education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

Arizona State University

Volume 33 Number 85 December 2, 2025 ISSN 1068-2341

Simulating the Statewide Scaleup of a Promising Teacher Education Initiative to Preempt its Unintended Consequences for Racial Inequity

Matthew Truwit
University of Louisville
Emanuele Bardelli
Santa Rosa City Schools
&
Matthew Ronfeldt
University of Michigan
United States

Citation: Truwit, M., Bardelli, E., & Ronfeldt, M. (2025). Simulating the statewide scaleup of a promising teacher education initiative to preempt its unintended consequences for racial inequity. *Education Policy Analysis Archives, 33*(85). https://doi.org/10.14507/epaa.33.8972

Abstract: Most evaluation research fails to adequately anticipate unintended consequences, thereby implicitly permitting the possibility of negative repercussions so long as they fall out of the purview of the policy under study. Simulating how otherwise promising initiatives might scale into educational policy, however, offers both researchers and policymakers a way to not only investigate but also preemptively address any such possible, pernicious side effects. We provide an illustrative example of this kind of forward-thinking evaluation research by generating hypothetical scenarios of the statewide implementation of an algorithmic teacher education initiative shown in prior research to have positive intended effects on improving the

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 21/10/2024
Revisions received: 11/9/2025
Accepted: 16/9/2025

quality of preservice teachers' clinical placements. By comparing these plausible implementation scenarios against the historical record of clinical placements that programs actually made, we are able to not only uncover the unintended but anticipated inequities that this initiative would likely introduce if adopted at scale but also proactively make algorithmic adjustments that prevent their occurrence without diminishing any intended positive impacts, all before causing real-world harm.
Keywords: unintended consequences; cooperating teachers; simulation; racial equity

Simulando la ampliación a nivel estatal de una prometedora iniciativa de formación docente para prevenir sus consecuencias no deseadas en la inequidad racial
Resumen: La mayoría de las investigaciones de evaluación no logran anticipar adecuadamente las consecuencias no deseadas, permitiendo implícitamente la posibilidad de repercusiones negativas siempre que queden fuera del alcance de la política analizada. Sin embargo, simular cómo iniciativas prometedoras podrían ampliarse a nivel de política educativa ofrece tanto a los investigadores como a los responsables de políticas una vía para no solo investigar, sino también abordar de manera preventiva dichos efectos secundarios perniciosos. En este estudio presentamos un ejemplo ilustrativo de este tipo de investigación evaluativa prospectiva, generando escenarios hipotéticos de la implementación estatal de una iniciativa algorítmica de formación docente que, según investigaciones previas, ha mostrado efectos positivos en la mejora de la calidad de las prácticas profesionales de los futuros docentes. Al comparar estos escenarios de implementación plausibles con los registros históricos de prácticas profesionales que los programas realmente realizaron, podemos no solo identificar las inequidades no deseadas pero previsibles que esta iniciativa introduciría si se adoptara a gran escala, sino también realizar ajustes algorítmicos proactivos que eviten su aparición sin disminuir los impactos positivos previstos, todo ello antes de causar daño en el mundo real.
Palabras-clave: consecuencias no deseadas; docentes tutores; simulación; equidad racial

Simulando a ampliação em escala estadual de uma iniciativa promissora de formação de professores para prevenir suas consequências não intencionais na desigualdade racial
Resumo: A maior parte das pesquisas de avaliação não consegue antecipar adequadamente as consequências não intencionais, permitindo implicitamente a possibilidade de repercussões negativas desde que estejam fora do escopo da política analisada. No entanto, simular como iniciativas promissoras poderiam ser ampliadas para o nível de política educacional oferece tanto a pesquisadores quanto a formuladores de políticas uma maneira de não apenas investigar, mas também abordar preventivamente esses possíveis efeitos colaterais nocivos. Este estudo apresenta um exemplo ilustrativo desse tipo de pesquisa avaliativa prospectiva, gerando cenários hipotéticos da implementação em todo o estado de uma iniciativa algorítmica de formação de professores que, em pesquisas anteriores, demonstrou efeitos positivos na melhoria da qualidade dos estágios de formação inicial docente. Ao comparar esses cenários plausíveis de implementação com os registros históricos dos estágios que os programas realmente realizaram, conseguimos não apenas identificar as desigualdades não intencionais, mas previsíveis, que essa iniciativa introduziria se fosse adotada em larga escala, como também fazer ajustes algorítmicos proativos que evitem sua ocorrência sem reduzir os impactos positivos pretendidos — tudo isso antes de causar danos no mundo real.
Palavras-chave: consequências não intencionais; professores cooperantes; simulação; equidade racial

Simulating the Statewide Scaleup of a Promising Teacher Education Initiative to Preempt its Unintended Consequences for Racial Inequity

For nearly as long as researchers have sought to measure the extent to which policy produces its desired impacts, there have been others who have focused on its side effects—its unintended consequences for actors or behaviors falling outside its purview (Bamberger et al., 2016; Hirschman, 1967; Merton, 1936). Investigations of side effects in education specifically have become quite common, as evidenced by studies of the unintended consequences of a wide range of contemporary policies like student testing (Smith & Rottenberg, 1991; Yeh, 2005), high-stakes teacher evaluation (Amrein-Beardsley & Collins, 2012; Hewitt, 2015; Lavigne, 2014), school choice (Lubienski, 2005), and affirmative action (Barrow et al., 2020). Typically, research on unintended consequences involves the retrospective (quasi-)experimental evaluation of the impacts of a policy across a spectrum of more distal and less obvious outcomes and/or individuals than those included in earlier evaluations of the same policy's narrower and more immediate intended effects. While these kinds of traditional evaluation studies may uncover new pathways that extend our understanding of the ways in which policy produces ripples of change, it is less clear that they have "a noticeable effect on the making and remaking of public policy" (Weiss, 1993, p. 98). Given the nature of its timing, for instance, rear-facing research can at best only hope to shape future policy, often years after any unintended consequences have had the opportunity to take root.

This raises the question of why researchers and policymakers fail to do a better job on the front end of preventing unintended consequences from ever occurring. The simplest and most common explanation is that these consequences were unforeseeable and could therefore only be studied in hindsight. Such phenomena were once (and better) characterized as *unanticipated* consequences, the original phrase used in the seminal work by sociologist Robert K. Merton (1936) to describe those impacts of purposive social action that were truly unenvisioned (and, as a result, unpreventable). However, some side effects can easily be predicted in advance. Scholars have traced Merton's original terminology through the subsequent literature to show how the conflation of the terms *unintended* and *unanticipated*—with the former becoming the predominant nomenclature for describing the unforeseen side effects of policy[1]—has ultimately masked the existence of the crucial subset of impacts that exist outside of the intentions but not necessarily the anticipation of policymakers (de Zwart, 2015; Parvin & Pollock, 2020). These "permitted outcomes" (de Zwart, 2015, p. 295) involve the underexamined consequences that policymakers were or should have been able to foresee but decided—whether due to indifference, risk tolerance, or the careful calculation of ethical and economic tradeoffs—were of less concern than the expected beneficial impacts they intended.

This historical conflation may point to a subtler and more pernicious explanation for the dearth of research on the unintended but still anticipated impacts of policy—namely, that the simpler yet less specific term "unintended consequences" continues to serve the purposes of both researchers and policymakers (de Zwart, 2015; Parvin & Pollock, 2020). For example, by failing to clarify whether and which side effects may have been anticipated, those who formulated a policy can avoid responsibility for any unintended repercussions that they knowingly permitted. Meanwhile, those evaluating the same policy get to sidestep the political challenge of assigning any explicit blame while also imbuing their work with the added importance that comes from "uncovering" effects that

[1] None of the studies of unintended consequences of education policy cited above use the original term "unanticipated consequences", much less consider how and whether any of the unintended outcomes studied are unanticipated or permitted.

were presumably unforeseeable. In essence, the imprecision of the phrase "unintended consequences" gives each party an easy way out of doing their best to craft policy with lower likelihoods of causing harm by discouraging them from assessing this possibility in advance.

This avenue of avoidance is of particular salience to the study of race and education policy. Scholars have argued that the prevalence and persistence of inequities across the school system, coupled with the "continued promotion of policies and practices that are known to be racially divisive," implies at minimum a "tacit intentionality" (Gillborn, 2005, p. 499)—and even an explicit antiblackness (Dumas, 2016)—among policymakers in their reproduction and entrenchment of a status quo privileging the White and wealthy. Abundant evidence exists of racial disparities in students' access to effective instruction (Clotfelter et al., 2023) and well-funded, high-quality schools (Babbs Hollett & Frankenberg, 2022; Weathers & Sosina, 2022), their experience of exclusionary discipline (Anderson & Ritter, 2017; Welsh & Little, 2018), and, as a result, in their educational outcomes (Gopalan, 2019; Ladson-Billings, 2006); similar evidence can be found of racial disparities in the evaluation (Campbell, 2020) and retention of teachers (Carver-Thomas et al., 2024). And yet the color-evasive style of policy that prevails within our system of education has not just fallen short of addressing but has instead perpetuated and even exacerbated these inequities in ways that, even if unintended, were not always unforeseeable (Gillborn, 2019; Goldin & Khasnabis, 2021; Wells, 2014).

However, if researchers were to sharpen their evaluations of the side effects of color-evasive education policy by better distinguishing unintended consequences that were able to be anticipated from those that were not, our understanding of the persistent patterns of racial disparities in schools could move from one centered around the passive reproduction of an inequitable status quo without any responsible agents—a point from which it is difficult to know where change can and should begin—to one recognizing its active, daily recreation on both individual and institutional levels, both of which share responsibility in its intentional disrupting (Leonardo, 2004). Put differently, more clearly defining and focusing on unintended but anticipated consequences in educational evaluations is a crucial first step in preempting the "racist outcomes of contemporary policy [that] may not be coldly calculated but … are far from accidental" (Gillborn, 2005, p. 499). By proactively considering a policy's likely "permitted outcomes" rather than retrospectively estimating its "unintended consequences," those who craft a policy would no longer enjoy the privilege of professing a desire to dismantle structural inequalities without actively embracing these values in its design; moreover, those evaluating it would be forced to engage their work in more critical conversations about policymaker accountability, tradeoffs across actors and outcomes, and the magnitude of potential harm, often to those already most marginalized (Parvin & Pollock, 2020).

To take this a step further, we believe that evaluations of permitted outcomes, if they are to have the greatest capacity to meaningfully shape and reshape policy for societal good, must be done both in direct collaboration with policymakers and in advance of policy implementation. Partnering with the designers and implementers of policy can help to minimize the subset of unintended outcomes that are unanticipated, as additional actors with more diverse knowledge, frameworks, and constituencies will be able to envision a wider set of outcomes for a broader set of stakeholders. Such a cooperative approach will also better facilitate the careful navigation of the challenging political dynamics that can arise during critical conversations around responsibility and intention (Oliver et al., 2020).

In addition, attempting to quantify and weigh the cost of any anticipated but unintended consequences *prior* to actual implementation positions the practice of evaluation as something fundamental to the design and enactment of policy, ensuring its impact on the actual policymaking process, rather than merely as regulation in hindsight that hopes to inform it in the future (Parvin &

Pollock, 2020). However, the proactive prognostication of side effects requires a different approach from the rear-facing, causal methods typically used to evaluate policy—one that need not rely on consequences to have already occurred. Instead, we advocate for the use of simulations, through which researchers can answer critical *what if* questions about the unintended but foreseeable consequences of a policy *before* its actual enactment. Unlike traditional approaches to evaluation, which tend to provide rigorous, context-specific estimates of an intervention's narrow and immediate impacts, simulations combine historical data with models of likely outcomes that allow for uncertainty while still closely imitating the real world to generate plausible impacts across a broad spectrum of possible implementations. The use of simulations to examine and estimate the unintended consequences of educational policy is uncommon, though not unprecedented. For instance, researchers have employed the approach to retrospectively interrogate the efficacy of federal legislation like the No Child Left Behind Act (Grissmer et al., 2014; Lee, 2004) and higher education policy around college admissions (Matos-Díaz & García, 2014). However, it is especially rare to see researchers engage in this kind of work when evaluating the hypothetical consequences of policy in advance of implementation (cf. Morris & Johnson (2024), who offer one such promising illustration of the potential impacts of statewide adoption of a minimum grading policy).

Engaging in this kind of forward-facing, collaborative evaluation of the anticipated but unintended and not-yet-observed outcomes of a policy is much easier said than done. Therefore, in this paper, we provide an exemplar of how to do so, drawing on our experience running simulations in partnership with the Tennessee Department of Education (TDoE) to preemptively investigate whether and how the statewide scaleup of an otherwise promising initiative could have unintentionally introduced or exacerbated racial inequities in the field of preservice teacher education. In doing so, we identify a way to refine one potential educational policy by anticipating and forestalling its inequitable side effects; more importantly, however, we illustrate the power of and establish some key principles for the kind of proactive, cooperative approach to evaluating unintended consequences that we believe offers unfulfilled promise in promoting equity for researchers and policymakers across fields.

Background

Pre-service student teachers (PSTs) traditionally conclude their preparation with student teaching, a clinical experience in a real classroom under the supervision of a certified teacher of record, usually referred to as a clinical mentor (CM) or cooperating teacher. Recent causal evidence has established a clear link between the instructional effectiveness of CMs and that of their PSTs, whereby PSTs become better teachers themselves when completing their clinical placements under CMs demonstrating higher-quality instruction (Goldhaber et al., 2022; Ronfeldt et al., 2020, 2025). One of the earliest studies in this vein was the Mentors Matter Recruitment (MMR) initiative, a field experiment designed and conducted over three consecutive years beginning in 2018-2019 under a research-practice partnership with the TDoE and several educator preparation programs (EPPs). The MMR initiative relied on an algorithm that leveraged state administrative data on observation ratings, value-added to student achievement scores, and years of classroom teaching to generate recommendation lists of the most instructionally effective and experienced teachers in each district and subject area who might serve as CMs (Ronfeldt et al., 2020). (We describe the algorithm and implementation of this initiative in greater detail below.) Sharing these lists with a group of randomly assigned districts resulted, as intended, in the recruitment of substantially more instructionally effective and experienced CMs by more than half of a standard deviation. Furthermore, the PSTs placed with these CMs reported feeling more prepared to teach at the end of their placements.

A subsequent study fully replicated these contrasts in CM effectiveness and experience, first experimentally within the same program and partner districts and then quasi-experimentally among a new set of EPPs using recommendation lists in the absence of significant state oversight and research partner support (Ronfeldt et al., 2025). In addition, extension analyses found that PSTs received better clinical evaluations (i.e., observation ratings during student teaching) when placed with CMs in districts that used recommendation lists for their recruitment, suggesting that the initiative also positively impacted PST instructional performance. Altogether, this body of causal evidence highlights the viability of the algorithm underlying the MMR initiative as a tool for achieving the intended consequence of increasing the instructional effectiveness and experience of the CM pool and, subsequently, improving the teaching of their PSTs. This, in turn, sparked the interest of the TDoE in expanding its use across the state.

At the same time, an emerging body of research about potential bias in observation ratings raised concerns among researchers, policymakers, and practitioners—both within our partnership and more broadly (e.g., Close et al., 2020)— about the unintended consequences of scaling up initiatives that, like MMR, depend on teacher evaluation data. Namely, prior work has found evidence that teachers receive lower ratings when they are observed in classrooms with higher proportions of students of color, lower-income students, and lower-achieving students, even after accounting for other indications of their instructional effectiveness like value-added (Campbell & Ronfeldt, 2018; Steinberg & Garrett, 2016; Steinberg & Sartain, 2020). Similar studies have shown that Black and male teachers also tend to receive lower scores than their otherwise comparable peers (Campbell, 2020; Grissom & Bartanen, 2022). This literature suggests that the use of observation ratings to recruit instructionally effective mentors to serve as CMs may unintentionally overidentify White, female teachers working with more privileged student populations.

Complicating matters further, studies across contexts suggest that existing procedures already preferentially select clinical placements along some of the same dimensions that observation ratings appear to privilege. Teachers who serve as CMs do tend to be more instructionally effective (in terms of observation ratings and value-added) and experienced than their colleagues (Bastian et al., 2022; Goldhaber et al., 2014; Krieg et al., 2020; Matsko et al., 2021; Ronfeldt et al., 2018); however, they are not demographically representative of the workforce at large, with evidence of the disproportionate recruitment of female and, in some studies, White teachers (Goldhaber et al., 2014; Krieg et al., 2016; Ronfeldt et al., 2018). Similarly, prior work has found that schools appear more likely to be selected as field placement schools (FPSs) when they have higher average student achievement (Krieg et al., 2016) and lower proportions of students receiving free or reduced-price lunch (Matsko et al., 2021; Ronfeldt, 2012, 2015). Findings around student racial composition are less consistent across settings, with conflicting evidence that schools hosting PSTs disproportionately serve White (Matsko et al., 2021; Ronfeldt, 2012) and underrepresented minority (Krieg et al., 2020; Ronfeldt, 2015) populations.

Together, these two lines of research, which have identified potential biases embedded in both observation ratings and traditional recruitment practices, raise the possibility that scaling up the MMR initiative may have unintended consequences for racial inequity. The underlying algorithm relies on observation ratings as a primary measure for identifying potential CMs, while the recommendation lists it generates merely complement and work within established recruitment procedures. If evaluation data and/or status quo practices privilege the selection of clinical placements under White and female CMs in FPSs with higher-achieving and Whiter student populations, then the statewide scaleup of the MMR initiative poses a foreseeable, if unintended, risk of introducing new and/or exacerbating existing inequities in the landscape of teacher preparation.

At the same time, it is not guaranteed that any policy that reduces the diversity of clinical placements would negatively impact racial equity more broadly construed. Prior research (e.g., Bolyard & Baker, 2021) has documented some skepticism that merely requiring diversity in PSTs' clinical placements can orient students' attitudes and beliefs toward racial justice, with the risk of reinforcing preexisting stereotypes toward students or teachers of color (see Grossman et al., 2012, for a review of some of this literature). As such, an exclusive commitment to the diversity of clinical placements—particularly when the alternative is oriented toward improving quality—could still have harmful ramifications for equity, especially if it comes at the expense of increasing CM effectiveness.

Given all this uncertainty, rather than make a judgment call about the potential tradeoffs, dive straight into implementation, and evaluate any side effects in hindsight, the TDoE decided to partner with us to simulate the scaleup of the MMR initiative and proactively predict whether its rollout across Tennessee might have these kinds of unintended but anticipated consequences. More specifically, as a team, we collectively sought to compare the characteristics of the CMs and FPSs that actually hosted PSTs with those that would have hypothetically been recruited if all programs in the state used the MMR algorithm. The following three research questions guided our work: Under plausible statewide implementation of the MMR algorithm, (1) to what extent would the average instructional effectiveness and experience of CMs have improved, and (2) in what other ways would the average characteristics of CMs and their FPSs have changed? Additionally, (3) if statewide implementation would have been less equitable than historical practices, can we adjust our algorithm to preempt the introduction or exacerbation of any biases? Our first research question serves as confirmation that the intended consequences of the policy as documented in prior work are replicated within our simulations; the second seeks to estimate its unintended yet anticipated consequences, while the third aims to discover the best way to forestall any potential side effects we ultimately foresee.

Methods

By comparing the characteristics of clinical placements hypothetically selected under statewide implementation of the MMR initiative with those actually observed in the historical data, we aim to estimate what would have plausibly occurred had all programs in the state had access to algorithmically generated, district- and subject-specific recommendation lists of the most promising teachers who could serve as CMs. To do this, we first prepare historical teacher and school data to identify not only those CMs and FPSs serving as clinical placements in 2017-2018 (the year prior to the introduction of the original MMR initiative) but also the eligible pool of all teachers who could have served in each district and subject area (and their schools).

Data

Historical Record of Clinical Mentors

Historical clinical mentor data from Tennessee include CM-PST linkages from all but five EPPs (86.0%), connecting 3,274 unique CMs with 2,259 unique PSTs—more than two thirds (70.5%) of all new teachers prepared in the state in the 2017-2018 academic year. To assess external validity, we explore the extent to which PSTs with and without CM linkages differ across observable characteristics in Table 1. Notably, the sample of PSTs for whom no CM linkages exist is markedly different from those for whom CMs are identified. Specifically, PSTs whose CMs and FPSs are unidentified are far less likely to be White and female and are overwhelmingly enrolled in graduate-level or alternative-certification job-embedded pathways to a teaching credential; they are also disproportionately likely to be pursuing a high-need

endorsement area. These differences stem, in part, from the fact that all five programs which did not report CM linkages to the TDoE—including a pair of HBCUs and a pair of exclusively alternative certification pathways—are located in the two largest urban areas in the state.

Table 1

PST Characteristics by Availability of CM Linkage Data

	Overall	w/o CM	w/ CM	Diff.	ES	*p*-value	*N*
Woman	0.767	0.673	0.806	0.133	0.319	<0.001	3,197
White	0.830	0.684	0.891	0.207	0.570	<0.001	3,170
Black	0.099	0.229	0.045	-0.184	0.643	<0.001	3,170
Other Race/Ethnicity	0.070	0.086	0.064	-0.023	0.089	0.023	3,170
GPA	3.472	3.368	3.504	0.136	0.374	<0.001	2,905
Job-Embedded	0.316	0.834	0.100	-0.734	2.274	<0.001	3,203
Internship	0.089	0.015	0.120	0.105	0.373	<0.001	3,203
Undergraduate	0.497	0.124	0.653	0.529	1.208	<0.001	3,203
High-Need Endorsement	0.259	0.324	0.232	-0.093	0.212	<0.001	3,203

Note. All variables are indicators except for GPA, which exists on a 4-point scale. X^2 (8, N = 3,203) = 2,952.95, $p < .001$.

The uneven availability of historical CM data could raise some concern about our claims of simulating "statewide" implementation and of generating unbiased estimates of the inequities that doing so may produce. The historical record of CMs (and their FPSs) serves as the comparison group against which we contrast the hypothetical placements plausibly produced under statewide implementation of the MMR initiative. Research suggests evidence of racial and gender homophily in placement practices (Krieg et al., 2020); given that the unobserved placements in our administrative data disproportionately involve male PSTs and PSTs of color attending EPPs in urban settings, the observed historical record may comprise a higher proportion of female and White CMs than the true pool of those who actually served, thereby perhaps contributing to the underestimation of any biases that occur when scaling up the intervention.

At the same time, though, as we expand upon later, the location and total number of identifiable placements in the historical record is also used to determine how many CMs were needed within each district and grade/subject area in our simulations. Consequently, our simulated list use should mechanically produce patterns of under-recruitment that mirror any lack of linkage coverage exhibited by the administrative data. For instance, even if unidentified CMs from disproportionately urban districts are excluded from the comparison group, the pool of CMs hypothetically recruited in our simulations also omits the same number from the same districts. This reciprocal relationship—coupled with our use of block fixed effects that restrict comparisons between the hypothetical pool and historical record to within the same district and subject area—should largely mitigate these concerns.

Pool of Hypothetical Clinical Mentors

The MMR algorithm works by first identifying and then ranking all teachers who could have potentially served as CMs. Identification involves determining the eligibility to mentor each PST, where eligibility indicates that a teacher (a) teaches a subject that corresponds with a PST's endorsement area, (b) works in a school district in which a PST could be placed, and (c) meets the state's qualification and experience requirements for eligibility. To determine (a), we lift PST

endorsements from the same administrative records for all program completers in the state that include the PST-CM linkages described above, while statewide data on teaching assignments come either from a personnel dataset, which provides information on teachers' annual assignment codes (e.g., Grade 9-12 Teacher) and tested subjects (e.g., English III), or a separate course-level datafile, which includes the names of the classes each teacher taught (e.g., AP English Language & Composition). These three variables collectively help identify the course content assigned to each teacher with varying levels of granularity and coverage. We use all three to generate a set of about 2,000 unique indicators (henceforth referred to as *teaching assignment indicators*) that take a value of one for teachers we observe teaching a given subject and zero otherwise. We draw on these teaching assignment indicators to construct a correspondence between the narrower subjects taught by CMs and the coarser endorsement areas that PSTs pursue in Stage 2 below.

With respect to part (b) of eligibility, the personnel file mentioned above also provides indication of all schools and districts in which each teacher is employed. Part (c) involves meeting state-determined criteria of qualification and experience.[2] These criteria are usually operationalized by programs as requiring clinical mentors to have at least three years of teaching experience and score "above expectations" on their yearly teacher evaluations (in addition to holding an active teacher license for the endorsement area of the PSTs who they would supervise). We pull data on years of teaching experience from the same personnel file, while the Tennessee evaluation system provides two sources of data on instructional effectiveness—observation ratings (ORs), theoretically available for all teachers, and value-added to student achievement (VAMs), available for the teachers in the state assigned to tested grades or subjects (less than half of all teachers). For each, we keep three prior years of evaluation metrics, ranging back to the 2014-2015 school year. These three measures—ORs, VAMs, and years of experience—are also the same sources of data that we use to ultimately rank all teachers who could have potentially served as CMs; we provide greater detail on how they are used algorithmically to generate recommendation lists below.

CM and FPS Characteristics

In order to explore how the hypothetical statewide implementation of the MMR initiative would have influenced placement practices along dimensions other than CM instructional effectiveness and experience, we incorporate several other sources of data. For CM characteristics, the same personnel dataset provides information on teachers' race/ethnicity and gender. Additionally, we use the identifiers for the school(s) of teachers' employment in 2017-2018 to merge a school-level datafile containing many potential FPS characteristics including average student demographics, achievement, and discipline as well as average teacher turnover and effectiveness. These variables collectively serve as our outcomes of interest for our second research question in the simulations described below.

Simulation Stages

We organize our simulation into three distinct but interconnected stages. These stages mimic as closely as possible the real-world CM recruitment process that EPPs follow when making clinical placements for PSTs, both traditionally and as modified by the original MMR initiative (which was designed in partnership with program leaders to closely map onto existing practices). First, we determine the total number of placements needed in each school district and

[2] Tennessee Educator Preparation Policy 5.504 articulates eligibility to serve as a clinical mentor as licensed teachers who "are qualified and experienced in their fields" (p. 39). See https://www.tn.gov/content/dam/tn/stateboardofeducation/documents/meetingfiles2/1-27-17_III_G_Educator_Preparation_Policy_5_504_Clean_Copy.pdf for more details.

for each PST endorsement area; second, we identify and rank all teachers in each school district by endorsement area block that meet the requirements to serve as CMs; and third, we simulate teachers' responses to the request to mentor a PST. Below, we provide more details on each stage of our simulation.

Stage 1: Calculate the Number of Requested Placements in Each Endorsement Area

Stage 1 involves estimating the total number and district-by-subject area distribution of CMs that all programs in the state would have needed to recruit with recommendation lists constructed using the MMR algorithm. We first calculate the sum of PST-CM linkages in 2017-2018 to identify the total number of clinical placements that all EPPs across the state had to make. However, these links do not specify the PST endorsement area for which each CM served or the district in which each PST was placed—two crucial pieces of information for making placements both via status quo recruitment practices, as identified in prior literature (e.g., St. John et al., 2021), and through the MMR algorithm, which conducts list-based recruitment independently within each district by endorsement area block. As a result, for each linkage involving a PST pursuing multiple endorsement areas and/or a CM working across multiple districts, we cannot clearly identify the district-by-endorsement area block in which PSTs historically requested placements; therefore, we cannot determine the corresponding district-by-endorsement area list from which a CM should be recruited when generating our hypothetical comparison.

To account for this uncertainty, we reshape the record of historical linkages to represent all possible district by endorsement area blocks for which a teacher who historically served as a CM could have been fulfilling a PST request. For example, we count a CM employed in two different districts hosting a PST pursuing both elementary and special education endorsements as four distinct historical placement requests (though, as discussed later, we weight our regressions to account for this duplication). This forces us to drop the few placements for which PSTs' endorsement area(s) (N = 12) and/or CMs' district(s) (N = 186) are unidentified. We complete Stage 1 by totaling the number of historical placements made in each district-by-endorsement area block (including these duplications).

Stage 2: Algorithmically Identify and Rank All Potential CMs in Each Endorsement Area

The second stage emulates what we have thus far called the MMR "algorithm," which diverges somewhat from how PSTs' field placements are traditionally made but mimics the process used in the original intervention. In this stage, we generate a ranked "recommendation list" for each district-by-endorsement area block across the state of the most promising (i.e., instructionally effective and experienced) teachers who could have potentially been recruited as CMs.

Indication of the district(s) of employment for all teachers is readily available. However, determining the eligible endorsement areas each teacher could supervise is more challenging, as the available information on the subject areas to which a teacher is assigned is much more granular than—and thus does not directly correspond to—PST endorsement areas. Therefore, we sequentially classify all CMs as eligible or ineligible for each endorsement area based on their values for the set of roughly 2,000 teaching assignment indicators we constructed (described above). This step corresponds to the status quo practice of schools, districts, and programs assessing—based upon the endorsement areas of each PST—which teachers in each district could serve as CMs.

More specifically, we use lasso-penalized linear probability models to identify the teaching assignments that most strongly indicate whether a teacher hosted a PST in each endorsement area. Intuitively, these models eliminate all indicators uncorrelated with having served as a CM for each specific endorsement area (e.g., shrinking the coefficient on teaching secondary biology to zero when predicting the likelihood of serving as a CM for elementary education endorsements) while

retaining those that offer the most predictive power. In this way, lasso regression helps produce sparser models (i.e., simpler and with fewer variables) that facilitate interpretation and minimize the risk of overfitting (i.e., poorly generalizing to a different sample), making it an especially appropriate and necessary technique when researchers have many highly correlated and often irrelevant predictors at hand.

Figuratively speaking, lasso can approach this task of cutting out less predictive variables with a machete or a scalpel (or anywhere in between), adjusting this severity by fine-tuning what is called a regularization hyperparameter (λ). To find the optimal level of λ across a range of potential values (i.e., picking the right size blade for the task), we use cross-validation. More specifically, we randomly divide our dataset of linked PSTs and historical CMs into 10 groups and iteratively hold out one at a time as a test set. Then, for each iteration, we train our lasso model on the collective remaining nine groups before estimating model fit (as measured by mean squared error) on the test set. We follow this procedure for each possible value of λ, taking as optimal the one that produces the lowest mean squared error averaged across all ten iterations. We then use our entire sample to run a final lasso model at this value of λ, thereby producing a set of coefficients for only those indicators that are collectively most helpful in determining a given PST endorsement area-CM teaching assignment correspondence and shrinking the coefficients for all unpredictive indicators to zero. We carry out this same process for each endorsement area to generate a set of final, trained linear probability models.

We then use these models to determine the best-performing threshold for determining CM eligibility to serve for a given endorsement area (and therefore whether a teacher should be included on a given recruitment list). More specifically, for this step, we estimate the predicted probability of having mentored a PST in each endorsement area using the coefficients from our trained models. We then consider a range of potential cut-offs, where only teachers with higher predicted probabilities would be classified as having been eligible. For each cut-off within this range, we plot the true positive rate (i.e., the proportion of actual CM-PST linkages predicted correctly by the model) against the false positive rate (i.e., the proportion of nonexistent linkages incorrectly predicted as linkages) to generate what is called a receiver operating characteristic curve. From this curve, we identify the specific value that maximizes true positives while minimizing false positives. Appendix Figure A1 displays the curve for elementary education endorsements—the most common endorsement area in our sample. The diagonal straight line represents a prediction based on a coin flip, with the area above this line improving on random prediction. We identify an optimal cutoff of 0.294 (corresponding to the point on the curve closest to (1,0), which indicates all true positives and no false positives).

Finally, we take the results of this process and apply them to all teachers across the state. That is, we again use the coefficients from our trained models to estimate the predicted probability of being eligible to mentor a PST in each endorsement area for *all* teachers; we then dichotomize these probabilities, classifying each mentor as eligible or ineligible depending on whether their value exceeds or falls below the optimal cutoff for a particular endorsement area. Drawing on the example above, we would include any teacher with a predicted probability of being able to serve as a CM for elementary education endorsements of at least 29.4%—based on their teaching assignment codes, their course assignments, and their tested subjects—as eligible for inclusion on the recommendation list for a given district. Note that for particularly uncommon endorsement areas, no teachers receive a predicted probability above the optimal cutoff. As a result, we further exclude from our sample the small number of placement requests (N = 178, approximately 4.0%) involving especially rare endorsement areas due to the difficulty of identifying eligible CMs; Panels A and B of Appendix

Table A1 report the list and relative frequencies of included and excluded endorsement areas respectively.

Lastly, we move to the step of ranking the teachers eligible for each list, ordering them from most experienced and effective to least. To accomplish this, we generate a single composite measure that approximates their promise or quality which we call the recommendation index. This index, developed in the original MMR intervention, draws on administrative data from the three years prior to recruitment, with the most recent year contributing half of the total score and the two preceding years a quarter each. Within each year, we take an average of teachers' standardized ORs (weighted at 40%), VAMs (also 40%), and years of experience (20%). In most cases (see exceptions below), teachers missing some elements of evaluation data (e.g., those in non-tested subjects without VAMs) remain included, with non-missing variables simply reweighted accordingly (for further detail on the calculation of the recommendation index, see Ronfeldt et al. [2020]). We then construct the actual lists by sorting all identified teachers by recommendation index score from highest to lowest within each district-by-endorsement area block. Our lists end when we run out of teachers who (a) are classified as eligible for a given endorsement area, (b) are employed in a given district, and (c) have at least three years of experience and an average OR in the prior year of at least a 4.0 (i.e., the most common operationalization of the minimum threshold criteria established by the state for serving as a CM). Stage 2 ends when we have generated recommendation lists for all district-by-endorsement area blocks in the state.

Importantly, this stage is the step at which the MMR initiative (and thus our simulation) and traditional recruitment practices mostly diverge. While recruitment lists are exclusively a function of instructional effectiveness and experience, prior literature on status quo recruitment practices (Krieg et al., 2020; Maier & Youngs, 2009)—as well as anecdotal evidence from ad hoc interviews with program leaders and district officials in Tennessee—reveal that CMs tend to be recruited based on a variety of potentially competing priorities. For instance, while teaching quality and coaching capacity are certainly important criteria in determining a placement, placement coordinators also report balancing these with additional concerns like making placements with strong relational fit between CM and PST, with logistical challenges around geographical proximity and determining the pool of available CMs, and with principal or district motivations that can include a desire to reward seniority, provide "help" to struggling teachers, and give every teacher a chance to serve (Ronfeldt et al., 2020).

It is also worth pointing out that this algorithm-based approach to recruitment inherently excludes some potentially promising teachers from being selected as CMs due to issues with data coverage. Specifically, we can only determine the eligibility of those teachers for whom teaching assignment data are available; additionally, we only generate a recommendation index value for teachers with evaluation data, as we opt not to include any teachers missing both ORs and VAMs (and therefore with only an identifiable amount of teaching experience). Teaching assignment data is missing districtwide for a small number of districts; because we cannot, therefore, identify any eligible teachers working in these districts to hypothetically serve as CMs, we exclude the 39 placements historically made in these districts from our analysis entirely. However, we retain in our comparison sample all other historical placements for which a CM is missing teaching assignment (N = 491) and/or evaluation data (N = 240) so as to compare the results of our simulation against the most comprehensive historical record available, even if these same CMs have no likelihood of being selected under hypothetical statewide implementation of the MMR algorithm due to missing data. While these historical placements often drop out of analyses—for example, placements missing evaluation data will be excluded when using instructional effectiveness as an outcome—they still

may contribute to the comparison group for outcomes from other datasets (e.g., CM and FPS demographics).

Stage 3: Extend the Offer to Serve as a CM and Account for the Likelihood of Acceptance

Our final stage aims to model the real possibility—both in status quo recruitment and in the MMR initiative—that teachers reject the offer to serve as a CM. Given that we are unable to determine whether a teacher would accept a request to serve, we run a Monte Carlo simulation to account for the stochastic nature of CM decision-making. Intuitively, this means that we run a series of random trials wherein we repeatedly simulate what might have happened if every program in the state had been provided recommendation lists and had used them as prescribed—offering the opportunity to serve as a CM to the first teacher on each list and working their way down until all placements were made—while accounting for uncertainty in how teachers might respond. By aggregating and summarizing the results of individual trials, we can estimate the plausible distribution of effects of hypothetical statewide use of these algorithmically generated lists based on randomly varying patterns of teacher responses to the request to serve as CMs.[3] Specifically, within each trial, we move from the top of the list (i.e., highest index value) down and randomly[4] predict whether each teacher accepts the offer to serve as a CM. This process stops when the total number of recruited CMs matches the total number of requested placements for each district-by-endorsement area block as determined in Stage 1 above or when a list runs out.[5] Stage 3 comes to a close when we have repeated our simulation 10,000 times, thereby generating a plausible distribution of the statewide effects of receiving recommendation lists.

In our preferred specification, we set the probability of CM acceptance to 75%. This number is based on the average proportion of teachers on recommendation lists provided to districts who were observed actually serving as CMs during the original MMR initiative. Previous work has found that teachers are very likely to accept the invitation to serve as a CM, with estimates of as high as a 95% acceptance rate (e.g., Ronfeldt et al., 2020). At the same time, other factors may drive this value

---

[3] It is important to note here that our simulation assumes that eligibility and acceptance to serve as a CM are independent across endorsement areas. This allows the possibility of a single mentor hosting PSTs in multiple endorsement areas (e.g., K-3 early childhood and K-5 elementary education). While historical records indicate that some CMs do host more than one placement, occurrences where a CM serves multiple PSTs—especially beyond two or three—are less common. To address this issue, we include a block (i.e., endorsement area by district) fixed effect in our models, assuming that, on average and all else equal, the possible bias arising from this independence assumption is accounted for at the block level.

[4] We use random here as in random variable—that is, a variable that can have more than one outcome given the same set of inputs (or, technically, a stochastic process). In our case, the input is receiving a request to serve as a CM; the output is accepting or declining this offer. As we are not aware of any empirical study that has attempted to determine which characteristics might predict the likelihood of a teacher accepting (or declining) the offer to serve as a CM—either under status quo practices or when using recommendation lists—we have decided to use the uninformative prior of a Bernoulli distribution that encodes the least amount of prior knowledge about the CM decision-making process.

[5] While our simulations seek to make the exact same number of placements within each district and endorsement area block as the historical record, in some cases, we fall short. This happens primarily for less common endorsement areas (e.g., ESL) or in smaller districts, where the number of PSTs historically placed with CMs in these blocks is close to or exceeds the number of teachers on the corresponding list. The exact number of how many placements this affects varies in our Monte Carlo simulations as a function of how many potential CMs randomly end up declining the offer to serve but is generally quite small; in our 100% acceptance models, our recommendation lists produce 4,443 placements (just 26 fewer than the 4,469 we retain from the historical record).

lower; for instance, since recruitment often takes place over the summer, lists must be developed using teacher assignment information from the prior school year and therefore do not always accurately reflect teacher employment and assignment information for the upcoming year. Therefore, we conceptualize the act of declining the invitation to serve as a CM in this stage as reflective of not only the probability of a CM turning down an offer but also of instances where teachers retire or move recruitment blocks from year to year (e.g., switching districts or teaching assignments).

As a robustness check, we also provide estimates from analyses that assume 100% mentor acceptance. These analyses are not Monte Carlo simulations; they inherently do not introduce any stochasticity and are therefore run only once. This alternative specification serves three purposes. First, it provides a mechanical upper bound for our first research question investigating the hypothetical impacts of statewide implementation of the MMR algorithm on the instructional quality of the mentors recruited, given that any mentor who declines will have a higher index score than the subsequent mentor (lower on the list) who accepts in their place. It also provides a theoretical upper bound[6] on the bias that statewide implementation might introduce, given prior evidence of ORs favoring White, female teachers employed in Whiter, wealthier, and higher-achieving school. Finally, it offers traditional measures of statistical significance, as results come from a single regression model per outcome measure rather than a distribution across 10,000 Monte Carlo simulation trials.

Sample

Our final comparison sample consists of 4,469 placements made across the state in 2017-2018. This historical record of the placements in which PSTs actually completed their student teaching serves as our point of reference in assessing whether and how the hypothetical statewide adoption of recommendation lists would have been an improvement over existing recruitment procedures, in terms of the instructional effectiveness and experience of CMs as well as any potential bias in CM and FPS characteristics.

Still, while it is critical to understand how the MMR intervention, when scaled across the state, compares to historical practices, it is not enough to stop there. Previous research has shown that existing structures for determining clinical placements may already result in the over-selection of certain kinds of teachers and schools as CMs and FPSs. Thus, the scaleup of an intervention that does no worse than the status quo is not necessarily a triumph if the status quo already involves considerable underrepresentation. Therefore, to assess the extent to which historical placement practices in Tennessee *already* privilege White, women teachers working in Whiter, wealthier, and higher-achieving schools, we take the preliminary step of comparing the historical record of placements with *all* eligible placements (i.e., every teacher on each list, regardless of their ranking) across a wide range of CM and FPS characteristics, intuitively restricting all comparisons to within the same district and endorsement area block through the inclusion of fixed effects.

Interestingly, our findings in Table 2 differ from prior research, as status quo recruitment practices in Tennessee in 2017-2018 do not necessarily appear biased across CM and FPS demographic characteristics. Compared to all other possible placements within the same district and endorsement area blocks, historical placements consist of similar proportions of

[6] We expect that estimates on these characteristics, while not monotonic by construction like the components of the recruitment index, will also be reduced, given evidence in prior literature that suggests measures of instructional quality are related to teacher and school characteristics. However, these changes inherently depend on the distributions of the characteristics of teachers (and schools) given an offer to serve as CM (and FPS).

White and women CMs teaching in FPSs with comparable compositions of students by race, socioeconomic status, and achievement. The only notable difference is that PSTs are placed, on average, in FPSs with higher staff turnover and lower average faculty instructional effectiveness than the average eligible FPS—a result in the opposite direction of what we might have expected. We believe that this divergence from previous literature stems from our inclusion of district by endorsement area fixed effects and our comparison group restriction to only those teachers and schools who meet eligibility criteria, which offer a more compelling counterfactual than, for instance, comparing to all teachers across the state.[7]

Table 2

Comparison of Historical Placements to All Eligible Placements

	(1)		(2)		
	All Eligible Placements		**Historical Difference**		
	Constant	**SE**	***ß***	**SE**	***N***
Panel A: CM Effectiveness					
Placement Index	0.396***	(0.004)	-0.215***	(0.034)	54,737
Observation Ratings (ORs)	4.398***	(0.001)	-0.082***	(0.013)	54,122
Value-Added (VAMs)	0.070***	(0.001)	0.005	(0.012)	29.043
Years of Experience	14.657***	(0.037)	-2.122***	(0.328)	54,545
Panel B: CM Characteristics					
Woman	0.820***	(0.001)	0.009	(0.010)	54,460
White	0.834***	(0.001)	0.008	(0.012)	53,840
Black	0.074***	(0.001)	-0.003	(0.005)	53,840
Asian	0.002***	(0.000)	0.003	(0.002)	53,840
Other Ethnic-Racial Identity	0.091***	(0.001)	-0.008	(0.013)	53,840
Panel C: FPS Student Characteristics					
White Student %	69.228***	(0.160)	-1.709	(1.340)	54,144
Black Student %	18.566***	(0.095)	0.644	(0.799)	54,144
Hispanic Student %	9.119***	(0.089)	1.067	(0.747)	54,144
Asian Student %	2.503***	(0.017)	-0.017	(0.146)	54,144
F/RPL Student %	33.063***	(0.157)	1.519	(1.320)	54,144
Student with Disability %	13.787***	(0.018)	-0.009	(0.149)	54,144
ELL Student %	0.665***	(0.013)	0.118	(0.113)	54,144
Proficient/Advanced Rate	42.690***	(0.104)	0.137	(0.879)	53,437
Suspension Rate	5.468***	(0.042)	0.312	(0.348)	49,833
Panel D: FPS Teacher Characteristics					
Teacher Turnover Rate	0.137***	(0.001)	0.014*	(0.005)	54,634
Average Teacher OR	4.174***	(0.002)	-0.081***	(0.017)	54,721
Average Teacher VAM	0.021***	(0.002)	-0.024	(0.012)	40,374

Note. Column (1) displays the average value of each characteristic among all eligible potential placements; Column (2) illustrates the difference between Column (1) and all historical placements. Block (i.e.,

[7] In these analyses, we also find surprising evidence of how much room for improvement historical recruitment practices have in terms of the instructional effectiveness and experience of the teachers serving as CMs. Placements were historically made with teachers who have about two fewer years of experience and lower ORs than all other eligible CMs in the same districts and endorsement areas, underscoring again the potential of the MMR initiative to improve the instructional quality of teachers recruited to serve as CMs.

district by endorsement area) fixed effects included, with each CM also weighted by the inverse of the number of blocks for which they historically or hypothetically served. Standard errors clustered by district. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Altogether, Table 2 suggests that business-as-usual clinical placement processes may, in fact, produce a representative pool of CMs and FPSs, at least for this particular year in Tennessee. At the same time, we note that this analysis does not necessarily prove or guarantee the absence of any evidence of inequity in status quo recruitment practices, as research suggests there may be inequities baked into the measures of effectiveness used as thresholds for determining the eligibility of teachers to serve as mentors. If such biases exist, then teachers of color or teachers working in classrooms with minoritized and marginalized populations may have lower likelihoods of being rated at or above the category of "Above Expectations" than their equally instructionally effective colleagues, thereby limiting the diversity of the pool of potential CMs and FPSs. Even were that not the case, an argument could be made that merely having a set of clinical placements demographically representative of the historical and eligible pools of teachers is insufficient. The diversity of the teacher workforce falls woefully short of that of the student population, and a growing body of literature has documented the potential benefits of teacher diversity for student academic and non-academic outcomes (for a review of the evidence, see Bristol & Martin-Fernandez, 2019, and Redding, 2019). Although we do not explicitly entertain this possibility in this paper, we think it valuable to point out that perhaps truly equitable practices would produce a group of CMs representative of the statewide population of students rather than teachers.

Analytic Approach

Although our simulation may be complex, our analytic approach is quite simple. Within each trial, we use linear regression to estimate the difference between historical placements and hypothetical placements made under plausible statewide implementation of recommendation lists for a variety of CM and FPS characteristics. Our regression model is

$$Y_{bi} = \beta_0 + \beta_1 \cdot List_{bi} + \delta_b + \varepsilon_{bi}$$

where Y_{bi} is a characteristic of the CM (e.g., gender) or FPS (e.g., percentage of White students) for PST i in recruitment block b (i.e., school district by endorsement area). $List_{bi}$ is an indicator variable that flags all "hypothetical" CMs recruited using each district-by-endorsement area recommendation list. β_1 is the coefficient of interest and estimates the difference in variable Y_{bi} between hypothetical CMs—teachers on a recommendation list who accepted the invitation to serve—and historical CMs. Note that, by the nature of how the sample is constructed within each trial, CMs who historically served *and* were hypothetically recruited via recommendation lists (and their respective FPSs) make no contribution to this coefficient; instead, impacts of receiving these lists are driven by differences between historical placements that would not have been made under the plausible statewide scaleup of the MMR algorithm and vice versa.

Our regression models also include a placement block fixed effect, δ_b, to account for possible differences in Y_{bi} across district-by-endorsement area blocks. This fixed effect effectively restricts our estimates of β_1 to leverage variation within placement blocks and ensures that we make comparisons of the characteristics of historical and hypothetical placements that are unaffected by differences across districts and/or endorsement areas. In addition, we cluster standard errors at the district level (ε_{bi}) to further account for any possible correlation in residual terms within each cluster. Finally, to account for the fact that some placements involve the same

pairings of CMs (i.e., those employed in multiple districts) and PSTs (i.e., those with multiple endorsement areas), we weight each observation by the inverse of the number of unique districts and endorsement areas captured in each CM-PST linkage; to return to our (much) earlier example, a CM employed in two different districts hosting a PST with both elementary and special education endorsements (resulting in four possible combinations) has each observation weighted at 0.25.

As we report in greater detail below, we find that hypothetical statewide use of our algorithm would have introduced certain small but significant inequities in the kinds of schools chosen as FPSs. Subsequently, then, we sought to find ways to adjust the algorithm underlying the MMR initiative to correct for these inequities, specifically focusing on ORs, given the body of research documenting their potential sensitivity to teacher and student characteristics. More specifically, we use a two-step regression-based approach to explore the extent to which preliminarily adjusting ORs for FPS characteristics can address any new inequities in the clinical placements that would have plausibly been made under statewide implementation. In the first step, we use the regression

$$OR_{bi} = \beta_0 + X_{bi}\Gamma + \varepsilon_{bi}$$

where OR_{bi} is the unadjusted weighted composite OR for teacher i in recruitment block b. X_{bi} is a vector of all available FPS observed characteristics (see Panels C and D of Table 2)—namely, compositional variables like the percentages of students in each available racial category (omitting the proportion of White students to avoid collinearity), eligible for free and reduced-price lunch, with disabilities, and identified as English language learners as well as school-average measures of achievement, discipline, staff turnover, and teacher instructional effectiveness. To account for varying levels of data coverage, we impute a value of zero and add an indicator for all observations with missing values for each variable.

After fitting this model, we calculate the residual terms, which thereby only include variation independent of the vector of included FPS characteristics. We then incorporate these residuals—in lieu of raw ORs—in the creation of composite recruitment index scores at the end of Stage 2 of our simulation. Using this adjusted index, we then re-rank the teachers in each endorsement area and district to construct alternative recommendation lists before simulating their use and estimating their effects following the otherwise unaltered approach described in the sections above.

Results

Table 3 reports estimates of the simulated impacts of statewide implementation of the targeted recommendation lists generated by the MMR algorithm on the characteristics of CMs and FPSs by comparing hypothetical and historical placements. Column (1) displays our preferred approach, wherein we account for the possibility that teachers could decline the invitation to serve as a CM by setting the probability of accepting the invitation to 75%; estimates in this approach come from the median of the distribution of all 10,000 Monte Carlo simulations and are accompanied by a 90% credible interval.[8] In Column (2), we report alternative estimates under the

[8] Note that we do not report traditional significance levels for the 75% acceptance models. This is because these estimates should be interpreted within a Bayesian framework of credible intervals rather than through traditional inferential statistics. For consistency's sake, we still mark coefficients of which the 90% credible

assumption of 100% acceptance; these results provide traditional measures of statistical significance as well as upper bounds for the simulated impacts of statewide implementation of the algorithm.

Table 3

Simulated Effects of Recommendation List Use on CM and FPS Characteristics

	(1) 75% Acceptance		(2) 100% Acceptance		
	Median *ß*	**90% CI**	***ß***	**SE**	***N***
Panel A: CM Effectiveness					
Placement Index	0.700†	[0.694, 0.705]	0.762***	(0.029)	8,641
Observation Ratings (ORs)	0.213†	[0.207, 0.218]	0.224***	(0.016)	8,490
Value-Added (VAMs)	0.120†	[0.111, 0.128]	0.141***	(0.023)	2,771
Years of Experience	8.197†	[8.066, 8.324]	9.041***	(0.542)	8,417
Panel B: CM Characteristics					
Woman	0.006†	[0.002, 0.111]	0.005	(0.008)	8,426
White	-0.002	[-0.005, 0.002]	-0.001	(0.010)	8.235
Black	0.001	[-0.003, 0.002]	-0.004	(0.004)	8.235
Asian	-0.002†	[-0.002, -0.001]	-0.001	(0.001)	8.235
Other Ethnic-Racial Identity	0.004†	[0.001, 0.006]	0.006	(0.012)	8.235
Panel C: FPS Student Characteristics					
White Student %	3.093†	[2.891, 3.301]	3.133*	(1.368)	8,614
Black Student %	-1.731†	[-1.906, -1.561]	-1.799*	(0.870)	8,614
Hispanic Student %	-1.532†	[-1.649, -1.420]	-1.533*	(0.682)	8,614
Asian Student %	0.186†	[0.154, 0.217]	0.211*	(0.106)	8,614
F/RPL Student %	-2.574†	[-2.768, -2.382]	-2.686*	(1.235)	8,614
Student with Disability %	-0.100†	[-0.171, -0.030]	-0.168	(0.140)	8,614
ELL Student %	-0.201†	[-0.218, -0.183]	-0.194*	(0.087)	8,614
Proficient/Advanced Rate	1.871†	[1.681, 2.060]	1.777*	(0.783)	8,389
Suspension Rate	-0.476†	[-0.548, -0.406]	-0.498*	(0.242)	7,833
Panel D: FPS Teacher Characteristics					
Teacher Turnover Rate	-0.013†	[-0.014, -0.012]	-0.014**	(0.005)	8,667
Average Teacher OR	0.108†	[0.105, 0.112]	0.114***	(0.020)	8,632
Average Teacher VAM	0.046†	[0.040, 0.053]	0.038	(0.024)	6,562

Note. Column 1 is estimated using a Monte Carlo simulation with 10,000 trials, where † indicates that the 90% credible interval does not cross 0; Column 2 is estimated using a single traditional linear regression model, where $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. Block (i.e., district by endorsement area) fixed effects included, with each CM also weighted by the inverse of the number of blocks for which they historically or hypothetically served. Standard errors clustered by district.

The first panel of Table 3 answers our first research question by replicating and confirming the intended impacts of a hypothetical statewide scaleup of the MMR initiative. We estimate that the CMs recruited by recommendation lists would have been consistently more instructionally effective and experienced than those teachers who historically served as CMs in 2017-2018 by more than two thirds of a standard deviation on the recommendation index. This contrast appears driven by both

interval does not cross zero, which is analogous to testing for statistical significance in traditional regression models.

effectiveness and experience; on average, hypothetical mentors had ORs that were about 0.20 points higher than their historical counterparts, as well as VAMs that were about 0.12 standard deviations higher and more than an additional eight years of teaching experience. These results contribute to the significant body of causal evidence that this algorithm, as intended, can identify and lead to the recruitment of more instructionally effective and experienced teachers. Notably, our results hold for both the ideal and the more realistic scenarios of list implementation; that is, even after accounting for the random 25% possibility that each teacher declines the invitation to serve as a CM, estimates of the effects of hypothetical statewide implementation on CM instructional effectiveness and experience are only reduced by about 10% and remain comparable to those observed via experimental contrast between treatment and control groups in prior work (Ronfeldt et al., 2020). Obtaining similar estimates to the results of a real-world field experiment not only corroborates the beneficial intended impacts of the initiative on a set of immediate and closely related measures but also provides face validity evidence for our simulations, suggesting that they closely approximate real-world recommendation list use.

We next attempt to anticipate the unintended consequences of scaled up implementation across a broader spectrum of outcomes, exploring other ways that the average characteristics of CMs and their FPSs might have changed under statewide adoption of the MMR initiative. In terms of gender (see Table 3 Panel B), we find that slightly more of the CMs recruited using recommendation lists were women than those who served historically, though this estimate is non-significant in our alternative (100% acceptance) specification. Moreover, it is quite small in magnitude (approximately half a percentage point) and is roughly equivalent to having recruited only about 25 more women to serve as CMs across the entire state. We also find that statewide implementation would not have significantly changed the racial or ethnic distribution of CMs and certainly not in a way that would have resulted in the over-recruitment of White CMs. These initial results suggest that scaling the MMR algorithm up as statewide policy would not appear to introduce practically meaningful gender or racial inequities in the kinds of teachers chosen to serve as CMs.

However, we find a different story when we examine FPS characteristics in Panels C and D of Table 3. On average, compared to historical placements, hypothetical placements made via recommendation lists were located in schools with significantly more White, Asian, and higher achieving students and fewer Black, Hispanic, and English language learners as well as significantly fewer students receiving free or reduced-price lunch and out-of-school suspensions. In addition, hypothetical placement schools had lower teacher turnover and higher average faculty instructional effectiveness. While these latter two impacts could be considered a benefit of the algorithm—resulting in the placement of PSTs in schools with better working conditions more conducive to their learning and development—the other results collectively reveal the possibility of small but significant biases in the algorithm underlying the MMR initiative in terms of the kinds of FPSs in which PSTs would have been placed.

Therefore, we explore possible corrections to the algorithm to address these biases in Table 4. In these analyses, we assess the impact of preliminarily (i.e., before ranking teachers) adjusting the OR component of the MMR recruitment index to account for FPS-level student and teacher characteristics. Compared to the hypothetical effects of statewide implementation produced under the original algorithm in Table 3, we find that this adjustment would have resulted in the selection of FPS sites that much more closely resemble the historical record, with at most very modest differences that do not appear practically meaningful. For example, under this version of the algorithm, PSTs would have been placed in FPSs with only about half a percentage point more White students (and half a percentage point fewer Hispanic students). The sole FPS characteristic in Panels C and D that retains a traditional level of statistical significance in our alternative specification

(Column 2; 100% acceptance) is a reduction in the average teacher turnover rate by roughly half a percentage point. And crucially, while nearly entirely alleviating the initiative's potential negative unintended consequences, this adjustment would lead to at most a very small reduction on its positive intended impacts. Estimates of the contrast in instructional effectiveness and experience (in Panel A of Table 4) are diminished by no more than 10% from those in Table 3, with CMs hypothetically recruited via this revised algorithm still scoring approximately two thirds of a standard deviation higher on the recommendation index than those who historically served.

Table 4

Simulated Effects of OR-Adjusted Recommendation List Use on CM and FPS Characteristics

	(1) 75% Acceptance		(2) 100% Acceptance		
	Median *ß*	**90% CI**	***ß***	**SE**	***N***
Panel A: CM Effectiveness					
Placement Index	0.654†	[0.648, 0.660]	0.721***	(0.026)	8,641
Observation Ratings (ORs)	0.140†	[0.133, 0.145]	0.150***	(0.013)	8,473
Value-Added (VAMs)	0.113†	[0.105, 0.121]	0.133***	(0.022)	2,821
Years of Experience	8.503†	[8.385, 8.628]	9.401***	(0.531)	8,417
Panel B: CM Characteristics					
Woman	0.008†	[0.004, 0.013]	0.008	(0.008)	8,431
White	-0.006†	[-0.010, -0.003]	-0.004	(0.010)	8.234
Black	0.006†	[0.003, 0.009]	0.003	(0.003)	8.234
Asian	-0.001†	[-0.002, -0.001]	-0.001	(0.001)	8.234
Other Ethnic-Racial Identity	0.002	[-0.001, 0.004]	0.002	(0.010)	8.234
Panel C: FPS Student Characteristics					
White Student %	0.615†	[0.399, 0.830]	0.617	(1.034)	8,623
Black Student %	0.010	[-0.177, 0.193]	0.003	(0.601)	8,623
Hispanic Student %	-0.543†	[-0.670, -0.423]	-0.528	(0.618)	8,623
Asian Student %	-0.062†	[-0.093, -0.032]	-0.075	(0.095)	8,623
F/RPL Student %	-0.049	[-0.245, 0.146]	0.004	(1.023)	8,623
Student with Disability %	0.062	[-0.004, 0.125]	0.070	(0.143)	8,623
ELL Student %	-0.087†	[-0.107, -0.069]	-0.084	(0.078)	8,623
Proficient/Advanced Rate	-0.485†	[-0.665, -0.302]	-0.472	(0.661)	8,414
Suspension Rate	0.351†	[0.264, 0.435]	0.383	(0.303)	7,850
Panel D: FPS Teacher Characteristics					
Teacher Turnover Rate	-0.006†	[-0.007, -0.006]	-0.006*	(0.003)	8,682
Average Teacher OR	0.010†	[0.006, 0.013]	0.009	(0.016)	8,632
Average Teacher VAM	0.022†	[0.014, 0.028]	0.022	(0.020)	6,569

Note. Column 1 is estimated using a Monte Carlo simulation with 10,000 trials, where † indicates that the 90% credible interval does not cross 0; Column 2 is estimated using a single traditional linear regression model, where $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. Block (i.e., district by endorsement area) fixed effects included, with each CM also weighted by the inverse of the number of blocks for which they historically or hypothetically served. Standard errors clustered by district.

Discussion

In the traditional approach to studying interventions being considered for policy implementation and scaleup, researchers often use experimental methods designed to evaluate "strategies which are well-defined and portable … on proximal and readily quantifiable outcomes" in a controlled setting (Oliver et al., 2020, p. 62). While these methods work well for identifying intended impacts, they often ignore unintended consequences. On the other hand, methods for investigating the side effects of policy typically involve retrospective evaluations of impacts on more distal and less easily measured outcomes and individuals. This approach is contingent on unintended consequences having already occurred and assumes, often naively, that any effects that researchers uncover will play a role in shaping future policy. For those side effects that are truly unforeseeable, waiting for them to take root may be necessary; however, where unintended consequences can be anticipated, a more proactive approach offers far greater promise to meaningfully inform policy.

Failing to anticipate how policy unintentionally perpetuates or exacerbates harms is inexcusable, particularly among researchers and policymakers concerned with the inequities that pervade our public system of education. Given overwhelming evidence of widespread and persistent disparities in the access, opportunities, and resources provided to students and teachers of color in American schools (Clotfelter et al., 2023; Gopalan, 2019; Ladson-Billings, 2006), and the ineffectiveness of color-evasive attempts to reduce their magnitude (Gillborn, 2019; Goldin & Khasnabis, 2021; Wells, 2014), it is difficult to claim that whatever unintended impacts a program or policy may have in reproducing or worsening racial inequity could not have been fathomed prior to its implementation. And yet continuing to conflate the predominant term of "unintended" with its more precise predecessor of "unanticipated" allows just that; when all unintended consequences are unanticipated, policymakers can avoid taking blame—and researchers can sidestep assigning it—for the foreseeable side effects they (at a minimum, implicitly) permitted.

We believe that running simulations on historical data provides an avenue to more carefully and proactively consider the unintended implications for (in)equity of transforming otherwise promising initiatives into statewide policy prior to adoption at scale. Intuitively, each trial in a simulation represents a possible implementation scenario that attempts to mimic rollout in the real world while accounting for the stochasticity that makes perfect prediction unattainable. Aggregating across these trials then provides a distribution of the plausible effects that policymakers could expect if they were to make policy a reality. In this paper, we offer an illustration of the promise of this approach by detailing how we collaborated with the TDoE to simulate the statewide scaleup of an otherwise auspicious initiative, investigating its likely unintended consequences for racial inequity and identifying possible refinements to forestall them in advance.

Lessons Learned about the Initiative Itself

Overall, the results of our simulations both reinforce the promise and reveal the potential unintended drawbacks of using algorithm-generated recommendation lists to identify and recruit high-quality teachers to serve as CMs. The hypothetical CMs recruited using these recommendation lists were consistently more instructionally effective (by more than 0.20 points on the state observational rubric and more than 0.10 standard deviations in VAMs) and experienced (by eight or nine years) than the teachers actually recruited by programs to serve as CMs in the past. Consistent with prior experimental results in real practice settings, this result suggests that these recommendation lists can serve as a potential policy lever to—as intended—meaningfully increase the instructional quality and experience of CMs, which, in past research, have been shown also to improve the preparedness and instructional performance of their pre-service teachers (Goldhaber et al., 2022; Ronfeldt et al., 2020, 2025). That said, it is likely that

any increase in PST effectiveness would be quite modest. Prior research in the same setting estimated that being placed with a CM whose ORs were a full point higher is associated with at most an increase of 0.10 in PSTs' ORs, which roughly equates to an additional half a year of teaching experience (Ronfeldt et al., 2018); as such, the statewide implementation of the MMR initiative might only offer the equivalent of a 0.02 point increase in PSTs' average ORs or an additional month of experience. At the same time, though, this kind of policy intervention is inexpensive and relatively easy to implement, especially compared to alternatives (e.g., professional development for CMs); and it requires EPPs to adjust their current program structures very little, if at all, making statewide implementation viable even with a minimal level of support from departments of education. Consequently, though any improvement in instructional effectiveness for an individual PST may be modest, this intervention as statewide policy has potential to impact thousands of PSTs and their students.

Crucially, though, we find that scaling up this intervention across the state without making any adjustments to the algorithm upon which it relies could have negative unintended consequences in terms of the equitable selection of school sites as FPSs. Though small in magnitude, estimates from our simulations suggest that expanding the use of recommendation lists statewide would likely have resulted in the selection of FPSs with Whiter, wealthier, and higher-achieving student populations than those of the schools where PSTs were historically placed. We hypothesized that these predicted disparities likely arose from the MMR algorithm's dependence upon ORs, which have been shown in prior literature to be higher for teachers working in schools with more White and higher-achieving students, even after accounting for alternative measures of instructional effectiveness (Campbell, 2020; Campbell & Ronfeldt, 2018; Grissom & Bartanen, 2022; Steinberg & Garrett, 2016; Steinberg & Sartain, 2020). Put simply, if structural inequities are baked into evaluation measures, then these structural inequities will find their way into initiatives that depend upon such data, and scaling up these initiatives risks their further institutionalization and reproduction.

Fortunately, our analyses also suggest that we can largely eliminate these inequities by adjusting the ORs used in the algorithm for school characteristics. Removing the variation in ORs explained by school characteristics before calculating our recommendation index substantially reduced, and in most cases eliminated entirely, any inequities in the FPSs that statewide implementation of the MMR intervention would have likely produced. Importantly, doing so does not come at the expense of the increase in mentor instructional effectiveness and experience that the algorithm achieves, nor does it have any negative repercussions for the racial and gender representativeness of teachers selected to serve as CMs. One of the clearest benefits of this revised algorithm, then, is its capacity to attend simultaneously to issues of placement diversity and CM quality, thereby mitigating any concerns as to how an exclusive orientation toward the former without complementary attention to the latter runs the risk of reinforcing preexisting racial stereotypes among PSTs (Bolyard & Baker, 2021).

Though the adjusted algorithm appears particularly promising, we caution that the current study is still not independently sufficient to warrant its widespread adoption without continued investigation and monitoring. While close to a hypothetical statewide implementation, our simulations are limited to only those endorsement areas, programs, districts, and teachers for whom data are available. In addition, we only consider the hypothetical impacts of scaling up this intervention in 2017-2018, largely due to the lack of consistent and comprehensive CM-PST linkages in other years prior to the introduction of the intervention among Tennessee EPPs. If our sample is biased by any omissions—or if any longitudinal changes in recruitment practices

make this particular year an anomaly—then our results may fall short of accurately portraying how the statewide implementation of this intervention would play out across different settings.

Lessons Learned about Simulating Statewide Scaleup

More broadly, we believe this study serves as an effective example of how to carefully and preemptively consider the unintended consequences of a policy intervention for racial equity by simulating its adoption at scale. While simulation-based research to evaluate education policy is not without precedent (Grissmer et al., 2014; Lee, 2004), it is rarely conducted prior to actual implementation (cf. Morris & Johnson, 2024) and/or with an explicit emphasis on demographic equity (cf. Matos-Díaz & García, 2014). Consequently, we take this opportunity to unpack several crucial takeaways for other researchers interested in similarly improving their evaluations of side effects to be more proactive and productive.

Foremost among them is the importance of having a strong, trusting partnership with policymakers who are open to acknowledging and shouldering their own potential culpability in producing and reproducing harm. Grappling with and explicitly labeling the subset of unintended consequences that can be anticipated requires that policymakers be willing to forgo their ostensible ignorance of the possible negative repercussions of the policy they enact, even if they could otherwise claim that such side effects fall outside of their purview.

There are countless challenges to forming strong partnerships with policymakers, many of which are a function of time (Booker et al., 2019; Conaway, 2020; Henrick et al., 2017). For example, developing the level of collaborative trust that can allow policymakers to engage in analysis and discussion of their accountability for continued inequity takes time, yet the same passage of time increasingly jeopardizes stability by increasing the likelihood of turnover. Researchers and policymakers also tend to operate on different timelines, and the slow, methodical exploration of unintended yet anticipated consequences may be difficult for elected or appointed officials who face pressures to make change more quickly, regardless of whether such change is based in evidence.

These temporal challenges underscore how collaboration also requires significant investment from all partners. Part of that investment is fiscal; because simulation work involves research that produces specific and highly contextualized hypothetical implementations rather than new and transferrable scientific knowledge, it is not as valued by academic journals and perhaps therefore not prioritized as highly by researchers. It may be necessary that policymakers, as the TDoE did in our case, be willing to financially support researchers to allocate their energy to these kinds of anticipatory policy analyses. However, much of the necessary investment also involves a non-pecuniary commitment from all parties to the value of the potential policy and to the importance of societal equity, thus ensuring that the time taken to preemptively vet its plausible statewide scaleup is not spent in vain. These analyses, for instance, had the good fortune of coming at the tail end of a long and productive research-practice partnership with substantial personnel stability and a shared consensus around the value of rigorous research, careful policy, and educational equity.

Finally, a good partnership is a broad one, involving not just state officials and academics but also community stakeholders from all the constituencies likely to be impacted by a policy, including practitioners, students, and families. The vantage points and histories of each different group of stakeholders can help bring to the partnership unique insights into anticipated forms of harm and promise. A larger and more diverse group of partners therefore increases the likelihood that each of a policy's likely unintended consequences actually be anticipated—and thus avoided—allowing for the preemptive evaluation of more potential side effects and minimizing the likelihood of harm. Relatedly, establishing a broad partnership across many

different stakeholders also pays dividends in improving simulation quality. Rich, detailed knowledge of the complexities of what real-world implementation of a policy would plausibly look like is necessary for the hypothetical scenarios of policy adoption generated by simulation to be accurate, and this knowledge comes best from the people responsible for turning policy into practice. In this particular case, our partnership directly included leaders at several EPPs and also indirectly drew on the perceptions of district officials, CMs, and PSTs via survey and interview. Including EPP leaders in our partnership ensured, for example, that the three stages of our simulations neatly mapped to the process of CM recruitment that programs typically used across the state, helping us construct a hypothetical that more closely reflected reality.

That said, it is impossible to perfectly predict and mimic what policy implementation would look like in reality, and researchers will always have to make decisions about how to model and simulate hypothetical scenarios that simplify or gloss over crucial complexity. For example, in our simulations, we treat CMs' acceptance of the offer to serve as a CM as a random process and assign all CMs—regardless of their characteristics—the same likelihood of 75%, based on calculations of the proportion of teachers on lists who ended up serving as CMs in previous iterations of the initiative. However, our simulations ignore any factors considered by district and program leaders when deciding to skip over a given teacher on a recommendation list, even though in practice we know that we encouraged programs to consider lists as guided supports rather than sequential prescriptions and advised them to pass over listed teachers if they had any reasons for doing so. Consequently, any relationship between the likelihood of being passed over by a program and a given CM or FPS characteristic has the potential to bias our results. We point this out not to undermine our own analysis but to illustrate how even the most sophisticated simulations may differ from the ways in which real-world implementation might unfold, while also highlighting how further leveraging collaboration among partners could have improved the accuracy and quality of our hypotheticals.

Taken together, these lessons underscore the importance of strong, broad, and stable partnerships for more intentionally designing and implementing equitable policy by proactively addressing potential adverse side effects. These negative consequences, while unintended, are often not difficult to anticipate, and with a collective commitment and investment on the part of researchers, practitioners, and policymakers, we can and must do better in proactively preempting their occurrence.

Acknowledgements

We appreciate the generous financial support that was provided for this research by the Institute of Education Sciences (IES), U.S. Department of Education through the Statewide, Longitudinal Data Systems Grant (PR/Award R372A150015). Emanuele Bardelli and Matthew Truwit also received pre-doctoral support from the Institute of Education Sciences (IES), U.S. Department of Education (PR/Awards R305B150012 and R305B200011). We are grateful to countless members of the TDoE for their work in building the foundation of our partnership, but especially Kevin Schaaf for his work with the implementation and management of the Mentors Matter Recruitment initiative. This research was supported in part through computational resources and services provided by Advanced Research Computing (ARC), a division of Information and Technology Services (ITS) at the University of Michigan, Ann Arbor. Notwithstanding any Tennessee Department of Education (TDoE) data or involvement in the creation of this research product, the TDoE does not guarantee the accuracy of this work or endorse the findings. Any errors

are the sole responsibility of the author(s). We have no known potential conflict of interest with respect to the research, authorship, and/or publication of this article.

References

Amrein-Beardsley, A., & Collins, C. (2012). The SAS education value-added assessment system (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives, 20*(12). https://doi.org/10.14507/epaa.v20n12.2012

Anderson, K. P., & Ritter, G. W. (2017). Disparate use of exclusionary discipline: Evidence on inequities in school discipline from a U.S. state. *Education Policy Analysis Archives, 25*(49). https://doi.org/10.14507/epaa.25.2787

Babbs Hollett, K., & Frankenberg, E. (2022). A critical analysis of racial disparities in ECE subsidy funding. *Education Policy Analysis Archives, 30*(14). https://doi.org/10.14507/epaa.30.7003

Bamberger, M., Tarsilla, M., & Hesse-Biber, S. (2016). Why so many "rigorous" evaluations fail to identify unintended consequences of development programs: How mixed methods can contribute. *Evaluation and Program Planning*, *55*, 155–162. https://doi.org/10.1016/j.evalprogplan.2016.01.001

Barrow, L., Sartain, L., & De La Torre, M. (2020). Increasing access to selective high schools through place-based affirmative action: Unintended consequences. *American Economic Journal: Applied Economics, 12*(4), 135-163. https://doi.org/10.1257/app.20170599

Bastian, K. C., Patterson, K. M., & Carpenter, D. (2022). Placed for success: Which teachers benefit from high-quality student teaching placements? *Educational Policy, 36*(7), 1583-1611. https://doi.org/10.1177/0895904820951126

Bolyard, C. S. & Baker, A. M. (2021). Diversity placements: Supporting the development of socially just teachers or reinforcing negative stereotypes? *Critical Questions in Education, 12*(3), 189-216.

Booker, L., Conaway, C., & Schwartz, N. (2019). *Five ways RPPs can fail and how to avoid them: Applying conceptual frameworks to improve RPPs.* William T. Grant Foundation. https://wtgrantfoundation.org/five-ways-rpps-can-fail-and-how-to-avoid-them-applying-conceptual-frameworks-to-improve-rpps

Bristol, T. J., & Martin-Fernandez, J. (2019). The added value of Latinx and Black teachers for Latinx and Black students: Implications for policy. *Policy Insights from the Behavioral and Brain Sciences*, *6*(2), 147-153. https://doi.org/10.1177/2372732219862573

Campbell, S. L. (2020). Ratings in black and white: A QuantCrit examination of race and gender in teacher evaluation reform. *Race Ethnicity and Education*, 1–19. https://doi.org/10.1080/13613324.2020.1842345

Campbell, S. L., & Ronfeldt, M. (2018). Observational evaluation of teachers: Measuring more than we bargained for? *American Educational Research Journal, 55*(6), 1233-1267. https://doi.org/10.3102/0002831218776216

Carver-Thomas, D., Bianco, M., Goings, R., & Hyler, M. E. (2024). Policies and practices for recruiting and retaining teachers of color. *Education Policy Analysis Archives, 32*(59). https://doi.org/10.14507/epaa.32.8123

Close, K. Amrein-Beardsley, A., & Collins, C. (2020). Putting teacher evaluation systems on the map: An overview of state's teacher evaluation systems post-Every Student Succeeds Act. *Education Policy Analysis Archives, 28*(58). https://doi.org/10.14507/epaa.28.5252

Clotfelter, C. T., Ladd, H. F., & Clifton, C. R. (2023). Racial differences in student access to high-quality teachers. *Education Finance and Policy, 18*(4), 738-752. https://doi.org/10.1162/edfp_a_00402

Conaway, C. (2020). Maximizing research use in the world we actually live in: Relationships, organizations, and interpretation. *Education Finance and Policy, 15*(1), 1-10. https://doi.org/10.1162/edfp_a_00299

de Zwart, F. (2015). Unintended but not unanticipated consequences. *Theory and Society*, *44*(3), 283–297. https://doi.org/10.1007/s11186-015-9247-6

Dumas, M. J. (2016). Against the dark: Antiblackness in education policy and discourse. *Theory into Practice*, *55*(1), 11–19. https://doi.org/10.1080/00405841.2016.1116852

Gillborn, D. (2005). Education policy as an act of white supremacy: Whiteness, critical race theory and education reform. *Journal of Education Policy*, *20*(4), 485–505. https://doi.org/10.1080/02680930500132346

Gillborn, D. (2019). Hiding in plain sight: Understanding and addressing whiteness and color-blind ideology in education. *Kappa Delta Pi Record*, *55*(3), 112–117. https://doi.org/10.1080/00228958.2019.1622376

Goldhaber, D., Krieg, J., & Theobald, R. (2014). Knocking on the door to the teaching profession? Modeling the entry of prospective teachers into the workforce. *Economics of Education Review, 43*, 106-124. https://doi.org/10.1016/j.econedurev.2014.10.003

Goldhaber, D., Ronfeldt, M., Cowan, J., Gratz, T., Bardelli, E., & Truwit, M. (2022). Room for improvement? Mentor teachers and the evolution of teacher preservice clinical evaluations. *American Educational Research Journal*, *59*(5), 1011–1048. https://doi.org/10.3102/00028312211066867

Goldin, S., & Khasnabis, D. (2021). In the pursuit of justice: Moving past color-evasive efforts. *Educational Forum*, *86*(1), 1–4. https://doi.org/10.1080/00131725.2022.1997307

Gopalan, M. (2019). Understanding the linkages between racial/ethnic discipline gaps and racial/ethnic achievement gaps in the United States. *Education Policy Analysis Archives, 27*(154). https://doi.org/10.14507/epaa.27.4469

Grissmer, D. W., Ober, D. R., & Beekman, J. A. (2014). Focusing on short-term achievement gains fails to produce long-term gains. *Education Policy Analysis Archives, 22*(5). http://doi.org/10.14507/epaa.v22n5.2014

Grissom, J. A., & Bartanen, B. (2022). Investigating race and gender biases in high-stakes teacher observations. *Journal of Policy Analysis and Management, 41*(1), 131-161. https://doi.org/10.1002/pam.22352

Grossman, P., Ronfeldt, M., & Cohen, J. J. (2012). The power of setting: The role of field experience in learning to teach. In K. R. Harris, S. Graham, T. Urdan, A. G. Bus, S. Major, & H. L. Swanson (Eds.), *APA educational psychology handbook, Vol. 3. Application to learning and teaching* (pp. 311–334). American Psychological Association. https://doi.org/10.1037/13275-023

Henrick, E. C., Cobb, P., Penuel, W. R., Jackson, K., & Clark, T. (2017). *Assessing research-practice partnerships: Five dimensions of effectiveness*. William T. Grant Foundation. https://rpp.wtgrantfoundation.org/wp-content/uploads/2019/09/Assessing-Research-Practice-Partnerships.pdf

Hewitt, K. K. (2015). Educator evaluation policy that incorporates EVAAS value-added measures: Undermined intentions and exacerbated inequities. *Education Policy Analysis Archives, 23*(76). https://doi.org/10.14507/epaa.v23.1968

Hirschman, A. O. (1967). *Development projects observed*. Brookings Institution Press.

Krieg, J. M., Goldhaber, D., & Theobald, R. (2020). Teacher candidate apprenticeships: Assessing the who and where of student teaching. *Journal of Teacher Education*, *71*(2), 218–232. https://doi.org/10.1177/0022487119858983

Krieg, J. M., Theobald, R., & Goldhaber, D. (2016). A foot in the door: Exploring the role of student teaching assignments in teachers' initial job placements. *Educational Evaluation and Policy Analysis*, *38*(2), 364–388. https://doi.org/10.3102/0162373716630739

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in U.S. schools. *Educational Researcher, 35*(7), 3-12. https://doi.org/10.3102/0013189X035007003

Lavigne, A. L. (2014). Exploring the intended and unintended consequences of high-stakes teacher evaluation on schools, teachers, and students. *Teachers College Record, 116*(1), 1-29. https://doi.org/10.1177/016146811411600103

Lee, J. (2004). How feasible is adequate yearly progress (AYP)? Simulations of school AYP "uniform averaging" and "safe harbor" under the No Child Left Behind Act. *Educational Evaluation and Policy Analysis, 12*(14). https://doi.org/10.14507/epaa.v12n14.2004

Leonardo, Z. (2004). The color of supremacy: Beyond the discourse of 'white privilege.' *Educational Philosophy and Theory*, *36*(2), 137–152. https://doi.org/10.1111/j.1469-5812.2004.00057.x

Lubienski, C. (2005). Public schools in marketized environments: Shifting incentives and unintended consequences of competition-based educational reforms. *American Journal of Education, 111*(4), 464-486. https://doi.org/10.1086/431180

Maier, A. & Youngs, P. (2009). Teacher preparation programs and teacher labor markets: How social capital may help explain teachers' career choices. *Journal of Teacher Education, 60*(4), 393-407. https://doi.org/10.1177/0022487109341149

Matos-Díaz, H. & García, D. (2014). Modeling college graduation GPA considering equity in admissions: Evidence from the University of Puerto Rico. *Education Policy Analysis Archives, 22*(96). http://doi.org/10.14507/epaa.v22n96.2014

Matsko, K. K., Ronfeldt, M., & Nolan, H. G. (2021). How different are they? Comparing teacher preparation offered by traditional, alternative, and residency pathways. *Journal of Teacher Education, 73*(3), 225-239. https://doi.org/10.1177/00224871211015976

Merton, R. K. (1936). The unanticipated consequences of purposive social action. *American Sociological Review*, *1*(6), 894–904. https://doi.org/10.2307/2084615

Morris, S. R., & Johnson, A. H. (2024). Could minimum grading enhance high school graduation rates and cost-effectiveness across Arkansas? *Education Policy Analysis Archives, 32*(40). https://doi.org/10.14507/epaa.32.8496

Oliver, K., Lorenc, T., & Tinkler, J. (2020). Evaluating unintended consequences: New insights into solving practical, ethical and political challenges of evaluation. *Evaluation*, *26*(1), 61–75. https://doi.org/10.1177/1356389019850847

Parvin, N., & Pollock, A. (2020). Unintended by design: On the political uses of "unintended consequences." *Engaging Science, Technology, and Society*, *6*, 320-327. https://doi.org/10.17351/ests2020.497

Redding, C. (2019). A teacher like me: A review of the effect of student–teacher racial/ethnic matching on teacher perceptions of students and student academic and behavioral outcomes. *Review of Educational Research*, *89*(4), 499-535. https://doi.org/10.3102/0034654319853545

Ronfeldt, M. (2012). Where should student teachers learn to teach? Effects of field placement school characteristics on teacher retention and effectiveness. *Educational Evaluation and Policy Analysis, 34*(1), 3-26. https://doi.org/10.3102/0162373711420865

Ronfeldt, M. (2015). Field placement schools and instructional effectiveness. *Journal of Teacher Education*, *66*(4), 304-320. https://doi.org/10.1177/0022487115592463

Ronfeldt, M., Bardelli, E., Truwit, M., Mullman, H., Schaaf, K., & Baker, J. C. (2020). Improving preservice teachers' feelings of preparedness to teach through recruitment of instructionally effective and experienced cooperating teachers: A randomized experiment. *Educational Evaluation and Policy Analysis*, *42*(4), 551–575. https://doi.org/10.3102/0162373720954183

Ronfeldt, M., Bardelli, E., Truwit, M., Schaaf, K., & Baker, J. C. (2025). Mentors Matter Recruitment replication & extension: Investigating effects across implementation years. *Journal of Research on Educational Effectiveness*, 1–24. https://doi.org/10.1080/19345747.2023.2287615

Ronfeldt, M., Brockman, S. L., & Campbell, S. L. (2018). Does cooperating teachers' instructional effectiveness improve preservice teachers' future performance? *Educational Researcher, 47*(7), 405-418. https://doi.org/10.3102/0013189X18782906

Smith, M. L., & Rottenberg, C. (1991). Unintended consequences of external testing in elementary schools. *Educational measurement: Issues and practice, 10*(4), 7-11. https://doi.org/10.1111/j.1745-3992.1991.tb00210.x

St. John, E., Goldhaber, D., Krieg, J., & Theobald, R. (2021). How the match gets made: Exploring student teacher placements across teacher education programs, districts, and schools. *Journal of Education Human Resources, 39*(3), 261-288. https://doi.org/10.3138/jehr-2020-0014

Steinberg, M. P., & Garrett, R. (2016). Classroom composition and measured teacher performance: What do teacher observation scores really measure? *Educational Evaluation and Policy Analysis*, *38*(2), 293–317. https://doi.org/10.3102/0162373715616249

Steinberg, M. P., & Sartain, L. (2020). What explains the race gap in teacher performance ratings? Evidence from Chicago public schools. *Educational Evaluation and Policy Analysis*, *43*(1), 60–82. https://doi.org/10.3102/0162373720970204

Weathers, E. S., & Sosina, V. E. (2022). Separate remains unequal: Contemporary segregation and racial disparities in school district revenue. *American Educational Research Journal, 59*(5), 905-938. https://doi.org/10.3102/00028312221079297

Weiss, C. H. (1993). Where politics and evaluation research meet. *American Journal of Evaluation*, *14*(1), 93–106. https://doi.org/10.1177/109821409301400119

Wells, A. S. (2014). *Seeing past the "colorblind" myth: Why education policymakers should address racial and ethnic inequality and support culturally diverse schools*. National Education Policy Center. http://nepc.colorado.edu/publication/seeing-past-the-colorblind-myth

Welsh, R. O., & Little, S. (2018). The school discipline dilemma: A comprehensive review of disparities and alternative approaches. *Review of Educational Research*, *88*(5), 752-794. https://doi.org/10.3102/0034654318791582

Yeh, S. S. (2005). Limiting the unintended consequences of high-stakes testing. *Education Policy Analysis Archives, 13*)43_. https://doi.org/10.14507/epaa.v13n43.2005

About the Authors

Matthew Truwit
University of Louisville
matthew.truwit@louisville.edu
https://orcid.org/0000-0003-3426-8083
Matthew Truwit is an assistant professor of evaluation in the Department of Educational Leadership, Evaluation, and Organizational Development in the College of Education and Human Development at the University of Louisville. His research is driven by a desire to understand, critically evaluate, and ultimately improve the ways in which policies both enable and constrain the teaching and learning that take place in schools.

Emanuele Bardelli
Santa Rosa City Schools
ebardelli@srcs.k12.ca.us
https://orcid.org/0000-0003-3383-9315
Emanuel Bardelli is Executive Director of Information and Evaluation for Santa Rosa City Schools. His research focuses on the development, implementation, and evaluation of local educational policies to advance equity.

Matthew Ronfeldt
University of Michigan
ronfeldt@umich.edu
https://orcid.org/0000-0003-1702-1812
Matthew Ronfeldt is a professor of educational studies at the University of Michigan School of Education. His research aims to improve teaching quality by focusing on preservice and inservice teacher education, teacher labor markets, the organizational contexts of schools, and the assessment of teaching and teacher preparation.

education policy analysis archives

Volume 33 Number 85 December 2, 2025 ISSN 1068-2341