by meeting calls for more "clearly articulated criteria for high quality studies" (Grossman, 2008), and for more explicit guidance for peer reviewers "about the kinds of things that should be present in empirical research for it to be published" (Zeichner, 2005).

The Need for Impact Evaluations

This article focuses on *impact* evaluations of teacher preparation practices -- elsewhere described as "effects of teacher education" research (Borko, Liston, & Whitcomb, 2007) and "effectiveness research" (Hill, Mancenido, & Loeb, 2021). Impact evaluations investigate the effects of specific teacher preparation practices on pre-service teachers' knowledge, skills, and/or dispositions, and/or on their (future) students' learning and development. This is different from *process* evaluations (also referred to as *program monitoring* or *implementation analysis*), which generally aims to investigate how certain teacher preparation practices are implemented, and/or the extent to which they are implemented as designed (Humphrey et al., 2016).

According to modern standards for making causal claims in social science research (King, Keohane, Verba, 1994; Mackie, 1974; Shadish, Cook, & Campbell, 2002), teacher

education researchers undertaking impact evaluations need to: (1) clearly identify a relationship between a teacher preparation practice and specific outcomes (i.e., when X changes, Y also changes); (2) ensure that the practice precedes the expected changes in the outcomes (i.e., X happens before Y); and, (3) reject any other reasonable explanations for changes in the outcome (i.e., nothing else can explain the changes in Y except X). Unfortunately, reviews of teacher preparation research over the past two decades consistently lament the lack of research that meet these standards, with most research in the field not designed to isolate the effects of specific teacher preparation practices from the many varied influences on pre-service teacher development (e.g., Wilson, Floden, & Ferrini-Mundy, 2002; Grossman, 2008).

This is problematic for both political and practical reasons. From a political standpoint, teacher preparation programs are increasingly challenged to demonstrate their impact on their graduates' outcomes using causal methods (Cochran-Smith, et al. 2017; Teacher Preparation Issues, 2016; Wineburg, 2006)[1]. As many "insider" critics of the field have expressed (Grossman, 2008; Zeichner, 2005), if teacher education researchers continue to provide few compelling responses to these challenges, we as a field will continue to lose "professional jurisdiction" over our work. This is because lack of causal evidence may continue to be interpreted as lack of impact and, therefore, utility of teacher education (rather than a lack of investment in generating causal evidence); and/or, because policymakers will continue to pay greater attention to (and we as a field will continue to be at the mercy of) research conducted by

[1] In some ways, these recent political pressures can be interpreted as a result of hotly-contested debates in the late 1990s / early 2000s over whether policymakers could better improve overall teaching quality by prioritising improving the quality of teacher preparation or the quality of incoming teacher candidates (e.g., Ballou & Podgursky, 2000; Darling-Hammond, et al., 2005). These debates suffered from a lack of rigorous causal evidence on the impact of teacher preparation, leading many states towards increased accountability pressures to incentivise programs to generate such evidence.

those outside our field (e.g., psychologists and economists)[2]. As Wilson, Floden, & Ferrini-Mundy (2002) warn: "unless we—as teacher educators and researchers—produce sound, robust measures of impact, others—policy makers and critics—will produce other, less appropriate measures" (p. 201).

There is also a practical imperative for increasing the number and improving the quality of impact evaluations of teacher preparation practices. Without much evidence on the teacher preparation practices that lead to more effective teachers, teacher educators have little concrete guidance when deciding how to best prepare their pre-service teachers (Grossman, 2008), and policymakers have little evidence upon which to base decision-making about accreditation requirements, accountability regimes, and grant programs. Recently there has been some guidance coming from program-level evaluations showing certain program characteristics may be particularly effective (e.g., Boyd et al., 2009). However, as Del Schalock et al. (2006) note, these studies "carry little explanatory power as to why, how, or what within teaching or teacher preparation account for relationships found and, thus, have limited utility in guiding or refining policy, practice, or research" (p.110).

The Challenges of Undertaking Impact Evaluations

Despite the political and practical imperatives, many teacher education researchers have acknowledged the difficulties of undertaking impact evaluations. Reasons include: lack of funding and allocated research time for scholars of teacher education (Cochran-Smith, 2004; Cochran-Smith et al., 2012; Feuer, et al., 2013); heavy teaching loads and mostly part-time or

[2] This is not to say that important research cannot or should not be undertaken by non-teacher educators. Rather, it is simply to acknowledge that teacher educators operate in a highly regulated and politically volatile environment where decisions about our work are often not guided by our own research (Mayer, 2021). Further, as Wilson, Floden, & Ferrini-Mundy (2002) warn: "unless we—as teacher educators and researchers—produce sound, robust measures of impact, others—policy makers and critics—will produce other, less appropriate measures" (p. 201).

adjunct status of practicing teacher educators (Borko, Liston, & Whitcomb, 2007; Nuttall et al., 2006; Zeichner, 2005); historical separation between research and practice amongst education school faculty (Goodlad, 1990; Labaree, 2004); limited doctoral training on evaluative methods given the dominance of interpretive and design-based research over the past few decades (Borko, Liston, Whitcomb, 2007; Grossman, 2008; Wilson, 2006; Zientek et al., 2008); the lack of teacher education research published in top-ranked peer-reviewed journals (Grossman, 2008; Wilson, Floden, & Ferrini-Mundy, 2002; Zeichner, 2005); lack of capacity to engage in multi-site and/or multi-disciplinary partnerships so as to undertake large-scale evaluations (Borko, Liston, & Whitcomb, 2007; Del Schalock, et al. 2006); the politicisation of teacher education that leads much research and practice to be reactive to political imperatives and the local teaching market (Grossman & McDonald, 2008; Cochran-Smith & Villegas, 2015).

Although these challenges are significant, I share Nuttall et al.'s concern that:

> these explanations have become something of a 'default setting', and do not go far enough in explicating the complex set of circumstances with which teacher education researchers must engage in order to attain the research quality and credibility necessary to influence initial teacher education policy, as well as teacher education practice (2006: p. 326).

As such, the aim of this article is to explicate -- and support teacher education researchers to address -- one such circumstance that hampers the "research quality and credibility" of impact evaluations in the field: the lack of research design and measurement approaches that meet modern standards for making causal claims.

Issues of Research Design and Measurement

Reviews of teacher education research stretching over the past two decades have raised concerns about the research design and measurement approaches of impact evaluations of teacher preparation practices (Grossman, 2008). For example, in the first comprehensive review

of research on effective teacher preparation, Wilson, Floden, & Ferrini-Mundy (2002) noted the limited empirical evidence base for many common teacher preparation practices and made the first major call for more rigorous impact evaluations in the field. Similarly, Cochran-Smith (2005) writes in her summary of the findings and recommendations of the *AERA Panel on Research and Teacher Education*: "In particular, we need more and better research on the outcomes of teacher education… This requires better data collection and analysis tools for studying outcomes and consistent use of these tools across individual studies" (p. 302). More recently, Sleeter (2014) reviewed the 196 articles published in four top-ranked teacher education journals in 2012, finding: "Only about 1% of the articles reported large-scale mixed-methods studies, only 6% examined the impact of teacher education on teaching practice and/or student learning, and only one did both" (p. 6).

While many teacher education researchers have raised concerns, no single review has summarized these concerns about research design and measurement in one place. As such, in what follows, I bring together themes from teacher education researchers over the past two decades[3] to identify the key issues of internal validity, measurement, external validity that hamper impact evaluations of teacher preparation practices.

Internal validity

Most studies of teacher preparation practices use small-scale, single-case study designs to investigate how a group of pre-service teachers attending a single teacher preparation program experience a specific teacher preparation practice unique to their context (Cochran-Smith &

[3] The reviews and research commentaries consulted to summarise these issues were: Borko, Liston, & Whitcomb (2007); Cochran-Smith & Zeichner (2005); Cochran-Smith & Villegas (2015); Cochran-Smith et al. (2012); Cochran-Smith et al. (2015); Cochran-Smith et al. (2016); Crawford & Tan (2018); Grossman (2008); Grossman & McDonald (2008); Hill, Mancenido, & Loeb (2021); Livingstone & Flores (2017); Mayer (2021); Menter et al. (2010); Nuttall et al. (2006); Sleeter (2014); Wilson (2006); Wilson, Floden, & Ferrini-Mundy (2002); Zientek et al. (2008).

Zeichner, 2005; Cochran-Smith et al., 2016; Grossman, 2008; Menter et al., 2010; Wilson, Floden, & Ferrini-Mundy,, 2002). This single-case approach often allows researchers to develop a deep and nuanced understanding of the dynamic interplay between a teacher preparation practice and pre-service teacher experiences; i.e., how are particular practices in this context enacted, received, perceived, and therefore, learned from?

However, this single-case approach does not allow researchers to identify the causal effects of teacher preparation practices. This is because single-case designs leave open many possible alternative explanations for observed changes in participant outcomes -- i.e., threats to internal validity (Murnane & Willett, 2010; Shadish, Cook, & Campbell, 2002). Three in particular have been flagged by prior reviews as rarely accounted for in impact evaluations: (1) pre-service teachers' natural learning over time (*maturation effects*); (2) pre-service teachers opting into (or out of) receiving a teacher preparation practice that is particularly suited towards them (or not) (*selection effects*); and, (3) improvements caused by other teacher preparation practices that occur at the same time (*history effects*)[4]. Without using a research design that can account for these threats to internal validity, impact evaluations cannot isolate whether study findings are due to the teacher preparation practice or any of these other explanations.

As Zeichner (2005) describes in his summary of the findings and recommendations of the *AERA Panel*, what is needed are research designs that allow for "*systematic analyses of distinct alternatives*" (p. 745), such as those that compare the relative effectiveness of two different teacher preparation practices, or those that compare a novel teacher preparation practice with 'business as usual' ('control'). Doing this will allow researchers to better "disentangle the influence of teacher characteristics from those of their preparation programs" (Zeichner &

[4] See Shadish, Cook, & Campbell (2002) for a full typology of threats to internal validity in causal research, not all of which have been noted by prior reviews of the field.

Conklin, 2005: p. 663). While practically challenging to implement in the context of teacher education, this includes using research designs that account for natural learning over time, such as through comparing the outcomes of a group of study participants who received the preparation practice with a group who did not. It also means using research designs that ensure participants are not able to choose to be part of (or not part of) the treatment group, and that the only difference in the experiences of participants during the study is that one group receives the teacher preparation practice.

Measurement

The second set of issues raised by prior reviews relate to measurement. One commonly raised is that impact evaluations of teacher preparation practices often do not use replicable measurement approaches, such as clearly defined and consistent rubric or rating schemes when measuring participant outcomes (Cochran-Smith et al., 2015; Floden & Meniketti, 2005; Grossman, 2008). Instead, teacher education researchers tend to prefer idiosyncratic measures usually developed through subjective inductive analysis of collected data. These measurement procedures are often not reported in detail. As Zeichner (2005) remarks: "many studies reviewed [by the *AERA Panel*] provide no information about how instruments used for data collection were developed and, validated, or how their reliability was assessed" (p.741). This makes it difficult to assess the validity of researchers' inferences during the peer review process, and/or design future replication studies to help build the evidence base.

A related issue raised by prior reviews is the lack of commonly used measures across impact evaluations. This limits the accumulation of knowledge across the field, because when most studies use idiosyncratic measures, researchers undertaking meta-analyses or literature reviews cannot confidently compare the relative effectiveness of different teacher preparation

practices, nor can they evaluate whether any differences in study findings across contexts are simply due to different measurement approaches (Zeichner, 2005). One contributing factor to this issue is that not enough researchers in the field are producing and openly providing quality measures for use by others (Grossman, 2008; Grossman & McDonald, 2008), particularly at the grain size needed for evaluations of teacher preparation practices (Zeichner, 2005). As Zientek et al. (2008) note, "In the present set of studies, surveys were the dominant mode for collecting data, but only about half (52%) of the researcher-developed surveys were made available" (p.212).

Another issue raised by prior reviews is that impact evaluations of teacher preparation practices rarely investigate effects on pre-service teachers' practice and student learning, preferring more proximal outcomes like pre-service teacher perceptions and understanding[5]. For example, in their review of research on preparing PSTs to teach science, Cochran-Smith, et al. (2016) note most studies "focused on relatively short-term changes in teacher candidates' understandings and beliefs in the context of science methods courses". Similarly, Menter et al. (2010) find in their review of teacher education research in the UK that 60.1% of the 446 studies reviewed use "reflection" as their primary method of data collection. However, as Wilson, Floden, & Ferrini-Mundy (2002) state: "Although it is important to know how teachers feel about the benefits of field experiences, attitude surveys do not answer questions about what prospective teachers actually learn". In other words, just because pre-service teachers learn it in

[5] Cochran-Smith (2004) notes that much of this can be attributed to field norms: "over the past two decades... Very little of [teacher preparation] research was designed to establish empirical linkages to pupils' learning, partly because teachers' knowledge, learning, and beliefs were assumed to be important outcomes of teacher preparation in and of themselves and partly because it was considered self-evident that teachers who knew more, taught better" (p.113). More recently, research has empirically demonstrated: (a) that there is a relationship between teacher knowledge, teacher practice, and student outcomes (e.g., Hill & Chin, 2018; Piasta et al., 2009), and (b) that teachers self-reported perceptions of the effectiveness of teacher professional development often do not match with observed changes in their knowledge, practice, and student outcomes (e.g., Copur-Gencturk & Thacker, 2020; Jacob, Hill, & Corey, 2017).

teacher preparation, doesn't mean that they will do it later on when they are teaching, or that their K-12 students will learn from how it is implemented (Clift & Brady, 2005; Diez, 2010; Gainsburg, 2012). For researchers to identify these longitudinal effects, they need to use measurement approaches that can allow for multiple outcome measures over time -- approaches that allow researchers to track pre-service teachers into their classrooms.

Finally, prior reviews have raised concerns about the extent to which measurement approaches mitigate participant and researcher bias. Two issues are particularly salient: first, the issue of *stakes* -- i.e., using outcome measures, such as course grades or licensure tests, that have stakes attached for participants. Measures like these may show inflated effects because participants may try harder than they would otherwise, and/or researchers may unknowingly bias their grading towards identifying effects. Second, the issue of *reactivity bias* -- i.e., participants changing their behavior to meet their or others' expectations about the effectiveness of the teacher preparation practice. Though admittedly practically challenging to do in the context of teacher education, not accounting for reactivity bias (e.g., by blinding participants to their treatment condition) can lead to potential placebo effects, and/or findings that may primarily be due to participants changing their behaviour as a result of being studied or their "desires to please the instructor" (Clift & Brady, 2005: p. 303).

External validity

The third set of issues raised by prior reviews relate to external validity -- i.e., the extent to which study findings may generalize to other teacher preparation programs and/or other pre-service teachers. This is a particular challenge for the field given research is often conducted within the specific context of a single teacher preparation program (Cochran-Smith et al., 2016; Sleeter, 2014). That said, designing and reporting on external validity is critically important

given there are thousands of diverse and dynamic teacher preparation programs across the US, and policymakers and teacher educators need to know whether and how study findings can scale to their/other contexts (Conaway & Goldhaber, 2020). Researchers also need to understand what works for whom and in what contexts in order to build a shared evidence base (Wilson, Floden, & Ferrini-Mundy, 2002).

When it comes to external validity, prior reviews of the field have primarily raised concerns about the lack of reporting for generalizability (Cochran-Smith et al., 2012; Grossman, 2008; Zeichner, 2005; Zientek et al., 2008). Two issues are particularly salient: (a) the lack of reporting on the representativeness of study participants (*participant generalizability*) -- e.g., by comparing the demographics of study participants to either the demographics of the teacher preparation program they are sampled from and/or the demographics of pre-service teachers in the US; and (b) the lack of discussion on the representativeness of study site(s) (*site generalizability*) -- e.g., by naming study site(s) and/or describing their location, size, program type, and certification level such that others can determine the relevance of findings to other teacher preparation contexts. It is important to note that more is not necessarily better when it comes to reporting on participants and sites. Rather, as Floden & Meniketti (2005) caution, teacher education researchers need to report study details in more standardized ways that enable us to compare across studies and build a shared evidence base:

> when courses [in teacher preparation programs] were described, the descriptions were idiosyncratic. Authors would mention particular texts or describe the academic tasks they posed to students, in ways that offered detail, but gave scant opportunity for linking to other studies. (p. 286)

Prior reviews have also raised issues regarding the extent to which studies are designed for scalability. Impact evaluations of teacher preparation practices are often conducted by their developers, who are often highly enthusiastic about, skilled in, and/or suited to the practice. This

can limit external validity as study findings may not replicate when teacher preparation practices are undertaken by other teacher educators. Unless study authors mitigate these potential *instructor effects* in their research design (e.g., by training instructors to deliver the teacher preparation practice), others cannot assess whether the practice can be scaled to other contexts. A related issue raised by prior reviews is the potential bias of researchers who study their own practices as teacher educators (*researcher bias*). Researchers may unknowingly make and/or overlook reporting on various research design and measurement choices that limit the generalizability of study findings to other contexts. As Wilson, Floden, & Ferrini-Mundy (2002) note: "Although local teacher educator researchers have valuable knowledge of the phenomenon under investigation, critics have the right to raise questions about the conflict of interest apparent for teacher educators doing research that validates the need for teacher education" (p.194).

The Present Study

To aid future researchers in addressing these issues of internal validity, measurement, and external validity, I systematically reviewed the research designs and measurement approaches of impact evaluations of teacher preparation practices conducted over the past two decades. I conducted an online and hand search of studies of teacher preparation practices to identify impact evaluations published in peer-reviewed journals between 2002-2019. I then coded studies based on whether they address the issues of internal validity, measurement, and external validity identified above. My aim was to both empirically assess the prevalence of these issues to help identify priorities and opportunities for the field, as well as highlight innovative and promising approaches that can be emulated by future researchers.

While no other study has focused exclusively on the research design and measurement approaches of impact evaluations of teacher preparation practices, three studies have undertaken

a similar methodology of assessing the state of the field to provide guidance for future researchers. First, Zientek et al. (2008) reviewed the reporting practices of all quantitative research studies cited in *Studying Teacher Education: The Report of the AERA Panel on Research and Teacher Education* (Cochran-Smith & Zeichner, 2005). The authors coded studies based on their statistical methods and the extent to which they report various tests or statistics as recommended by the AERA Standards. Second, Sleeter (2014) investigated the extent to which teacher education research is "designed to influence policy" by reviewing 196 articles published in four top-ranked journals in the field in 2012. Sleeter (2014) categorized studies based on research design and the extent to which they connected teacher preparation policies and practices to impacts on teachers. Finally, Crawford & Tan (2019) investigated the extent to which teacher education research uses mixed methods by reviewing articles published in *Journal of Teacher Education* and *Teaching and Teacher Education* between 2010-2016. The authors coded empirical studies based on whether they used qualitative, quantitative, or mixed methods.

In all three of these studies, the aim in assessing the state of the field was to highlight specific opportunities to advance research -- whether that be more standardised reporting practices in quantitative studies, more research designed to influence policy, or more mixed-methods studies. In this paper, I seek to do this same, but with a focus on supporting teacher education researchers to conduct more rigorous impact evaluations of teacher preparation practices[6].

Method

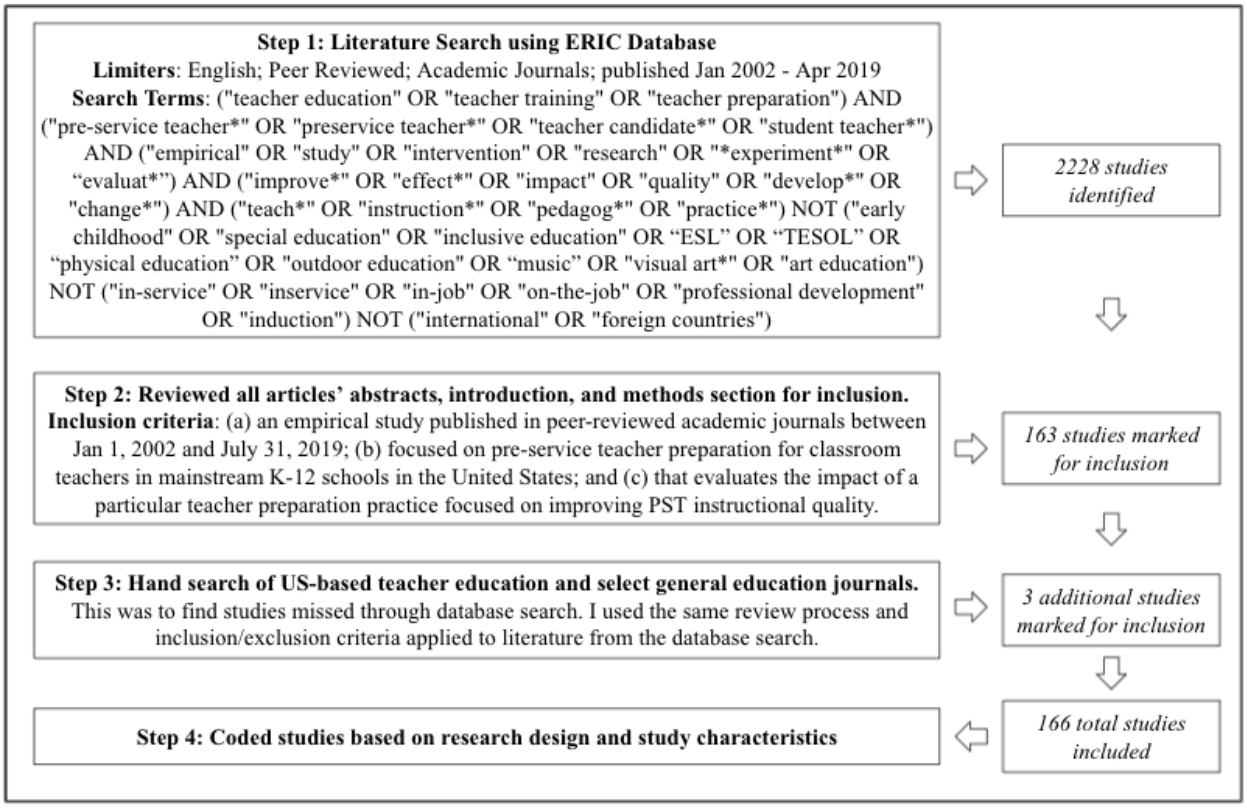
[6] While there is a small overlap in the sample of studies reviewed in this paper and those reviewed by Zientek et al. (2008), Sleeter (2014), and Crawford & Tan (2019), my review analyses these studies differently as my focus is on research design and measurement approaches.

Guided by general principles for undertaking systematic literature reviews (Kennedy, 2007; Polanin, Maynard & Dell, 2017; Slavin, 1986), I first defined my inclusion criteria, and then undertook a four-step process for identifying and assessing studies. This process is summarized in Figure 1 and detailed below.

Inclusion Criteria

I defined the scope of my review as *impact evaluations* of *teacher education practices* focused on *improving K-12 pre-service teachers' instructional quality* that were *published in peer-reviewed journals* between *January, 2002 and April, 2019*. I chose to begin from 2002 because this was the year of the first major call for more rigorous evaluative research in the field (Wilson, Floden, & Ferrini-Mundy, 2002). I also chose 2002 because it was the beginning of increased support for more rigorous impact evaluations in education through the establishment of the Institute of Education Sciences (Whitehurst, 2003). I also only included studies conducted within the US and that had been published in peer reviewed journals. I did this in order to investigate the development of the field since 2002, and to ensure future reviews can more easily replicate my approach and/or compare findings.

Figure 1. Review Procedure



Impact Evaluations

I defined *impact evaluations* as studies intending to investigate the *causal effects* of a specific teacher preparation practice on one or more outcomes. I included studies where authors used causal language like "evaluate", "impact", "effective[ness]", "change", "influence", or "develop" when describing the purpose of the study in their introduction, rationale, research questions, or methods sections. I excluded studies that simply described the implementation of a teacher preparation practice and/or studies that only sought to understand how pre-service teachers experienced a teacher preparation practice. There was no minimum sample size necessary for inclusion; and both quantitative and qualitative studies were included.

Teacher Preparation Practices

I defined teacher preparation practices as either: (1) a course or set of courses organized around a particular theme within a teacher preparation program (e.g. a redesigned science

methods course focused on inquiry; a new course for elementary teachers on sustainability); (2) an activity, pedagogy, or process implemented within a course in a teacher preparation program (e.g. the use of peer feedback on assignments; a short module on assessment within a math methods course); (3) a way of organizing a program component like the practicum (e.g. being assigned a certain type of mentors during practicum; being assigned to practicum placements in pairs); or (4) an experience organized by teacher educators for the purpose of developing pre-service teachers' skills, knowledge, or dispositions (e.g. undertaking a workshop on behavior management; a penpal arrangement between pre-service teachers and Grade 4 students). Based on this definition, I excluded studies that evaluated the effects of whole programs, and those that evaluated a bundle of practices related to a single philosophical approach.

Instructional Quality

Given the wide range of teacher preparation practices studied in the field and my limited resources, I followed the lead of Wilson, Floden, & Ferrini-Mundy (2002) and chose to focus only on teacher preparation practices directly aimed at improving pre-service teachers' *subject matter knowledge* and *pedagogical skills*, as well as teacher preparation practices conducted during or to improve the *clinical experience*. This meant that I included studies of teacher preparation practices that developed pre-service teachers' disciplinary content or pedagogical knowledge or skills, as well as their knowledge and skills in using certain general instructional moves (e.g., for classroom management); assessing and providing feedback to students; developing unit plans and instructional materials; and/or addressing diverse academic and behavioural needs (e.g., supporting English language learners and students with Individualized Education Plans)[7]. Studies did not need to directly measure instructional quality to be included

[7] This focus meant I excluded teacher preparation practices focused on domains more distal to instructional quality such as: family and community engagement, general teaching self-efficacy, skills and self-efficacy in using

(e.g., they could use changes in pre-service teacher knowledge or teacher self-efficacy as the primary outcome).

Step 1 & 2: Online Search and Inclusion

I conducted an online search in May 2019 of the primary US education research database (ERIC) using the limiters and search terms described in Figure 1. This returned an initial pool of 2228 studies. The abstract, introduction, research questions, and methods sections of the studies were reviewed based on the criteria described above to determine whether the study should be included in the review.

To ensure reproducibility, three research assistants were trained on and suggested clarifications to the criteria until independent percent agreement reached above 80%. 600 studies were double-coded in this manner. The remaining studies were coded for inclusion by me. To safeguard against accidental exclusion, two research assistants double-coded 100 randomly selected studies that had been coded only by me. No studies were found to be accidentally excluded.

Based on this review, 2065 of these studies were excluded and 163 were marked for inclusion. This rate of exclusion may seem large but is similar to other systematic reviews of teacher preparation research. For example, Sleeter (2014) analysed the 196 articles published in 2012 in four leading teacher education journals internationally and found "only 6% to examine the impact of teacher education on teaching practice and/or student learning".

Step 3: Hand Search

technology in the classroom, attitudes towards diversity, equity, and inclusion. Although research has shown that these domains are related to instructional quality, I chose to exclude these studies to ensure a manageable sample. I posit that we may be able to assume results from my sample of studies may generalize to studies focused on these other domains. I leave to future reviews to test whether this is empirically true.

Despite the wide online search parameters, I was concerned that key studies may have been excluded. As such, two research assistants undertook hand searches of the *Journal of Teacher Education*, *Teachers and Teaching*, *Teaching and Teacher Education*, *Educational Researcher*, *American Educational Research Journal*, and *Educational Evaluation and Policy Analysis* to identify any studies that were missed. Any studies flagged for inclusion were discussed with me and a consensus decision made. An additional 3 studies were identified through this process.

Characteristics of included studies

A total of 166 studies were included in this review. A reference list is presented in Online Appendix A, and a summary of study characteristics is presented in Table 1. Briefly, most studies evaluated a pedagogy or assessment within a course (54.22%; n=90) or a particular course or set of courses (29.52% n=49). The average number of study participants was 56.5, with 65.4% of studies having 50 or fewer participants, and 85.19% of studies having 100 or fewer participants. The modal study was of pre-service teachers at the undergraduate level (65.05%; n=108), attending traditional preparation programs (57.83%; n=96), and seeking certification at the elementary level (66.27%; n=110). The most common form of data collected was surveys (65.66%; n=109), followed by teacher educators' observations or reflections of pre-service teachers' participation in the teacher preparation practice (41.57%; n=69), pre-service teachers' written reflections or journals (34.94%; n=58), and interviews (34.94%; n=58). Finally, 39.76% (n=66) of studies used statistical analyses on at least one outcome measure.

Table 1. Characteristics of Included Studies

Code	% (n)
Type of teacher preparation practice evaluated	
• A course or set of courses	29.52% (49)
• A pedagogy or assessment within a course	54.22% (90)
• A structuring of a program component	9.64% (16)
• A short program, workshop, or learning experience	6.63% (11)
Study sample	
• Number of participants	Mean = 56.5, Median = 32, Min = 3, Max = 542
• Studies that did not report number of participants	2.41% (4)
Level of teacher preparation program attended by participants	
• Undergraduate	65.06% (108)
• Postgraduate	27.71% (46)
• Not reported	19.88% (33)
Type of teacher preparation program attended by participants	
• Traditional	57.83% (96)
• Alternative	6.02% (10)
• Not reported	39.76% (66)
Certification level of participants	
• Elementary	66.27% (110)
• Middle/high school	36.14% (60)
• Not reported	9.04% (15)
Types of data collected	
• Survey	65.66% (109)
• Teacher educator's observations or reflections	41.57% (69)
• Written reflections or journals	34.94% (58)
• Interviews	34.94% (58)
• Course assessment (e.g., test; assignment)	27.71% (46)
• Observations of the pre-service teacher's instruction (e.g., in practicum settings and simulated environments)	15.06% (25)
• External task (e.g., EdTPA)	3.01% (5)
• Other	6.63% (11)
Quantitative analysis	
• Uses statistical analysis	39.76% (66)

Step 4: Coding

Coding Scheme

I developed my coding scheme in partnership with another researcher with experience conducting systematic literature reviews of impact evaluations of teacher professional development. As no prior review had already identified a set of issues with research design and measurement that could be used as a coding scheme, we developed codes by first establishing a shared understanding of issues raised by prior reviews of the field (as discussed above). We then used our shared understanding to curate and simplify the taxonomy of issues related to internal validity, measurement, and external validity described in Campbell & Stanley (1963) and elaborated by Shadish, Cook, & Campbell (2002). We then added to these codes based on our broad reading of research in program evaluation, causal inference, and research design (e.g. King, Keohane, & Verba, 1994; Murnane & Willett, 2010; Institute of Education Sciences, 2020) as well as our understanding of the contexts of teacher education as researchers and practitioners. We then iteratively refined and clarified this coding scheme using a random sample of included studies.

It is important to note that these codes were not developed to reify a single research design, such as a randomized controlled trial (RCT). I do not subscribe to the view that RCTs are the only way to generate evidence of causation nor that a particular research design is "rigorous" and others are not. This is why I began this review by discussing issues with impact evaluations raised by prior reviews rather than discussing the prevalence of any one particular design. I wanted to ensure that I approached my review from the problems we seek to address (i.e., the issues of internal validity, measurement, and external validity that teacher education researchers

need to overcome), rather than from some preconceived belief that any one particular research design or any particular measurement approach -- qualitative or quantitative -- is the best.

The full coding scheme is presented in Online Appendix B, with key codes relevant to the present study described in the following sections and summarized in Table 2.

Internal Validity. I used three codes to assess whether studies addressed issues of internal validity raised by prior reviews: (a) whether the study used a comparison groups design to account for natural learning over time or 'business as usual' (*maturation effects*)[8]; (b) whether participants were able to choose to be part of the treatment group (or not) (*selection effects*); and (c) whether the only difference between comparison groups was that one group received the teacher preparation practice (*history effects*).

Measurement. I used three sets of codes to assess whether studies addressed issues of measurement. The first set of codes related to whether the study used replicable measures. There were two codes: (1a) whether participant outcomes were measured using a clearly defined and consistent rubric or rating scheme (*replicable measures*); and (1b) whether measures had previously been developed and used by other researchers to enable comparison of results with other studies (*established measures*). The second set of codes related to whether the study measured teacher practice and/or student outcomes. There were two codes: (2a) whether the study collects outcome measures of pre-service teachers' practice and/or student learning (i.e., not just pre-service teachers' perceptions or participation records) (*practice measures*); and (2b) whether the study collects multiple outcome measures over time to assess the persistence of

[8] Comparison groups designs are the most prevalent but not the only approach to addressing maturation effects. Other approaches include single-case designs (Kratochwill et al., 2010) and -- as discussed in the results section -- using unrelated dependent variables (e.g., Morris & Hiebert, 2017). For a comprehensive discussion of alternative designs, see Shadish, Cook, & Campbell (2002). I chose to focus on comparison groups designs as they are the most simple and common way to account for maturation effects in education research.

effects (*longitudinal measures*). The third set of codes related to whether the study addressed issues of participant and researcher bias. There were two codes: (3a) whether the study did not use outcome measures that had stakes attached for participants (*low stakes*); and (3b) whether study authors blinded participants to their treatment condition and/or discuss the effects of participants changing their behavior based on their expectations (*mitigated reactivity bias*).

External Validity. I used two sets of codes to assess whether studies addressed issues of external validity. The first set of codes related to whether the study discussed the representativeness of the study sample. There were two codes: (1a) whether the demographics of study participants were compared to either the demographics of the teacher preparation program they are sampled from and/or the demographics of pre-service teachers in the US (*participant generalizability*); and (1b) whether study authors named the study site(s) (i.e., the teacher preparation program in which the study took place), and/or described the location, size, program type, and certification level of study site(s) (*site generalizability*)[9]. The second set of codes related to whether the study was designed for scalability. There were two codes: (2a) whether multiple instructors were trained to deliver the teacher preparation practice and/or study authors discussed potential issues of having instructors highly enthusiastic about, skilled in, or suited to the teacher preparation practice (*instructor effects*); and (2b) whether the study was undertaken in one of the study authors' courses (*researcher bias*).

Coding Process

All included studies were independently coded by myself and at least one other trained research assistant. Any discrepant codes were discussed and consensus reached. Following all

[9] Another factor "*selectivity"* was considered as part of this code, however, it was not included as very few studies reported on it and there is no standard way for reporting across undergraduate and graduate teacher preparation programs (e.g., acceptance rate, average achievement on standardized tests like the SAT/ACT/GRE, undergraduate GPA).

coding, I calculated summary statistics in order to describe the study characteristics, research designs, and measurement approaches of included studies. I ran correlational analyses and trend analyses to determine any relationships between codes, and whether any types of studies or methods had become more or less prevalent over time. I found nothing meaningful and so do not present these results below, however, they are available upon request. I then went back to the individual codes to identify research designs and measurement approaches that addressed particular issues in innovative ways and could be used as guidance for future researchers. Finally, I synthesised the coding framework, the summary statistics, and the identified noteworthy studies to develop a common checklist of considerations for designing and reporting on impact evaluations of teacher preparation practices.

Table 2. Percent of Impact Evaluations Addressing Issues of Internal Validity, External Validity, and Measurement

Code	% (n)
Internal validity	
Study authors used a research design that systematically analyses distinct alternatives	
● A *comparison groups design* was used to account for natural learning over time or 'business as usual' (*Maturation effects*)	25.9% (43)
● Participants are not able to choose to be part of the treatment group (or choose not to be a part of it). (*Selection effects*)	Of the 43 studies that used comparison groups, 65.12% (28)
● The only difference between comparison groups was that one group received the teacher preparation practice (*History effects*)	Of the 43 studies that use comparison groups, 55.81% (24)
Measurement	
Study authors used replicable measures	
● Participant outcomes were measured using a clearly defined and consistent rubric or rating scheme (*Replicable measures*)	45.18% (75)
● Study authors used measures developed by other researchers to enable comparison of results with other studies (*Established measures*)	18.67% (31)
Study authors measured teacher practice and/or student outcomes	
● Study authors collected outcome measures of pre-service teachers' practice and/or student learning (*Practice measures*)	25.3% (42)
● Study authors collected multiple outcome measures over time to assess the persistence of effects (*Longitudinal measures*)	6.02% (10)

Study authors mitigated participant and researcher bias	
● Study authors do not use outcome measures that have stakes attached for participants (e.g., course grades, licensure tests) (*Low stakes*)	64.46% (107)
● Study authors blinded participants to their treatment condition and/or discussed potential biases from participants changing their behavior based on their expectations (*Mitigated reactivity bias*)	4.22% (7)
External validity	
Study authors discussed the representativeness of the study sample	
● The demographics of study participants were compared to *either* the demographics of the teacher preparation program they are sampled from and/or the demographics of pre-service teachers in the US. (*Participant generalizability*)	13.25% (22)
● The study site(s) were named and/or their location, size, program type, and certification level were described. (*Site generalizability*)	21.69% (36)
Study authors designed the study for scalability	
● Multiple instructors were trained to deliver the teacher preparation practice and/or study authors discussed potential issues of having instructors highly enthusiastic about, skilled in, or suited to it. (*Instructor effects*)	Of the 158 studies that evaluated a practice delivered by an instructor, 10.13% (16)
● The study was not undertaken in one of the study authors' courses. (*Researcher bias*)	Of the 92 studies undertaken in courses and that reported course instructors, 7.61% (7)

Results

In this section, I present my findings on the extent to which the impact evaluations reviewed addressed issues of internal validity, measurement, and external validity. My general finding is that all issues raised by prior reviews remain prevalent in the field, but that there are bright spots in the literature. I highlight these studies as guidance for future researchers, and then present a checklist of considerations for those seeking to undertake more rigorous impact evaluations of teacher preparation practices.

Internal Validity

Systematic Analyses of Distinct Alternatives

Of the 166 studies reviewed, only 26.67% (n=43) addressed maturation effects by using a comparison groups design. Though not initially accounted for in my coding scheme, an additional 3 studies addressed maturation effects by using within-person comparisons of

intended vs unrelated outcomes (1.81%; n=3; design described below). Given only comparison group designs can account for selection or history effects, I restricted analysis of those issues to only the 43 studies that used a comparison group design. Most of the 43 studies addressed selection effects by not allowing participants to choose whether they received the teacher preparation practice, either through randomization (46.51%; n=20), or by comparing cohorts of pre-service teachers over multiple semesters or years (18.60%; n=8). The remaining studies either allowed pre-service teachers to select which group to be in (11.63%; n=5), or did not report how pre-service teachers were assigned to groups (23.26%; n=10). Finally, of the 43 studies that used comparison groups, only 55.81% (n=24) addressed history effects by ensuring that the only difference between comparison groups was that one group received the teacher preparation practice. A summary of research designs used by studies is presented in Table 3. These results empirically affirm concerns raised by prior reviews that very few studies are designed to rigorously evaluate the causal effects of teacher preparation practices.

Table 3. Types of Research Designs Used

Code	% (n)
Type of Research Design	
• Posttest-only single-case	28.31% (47)
• Pretest-posttest single-case	43.98% (73)
• Posttest-only comparison group	10.24% (17)
• Pretest-posttest comparison group	15.66% (23)
• Within-person comparison (i.e., intended vs unrelated outcome)	1.81% (3)
Method of assigning groups for studies using a comparison group design (n=43)	
• Random assignment	46.51% (20)
• Cohort-/time-based	18.60% (8)
• Participants chose group	11.63% (5)
• Did not report	23.26% (10)

Many teacher education researchers have acknowledged the practical challenges of systematically analysing distinct alternatives given the context of teacher education: i.e., teacher

preparation programs are and operate within highly institutionalized organizations, and so researchers often have little control over what, when and how pre-service teachers learn (Boyd et al., 2008; Labaree, 2004). Despite this, some studies reviewed still managed to address issues of maturation effects, selection bias, and history effects by leveraging program structures. For example, Kopcha & Alger (2011) leveraged the fact that pre-service teachers in their program were "placed randomly into one of four cohorts at the beginning of the year" and so assigned treatment at the cohort level. Similarly, Mahalingappa, Hughes, & Polat (2018) randomly assigned the teacher preparation practice being evaluated to one of two sections of a course. The authors reported a minimal likely impact of selection bias given: (a) students had no foreknowledge of which section would be assigned the teacher preparation practice; (b) students enroll in sections primarily based on their schedule; and (c) researchers picked the section to be treated without knowledge of who was enrolled (see also Hanuscin & Zangori, 2016 for a similar design).

Some studies creatively analysed distinct alternatives by comparing groups of pre-service teachers across years or semesters, leveraging changes outside the control of study participants. For example, Matthews and Seaman (2007) evaluated the effects of a required specialised math content course by comparing cohorts the year before and after the course was introduced (see also Mizell & Cates, 2004; Smith et al., 2012). Similarly, Yadav et al. (2014) evaluated the effects of a one-week computational thinking module implemented in a required educational psychology course by comparing outcomes with students who took the same course (sans module) in the previous semester.

Some researchers argue that given the high-stakes nature of teacher preparation, it may be unethical to use comparison group designs if some pre-service teachers are not receiving a

teacher preparation practice that may be beneficial (Sikes, 2006). One way some studies dealt with this issue was to use a counterbalanced design, where one group receives one teacher preparation practice while another group receives a different one; then after a period of time, they switch. For example, Bulunuz & Jarrett (2009) randomly assigned one group of students to one teacher preparation practice (learning stations) for three topics (A, B, C) and a different practice (scaffolded reading) for three different topics (D, E, F); the other group of students was assigned the opposite (i.e., learning stations for topics D, E, F and scaffolded reading for topics A, B, C). Participant outcomes were measured after each set of topics.

Another way studies dealt with ethical concerns while still addressing issues of internal validity was by using within-person comparisons of intended vs unrelated outcomes. Pioneered by Morris & Hiebert (2017) and then replicated by Suppa, DiNapoli, & Mixell (2018) and Hiebert et al. (2019), these studies measure the effects of teacher preparation mathematics courses years after graduation. The studies do this by comparing graduate teachers' skills and knowledge across multiple mathematical content areas, some of which were explicitly taught during teacher preparation courses and some which were not. For example, Hiebert et al. (2019) test graduates on their conceptual understanding and ability to assess students' work in the topics of multiplying whole numbers, subtracting fractions, and dividing fractions (topics taught during their first-year mathematics courses) as well as finding the mean (a topic not taught throughout the teacher preparation program). To mitigate history effects (i.e., the effects of any other teacher preparation practices or other experiences that could lead to differences in knowledge and skills across different mathematics topics), Hiebert et al. (2019) collect data from graduates about other educational experiences, and then systematically test whether those experiences are driving

effects (e.g., other experiences during the preparation program; in-school professional development; the foci of school curricula).

Measurement

Reproducible Measures

Of the 166 studies reviewed, only 45.18% (n=75) measured participant outcomes -- whether qualitatively or quantitatively -- using a clearly defined and consistent rubric or rating scheme. For example, Isabelle & de Groot (2008) established a three-level rubric to analyse the depth of students' written open-ended responses to questions requiring scientific explanations. Bartell et al. (2013) developed a rubric for measuring pre-service teachers' conceptual understanding of mathematical content and their evaluations of children's mathematical understanding. Foley et al. (2017) and Soh, Fowler, & Zygielbaum (2007) coded the complexity of pre-service teachers' concept maps of their disciplinary content knowledge. The remaining studies reviewed (54.82%; n=91) used idiosyncratic measures developed through subjective inductive analysis of collected data.

Of studies reviewed, only 18.67% (n=31) used measures developed by other researchers to enable comparison of results across studies. Some study authors used measures wholesale (e.g., Mathematics Teaching Efficacy Beliefs Instrument (MTEBI); Science Teaching Efficacy Beliefs Instrument (STEBI)); while others adapted or augmented existing measures to accommodate additional outcomes. Using previously established measures allowed some researchers to demonstrate how their study was building on previous findings and developing a shared evidence base. For example, Wilkins & Brand (2004) used the Mathematics Belief Instrument (MBI) developed by Hart (2002). The authors noted how their study findings "provide additional support for the reliability of the initial quantitative findings related to the

MBI presented by Hart (2002)", and that this replication was important given their study "was conducted with a larger sample of preservice teachers". Similarly, Lambert & Bleicher (2013) used an established measure of people's perspectives on climate change, and then compared pre-service teachers' knowledge and beliefs about climate change with the broader US adult population. This comparison allowed them to discuss the magnitude of their study findings in relation to the broader US population.

In most cases, however, study authors did not draw connections between studies despite using measures developed by other researchers. Sometimes this was because no other studies had used the same measure in the context of teacher preparation, as noted by Kopcha & Alger (2011) and McGee & Colby (2014). However, in most cases study authors used established measures such as MTEBI or STEBI but overlooked the opportunity to compare the magnitude of impacts and/or to shed light on similarities across study populations.

Teacher Practice and/or Student Learning

Of the 166 studies reviewed, only 25.3% (n=42) evaluated the effect of a teacher preparation practice on pre-service teachers' practice and/or their students' outcomes. The remaining studies (74.7%; n=124) only collected and analysed data on participants' perceptions of the teacher preparation practice and/or on tests and surveys of their skills, knowledge, or beliefs.

Forty studies collected and analysed data on pre-service teachers' practice. Of these, 37.5% (n=15) used pre-service teachers' self-reports, 42.5% (n=17) used direct observations of participants' instruction, and 20% (n=8) used both. The primary way studies collected data on teacher practice was through pre-service teachers' school placements, whether concurrent to or soon after the teacher preparation practice. For example, Girod & Girod (2006) timed their Web-

based simulation experience in between already-scheduled mentor lesson evaluations of pre-service teachers. This allowed them to evaluate the effects of the simulation using data on pre-service teachers' practice. Some study authors triangulated practice data with other artifacts, such as lesson plans and instructional materials. These artifacts allowed researchers to track impact from teacher preparation practice, to pre-service teacher learning, to pre-service teacher enactment. For example, Bangel et al. (2010) analysed both lesson plans and lesson observations to observe whether what was taught ended up implemented in both instructional planning and instructional practice (see also Welsh & Schaffer, 2017).

One challenge with collecting teacher practice data is that it is resource intensive; teacher education researchers often do not have the research funding to hire raters to conduct classroom observations, nor the time to do it themselves. Some studies overcame this challenge by drawing on data already collected through state licensure requirement such as EdTPA (e.g., Santagata & Yeh, 2014). Other studies tasked pre-service teachers to collect data themselves, usually through course structures like assignments or online discussions. For example, Welsh & Schaffer (2017) used data from four course assignments spread across a semester; each assignment required students to provide a written lesson plan, a 10–15 minute video segment of their teaching the lesson, and a two- or three-page written reflection. Similarly Caughlan et al. (2013) used data from a course discussion board where pre-service teachers were required to post and reflect on multiple 5-minute videos of them teaching. However, Caughlan et al. (2013) admitted that because videos were chosen by pre-service teachers, "any measured change over time is partly an artifact of why each candidate chose each video".

Another challenge with collecting teacher practice data from practicum is that the quality of mentor teachers and placement school may significantly influence pre-service teachers'

practice. This means that any findings could actually reflect mentor teacher or placement school quality rather than the effects of the teacher preparation practice. One study, Kopcha & Alger (2011), addressed this issue by collecting two survey measures of practicum support (*Learning to Teach Scale*; *Learning Climate Questionnaire*), and then testing for any relationship between these variables and the outcome.

Only five studies reviewed (3%) collected data on student outcomes. A summary of these studies is presented in Table 4; they all leveraged distinct course structures that allowed for collection and analysis of student outcomes during practicum placements. Three studies collected perception data (i.e., pre-service teachers' observations, or students' self-reports); two studies collected student learning or behavioral data. For example, Stover, Yearta, & Sease (2014) collected and analysed blog posts that students wrote as part of an electronic pen pal arrangement with pre-service teachers (which was the teacher preparation practice being evaluated). Allen & Blackston (2003) collected multiple direct observations of students' target behaviors. Although pre-service teachers collected and quantitatively coded this observational data, the study authors sent independent observers to each student on three occasions to test for rater reliability.

Table 4. Summary of Studies Collecting Student Outcome Measures

Study	Teacher preparation practice evaluated	Student outcome measure
Allen & Blackston (2003)	Training pre-service teachers in "the use of collaboratively developed intervention scripts… to address academic or behavioral concerns about a target student"	Graphical analysis of trends in students' target behaviours. Student target behaviours were observed and quantitatively coded by pre-service teachers. Independent observers rated a sample of students' target behaviours to assess pre-service teachers' rater reliability.
Bullough et al. (2003)	A peer teaching model, where two student teachers work with one mentor during a student teaching placement	Thematic analysis of data from focus groups of students from participants' classrooms. Students were "asked questions about the advantages and disadvantages of having multiple student teachers".
Stover et al. (2014)	Pre-service teachers blogging with fifth graders ("electronic pen pals") about commonly read texts	Thematic analysis of student blog posts.

Thomas & Sondergeld (2015)	A "scaffolded approach" to teaching pre-service teachers to provide effective feedback on student work	Quantitative analysis of students' perceptions of feedback before and after working on a project with pre-service teachers (who provided them with written feedback at multiple timepoints). Thematic analysis of students' written responses regarding their feelings about receiving feedback on their work and how teachers could make their feedback to students more meaningful.
Weinburgh (2007)	A multi-week investigation of mealworms in an elementary science methods course. As part of this investigation, pre-service teachers brought mealworms to their school placement and taught children about them.	Thematic analysis of pre-service teachers' observations of any changes observed in the organisms and reactions of the children in their school placement.

Finally, of the studies reviewed only 6.02% (n=10) collected outcome measures over multiple time points to assess the persistence of effects (*longitudinal measures*). Three studies of these ten studies followed pre-service teachers across multiple semesters to assess the impact of content or methods courses on pre-service teaching practice during their practicum. For example, in addition to the data they collected during and directly after the science methods course being evaluated, Windschitl, Thompson, Braaten (2008) also conducted multiple lesson observations and debriefs with study participants in their practicum placements a year later (see also Cartwright, Smith, & Hallar, 2014; Ragland, 2016). The remaining seven studies ran the same posttest outcome measure (or a parallel form) multiple times (Hiebert et al., 2019; Isabelle & de Groot, 2008; Morris & Hiebert, 2017; Rodriguez-Valls & Ponce, 2013; Sayeski et al., 2015; Sayeski et al., 2017; Suppa, DiNapoli, & Mixell, 2018). This allowed study authors to assess whether pre-service teachers retained the knowledge and skills that they learned from the teacher preparation practice over time. In four studies, the timespan was relatively short (2-4 weeks) and only occurred once (to guard against participants simply remembering the answers, Sayeski et al. (2015) mixed up the order of test items). However, three studies collected data over multiple years, testing teachers on their pedagogical content knowledge yearly for 2-4 years following graduation (Hiebert et al., 2019; Morris & Hiebert, 2017; Suppa, DiNapoli, & Mixell, 2018).

Participant and Researcher Bias

Of studies reviewed, 35.54% (n=59) used outcome measures that had stakes attached for participants, such as course assessments, course evaluations, participation grades, or high-stakes credential tests. As noted above, these measures are prone to bias as participants may have incentives to inflate their scores, or researchers may unknowingly bias grades in favor of finding effects. That said, using these measures is sometimes necessary given that pre-service teachers have limited capacity to undertake extra work outside class, and collecting additional data is resource intensive. Some researchers mitigated bias when using outcome measures with stakes by simply re-analysing the raw data using a different observation rubric and/or with a different focus. For example, Santagata & Yeh (2014) re-analyse videos of participants teaching that were initially collected and submitted for the Performance Assessment for California Teachers Teaching Event (PACT), a high-stakes performance assessment tied to licensure requirements in California. Regardless of outcome measure, many study authors took extra measures to minimize the issue of stakes. For example, some researchers ensured that those rating and analysing the participant data were blind to treatment status (e.g., James & Scharmann, 2007; Schussler et al., 2017; Sayeski et al., 2017); and other researchers who doubled as course instructors explicitly informed participants that any responses to surveys or interviews would have no influence on grades (e.g., Morrison and McDuffie, 2009).

Of the 166 studies reviewed, only 4.22% (n=7) mitigated reactivity bias by blinding participants to their treatment condition and/or discussing potential biases from participants changing their behavior based on their or others' expectations. Studies did this differently depending on the practice being evaluated. Fives & Barnes (2017) evaluated the impact of providing pre-service teachers with additional guidance when developing assessments. As

participants entered the room where the study was being conducted, they randomly received a package of papers: some included the additional guidance, some did not. As participants completed the task individually and were not told the purpose of the study, they were blind to the fact that some participants were receiving additional guidance. Giebelhaus & Bowman (2002) randomly assigned the practicum mentors of one group of pre-service teachers to additional training. Pre-service teachers were not advised about whether their practicum mentor was given the additional training. Crespo & Sinclair (2008) and Olson et al. (2016) evaluated the impact of various within-course activities on pre-service teachers' knowledge and skills. To address issues of reactivity, participants in both studies were only advised after the course concluded that they were part of a study and asked for consent. Finally, the remaining three studies used the within-person comparison design described above, where researchers purposefully did not tell participants about the research design (i.e., to ensure that they did not respond differently based on the knowledge that they had only been taught some of the content being tested).

External Validity

Representativeness

Although nearly all studies reviewed provided some details about participant demographics and study site contextual factors, few discussed the representativeness of their study participants or sites so as to enable others to determine for whom and in what contexts findings may (or may not) generalize. Of the 166 studies reviewed, only 13.25% (n=22) compared study participant demographics to *either* the demographics of the teacher preparation program they were sampled from and/or the demographics of pre-service teachers in the US. Most discussions of participant representativeness were fairly straightforward. For example, Foley et al. (2017) evaluated a science methods course by randomly selecting 234 of the total

687 pre-service teachers who took the required course sometime between 2012-14. They then compared the sample with the broader population in terms of race and gender. Some studies, however, addressed the issue of representativeness through purposive sampling. For example, to ensure that their evaluation investigated effects across various levels of knowledge and abilities, Santagata & Yeh (2014) randomly selected four participants who scored near the average on a standardized teacher performance assessment (PACT), four who scored 1 SD above the mean, and four who scored 1 SD below (see also Matkins & Bell, 2007).

21.69% (n=36) of studies reviewed either named the study site(s) and/or described their location, size, type, and certification level to enable others to determine the relevance of study findings to other teacher preparation contexts. Addressing this issue is challenging given ethical concerns and institutional policies can limit disclosure of contextual information lest participants become identifiable (Sikes, 2006). This is perhaps why studies with larger samples and/or conducted across multiple different institutions were more likely to report teacher preparation program characteristics or to identify the programs in which the study was conducted.

Scalability

158 studies (95.18%) evaluated teacher preparation practices that were administered by instructors (e.g., courses, pedagogies, or experiences). As noted above, this raises the question of scalability: unless studies are designed to address issues of external validity like instructor effects and researcher bias, then findings may not generalize. Of the 158 studies, only 10.13% (n=16) either trained multiple instructors to deliver the teacher preparation practice and/or discussed issues of using instructors that are highly enthusiastic about, skilled in, or suited to it. In most cases, these studies were multi-site evaluations where multiple instructors implemented the same practice across different course sections and/or teacher preparation programs. For example,

Olson et al. (2016) evaluated how three instructors used various video cases across nine course sections of elementary science methods courses over the span of four semesters. To reduce instructor effects, study authors ensured all instructors taught at least one section across all study conditions, and that sections "used the same course syllabi, objectives, goals, and reading materials to ensure continuity in course content and sequence". Some studies with multiple instructors collected additional implementation data to ensure fidelity (e.g., Bell, Matkins, & Gansneder, 2011). Further, some studies ensured at least one study author did not have any interactions with participants to ensure an "outsider" perspective for data collection and analysis (e.g., Bennet & Hart, 2014; Hart & Bennett, 2013; Hoppey & Mickelson, 2017).

Of the 139 studies that were undertaken within courses, 61.15% (n=85) reported that at least one study author was also a course instructor. As noted above, evaluating one's own practices is often fraught as study authors may unknowingly bias data collection and analysis towards identifying positive findings. That said, evaluating one's own practices is often necessary given the funding, release time, and resources available to teacher education researchers. In these cases, some researchers guarded against issues of researcher bias by ensuring that data collection and analysis was not conducted by course instructors. For example, Hanuscin & Zangori (2016) note that "the consent process was conducted by another individual and the instructor did not know whether or not students consented until the conclusion of the semester". Similarly, Laframboise & Shea (2009) and Duncan, Pilitsis, & Piegaroal. (2010) collaborated with graduate students who conducted interviews with participants and did not share data or preliminary findings with the first author/course instructor until after the course concluded.

Checklist of Considerations

To aid future researchers in developing and reporting on impact evaluations of teacher preparation practices, I present a checklist of considerations in Appendix A. This checklist largely reflects the issues with research design and measurement described above[10]. For each consideration on the checklist, researchers are provided with a simple yes/no question to assess whether they are addressing the issue in their impact evaluation. If the answer is no, then researchers are presented with a follow-up open-ended question to assist them in mitigating the underlying problem. Teacher education researchers can use this checklist at any point during the development, implementation, and write-up of their impact evaluations. Given the complexity of impact evaluations, the aim is to support researchers by presenting the wide range of considerations related to internal validity, measurement, and external validity in a concise and consolidated way (see also Hill, Mancenido, & Loeb, 2021). Peer reviewers, research funders, and policymakers can also use this checklist to demand greater transparency and accountability for more rigorous causal research methods in the field (Zeichner, 2005; Zientek et al., 2008).

Discussion and Conclusion

If we want high quality instruction for all students, we need well prepared teachers. To have well prepared teachers, we need to identify the most effective teacher preparation practices and get policymakers and teacher educators to implement them at scale. To identify the most effective teacher preparation practices and provide clear guidance to policymakers and teacher educators about what works, we need to increase the quantity and rigor of our impact evaluations. To increase the quantity and rigor of our impact evaluations, we need to make tractable and identify ways of addressing issues of research design and measurement that have

[10] Some considerations listed have not been discussed in this review; this is because they have not been raised by the field as a general issue, however are still important considerations about research design and measurement.

been raised by teacher education researchers over the past two decades. This has been the aim of this review.

Like prior reviews, I find that we still have a lot of work to do as a field. Of the 166 studies reviewed, few used research designs that addressed issues of internal validity like maturation effects (27.71%; n=46), selection effects (12.04%; n=20), and history effects (14.46%; n=24). Further, few studies measured the impact of teacher preparation practices on pre-service teachers' instruction, whether observed or self-reported (24.01%; n=40), and even fewer measured the impact on student outcomes (3.01%; n=5). Less than half of the studies (45.18%; n=75) measured outcomes using reproducible measures, and only 18.67% (n=31) used established measures allowing for comparison of results across the field. And finally, few studies discussed the representativeness of their study participants (13.25%; n=25) or study site(s) (21.69%; n=36) to support others in identifying for whom and in what contexts study findings may generalize.

These findings may be skewed by my narrow inclusion criteria (i.e., impact evaluations of teacher preparation practices focused on improving instructional quality) or search procedure (e.g., online search only through ERIC; hand search of only four top-rated journals). For example, I excluded studies that made causal claims in their discussion or conclusion sections (e.g., "these findings show that this practice improves pre-service teachers' teaching skills"), but did not explicitly use causal language in their framing (e.g., "the aim of this study is to understand how pre-service teachers experience this practice"). Although I did this because I did not want to mischaracterize studies, it means that the number of studies reviewed underrepresents how many impact evaluations have been undertaken *in practice*. It also means that my findings may be inflated as studies that do not characterise themselves as impact

evaluations but then make causal claims may be more likely to exhibit issues of internal validity, measurement, and external validity.

Improving the rigor of impact evaluations is challenging but not impossible. Above, I have highlighted a number of studies that have demonstrated innovative and promising approaches for addressing various issues of research design and measurement. I have also presented a checklist of considerations in Appendix A to provide teacher education researchers with more specific guidance on conducting impact evaluations of teacher education. In the sections below, I make more general recommendations about how we can improve as a field.

Focus on Causal Research

Despite being warned by prior reviews about the limited number of impact evaluations, and despite the wide-ranging search terms and thousands of studies screened, I remain struck that only 166 studies were identified for inclusion in this review. As others have noted, this dearth of studies can be partly attributed to teacher education researchers' lack of dedicated funding and time (Cochran-Smith, 2004; Zeichner, 2005; Sleeter, 2012). However, calls for more impact evaluations of teacher preparation practices have been ongoing since 2002, and since then there has been a proliferation of causal research across education research broadly and cognate subfields (e.g., teacher professional development) as well as a significant increase in professional support to undertake this work (e.g., funding through the IES and NSF; establishment of societies and organisations like the *Society for Research on Educational Effectiveness*). Given these advancements elsewhere yet limited development within teacher education, we must own up to our deprivileging of causal work. We must acknowledge that over and above the many challenges of doing this work, many teacher education researchers have chosen not to undertake impact evaluations. Again, this is not to say that the dominant research traditions of the field over

the past two decades have not provided useful insights (Borko, Liston, & Whitcomb, 2007; Cochran-Smith et al., 2016). Rather, it is simply to acknowledge that there continue to remain a number of answerable yet unanswered questions regarding the most effective teacher preparation practices, and that policymakers, teacher educators, and researchers will benefit if we shift our collective focus towards them.

Addressing this at a field-level requires clarifying our theories of action about how various types of research improve teacher preparation, and then re-aligning incentives so that teacher education researchers are more likely to choose to undertake this challenging yet important work. Teacher educators *are* interested in improving their practice. However, faced with the urgency and necessity of the day-to-day of teacher preparation, it is easier and more efficient for teacher education researchers to undertake localised descriptive investigations of one's own practices, rather than more systematic and reproducible impact evaluations of distinct alternatives. But these localised descriptive investigations cannot be the basis for policymakers' decision-making on system-wide policies because they do not give insights into what practices are effective for whom and in what contexts. Only impact evaluations can help policymakers decide whether mandating certain program structures through accreditation requirements, or investing in grant programs that incentivize teacher educators to implement certain pedagogies will improve pre-service teacher learning at scale. However, while policymakers are interested in generating more causal evidence to inform their decision-making, they cannot undertake impact evaluations themselves because they do not run teacher preparation programs and cannot control what teacher educators do in their classrooms.

On a practical level, then, one way forward is to focus on ensuring that doctoral students and early career teacher education researchers have adequate training and support in not just

qualitative and quantitative research methods, but also in research design and measurement (Borko Liston, & Whitcomb, 2007; Grossman, 2008; Wilson, 2006). As others have noted: "That many teacher educators have specialized in research methods most attuned for interpretive and practitioner studies means that as a field, teacher education has less ability to design studies that speak to policymakers' concerns and reflect teacher educators' deep knowledge about learning to teach" (Borko, Liston, & Whitcomb, 2007: p.9). Addressing this gap in training involves going beyond the standard prescription that 'research methods should be tailored to specific research questions', and providing dedicated courses that help researchers-in-training to identify causal questions and design studies that effectively answer them.

Measure Teacher Practice and Student Outcomes

Another recommendation stemming from this review is the need to develop more measures of teaching practice and student outcomes validated for use in the context of teacher education. This includes developing measures that go beyond general assessments of instructional quality, and can detect improvements in the specific domains or parts of instruction targeted by teacher preparation practices. Many researchers are already working on this, including through developing teaching simulations that require observational measures of discrete elements of instruction (Cohen et al., 2020; Shaughnessy & Boerst, 2018); or developing rubrics and rating schemes for assessing teaching artifacts, such as science teaching portfolios (Kloser et al., 2017; Martínez et al., 2012). Importantly, many of these measures are also starting to be publicly shared through repositories like EdInstruments (http://edinstruments.com/) to make it easier for researchers to find and use. Despite these developments, more needs to be done. More measures need to be developed, particularly in non-STEM teaching areas; and teacher education researchers need more encouragement to adopt and use established measures.

That said, even if we had better measures, teacher education researchers must still overcome the practical challenge of actually collecting data on teacher practice and student outcomes. Researchers could provide teachers with recording equipment and task them to collect video of themselves teaching; or alternatively, when teachers are asked to provide samples of instructional materials, they could also provide copies of strong, average, and weak student responses collected through their practicum. If this is done across multiple time points, researchers could assess improvements in teachers' practice and associated impact on student learning. Work like this has already gained some traction in the literature: Waggoner, et al. (2015) discuss how their program tasks all pre-service teachers to collect student-level pre- and post-test data for a unit they teach during practicum; and Brady, et al. (2016) describe how student data from practica can be used to develop "value-added" measures that generate evidence of program effectiveness.

Another possibility is developing standardized instruments and procedures that can be used to evaluate teacher preparation practices that are not discipline-specific. For example, to get around the challenge of study participants being prepared for different subject areas and grade levels, Metcalf & Cruikshank (1991) developed thirty 15-minute "Reflective Teaching" lesson plans which could be prepared for and implemented by participants and then scored by trained raters using a common rubric. Similarly, Jacoby-Senghor, Sinclair, & Shelton (2016) provided participants with materials on Byzantine history and gave them 20-minutes to prepare a seven-minute lesson. Participants taught their lesson to someone who then took a content test. This enabled the authors to measure both instructional quality and student learning. Studies like this offer the prospect of being able to evaluate teacher preparation practices across the whole causal

pathway from what teacher educators do, to what pre-service teachers' learn, to how these teachers teach, to what their students learn.

Clear Guidance for Reporting on Studies

Like prior reviews (Cochran-Smith et al., 2012; Zientek et al., 2008; Zeichner, 2005), I find that studies generally underreport information about study participants, study site(s), and other contextual factors that influence how teacher preparation practices are implemented. This makes it difficult to determine to whom and in what contexts study findings generalize. I recommend, like others have before, that the field invest in developing shared vocabulary and norms for reporting (Borko, Liston & Whitcomb, 2007; Grossman & McDonald, 2008; Zeichner, 2005). This must go beyond the general prescriptions from the AERA Standards (2006; e.g., "The social, historical, or cultural context of the phenomena studied should also be described"), and provide clarity to future researchers about what exactly to report and how. This will be challenging given the diversity of teacher education contexts and the myriad influences that contribute to pre-service teacher development. That said, without clear guidelines for reporting, teacher education researchers will continue to provide either too little (or in some cases, too much) detail, making it difficult for policymakers, teacher educators, and other researchers to compare findings across studies, and determine the relevance of findings across contexts.

Developing shared vocabulary and norms is a significant task, but can be made more tractable if teacher education researchers begin to undertake more linked small-scale studies (Grossman, 2008). As Sleeter (2014) describes: "researchers in different geographic locations... with careful planning and coordinating, could carry out linked small-scale studies that ask the same questions and use the same methodology". This sort of collaboration will require

developing a shared vocabulary and norms for describing the teacher preparation practice (so it can be implemented with fidelity) and reporting on site-specific contextual factors (so that researchers across institutions can identify and/or control for potential influences). From these efforts, other researchers could then conduct replication studies, adding more or less context depending on what they find is important for understanding the generalizability of findings across contexts. These sorts of "chains of inquiry around particular questions and consistently defined outcomes" (Zeichner, 2005: p.742) will enable teacher education researchers to achieve our collective goal of developing a shared evidence base.

Appendix A

Considerations for designing and reporting on impact evaluations of teacher preparation practices

It is challenging but necessary for teacher education researchers to evaluate the causal impacts of teacher preparation practices on pre-service teacher learning and/or student outcomes. The following list of considerations aims to support teacher education researchers in these efforts. Given the diverse and contingent contextual realities of teacher education programs, it is understandable that not all considerations may be able to be fully addressed by every study. As such, considerations have been listed in a loose hierarchy to help teacher education researchers prioritise their efforts.

Consideration	(Y/N)	If N, then…
Internal validity		
1. The research design systematically analyses distinct alternatives		
● A comparison groups design is used to account for natural learning over time or 'business as usual'		If N, how will you know what would have happened to participants had they not received the teacher preparation practice? For example, how will you know that improvements aren't just due to natural learning over time?
● Participants are not able to choose to be part of the treatment group (or choose not to be a part of it).		If N, how will you know that the effects aren't caused by participants choosing to receive the treatment (or not) because it is better for them?
● During the study, the only difference between the treatment group and the control group is that the former receive the teacher preparation practice being evaluated.		If N, how will you know that the effects are due to the treatment and no other teacher preparation practice?
Measurement		
1. Outcome measures are replicable		
● Participant outcomes are measured using a clearly defined and consistent rubric or rating scheme		If N, how can you ensure that other researchers analysing the data would come to the same conclusions about participant outcomes?

● At least one measure used for this study has also been used in other studies	If N, how will you compare your study findings with other researchers'? How will you know how big or small any detected effects are?
2. Outcome measures are fit-for-purpose	
● Outcome measures accurately assess the study's theoretically-defined constructs of interest.	If N, how will you know that any detected effects reflect changes in the target skill, knowledge, or disposition (rather than changes in some other related but not desired construct)?
● Outcome measures can meaningfully differentiate between study participants' performance.	If N, how will you identify differences in outcomes between participants and ensure that any lack of study findings aren't due to noise?
● Outcome measures are fine-grained enough to detect study participants' growth (e.g., changes in the outcome due to the teacher preparation practice).	If N, how will you know that your outcome measures are able to detect any hypothesised effects?
3. Effects on teacher practice and/or student learning are measured	
● Outcome measures of pre-service teachers' practice and/or student learning are collected	If N, how will you know that the teacher preparation practice has effects on teacher practice and/or student learning?
● Outcome measures are collected over time	If N, how will you know whether effects persist?
4. Participant and researcher bias is mitigated	
● Outcome measures and how they are scored mitigates potential researcher or participant bias (e.g., multiple raters blind to treatment status are used)	If N, how will you know that findings are not influenced by the context in which outcomes are collected or the biases held by participants or researchers? For example, how will you know that any self-report outcomes are not influenced by participants' desirability or recall bias? Or that researchers are not scoring outcomes in a way that are biased towards finding an effect?
● Participants are blind to treatment status and/or not given undue expectations about	If N, how will you know that any effects detected are not due to participants changing

the effects of the teacher preparation practice (e.g. they are not told to expect a particular outcome, or given cues to behave differently than they would ordinarily).	their behaviour as a result of being studied? For example, how will you know that any effects detected are not placebo effects?
External validity	
1. The representativeness of the study sample is described	
● The demographics of study participants are compared to either the demographics of the teacher preparation program they are sampled from and/or the demographics of pre-service teachers in the US.	If N, how will you identify the broader population that study findings may generalize to?
● The study site(s) were named and/or their location, size, program type, and certification level were described.	If N, how will you identify potential contextual factors that may be influencing your results so that you and others can determine whether your study findings are applicable to other contexts?
2. The study is designed for scalability	
● Multiple instructors are trained to deliver the teacher preparation practice.	If N, how will you know that study findings can be replicated by other teacher educators in other contexts?
● The teacher preparation practice being evaluated is not being delivered by one of the study authors in one of their courses.	If N, how will you know that study findings are not due to the specific behaviours or context of the study authors? For example, how will you know that findings may not just be limited to someone highly knowledgeable, enthusiastic, and/or suited to the teacher preparation practice?

References

Allen, S. J., & Blackston, A. R. (2003). Training preservice teachers in collaborative problem solving: An investigation of the impact on teacher and student behavior change in real-world settings. *School Psychology Quarterly*, *18*(1), 22–51. https://doi.org/10.1521/scpq.18.1.22.20878

American Educational Research Association. (2006). Standards for Reporting on Empirical Social Science Research in AERA Publications. *Educational Researcher*, *35*(6), 33–40. https://doi.org/10.3102/0013189X035006033

Ballou, D., & Podgursky, M. (2000). Reforming teacher preparation and licensing: What is the evidence? *Teachers College Record*, *102*(1), 5–27.

Bangel, N. J., Moon, S. M., & Capobianco, B. M. (2010). Preservice Teachers' Perceptions and Experiences in a Gifted Education Training Model. *Gifted Child Quarterly*, *54*(3), 209–221. https://doi.org/10.1177/0016986210369257

Bartell, T. G., Webel, C., Bowen, B., & Dyson, N. (2013). Prospective teacher learning: Recognizing evidence of conceptual understanding. *Journal of Mathematics Teacher Education*, *16*(1), 57–79. https://doi.org/10.1007/s10857-012-9205-4

Bennett, S., & Hart, S. (2014). Addressing the 'Shift': Preparing Preservice Secondary Teachers for the Common Core. *Reading Horizons: A Journal of Literacy and Language Arts*, *53*(4).

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

Boyd, D., Grossman, P. L., Hammerness, K., Lankford, R. H., Loeb, S., McDonald, M., Reininger, M., Ronfeldt, M., & Wyckoff, J. (2008). Surveying the Landscape of Teacher Education in New York City: Constrained Variation and the Challenge of Innovation. *Educational Evaluation and Policy Analysis*, *30*(4), 319–343. https://doi.org/10.3102/0162373708322737

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.

Brady, M. P., Miller, K., McCormick, J., & Heiser, L. A. (2018). A Rational and Manageable Value-Added Model for Teacher Preparation Programs. *Educational Policy*, *32*(5), 728–750. https://doi.org/10.1177/0895904816673741

Bullough, R. V., Young, J., Birrell, J. R., Cecil Clark, D., Winston Egan, M., Erickson, L., Frankovich, M., Brunetti, J., & Welling, M. (2003). Teaching with a peer: A comparison of two models of student teaching. *Teaching and Teacher Education*, *19*(1), 57–73. https://doi.org/10.1016/S0742-051X(02)00094-X

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

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Rand McNally & Company.

Cartwright, T., Smith, S., & Hallar, B. (2014). Confronting Barriers to Teaching Elementary Science: After-School Science Teaching Experiences for Preservice Teachers. *Teacher Education & Practice*, *27*(2/3), 464–487.

Caughlan, S., Juzwik, M. M., Borsheim-Black, C., Kelly, S., & Fine, J. G. (2013). English Teacher Candidates Developing Dialogically Organized Instructional Practices. *Research in the Teaching of English*, *47*(3), 212–246.

Clift, R. T., & Brady, P. (2005). Research on Methods Courses and Field Experiences. In M. Cochran-Smith & K. Zeichner (Eds.), *Studying Teacher Education: The report of the AERA Panel on Research and Teacher Education* (pp. 309–424). Lawrence Erlbaum Associates, Inc.

Cochran-Smith, M. (2004). Ask a Different Question, Get a Different Answer: The Research Base for Teacher Education. *Journal of Teacher Education*, *55*(2), 111–115. https://doi.org/10.1177/0022487104262971

Cochran-Smith, M. (2005). Studying Teacher Education: What We Know and Need to Know. *Journal of Teacher Education*, *56*(4), 301–306. https://doi.org/10.1177/0022487105280116

Cochran-Smith, M., Baker, M., Burton, S., Chang, W.-C., Carney, M. C., Fernández, M. B., Keefe, E. S., Miller, A. F., & Sánchez, J. G. (2017). The accountability era in US teacher education: Looking back, looking forward. *European Journal of Teacher Education*, *40*(5), 572–588. https://doi.org/10.1080/02619768.2017.1385061

Cochran-Smith, M., Cannady, M., Mceachern, K. P., Mitchell, K., & Piazza, P. (2012). Teachers' Education and Outcomes: Mapping the Research Terrain. *Teachers College Record*, *114*(10), 1–49.

Cochran-Smith, M., & Villegas, A. M. (2015). Framing Teacher Preparation Research: An Overview of the Field, Part 1. *Journal of Teacher Education*, *66*(1), 7–20. https://doi.org/10.1177/0022487114549072

Cochran-Smith, M., Villegas, A. M., Abrams, L., Chavez-Moreno, L., Mills, T., & Stern, R. (2015). Critiquing Teacher Preparation Research: An Overview of the Field, Part II. *Journal of Teacher Education*, *66*(2), 109–121. https://doi.org/10.1177/0022487114558268

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. A. Bell (Eds.), *Handbook of Research on Teaching* (Fifth, pp. 439–547). American Educational Research Association. https://doi.org/10.3102/978-0-935302-48-6_7

Cochran-Smith, M., & Zeichner, K. M. (2005). *Studying teacher education: The report of the AERA Panel on research and teacher education*. Lawrence Erlbaum Associates, Inc. https://doi.org/10.4324/9780203864043

Cohen, J., Wong, V., Krishnamachari, A., & Berlin, R. (2020). Teacher Coaching in a Simulated Environment. *Educational Evaluation and Policy Analysis*, *42*(2), 208–231. https://doi.org/10.3102/0162373720906217

Conaway, C., & Goldhaber, D. (2020). Appropriate Standards of Evidence for Education Policy Decision Making. *Education Finance and Policy*, *15*(2), 383–396. https://doi.org/10.1162/edfp_a_00301

Copur-Gencturk, Y., Cimpian, J. R., Lubienski, S. T., & Thacker, I. (2020). Teachers' Bias Against the Mathematical Ability of Female, Black, and Hispanic Students. *Educational Researcher*, *49*(1), 30–43. https://doi.org/10.3102/0013189X19890577

Crawford, R., & Tan, H. (2019). Responding to evidence-based practice: An examination of mixed methods research in teacher education. *The Australian Educational Researcher*, *46*(5), 775–797. https://doi.org/10.1007/s13384-019-00310-w

Crespo, S., & Sinclair, N. (2008). What makes a problem mathematically interesting? Inviting prospective teachers to pose better problems. *Journal of Mathematics Teacher Education*, *11*(5), 395–415. https://doi.org/10.1007/s10857-008-9081-0

Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Vasquez Heilig, J. (2005). Does Teacher Preparation Matter? Evidence about Teacher Certification, Teach for America, and Teacher Effectiveness. *Education Policy Analysis Archives*, *13*(42). https://doi.org/10.14507/epaa.v13n42.2005

Del Schalock, H., Schalock, M. D., & Ayres, R. (2006). Scaling Up Research in Teacher Education: New Demands on Theory, Measurement, and Design. *Journal of Teacher Education*, *57*(2), 102–119. https://doi.org/10.1177/0022487105285615

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

Duncan, R. G., Pilitsis, V., & Piegaro, M. (2010). Development of Preservice Teachers' Ability to Critique and Adapt Inquiry-based Instructional Materials. *Journal of Science Teacher Education*, *21*(1), 81–102. https://doi.org/10.1007/s10972-009-9153-8

Feuer, M. J., Floden, R. E., Chudowsky, N., & Ahn, J. (2013). *Evaluation of Teacher Preparation Programs: Purposes, Methods, and Policy Options.* National Academy of Education. https://files.eric.ed.gov/fulltext/ED565694.pdf

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

Floden, R., & Meniketti, M. (2005). Research on the Effects of Coursework in the Arts and Sciences and in the Foundations of Education. In M. Cochran-Smith & K. Zeichner (Eds.), *Studying Teacher Education: The report of the AERA Panel on Research and Teacher Education* (pp. 261–308). Lawrence Erlbaum Associates, Inc.

Foley, R. W., Archambault, L. M., Hale, A. E., & Dong, H.-K. (2017). Learning Outcomes in Sustainability Education Among Future Elementary School Teachers. *Journal of Education for Sustainable Development*, *11*(1), 33–51. https://doi.org/10.1177/0973408217725861

Forzani, F. M. (2014). Understanding "Core Practices" and "Practice-Based" Teacher Education: Learning From the Past. *Journal of Teacher Education*, *65*(4), 357–368. https://doi.org/10.1177/0022487114533800

Gainsburg, J. (2012). Why new mathematics teachers do or don't use practices emphasized in their credential program. *Journal of Mathematics Teacher Education*, *15*(5), 359–379. https://doi.org/10.1007/s10857-012-9208-1

Giebelhaus, C. R., & Bowman, C. L. (2002). Teaching Mentors: Is It Worth the Effort? *The Journal of Educational Research*, *95*(4), 246–254. https://doi.org/10.1080/00220670209596597

Girod, M., & Girod, G. (2006). Exploring the Efficacy of the Cook School District Simulation. *Journal of Teacher Education*, *57*(5), 481–497. https://doi.org/10.1177/0022487106293742

Goodlad, J. I. (1990). *Teachers for our nation's schools* (1st ed.). Jossey-Bass.

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*, *45*(1), 184–205. https://doi.org/10.3102/0002831207312906

Hanuscin, D. L., & Zangori, L. (2016). Developing Practical Knowledge of the Next Generation Science Standards in Elementary Science Teacher Education. *Journal of Science Teacher Education*, *27*(8), 799–818. https://doi.org/10.1007/s10972-016-9489-9

Hart, L. C. (2002). Preservice Teachers' Beliefs and Practice After Participating in an Integrated Content/Methods Course. *School Science and Mathematics*, *102*(1), 4–14. https://doi.org/10.1111/j.1949-8594.2002.tb18191.x

Hart, S. M., & Bennett, S. M. (2013). Disciplinary Literacy Pedagogy Development of STEM Preservice Teachers. *Teacher Education & Practice*, *26*(2), 221–241.

Hiebert, J., Berk, D., Miller, E., Gallivan, H., & Meikle, E. (2019). Relationships Between Opportunity to Learn Mathematics in Teacher Preparation and Graduates' Knowledge for Teaching Mathematics. *Journal for Research in Mathematics Education*, *50*(1), 23–50. https://doi.org/10.5951/jresematheduc.50.1.0023

Hill, H. C., & Chin, M. (2018). Connections Between Teachers' Knowledge of Students, Instruction, and Achievement Outcomes. *American Educational Research Journal*, *55*(5), 1076–1112. https://doi.org/10.3102/0002831218769614

Hill, H. C., Mancenido, Z., & Loeb, S. (2021). Effectiveness Research for Teacher Education. (EdWorkingPaper: 20-252). In *EdWorkingPapers.com*. https://doi.org/10.26300/zhhb-j781

Hoppey, D., & Mickelson, A. M. (2017). Partnership and Coteaching: Preparing Preservice Teachers to Improve Outcomes for Students with Disabilities. *Action in Teacher Education*, *39*(2), 187–202. https://doi.org/10.1080/01626620.2016.1273149

Institute of Education Sciences. (2020). *What Works Clearinghouse™ Standards Handbook Version 4.0*. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

Isabelle, A. D., & de Groot, C. (2008). Alternate Conceptions of Preservice Elementary Teachers: The Itakura Method. *Journal of Science Teacher Education*, *19*(5), 417–435. https://doi.org/10.1007/s10972-008-9105-8

Jacob, R., Hill, H., & Corey, D. (2017). The Impact of a Professional Development Program on Teachers' Mathematical Knowledge for Teaching, Instruction, and Student Achievement. *Journal of Research on Educational Effectiveness*, *10*(2), 379–407. https://doi.org/10.1080/19345747.2016.1273411

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

James, M. C., & Scharmann, L. C. (2007). Using analogies to improve the teaching performance of preservice teachers. *Journal of Research in Science Teaching*, *44*(4), 565–585. https://doi.org/10.1002/tea.20167

Kennedy, M. M. (2007). Defining a Literature. *Educational Researcher*, *36*(3), 139–147. https://doi.org/10.3102/0013189X07299197

King, G., Keohane, R. O., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*. Princeton University Press.

Kloser, M., Borko, H., Martínez, J. F., Stecher, B., & Luskin, R. (2017). Evidence of Middle School Science Assessment Practice From Classroom-Based Portfolios. *Science Education*, *101*(2), 209–231. https://doi.org/10.1002/sce.21256

Kopcha, T. J., & Alger, C. (2011). The impact of technology-enhanced student teacher supervision on student teacher knowledge, performance, and self-efficacy during the field experience. *Journal of Educational Computing Research*, *45*(1), 49–73. psyh. https://doi.org/10.2190/EC.45.1.c

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). Single-case designs technical documentation. *What Works Clearinghouse*. https://files.eric.ed.gov/fulltext/ED510743.pdf

Labaree, D. F. (2004). *The Trouble with Ed Schools*. Yale University Press.

Laframboise, K. L., & Shea, K. (2009). Developing Understanding of Research-based Pedagogy with Preservice Teachers: An Instrumental Case Study. *The Qualitative Report*, *14*(1), 105–128.

Lambert, J. L., & Bleicher, R. E. (2013). Climate Change in the Preservice Teacher's Mind. *Journal of Science Teacher Education*, *24*(6), 999–1022. https://doi.org/10.1007/s10972-013-9344-1

Livingston, K., & Flores, M. A. (2017). Trends in teacher education: A review of papers published in the European journal of teacher education over 40 years. *European Journal of Teacher Education*, *40*(5), 551–560. https://doi.org/10.1080/02619768.2017.1387970

Mackie, J. L. (1974). *The cement of the universe: A study of causation*. Clarendon Press.

Mahalingappa, L., Hughes, E. M., & Polat, N. (2018). Developing Preservice Teachers' Self-Efficacy and Knowledge through Online Experiences with English Language Learners. *Language and Education*, *32*(2), 127–146.

Martínez, J. F., Borko, H., & Stecher, B. M. (2012). Measuring instructional practice in science using classroom artifacts: Lessons learned from two validation studies. *Journal of Research in Science Teaching*, *49*(1), 38–67. https://doi.org/10.1002/tea.20447

Matkins, J. J., & Bell, R. L. (2007). Awakening the Scientist Inside: Global Climate Change and the Nature of Science in an Elementary Science Methods Course. *Journal of Science Teacher Education*, *18*(2), 137–163. https://doi.org/10.1007/s10972-006-9033-4

Matthews, M. E., & Seaman, W. I. (2007). The Effects of Different Undergraduate Mathematics Courses on the Content Knowledge and Attitude towards Mathematics of Preservice Elementary Teachers. *Issues in the Undergraduate Mathematics Preparation of School Teachers*, *1*.

Mayer, D. (2021). The connections and disconnections between teacher education policy and research: Reframing evidence. *Oxford Review of Education*, *47*(1), 120–134. https://doi.org/10.1080/03054985.2020.1842179

McGee, J., & Colby, S. (2014). Impact of an Assessment Course on Teacher Candidates' Assessment Literacy. *Action in Teacher Education*, *36*(5–6), 522–532. https://doi.org/10.1080/01626620.2014.977753

Menter, I., Hulme, M., Murray, J., Campbell, A., Hextall, I., Jones, M., Mahony, P., Procter, R., & Wall, K. (2010). Teacher education research in the UK: The state of the art. *Swiss Journal of Educational Research*, *32*(1), 121–142. https://doi.org/10.24452/sjer.32.1.4829

Metcalf, K. K., & Cruickshank, D. R. (1991). Can Teachers Be Trained to Make Clear Presentations? *The Journal of Educational Research*, *85*(2), 107–116. https://doi.org/10.1080/00220671.1991.10702820

Mizell, J. A., & Cates, J. (2004). The Impact of Additional Content Courses on Teacher Candidates' Beliefs regarding Mathematics Content and Pedagogy. *Issues in the Undergraduate Mathematics Preparation of School Teachers*, *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

Morrison, J., & McDuffie, A. R. (2009). Connecting Science and Mathematics: Using Inquiry Investigations to Learn About Data Collection, Analysis, and Display. *School Science and Mathematics*, *109*(1), 31–44. https://doi.org/10.1111/j.1949-8594.2009.tb17860.x

Murnane, R. J., & Willett, J. B. (2010). *Methods Matter: Improving Causal Inference in Educational and Social Science Research*. Oxford University Press.

Nokes, J. D., Bullough, R. V., Egan, W. M., Birrell, J. R., & Merrell Hansen, J. (2008). The paired-placement of student teachers: An alternative to traditional placements in secondary schools. *Teaching and Teacher Education*, *24*(8), 2168–2177. https://doi.org/10.1016/j.tate.2008.05.001

Nuttall, J., Murray, S., Seddon, T., & Mitchell, J. (2006). Changing Research Contexts in Teacher Education in Australia: Charting new directions. *Asia-Pacific Journal of Teacher Education*, *34*(3), 321–332. https://doi.org/10.1080/13598660600927224

Olson, J. K., Bruxvoort, C. N., & Vande Haar, A. J. (2016). The impact of video case content on preservice elementary teachers' decision-making and conceptions of effective science teaching. *Journal of Research in Science Teaching*, *53*(10), 1500–1523. https://doi.org/10.1002/tea.21335

Piasta, S. B., Connor, C. M., Fishman, B. J., & Morrison, F. J. (2009). Teachers' Knowledge of Literacy Concepts, Classroom Practices, and Student Reading Growth. *Scientific Studies of Reading*, *13*(3), 224–248. https://doi.org/10.1080/10888430902851364

Polanin, J. R., Maynard, B. R., & Dell, N. A. (2017). Overviews in Education Research: A Systematic Review and Analysis. *Review of Educational Research*, *87*(1), 172–203. https://doi.org/10.3102/0034654316631117

Ragland, R. (2016). Implementing an Evidence-Based Reflective Teaching Cycle: Using Scholarly Research in Curriculum Design. *Mid-Western Educational Researcher*, *28*(3), 196–217.

Rodriguez-Valls, F., & Ponce, G. (2013). Classroom, the We Space: Developing Student-Centered Practices for Second Language Learner (SLL) Students. *Education Policy Analysis Archives*, *21*, 55. https://doi.org/10.14507/epaa.v21n55.2013

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

Sayeski, K. L., Earle, G. A., Eslinger, R. P., & Whitenton, J. N. (2017). Teacher candidates' mastery of phoneme-grapheme correspondence: Massed versus distributed practice in teacher education. *Annals of Dyslexia*, *67*(1), 26–41. https://doi.org/10.1007/s11881-016-0126-2

Sayeski, K. L., Kennedy, M. J., Irala, S. de, Clinton, E., Hamel, M., & Thomas, K. (2015). The Efficacy of Multimedia Modules for Teaching Basic Literacy-Related Concepts. *Exceptionality*, *23*(4), 237–257. https://doi.org/10.1080/09362835.2015.1064414

Schussler, D., Frank, J., Lee, T.-K., & Mahfouz, J. (2017). Using Virtual Role-Play to Enhance Teacher Candidates' Skills in Responding to Bullying. *Journal of Technology and Teacher Education*, *25*(1), 30.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Houghton Mifflin.

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

Sikes, P. (2006). On dodgy ground? Problematics and ethics in educational research. *International Journal of Research & Method in Education*, *29*(1), 105–117. https://doi.org/10.1080/01406720500537502

Slavin, R. E. (1986). Best-Evidence Synthesis: An Alternative to Meta-Analytic and Traditional Reviews. *Educational Researcher*, *15*(9), 5–11. https://doi.org/10.3102/0013189X015009005

Sleeter, C. (2014). Toward Teacher Education Research That Informs Policy. *Educational Researcher*, *43*(3), 146–153. https://doi.org/10.3102/0013189X14528752

Smith, M. E., Swars, S. L., Smith, S. Z., Hart, L. C., & Haardorfer, R. (2012). Effects of an Additional Mathematics Content Course on Elementary Teachers' Mathematical Beliefs and Knowledge for Teaching. *Action in Teacher Education*, *34*(4), 336–348. eric.

Soh, L.-K., Fowler, D., & Zygielbaum, A. I. (2007). The Impact of the Affinity Learning Authoring Tool on Student Learning. *Journal of Educational Technology Systems*, *36*(1), 29–62. https://doi.org/10.2190/ET.36.1.d

Stover, K., Yearta, L. S., & Sease, R. (2014). "Experience Is the Best Tool for Teachers": Blogging to Provide Preservice Educators with Authentic Teaching Opportunities. *Journal of Language and Literacy Education*, *10*(2), 99–117.

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.

Suppa, S., DiNapoli, J., & Mixell, R. (2018). Teacher Preparation "Does" Matter: Relationships between Elementary Mathematics Content Courses and Graduates' Analyses of Teaching. *Mathematics Teacher Education and Development*, *20*(2), 25–57.

Thomas, A. F., & Sondergeld, T. (2015). Investigating the Impact of Feedback Instruction: Partnering Preservice Teachers with Middle School Students to Provide Digital, Scaffolded Feedback. *Journal of the Scholarship of Teaching and Learning*, *15*(4), 83–109.

Waggoner, J., Carroll, J. B., Merk, H., & Weitzel, B. N. (2015). Critics and Critical Analysis: Lessons from 19,000 P-12 Students in Candidates' Classrooms. *Teacher Education Quarterly*, *42*(1), 97–108.

Weinburgh, M. (2007). The Effect of Tenebrio obscurus on Elementary Preservice Teachers' Content Knowledge, Attitudes, and Self-efficacy. *Journal of Science Teacher Education*, *18*(6), 801–815. https://doi.org/10.1007/s10972-007-9073-4

Welsh, K. A., & Schaffer, C. (2017). Developing the Effective Teaching Skills of Teacher Candidates during Early Field Experiences. *Educational Forum*, *81*(3), 301–321.

Whitehurst, G. J. R. (2003). The Institute of Education Sciences: New Wine, New Bottles. *American Educational Research Association 2003 Annual Meeting Presidential Invited Session*.

Wilkins, J. L. M., & Brand, B. R. (2004). Change in Preservice Teachers' Beliefs: An Evaluation of a Mathematics Methods Course. *School Science and Mathematics*, *104*(5), 226–232. https://doi.org/10.1111/j.1949-8594.2004.tb18245.x

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

Wilson, S. M., Floden, R. E., & Ferrini-Mundy, J. (2002). Teacher Preparation Research: An Insider's View from the Outside. *Journal of Teacher Education*, *53*(3), 190–204. https://doi.org/10.1177/0022487102053003002

Windschitl, M., Thompson, J., & Braaten, M. (2008). How Novice Science Teachers Appropriate Epistemic Discourses Around Model-Based Inquiry for Use in Classrooms. *Cognition and Instruction*, *26*(3), 310–378. https://doi.org/10.1080/07370000802177193

Wineburg, M. S. (2006). Evidence in Teacher Preparation: Establishing a Framework for Accountability. *Journal of Teacher Education*, *57*(1), 51–64. https://doi.org/10.1177/0022487105284475

Yadav, A., Mayfield, C., Zhou, N., Hambrusch, S., & Korb, J. T. (2014). Computational Thinking in Elementary and Secondary Teacher Education. *Trans. Comput. Educ.*, *14*(1), 5:1-5:16. https://doi.org/10.1145/2576872

Zeichner, K. (2005). A research agenda for teacher education. In M. Cochran-Smith & K. Zeichner (Eds.), *Studying Teacher Education: The report of the AERA Panel on Research and Teacher Education* (pp. 737–759). Lawrence Erlbaum Associates, Inc.

Zeichner, K., & Conklin, H. G. (2005). Teacher Education Programs. In M. Cochran-Smith & K. Zeichner (Eds.), *Studying Teacher Education: The report of the AERA Panel on Research and Teacher Education* (pp. 645–735). Lawrence Erlbaum Associates, Inc.

Zientek, L. R., Capraro, M. M., & Capraro, R. M. (2008). Reporting Practices in Quantitative Teacher Education Research: One Look at the Evidence Cited in the AERA Panel Report. *Educational Researcher*, *37*(4), 208–216. https://doi.org/10.3102/0013189X08319762

Online Appendix A
List of Included Studies

- Abendroth, M., Golzy, J. B., & O'Connor, E. A. (2011). Self-Created Youtube Recordings of Microteachings: Their Effects upon Candidates' Readiness for Teaching and Instructors' Assessment. Journal of Educational Technology Systems, 40(2), 141–159.
- Adams, J. D., & Gupta, P. (2017). Informal science institutions and learning to teach: An examination of identity, agency, and affordances. Journal of Research in Science Teaching, 54(1), 121–138.
- Adler, R. F., & Kim, H. (2018). Enhancing future K-8 teachers' computational thinking skills through modeling and simulations. Education and Information Technologies, 23(4), 1501–1514.
- Adler, R. F., & Kim, H. (2018). Enhancing future K-8 teachers' computational thinking skills through modeling and simulations. Education and Information Technologies, 23(4), 1501–1514.
- Akcay, H., & Yager, R. (2010). Accomplishing the Visions for Teacher Education Programs Advocated in the National Science Education Standards. Journal of Science Teacher Education, 21(6), 643–664.
- Allen, S. J., & Blackston, A. R. (2003). Training preservice teachers in collaborative problem solving: An investigation of the impact on teacher and student behavior change in real-world settings. School Psychology Quarterly, 18(1), 22.
- Ambrose, R. (2004). Initiating Change in Prospective Elementary School Teachers' Orientations to Mathematics Teaching by Building on Beliefs. Journal of Mathematics Teacher Education, 7(2), 91–119.
- Anhalt, C. O., & Cortez, R. (2016). Developing understanding of mathematical modeling in secondary teacher preparation. Journal of Mathematics Teacher Education, 19(6), 523–545.
- Artzt, A. F., Sultan, A., Curcio, F. R., & Gurl, T. (2012). A capstone mathematics course for prospective secondary mathematics teachers. Journal of Mathematics Teacher Education, 15(3), 251–262.
- Bahr, D., Monroe, E. E., & Shaha, S. H. (2013). Examining Preservice Teacher Belief Changes in the Context of Coordinated Mathematics Methods Coursework and Classroom Experiences: Coordinated Methods and Classroom Experiences. School Science and Mathematics, 113(3), 144–155.
- Baker, R. S., & Milner, J. O. (2006). Complexities of Collaboration: Intensity of Mentors' Responses to Paired and Single Student Teachers. Action in Teacher Education, 28(3), 61–72.
- Bangel, N. J., Moon, S. M., & Capobianco, B. M. (2010). Preservice Teachers' Perceptions and Experiences in a Gifted Education Training Model. Gifted Child

Quarterly, 54(3), 209–221.
- Bangert, A., & Kelting-Gibson, L. (2006). Teaching Principles of Assessment Literacy through Teacher Work Sample Methodology. Teacher Education and Practice, 19(3), 351–364.
- Barlow, A. T., & Reddish, J. M. (2005). Teacher Candidates' Conceptual Understandings of Mathematics Concepts. Mid-Western Educational Researcher, 18(4), 19–24.
- Bartell, T. G., Webel, C., Bowen, B., & Dyson, N. (2013). Prospective teacher learning: Recognizing evidence of conceptual understanding. Journal of Mathematics Teacher Education, 16(1), 57–79.
- Baxter, B. K., Jenkins, C. C., Southerland, S. A., & Wilson, P. (2004). Using a Multilevel Assessment Scheme in Reforming Science Methods Courses. Journal of Science Teacher Education, 15(3), 211–232.
- Baylor, A. L. (2002). Expanding preservice teachers' metacognitive awareness of instructional planning through pedagogical agents. Educational Technology, Research and Development; New York, 50(2), 5.
- 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.
- Bell, R. L., Matkins, J. J., & Gansneder, B. M. (2011). Impacts of contextual and explicit instruction on preservice elementary teachers' understandings of the nature of science. Journal of Research in Science Teaching, 48(4), 414–436.
- Benken, B. M., & Brown, N. (2008). Integrating Teacher Candidates' Conceptions of Mathematics, Teaching, and Learning: A Cross-University Collaboration. Issues in the Undergraduate Mathematics Preparation of School Teachers, 1-15
- Bennett, S., & Hart, S. (2014). Addressing the 'Shift': Preparing Preservice Secondary Teachers for the Common Core. Reading Horizons: A Journal of Literacy and Language Arts, 53(4).
- Bickmore, B. R., Thompson, K. R., Grandy, D. A., & Tomlin, T. (2009). Science As Storytelling for Teaching the Nature of Science and the Science-Religion Interface. Journal of Geoscience Education, 57(3), 178–190.
- Bintz, W., & Shake, M. (2005). From University to Classrooms: A Preservice Teachers' Writing Portfolio Program and its Impact on Instruction in Teaching Strategies for Writing Portfolios in the Classroom. Reading Horizons, 45(3), 18.
- Birrell, J. R., & Bullough, R. V. (2005). Teaching with a Peer: A Follow-Up Study of the 1st Year of Teaching. Action in Teacher Education, 27(1), 72–81.
- Bischoff, P. J. (2006). The role of knowledge structures in the ability of preservice elementary teachers to diagnose a child's understanding of molecular kinetics. Science Education, 90(5), 936–951.
- Bischoff, P. J., & Golden, C. F. (2003). Exploring the Role of Individual and Socially Constructed Knowledge Mobilization Tasks in Revealing Preservice Elementary

Teachers' Understandings of a Triangle Fraction Task. School Science and Mathematics, 103(6), 266–273.
- Boesdorfer, S. B., & Asprey, L. M. (2017). Exploratory Study of the Teaching Practices of Novice Science Teachers Who Participated in Undergraduate Science Education Research. Electronic Journal of Science Education, 21(3), 21–45.
- 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.
- Britner, S. L., & Finson, K. D. (2005). Preservice teachers' reflections on their growth in an inquiry-oriented science pedagogy course. Journal of Elementary Science Education, 17(1), 39.
- Buchanan, L. B. (2015). Fostering Historical Thinking toward Civil Rights Movement Counter-Narratives: Documentary Film in Elementary Social Studies. The Social Studies, 106(2), 47–56.
- Buck, G. A., & Cordes, J. G. (2005). An Action Research Project on Preparing Teachers to Meet the Needs of Underserved Student Populations. Journal of Science Teacher Education, 16(1), 43–64.
- Bullough, R. V., Young, J., Birrell, J. R., Cecil Clark, D., Winston Egan, M., Erickson, L., … Welling, M. (2003). Teaching with a peer: A comparison of two models of student teaching. Teaching and Teacher Education, 19(1), 57–73.
- 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.
- Cardetti, F., & Truxaw, M. P. (2014). Toward Improving the Mathematics Preparation of Elementary Preservice Teachers. School Science and Mathematics, 114(1), 1–9.
- Carrier, S. J. (2013). Elementary Preservice Teachers' Science Vocabulary: Knowledge and Application. Journal of Science Teacher Education, 24(2), 405–425.
- Cartwright, T., Smith, S., & Hallar, B. (2014). Confronting Barriers to Teaching Elementary Science: After-School Science Teaching Experiences for Preservice Teachers. Teacher Education & Practice, 27(2/3), 464–487.
- Casey, S. A., Albert, J., & Ross, A. (2018). Developing Knowledge for Teaching Graphing of Bivariate Categorical Data. Journal of Statistics Education, 26(3), 197–213.
- Castro, A. M. (2006). Preparing Elementary Preservice Teachers to Use Mathematics Curriculum Materials. Mathematics Educator, 16(2), 14–24.
- Caughlan, S., Juzwik, M. M., Borsheim-Black, C., Kelly, S., & Fine, J. G. (2013). English Teacher Candidates Developing Dialogically Organized Instructional Practices. Research in the Teaching of English, 47(3), 212–246.
- Certo, J. L., Apol, L., Wibbens, E., & Hawkins, L. K. (2012). Living the Poet's Life: Using an Aesthetic Approach to Poetry to Enhance Preservice Teachers' Poetry Experiences and Dispositions. English Education, 44(2), 102–146.

- Chamberlin, M. T., & Candelaria, M. S. (2018). Learning from Teaching Teachers: A Lesson Experiment in Area and Volume with Prospective Teachers. Mathematics Teacher Education and Development, 20(1), 86–111.
- Chamberlin, M. T., & Powers, R. A. (2007). Selecting from Three Curricula for a Preservice Elementary Teacher Geometry Course. Issues in the Undergraduate Mathematics Preparation of School Teachers, 4.
- Chizhik, E. W., Chizhik, A. W., Close, C., & Gallego, M. (2017). SMILE (Shared Mentoring in Instructional Learning Environments): Effectiveness of a Lesson-Study Approach to Student-Teaching Supervision on a Teacher-Education Performance Assessment. Teacher Education Quarterly, 44(2), 27–47.
- Clark, K. M. (2012). History of mathematics: Illuminating understanding of school mathematics concepts for prospective mathematics teachers. Educational Studies in Mathematics, 81(1), 67–84.
- Cohen, M. D., & Nath, J. L. (2006). Paired Placements for Early Field Experiences. Teacher Education and Practice, 19(1), 24–40.
- Colwell, J. (2016). Examining Preservice Teachers' Beliefs about Disciplinary Literacy in History through a Blog Project. Action in Teacher Education, 38(1), 34–48.
- Conner, A., Edenfield, K. W., Gleason, B. W., & Ersoz, F. A. (2011). Impact of a content and methods course sequence on prospective secondary mathematics teachers' beliefs. Journal of Mathematics Teacher Education, 14(6), 483–504.
- Courtney, A., & King, F. B. (2009). Online Dialogue Discussion: A Tool to Support Pre-Service Teacher Candidates' Understanding of Literacy Teaching and Practice2. Contemporary Issues in Technology and Teacher Education, 9(3), 226–256.
- Crespo, S., & Nicol, C. (2006). Challenging Preservice Teachers' Mathematical Understanding: The Case of Division by Zero. School Science and Mathematics, 106(2), 84–97.
- Crespo, S., & Sinclair, N. (2008). What makes a problem mathematically interesting? Inviting prospective teachers to pose better problems. Journal of Mathematics Teacher Education, 11(5), 395–415.
- Curcio, R., & Adams, A. (2019). The Development of Mentoring Partnerships: How a Shared Learning Experience Enhanced the Final Internship. SRATE Journal, 28(1), 8.
- Cushmann, C. A., & Kemp, A. T. (2012). The Effects of Clinical Experiences on the Understanding of Classroom Management Techniques. Journal of Inquiry & Action in Education, 4(3), 44–58.
- Dass, P. M. (2005). Using a Science/Technology/Society Approach To Prepare Reform-Oriented Science Teachers: 14(1), 14.
- Dee, A. (2012). Collaborative Clinical Practice: An Alternate Field Experience. Issues in Teacher Education, 21(2), 147–163.
- DelleBovi, B. M. (2012). Literacy instruction: From assignment to assessment. Assessing Writing, 17(4), 271–292.

- DeMink-Carthew, J. (2017). Reform-Oriented Collaborative Inquiry as a Pedagogy for Student Teaching in Middle School. Middle Grades Research Journal, 11(1), 13–27.
- Dempsey, M. S., PytlikZillig, L. M., & Bruning, R. H. (2009). Helping preservice teachers learn to assess writing: Practice and feedback in a Web-based environment. Assessing Writing, 14(1), 38–61.
- Deniz, H. (2011). Examination of Changes in Prospective Elementary Teachers' Epistemological Beliefs in Science and Exploration of Factors Meditating That Change. Journal of Science Education and Technology, 20(6), 750–760.
- Duncan, R. G., Pilitsis, V., & Piegaro, M. (2010). Development of Preservice Teachers' Ability to Critique and Adapt Inquiry-based Instructional Materials. Journal of Science Teacher Education, 21(1), 81–102.
- Duran, L. B., McArthur, J., & Hook, S. V. (2004). Undergraduate Students' Perceptions of an Inquiry-Based Physics Course. Journal of Science Teacher Education, 15(2), 155–171.
- Edwards, C. J., Carr, S., & Siegel, W. (2006). Influences of Experiences and Training on Effective Teaching Practices to Meet the Needs of Diverse Learners in Schools. Education, 126(3), 580–592.
- Ely, E., Alves, K. D., Dolenc, N. R., Sebolt, S., & Walton, E. A. (2018). Classroom Simulation to Prepare Teachers to Use Evidence-Based Comprehension Practices. Journal of Digital Learning in Teacher Education, 34(2), 71–87.
- Fives, H., & Barnes, N. (2017). Informed and Uninformed Naïve Assessment Constructors' Strategies for Item Selection. Journal of Teacher Education, 68(1), 85–101.
- Foley, R. W., Archambault, L. M., Hale, A. E., & Dong, H.-K. (2017). Learning Outcomes in Sustainability Education Among Future Elementary School Teachers. Journal of Education for Sustainable Development, 11(1), 33–51.
- Giebelhaus, C. R., & Bowman, C. L. (2002). Teaching Mentors: Is It Worth the Effort? The Journal of Educational Research, 95(4), 246–254.
- Girod, M., & Girod, G. (2006). Exploring the Efficacy of the Cook School District Simulation. Journal of Teacher Education, 57(5), 481–497.
- Goodson-Espy, T., Cifarelli, V. V., Pugalee, D., Lynch-Davis, K., Morge, S., & Salinas, T. (2014). Applying NAEP to Improve Mathematics Content and Methods Courses for Preservice Elementary and Middle School Teachers. School Science and Mathematics, 114(8), 392–404.
- Groth, R. E., Bergner, J. A., & Burgess, C. R. (2016). An Exploration of Prospective Teachers' Learning of Clinical Interview Techniques. Mathematics Teacher Education and Development, 18(2), 48–71.
- Groves, F. H., & Pugh, A. F. (2002). Cognitive Illusions as Hindrances to Learning Complex Environmental Issues. Journal of Science Education and Technology, 11(4), 381–390.
- Hanuscin, D. L., & Zangori, L. (2016). Developing Practical Knowledge of the Next

Generation Science Standards in Elementary Science Teacher Education. Journal of Science Teacher Education, 27(8), 799–818.
- Hart, L. C. (2002). Preservice Teachers' Beliefs and Practice After Participating in an Integrated Content/Methods Course. School Science and Mathematics, 102(1), 4–14.
- Hart, S. M. & Bennett, S. M. (2013). Disciplinary Literacy Pedagogy Development of STEM Preservice Teachers. Teacher Education & Practice, 26(2), 221–241.
- Hawkins, S., & Park Rogers, M. (2016). Tools for Reflection: Video-Based Reflection Within a Preservice Community of Practice. Journal of Science Teacher Education, 27(4), 415–437.
- Hestness, E., Randy McGinnis, J., Riedinger, K., & Marbach-Ad, G. (2011). A Study of Teacher Candidates' Experiences Investigating Global Climate Change Within an Elementary Science Methods Course. Journal of Science Teacher Education, 22(4), 351–369.
- Hiebert, J., Berk, D., Miller, E., Gallivan, H., & Meikle, E. (2019). Relationships Between Opportunity to Learn Mathematics in Teacher Preparation and Graduates' Knowledge for Teaching Mathematics. Journal for Research in Mathematics Education, 50(1), 23.
- Hoppey, D., & Mickelson, A. M. (2017). Partnership and Coteaching: Preparing Preservice Teachers to Improve Outcomes for Students with Disabilities. Action in Teacher Education, 39(2), 187–202.
- Isabelle, A. D., & de Groot, C. (2008). Alternate Conceptions of Preservice Elementary Teachers: The Itakura Method. Journal of Science Teacher Education, 19(5), 417–435.
- Iwai, Y. (2019). Culturally Responsive Teaching in a Global Era: Using the Genres of Multicultural Literature. The Educational Forum, 83(1), 13–27.
- Jackson, C., Mohr-Schroeder, M., & Little, D. (2014). Using Informal Learning Environments to Prepare Preservice Teachers. Teacher Education & Practice, 27(2/3), 445–463.
- Jacobbe, T., Ross, D. D., Caron, D. A., Barko, T., & Busi, R. (2014). Connecting Theory and Practice: Preservice Teachers' Construction of Practical Tools for Teaching Mathematics. Teacher Education and Practice, 27, 332–355.
- James, M. C., & Scharmann, L. C. (2007). Using analogies to improve the teaching performance of preservice teachers. Journal of Research in Science Teaching, 44(4), 565–585.
- Jenkins, O. F. (2010). Developing teachers' knowledge of students as learners of mathematics through structured interviews. Journal of Mathematics Teacher Education, 13(2), 141–154.
- Jett, C. C. (2014). Using Mathematics Literature with Prospective Secondary Mathematics Teachers. Journal of Mathematics Education at Teachers College, 5(2), 49-53.
- Jo, I. (2016). Future Teachers' Dispositions Toward Teaching With Geospatial

Technologies. Contemporary Issues in Technology and Teacher Education, 16(3), 310–327.
- Jung, M. L., & Tonso, K. L. (2006). Elementary preservice teachers learning to teach science in science museums and nature centers: A novel program's impact on science knowledge, science pedagogy, and confidence teaching. Journal of Elementary Science Education, 18(1), 15–31.
- Kattoula, E., Verma, G., & Martin-Hansen, L. (2009). Fostering Preservice Teachers' "Nature of Science" Understandings in a Physics Course. Journal of College Science Teaching, 39(1), 18–26.
- Kaya, E., Newley, A., Deniz, H., Yesilyurt, E., & Newley, P. (2017). Introducing Engineering Design to a Science Teaching Methods Course Through Educational Robotics and Exploring Changes in Views of Preservice Elementary Teachers. Journal of College Science Teaching, 47(2), 66–75.
- Kenney, R., Shoffner, M., & Norris, D. (2014). Reflecting on the Use of Writing to Promote Mathematical Learning: An Examination of Preservice Mathematics Teachers' Perspectives. The Teacher Educator, 49(1), 28–43.
- Kinach, B. M. (2002). A cognitive strategy for developing pedagogical content knowledge in the secondary mathematics methods course: Toward a model of effective practice. Teaching and Teacher Education, 18(1), 51–71.
- Kopcha, T. J., & Alger, C. (2011). The Impact of Technology- Enhanced Student Teacher Supervision on Student Teacher Knowledge, Performance, and Self-Efficacy during the Field Experience. Journal of Educational Computing Research, 45(1), 49–73.
- Kucan, L., Palincsar, A. S., Busse, T., Heisey, N., Klingelhofer, R., Rimbey, M., & Schutz, K. (2011). Applying the Grossman et al. Theoretical Framework: The Case of Reading. Teachers College Record, 113(12), 2897–2921.
- Laframboise, K. L., & Shea, K. (2009). Developing Understanding of Research-based Pedagogy with Preservice Teachers: An Instrumental Case Study. The Qualitative Report, 14(1), 105–128.
- Lambert, J. L., & Bleicher, R. E. (2013). Climate Change in the Preservice Teacher's Mind. Journal of Science Teacher Education, 24(6), 999–1022.
- Lee, C. K.-P. (2012). An evaluation of an elementary science methods course with respect to preservice teacher's pedagogical. 13(2), 19.
- Lee, H.-J., Özgün-Koca, S. A., Meagher, M., & Edwards, M. T. (2018). Examining the Impact of a Framework to Support Prospective Secondary Teachers' Transition from "Doer" to "Teacher" of Mathematics. Mathematics Teacher Education and Development, 20(1), 112–131.
- Lewis, E., Dema, O., & Harshbarger, D. (2014). Preparation for Practice: Elementary Preservice Teachers Learning and Using Scientific Classroom Discourse Community Instructional Strategies. School Science and Mathematics, 114(4), 154–165.
- Lin, C.-Y. (2009). A comparison study of web-based and traditional instruction on pre-

service teachers' knowledge of fractions. Contemporary Issues in Technology and Teacher Education, 9(3), 257–279.
- Loverude, M. E., Gonzalez, B. L., & Nanes, R. (2011). Inquiry-based course in physics and chemistry for preservice K-8 teachers. Physical Review Special Topics - Physics Education Research, 7(1), 010106.
- Lubinski, C. A., & Otto, A. D. (2004). Preparing K-8 Preservice Teachers in a Content Course for Standards-Based Mathematics Pedagogy. School Science and Mathematics, 104(7), 336–350.
- Lux, N. J. (2013). Technology-Focused Early Field Experiences in Preservice Teacher Education. Journal of Digital Learning in Teacher Education, 29(3), 82–88.
- Mahalingappa, L., Hughes, E. M., & Polat, N. (2018). Developing preservice teachers' self-efficacy and knowledge through online experiences with English language learners. Language and Education, 32(2), 127–146.
- Matkins, J. J., & Bell, R. L. (2007). Awakening the Scientist Inside: Global Climate Change and the Nature of Science in an Elementary Science Methods Course. Journal of Science Teacher Education, 18(2), 137–163.
- Matthews, M. E., & Seaman, W. I. (2007). The Effects of Different Undergraduate Mathematics Courses on the Content Knowledge and Attitude towards Mathematics of Preservice Elementary Teachers. Issues in the Undergraduate Mathematics Preparation of School Teachers, 1.
- McCall, M. (2017). Elementary Preservice Science Teaching Efficacy and Attitude toward Science: Can a College Science Course Make a Difference? Electronic Journal of Science Education, 21(6), 1–11.
- McCulloch, A., Lovett, J., & Edgington, C. (2019). Designing to Provoke Disorienting Dilemmas: Transforming Preservice Teachers' Understanding of Function Using a Vending Machine Applet. Contemporary Issues in Technology and Teacher Education, 19(1), 4–22.
- McGee, J., & Colby, S. (2014). Impact of an Assessment Course on Teacher Candidates' Assessment Literacy. Action in Teacher Education, 36(5–6), 522–532.
- Merk, H., Waggoner, J., & Carroll, J. (2013). Co-Learning: Maximizing Learning in Clinical Experiences. AILACTE Journal, 10(1), 79–95.
- Mitchell, K., Homza, A., & Ngo, S. (2012). Reading Aloud with Bilingual Learners: A Fieldwork Project and Its Impact on Mainstream Teacher Candidates. Action in Teacher Education, 34(3), 276–294.
- Mizell, J. A., & Cates, J. (2004). The Impact of Additional Content Courses on Teacher Candidates' Beliefs regarding Mathematics Content and Pedagogy. Issues in the Undergraduate Mathematics Preparation of School Teachers, 4.
- Moreno, R., & Abercrombie, S. (2010). Promoting Awareness of Learner Diversity in Prospective Teachers: Signaling Individual and Group Differences within Virtual Classroom Cases. Journal of Technology and Teacher Education, 18(1), 111–130.

- 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.
- Morrison, J., & McDuffie, A. R. (2009). Connecting Science and Mathematics: Using Inquiry Investigations to Learn About Data Collection, Analysis, and Display. School Science and Mathematics, 109(1), 31–44.
- Morrone, A. S., Harkness, S. S., D'Ambrosio, B., & Caulfield, R. (2004). Patterns of Instructional Discourse that Promote the Perception of Mastery Goals in a Social Constructivist Mathematics Course. Educational Studies in Mathematics, 56(1), 19–38.
- Mostofo, J., & Zambo, R. (2015). Improving instruction in the mathematics methods classroom through action research. Educational Action Research, 23(4), 497–513.
- Nasir, A., & Heineke, A. J. (2014). Teacher Candidates and Latina/o English Learners at Fenton Elementary School: The Role of Early Clinical Experiences in Urban Teacher Education. Association of Mexican American Educators Journal, 8(1).
- Nesmith, S. M., Purdum-Cassidy, B., Cooper, S., & Rogers, R. D. (2017). Love It, Like It, or Leave It — Elementary Preservice Teachers' Field-Based Perspectives toward the Integration of Literature in Mathematics. Action in Teacher Education, 39(3), 321–339.
- Nokes, J. D. (2010). Preparing Novice History Teachers to Meet Students' Literacy Needs. Reading Psychology, 31(6), 493–523.
- Novak, E., & Wisdom, S. (2018). Effects of 3D Printing Project-based Learning on Preservice Elementary Teachers' Science Attitudes, Science Content Knowledge, and Anxiety About Teaching Science. Journal of Science Education and Technology, 27(5), 412–432.
- Olson, J. K., Bruxvoort, C. N., & Vande Haar, A. J. (2016). The impact of video case content on preservice elementary teachers' decision-making and conceptions of effective science teaching. Journal of Research in Science Teaching, 53(10), 1500–1523.
- Otto, C. A., Luera, G. R., & Everett, S. A. (2009). An Innovative Course Featuring Action Research Integrated with Unifying Science Themes. Journal of Science Teacher Education, 20(6), 537.
- Ozogul, G., & Sullivan, H. (2009). Student performance and attitudes under formative evaluation by teacher, self and peer evaluators. Educational Technology Research and Development, 57(3), 393–410.
- Petrosino Jr., A., & Mann, M. (2018). Data Modeling for Preservice Teachers and Everyone Else. Journal of College Science Teaching, 47(3).
- Phillippo, K., & Blosser, A. (2017). Stable roles, changed skills: Teacher candidate responses to instruction about adolescent psychosocial support practices. Advances in School Mental Health Promotion, 10(1), 5–25.
- Piro, J. S., & Hutchinson, C. J. (2014). Using a Data Chat to Teach Instructional Interventions: Student Perceptions of Data Literacy in an Assessment Course. The New Educator, 10(2), 95–111.

- Plummer, J. D., Zahm, V. M., & Rice, R. (2010). Inquiry and Astronomy: Preservice Teachers' Investigations of Celestial Motion. Journal of Science Teacher Education, 21(4), 471–493.
- Pratt, S. S. (2018). Area models to image integer and binomial multiplication. Investigations in Mathematics Learning, 10(2), 85–105.
- Ragland, R. (2016). Implementing an Evidence-Based Reflective Teaching Cycle: Using Scholarly Research in Curriculum Design. Mid-Western Educational Researcher, 28(3), 196–217.
- Rhine, S., Harrington, R., & Olszewski, B. (2015). The Role of Technology in Increasing Preservice Teachers' Anticipation of Students' Thinking in Algebra. Contemporary Issues in Technology and Teacher Education, 15(2), 85–105.
- Rigelman, N. M., & Ruben, B. (2012). Creating foundations for collaboration in schools: Utilizing professional learning communities to support teacher candidate learning and visions of teaching. Teaching and Teacher Education, 28(7), 979–989.
- Rinke, C. R., Gladstone-Brown, W., Kinlaw, C. R., & Cappiello, J. (2016). Characterizing STEM Teacher Education: Affordances and Constraints of Explicit STEM Preparation for Elementary Teachers: Characterizing STEM Teacher Education. School Science and Mathematics, 116(6), 300–309.
- Ro, J. M., Magiera, K., Gradel, K., & Simmons, R. (2013). Blog-Based Support for Preservice Teachers in an Afterschool Tutoring Program. Journal of Educational Technology Systems, 42(1), 69–84.
- Rodriguez-Valls, F., & Ponce, G. (2013). Classroom, the We Space: Developing Student-Centered Practices for Second Language Learner (SLL) Students. Education Policy Analysis Archives, 21, 55.
- Rosaen, C. L., Lundeberg, M., Terpstra, M., Cooper, M., Niu, R., & Fu, J. (2010). Constructing videocases to help novices learn to facilitate discussions in science and English: How does subject matter matter? Teachers and Teaching, 16(4), 507–524.
- Sabel, J. L., Forbes, C. T., & Zangori, L. (2015). Promoting Prospective Elementary Teachers' Learning to Use Formative Assessment for Life Science Instruction. Journal of Science Teacher Education, 26(4), 419–445.
- Salajan, F. D., Nyachwaya, J. M., Hoffman, J. G., & Hill, B. D. (2016). Improving Teacher Candidates' Lesson Planning Competencies Through Peer Review in a Wiki Environment. The Teacher Educator, 51(3), 185–210.
- Salter, I., & Atkins, L. (2013). Student-Generated Scientific Inquiry for Elementary Education Undergraduates: Course Development, Outcomes and Implications. Journal of Science Teacher Education, 24(1), 157–177.
- 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.
- Santau, A. O., Maerten-Rivera, J. L., Bovis, S., & Orend, J. (2014). A Mile Wide or an

Inch Deep? Improving Elementary Preservice Teachers' Science Content Knowledge Within the Context of a Science Methods Course. Journal of Science Teacher Education, 25(8), 953–976.

- Santoyo, C., & Zhang, S. (2016). Secondary Teacher Candidates' Lesson Planning Learning. Teacher Education Quarterly, 43(2), 3–27.
- Sayeski, K. L., Earle, G. A., Eslinger, R. P., & Whitenton, J. N. (2017). Teacher candidates' mastery of phoneme-grapheme correspondence: Massed versus distributed practice in teacher education. Annals of Dyslexia, 67(1), 26–41.
- Sayeski, K. L., Kennedy, M. J., Irala, S. de, Clinton, E., Hamel, M., & Thomas, K. (2015). The Efficacy of Multimedia Modules for Teaching Basic Literacy-Related Concepts. Exceptionality, 23(4), 237–257.
- Schussler, D., Frank, J., Lee, T.-K., & Mahfouz, J. (2017). Using Virtual Role-Play to Enhance Teacher Candidates' Skills in Responding to Bullying. Journal of Technology and Teacher Education, 25(1), 30.
- Scott, C. (2016). Using Citizen Science to Engage Preservice Elementary Educators in Scientific Fieldwork. Journal of College Science Teaching, 46(2), 37–41.
- Seung, E., Bryan, L. A., & Butler, M. B. (2009). Improving Preservice Middle Grades Science Teachers' Understanding of the Nature of Science Using Three Instructional Approaches. Journal of Science Teacher Education, 20(2), 157–177.
- Shin, E.-K., Wilkins, E. A., & Ainsworth, J. (2007). The Nature and Effectiveness of Peer Feedback during an Early Clinical Experience in an Elementary Education Program. Action in Teacher Education, 28(4), 40–52.
- Smith, M. E., Swars, S. L., Smith, S. Z., Hart, L. C., & Haardörfer, R. (2012). Effects of an Additional Mathematics Content Course on Elementary Teachers' Mathematical Beliefs and Knowledge for Teaching. Action in Teacher Education, 34(4), 336–348.
- Soh, L.-K., Fowler, D., & Zygielbaum, A. I. (2007). The Impact of the Affinity Learning Authoring Tool on Student Learning. Journal of Educational Technology Systems, 36(1), 29–62.
- Steele, M. D., Hillen, A. F., & Smith, M. S. (2013). Developing mathematical knowledge for teaching in a methods course: The case of function. Journal of Mathematics Teacher Education, 16(6), 451–482.
- Stover, K., Yearta, L. S., & Sease, R. (2014). "Experience Is the Best Tool for Teachers": Blogging to Provide Preservice Educators with Authentic Teaching Opportunities. Journal of Language and Literacy Education, 10(2), 99–117.
- Sunal, C. S., & Sunal, D. W. (2003). Teacher Candidates' Conceptualization of Guided Inquiry and Lesson Planning in Social Studies Following Web-Assisted Instruction. Theory & Research in Social Education, 31(2), 243–264.
- Suppa, S., DiNapoli, J., & Mixell, R. (2018). Teacher Preparation "Does" Matter: Relationships between Elementary Mathematics Content Courses and Graduates' Analyses of Teaching. Mathematics Teacher Education and Development, 20(2), 25–57.

- Tessier, J. (2010). An Inquiry-Based Biology Laboratory Improves Preservice Elementary Teachers' Attitudes About Science. Journal of College Science Teaching; Washington, 39(6), 84–90.
- Thanheiser, E. (2015). Developing prospective teachers' conceptions with well-designed tasks: Explaining successes and analyzing conceptual difficulties. Journal of Mathematics Teacher Education, 18(2), 141–172.
- Thomas, A. F., & Sondergeld, T. (2015). Investigating the Impact of Feedback Instruction: Partnering Preservice Teachers with Middle School Students to Provide Digital, Scaffolded Feedback. Journal of the Scholarship of Teaching and Learning, 15(4), 83–109.
- Trundle, K. C., Atwood, R. K., & Christopher, J. E. (2006). Preservice Elementary Teachers' Knowledge of Observable Moon Phases and Pattern of Change in Phases. Journal of Science Teacher Education, 17(2), 87–101.
- Uribe, S. N., & Vaughan, M. (2017). Facilitating student learning in distance education: A case study on the development and implementation of a multifaceted feedback system. Distance Education, 38(3), 288–301.
- Van Eck, R. N., Guy, M., Young, T., Winger, A. T., & Brewster, S. (2015). Project NEO: A Video Game to Promote STEM Competency for Preservice Elementary Teachers. Technology, Knowledge and Learning, 20(3), 277–297.
- Weinburgh, M. (2007). The Effect of Tenebrio obscurus on Elementary Preservice Teachers' Content Knowledge, Attitudes, and Self-efficacy. Journal of Science Teacher Education, 18(6), 801–815.
- Welsh, K. A., & Schaffer, C. (2017). Developing the Effective Teaching Skills of Teacher Candidates During Early Field Experiences. The Educational Forum, 81(3), 301–321.
- Westrick, J. M., & Morris, G. A. (2016). Teacher education pedagogy: Disrupting the apprenticeship of observation. Teaching Education, 27(2), 156–172.
- Wilkins, J. L. M., & Brand, B. R. (2004). Change in Preservice Teachers' Beliefs: An Evaluation of a Mathematics Methods Course. School Science and Mathematics, 104(5), 226–232.
- Windschitl, M., & Thompson, J. (2006). Transcending Simple Forms of School Science Investigation:The Impact of Preservice Instruction on Teachers' Understandings of Model-Based Inquiry. American Educational Research Journal, 43(4), 783–835.
- Windschitl, M., Thompson, J., & Braaten, M. (2008). How Novice Science Teachers Appropriate Epistemic Discourses Around Model-Based Inquiry for Use in Classrooms. Cognition and Instruction, 26(3), 310–378.
- Yadav, A., Mayfield, C., Zhou, N., Hambrusch, S., & Korb, J. T. (2014). Computational Thinking in Elementary and Secondary Teacher Education. Trans. Comput. Educ., 14(1), 5:1–5:16.
- Yeh, C., & Santagata, R. (2015). Preservice Teachers' Learning to Generate Evidence-

Based Hypotheses About the Impact of Mathematics Teaching on Learning. Journal of Teacher Education, 66(1), 21–34.

- Zhang, S., & Stephens, V. (2013). Learning to Teach English-Language Learners in Mainstreamed Secondary Classrooms. Teacher Education and Practice, 26(1), 99–116.

Online Appendix B
Full Coding Scheme

Study characteristics

Level 2 code	Level 1 code	Response options
Study sample	Number of participants	Numerical
	Did not report participants	1=Yes, 0=No
	Number of comparison groups (*1 if single-case*)	Numerical
	Number of study sites (*naturally occuring sites where the treatment is delivered, such as courses, TEPs and/or higher education institution (i.e. number of different places where treatment administration might be reasonably said to differ in some way because of context)*)	Numerical
Certification Level	Elementary	1=Yes, 0=No
	Secondary (*incl. middle years*)	1=Yes, 0=No
	Level not reported	1=Yes, 0=No
Program Type	Traditional (*incl. MAT programs*)	1=Yes, 0=No
	Alternative (*based in schools; whole program concurrently teacher of record and taking coursework*)	1=Yes, 0=No
	Type not reported	1=Yes, 0=No
Program Level	Undergraduate	1=Yes, 0=No
	Postgraduate	1=Yes, 0=No
	Level not reported	1=Yes, 0=No
Type of teacher preparation practice (select one)	A course or set of courses organized around a particular theme within a TEP	1=Yes, 0=No
	An activity, pedagogy, or process implemented within a course in a TEP (*i.e. less than 100% of the course is being analysed*)	1=Yes, 0=No
	An experience organized by teacher educators (outside a course) for PSTs for the purpose of development	1=Yes, 0=No

	A way of organizing a TEP component (*e.g. assigning certain types of mentors or peers in practicum*)	1=Yes, 0=No
	Other	1=Yes, 0=No
Outcome measure type	All treatment/comparison groups received at least some training related to the primary outcome variable (*treatment-dependent -- i.e., using a new curriculum or doing a very specific kind of instructional move that the comparison group did not learn/do*)	1=Yes, 0=No, 98=Cannot be determined (because no reporting of what the comparison group did), 99=N/A (there is no comparison group)
	Primary outcome measures are something that the treated group were doing as part of the treatment *(e.g., if the treatment is writing fieldnotes, and the outcome variable is fieldnotes)*	1=Yes, 0=No, 98=Cannot be determined
	Primary outcome measures are developed by researchers external to the project *(i.e. uses a measure that has not been developed by the researchers for the purpose of the evaluation; references a different researcher who has developed the outcome measure.)*	1=Yes, 0=No
Type of data collected	Evidence of participation in the intervention *(e.g. frequency of participation, the type of participation, implementation analysis)*	1=Yes, 0=No
	PSTs perspectives of how good the intervention was/how much they think they learned from it	1=Yes, 0=No
	Observed PSTs skill/knowledge/belief learning from the treatment *(e.g. survey of teacher beliefs, aptitude on assignments/tasks)*	1=Yes, 0=No
	Self-reported PST instructional practice *(e.g. actual frequency/likelihood of implementing a particular practice)*	1=Yes, 0=No
	Classroom observations of PST instructional practice	1=Yes, 0=No
	Teachers' self report of student learning AND/OR students' self-report of own learning	1=Yes, 0=No
	K-12 students' perceptions of PSTs	1=Yes, 0=No
	Observed K-12 student learning (e.g. test scores or samples of work)	1=Yes, 0=No
	Other	1=Yes, 0=No
Types of outcome measures	Survey of PSTs (Knowledge/Skills/Beliefs)	1=Yes, 0=No
	PSTs' Ongoing Reflections/Journals (on their learning from the treatment)	1=Yes, 0=No

	PSTs' course/classroom participation (e.g. audio/video recording of teacher education classroom, PSTs participation on online discussion forums, teacher educator notes on PSTs participation)	1=Yes, 0=No
	Interviews with PSTs	1=Yes, 0=No
	Lesson Observation of PSTs teaching (whether in a simulated environment or a K-12 classroom)	1=Yes, 0=No
	PSTs' Course Assessment/Test/Assignment	1=Yes, 0=No
	PSTs' external performance task (e.g. edTPA for licensure)	1=Yes, 0=No
Quantitative analysis	Uses statistical analyses	1=Yes, 0=No
	Presents power analysis	1=Yes, 0=No

Research design

Level 2 code	**Level 1 code**	**Response options**
Comparison	Between groups (e.g. one group who gets and one who doesn't get the treatment)	1=Yes, 0=No
	Within person (intended vs unrelated outcome)	1=Yes, 0=No
	Within person (pretest vs posttest)	1=Yes, 0=No
	Simulated/Estimated	1=Yes, 0=No
	No Comparison	1=Yes, 0=No
Assignment	Random	1=Yes, 0=No
	Score-Based	1=Yes, 0=No
	Time (cohort to cohort)	1=Yes, 0=No
	Matching (on demographics/covariates using statistical methods)	1=Yes, 0=No
	Non-random/Self-Selection/Convenience Sample	1=Yes, 0=No

Pretest	Administered multiple times (over a period of time)	1=Yes, 0=No
	Single administration	1=Yes, 0=No
	None	1=Yes, 0=No
Pretest type	One treatment group does not receive pretest	1=Yes, 0=No
	What participants are asked (to do) in pretest is different from posttest (i.e. more or different outcomes in the posttest)	1=Yes, 0=No
	What participants are asked (to do) in pretest is the same as in the posttest	1=Yes, 0=No
	Retrospective pretest (e.g. "what was your level before you took the test? what is your level now?")	1=Yes, 0=No
Posttest (*measures after the treatment*)	Administered multiple times (over a period of time) to test retention	1=Yes, 0=No
	Single	1=Yes, 0=No
	None	1=Yes, 0=No
Type of treatment	Switching/Reversed *(counterbalanced design)*	1=Yes, 0=No
	Repeated/Removed (*ie. there is a period where the treatment is not administered and an outcome is measured (removed); subsequently, there is a period where the treatment is re-administered (repeated)*)	1=Yes, 0=No
	Single (*treatment happened once*)	1=Yes, 0=No

Issues (Threats to validity)

Level 2 code	**Level 1 code**	**Response options**
Internal validity	During the study, the experiences of the treatment and comparison groups only differed in that the treatment group received the treatment. *(History effects)*	1=Yes, 0=No, 99=N/A (no comparison group)
	Whether treatment and comparison groups could interact and influence each others' behaviours during the intervention/before the posttest is discussed/reported.	1=Yes, 0=No, 99=N/A (No comparison group)

	Whether participants moved from treatment to comparison group (or vice versa, or not) is discussed/reported.	1=Yes, 0=No, 99=N/A (No comparison group)
	Attrition from the experiment is discussed (AND/OR tested for bias).	1=Yes, 0=No (Code 1 if even slightly mentioned)
Measurement	The primary outcome was measured using a rubric/rating scheme that was the same across time and groups.	1=Yes, 0=No
	Multiple raters are used (to ensure reliability of the outcome measure).	1=Yes, 0=No, 99=N/A (Outcome measure isn't subjectively rated)
	The influence of stakes being attached to performance on the outcome (e.g. course grades, licensure, etc.) is discussed/reported.	1=Yes, 0=No
	The influence of participants knowing the theory of change of the treatment (i.e. having expectations of its effect) is discussed/reported.	1=Yes, 0=No
	Treatment is blinded (participants do not know if they are part of the treatment or control condition; code 1 if reasonably sure from how they describe the intervention)	1=Yes, 0=No or not reported, 99=N/A (No control condition)
External validity	How study participants are similar/different from the broader TEP population is described. (If all PSTs in a TEP are in the study or if mandatory to participate, then 1 -- does not count if they just say 'PSTs who volunteered were included)	1=Yes, 0=No
	Details of the site of the study (usually the TEP) are reported such that a "typical" site can be identified. (*NOTE: 1 if the TEP is named; if not named, all the following must be reported: Location of program (state and urbanicity); Size of program (approximate number of students), Type of program (alternative/traditional), Level of program (undergrad/postgrad)*).	1=Yes, 0=No
	The effects of the instructor as different to the effects of the treatment are discussed/reported/dealt with (*EITHER researchers discuss the issue of the treatment being delivered by an instructor who may be highly enthusiastic, skilled, or suited to the treatment; OR the treatment is delivered by a number of trained facilitators*).	1=Yes, 0=No, 99=N/A (The treatment isn't delivered by an instructor)
	The research was undertaken in the authors' own teacher education course.	1=Yes, 0=No, 99=Unsure/Not reported
	The analysis is clustered by site.	1=Yes, 0=No, 99=N/A (Only

		1 site)