Article

Using Mixed Methods to Explore Variations in Impact Within RCTs: The Case of Project COMPASS

Journal of Mixed Methods Research
2022, Vol. 16(4) 478–499

DOI: 10.1177/15586898211033144
journals.sagepub.com/home/mmr
SAGE

Julie A. Edmunds[1], Dora Gicheva[2], Beth Thrift[1], and Marie Hull[2]

Abstract

Randomized controlled trials (RCTs) in education are common as the design allows for an unbiased estimate of the overall impact of a program. As more RCTs are completed, researchers are also noting that an overall average impact may mask substantial variation across sites or groups of individuals. Mixed methods can provide insight and help in unpacking some of the reasons for these variations in impact. This article contributes to the field of mixed methods research by integrating mixed methods into a recently developed conceptual framework for understanding variations in impact. We model the use of this approach within the context of an RCT for online courses that found differences in impact across courses.

Keywords

randomized controlled trials, variations in impact, technology, online courses

Pushed by support from federal funding, more and more educational studies are using a randomized controlled trial (RCT) to assess program impact (Connolly et al., 2018). A RCT allows for an unbiased estimate of the impact of the program because it compares results for two groups of individuals (or clusters) created by random assignment or chance: the treatment group and the control group. In well-conducted RCTs, the difference in outcomes between these two groups can be considered as a strong estimate of the overall program impact. RCTs can provide an overall average estimate of the program's impact; however, as researchers have shown, a program or project may have impacts that can vary across settings or groups of individuals (Bitler et al., 2006; McCormick et al, 2016; Raudenbush & Bloom, 2015; Weiss et al., 2017). For example, a federally funded RCT of charter schools found that there were overall no impacts on student mathematics achievement; however, impacts varied substantially by school ranging from a statistically significant positive impact of more than .60 standard deviations to a statistically significant negative impact of approximately 0.80 standard deviations (Gleason et al.,

[1]SERVE Center at University of North Carolina at Greensboro, NC, USA
[2]University of North Carolina at Greensboro, NC, USA

Corresponding Author:
Julie A. Edmunds, SERVE Center at University of North Carolina at Greensboro, 2634 Durham-Chapel Hill Boulevard, Suite 208, Greensboro, NC 27707, USA.
Email: jedmunds@serve.org

2010). Such a variation in impacts could occur within clusters; for example, individuals within charter schools could respond differently to the intervention as a result of their background characteristics. These variations could also occur between clusters as different schools may choose to implement strategies or approaches differently. Recognizing that interventions may have variations in impact, researchers have thus moved from answering only the question about what works to understanding ''what works for whom in which circumstances?'' (Nielsen & Miraglia, 2016) and to attempting to understand the mechanisms underlying these variations in impacts.

This article builds on and contributes to recent work that has articulated how mixed methods can contribute to a deeper understanding of causality. We specifically build on a recent *JMMR* article (Johnson et al., 2019) that makes the case that quantitative research is more focused on ''general causation,'' where researchers attempt to find causal patterns across large groups. In contrast, qualitative research is focused on ''singular or local causation'' or causation that occurs in individual settings. Having both ''general and local causal understanding is important for truly understanding the phenomena under scrutiny'' (Johnson et al., 2019, p. 145) and ''mixed methods researchers should examine and combine evidence of local causation with evidence of general causation (they are complementary)'' (Johnson et al., 2019, p. 155). In particular, researchers believe that mixed methods can help contextualize experimental findings and provide evidence of both general and singular causation to understand why impacts may vary across individuals or sites (Bamberger et al., 2010).

This article contributes to the discussion by adapting and utilizing a conceptual framework for studying variations in impact that was originally developed with a primarily quantitative lens (Weiss et al., 2014). We first describe this framework and then demonstrate how this framework can be applied to explore variations in impact using both quantitative and qualitative methods. We apply this method within the context of an RCT of an intervention to redesign online courses (Project COMPASS) in a community college setting. The evaluation found that the intervention had overall (on average) statistically significant impacts on reducing withdrawals and positive, although not statistically significant, impacts on successful course completion. The study also found variations in impact in the two courses in which the program was implemented with significant positive effects in one course and null or negative effects in the other (Edmunds et al., 2021). In this article, we describe how we applied a conceptual framework to explore these variations in impact across the two courses and to explore some of the potential mechanisms that may underly the variations we do see.

We suggest that using a conceptual framework, such as the one we propose, will allow researchers to more purposefully plan for a systematic examination of variations in impact across individuals and sites. We also argue that intentionally incorporating MMR into this framework will allow for a richer and deeper comprehension of the potential sources of these variations in impact. Understanding this variation can provide useful information to practitioners when they seek to implement evidence-based interventions in multiple settings. It is also important to note that, although we apply the approach to RCTs in this article, we believe it could be equally well applied to other causal impact designs, including quasi-experimental impact studies.

Theoretical Framework

As noted above, traditional experimental studies are very good at determining an estimate of an intervention's average treatment effect. Most are not designed, however, to understand why interventions work and why there might be variations in impact that occur across settings or across individuals. Mixed methods research (MMR) is a natural way to explore that variation.

This section describes the role MMR can play in causal research designs and then explicates a conceptual framework for more systematically exploring variations in impact.

Mixed Methods and Causal Design

Health sciences and other fields are perhaps ahead of education in incorporating qualitative methods into RCTs (see, e.g., O'Cathain, 2018; O'Cathain et al., 2014). However, research that combines methods is growing in prevalence in the field of education (Gorard & Taylor, 2004). Using two case studies of educational evaluations, Hanley et al. (2016) illustrated how mixed methods can provide a more complete picture of why an intervention succeeded, which helps educators make decisions about whether and how to implement an intervention in a different context. The benefits of integrating a mixed methods approach within an RCT have also been illustrated in recent research published in this journal (see, e.g., Mannell et al., 2021; Van Scoy et al., 2020).

Our article builds on a recent publication by Johnson et al. (2019), in which they explored different types of causal research and the role of mixed methods within a causal research framework. Their article makes the argument for a mosaic view of causation that integrates different views of causality, an approach that is aligned with the integration of different methods in MMR. A core distinction made in the article is between general causation and singular/local causation.

General causation is the attempt to make causal statements that apply to a broad population, such as ''Vaccines prevent mumps'' or ''Participating in Reading Recovery results in an increase in phonics awareness.'' Experimental designs, including RCTs that are conducted exclusively with quantitative or quantitized data, are intended to identify general causation or ''causal description'' (Johnson et al., 2019; Shadish et al., 2002). These analyses are often called ''variable-oriented'' analyses (Cragun et al., 2016; Porta, 2008) because they focus on the relationship between quantitative variables.

General causation is contrasted with singular or local causation, which focuses on causality within individual instances. An example given by Johnson et al. (2019) is ''That particular iceberg caused the Titanic to sink'' (p. 145). These analyses are often called ''case-oriented analyses'' (in contrast to variable-oriented analyses; Cragun et al., 2016; Porta, 2008) and the idea is that individual instances of causation can be identified. Qualitative data, with their focus on the particular (or singular) are well suited for looking at this type of causation (Johnson et al., 2019).

These concepts of general and singular causation are a useful frame for thinking about the role of mixed methods in understanding variations in impact. As one researcher argued,

> Mixed methods can help provide detailed contextual analysis. . . . While there are useful quantitative techniques designed to deal with treatment heterogeneity, qualitative methods can also be a strong aid to understanding how the treatment may have varied across the target population. (Bamberger et al., 2010, p. 7)

Researchers have long used MMR to explore variations in individual experiences (Morrison et al., 2014; Sammons et al., 2007). Researchers have also made the case for embedding qualitative research within the structure of an RCT to develop a much richer understanding of participants' experiences or to understand the context of implementation (Plano Clark et al., 2013). Additionally, researchers have used qualitative data to improve the quality of certain aspects of the RCT such as recruitment and randomization procedures (de Salis et al., 2008; Donovan et al., 2002). We have seen less frequent use of MMR in systematically exploring the variations in impacts that are often exhibited in RCTs. This article places MMR within the context of a methodological framework for examining variations in impact.

The Four C's: A Methodological Framework for Examining Variations in Impact

We build on a recently developed conceptual framework for studying the sources of variation in program effects from Weiss et al. (2014). They argued that key sources of variation could be represented by what they called ''the three Cs'': (1) Treatment **C**ontrast, (2) Client **C**haracteristics, and (3) Program **C**ontext. In their article, they also discussed Program **C**haracteristics, but did not define it as fourth ''C.'' To guide the reader more easily in this article, we describe Program Characteristics as the fourth C and thus call the framework the Four C's. All content within the Three C's and the Four C's framework are the same; the difference is only in the title. In this section, we provide an overview of this framework, considering how mixed methods may be integrated into each of these ''C's.'' An example of how this plays out is also shown in the data sources table.

The First "C": The Treatment Contrast. The Treatment **C**ontrast, the first ''C'' reflects the difference in program services experienced between the treatment and control group and includes four dimensions: (1) content, (2) quantity, (3) quality, and (4) the delivery mode of the services (Weiss et al., 2014). The authors recommended that these different dimensions be described as much as possible in both the treatment and control groups. Both quantitative and qualitative data could be utilized to explore these dimensions.

The implementation of program services may impact outcomes directly, but most often through mediators or immediate outcomes that are hypothesized to lead to later outcomes. Weiss et al. (2014) recommended that data also be collected on mediators in both the treatment and control groups to allow for additional exploration of variation in impacts. These analyses may be quantitative in nature using path modeling techniques. Qualitative analyses might include interviews with participants in both the treatment and control groups to understand connections between the mediators and the outcomes. Open-ended questions on surveys could also provide useful information.

The Second "C": Participant Characteristics. Participant (or client) **C**haracteristics, the second ''C,'' reflect the background characteristics that individuals bring to an intervention and can be considered as moderators that might influence variations in impact (Weiss et al., 2014). Examining these characteristics quantitatively—such as subgroup analyses by gender, race/ethnicity, previous academic performance, or level of risk—can help describe which populations may benefit more from the intervention (Bloom & Michalopoulos, 2010) but they will not necessarily inform the mechanisms by which certain groups might receive higher benefits (Weiss et al., 2014). More ethnographic qualitative data that explore the experiences of specific subgroups of participants could provide insight on possible mechanisms.

The Third "C": Program Context. The third ''C'' is also thought of as a moderator and is Program **C**ontext, which reflects the broader policy and environmental context within which the program operates. Programs may function differently when they are in different contexts, such as in urban vs. rural areas or when different types of policies are in effect (Weiss et al., 2014). Researchers could collect quantitative demographic data that might help explain differences in impact; they could also collect qualitative documents that delineate policies that are in place. Researchers could also use needs assessment strategies or draw on procedures used in situation analyses (Evens & Handelman, 2006) to examine the different settings in which programs are implemented.

The Fourth "C": Program Characteristics. The three C's described above are sources of variation that are closer or more proximal to the actual outcomes. Weiss et al. (2014) noted that there are

also various Program-level Factors or **C**haracteristics that can also affect program-level implementation and thus affect the extent to which there is a contrast in the services received by the treatment and control groups. They did not call this a fourth C, but for the reader's ease, we describe it as such throughout this article, describing this framework as the Four C's. These factors include, for example, the specificity of the program intervention, the characteristics of the implementing organizations (e.g., leadership, professional capacity, school climate) and the existence of an external monitor. Other program characteristics might be related to the extent to which there are differences in cultural context and perspectives between the implementers and the intended beneficiaries of the programs; these issues might be particularly important to consider with evaluations of interventions in developing countries or in other low-income or culturally unique settings. Documenting these factors can provide insight into some of the potential causes of variation in impact.

To further explicate the Four C's, we apply this framework to a RCT of an intervention. We first briefly describe the intervention and the study design and then move to discussing how we apply the Four C's Framework.

Studying the Impact of Project COMPASS

Project COMPASS was a development project funded under the U.S. Department of Education's First in the World competition. The goals of the project were to increase the number of students, particularly students of color, completing online courses and improve the academic performance of those students, with the ultimate goal of increasing the percentage of students who remained in postsecondary education. The project sought to achieve these outcomes by redesigning the delivery of a core set of online courses so that they incorporate a variety of technologies and strategies that increase the quality of the online learning experience.

Project COMPASS was structured around the Community of Inquiry (COI), a conceptual framework for online instruction incorporating three core components of the online experience: (1) teaching presence, (2) social presence, and (3) cognitive presence (Arbaugh, 2007; Garrison et al., 2001). Project COMPASS sought to increase these various types of online presence by implementing a set of ''High Tech'' and ''High Touch'' practices. High Tech practices involved the use of a key set of technologies such as web conferencing, web messaging with automated features, video presentations, video chat, and desktop sharing. As part of the project, treatment instructors were trained in the use of these technologies and in the use of High Touch strategies that are designed to improve student–teacher interactions. There were also strategies specifically aimed at meeting the needs of minority students, such as incorporating more minority figures into the course and live streaming events showing minority role models. Figure 1 provides an overview of the project components and the expected outcomes.

The project was implemented in the fall of 2017 and the spring of 2018 in courses in two different subject areas: Course A[1] and Course B, two introductory online courses taken by many students at Wake Tech. Appendix Table A 1 shows a comparison of some features of the two courses, the instructors and the student body.

The project's evaluation evolved as it was implemented and ended up consisting of two phases. The first phase, representing the original planned scope of the evaluation, utilized an embedded design (Plano Clark et al., 2013) in which the dominant methodology was an RCT. Under the RCT, students who enrolled in the online courses were randomly assigned by the researchers to course sections taught by treatment instructors who had been trained in the intervention or control instructors who used their normal instructional practices. Treatment instructors received extensive professional development and supplemental resources while control teachers did not. The college delayed any institution-wide information sharing until the project

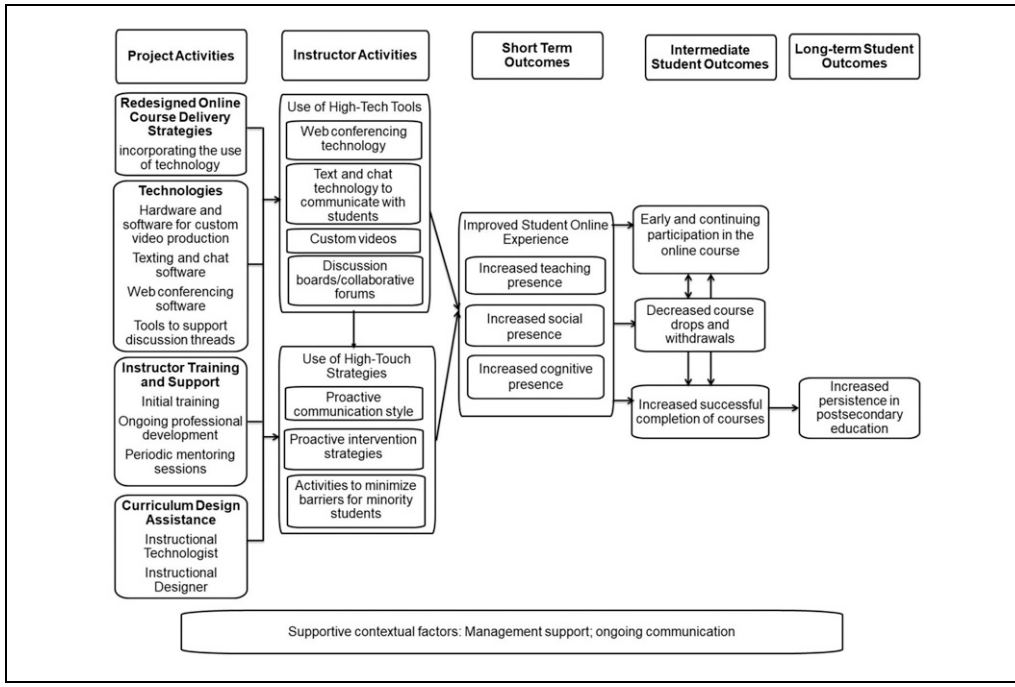

Figure 1. Project COMPASS logic model.

was over; however, it was possible that individual teachers may have interacted around the strategies and we would have had no knowledge of this. Students were not aware of the experiment and were only given the option to register for sections of the same class with the instructor to be announced. All students who registered for the class were included in the study, with the exception of a small group who were dropped administratively for nonpayment prior to the start of the semester. More details about the randomization procedure are available in Gicheva et al., 2020. The full sample described in this article included a total of 1,943 students, 1,032 students who enrolled in Course A and 911 students who enrolled in Course B, all of whom were randomly assigned to sections taught by either treatment or control instructors.

The study design was reviewed externally by an evaluation technical assistance team supported by the funding agency prior to the start of the project implementation and was assessed to have the potential to meet the study design standards of the What Works Clearinghouse (Institute of Education Sciences, 2017), a federally supported effort to identify high-quality studies in education. The study was monitored by the institutional review board of the University of North Carolina at Greensboro and students were asked to actively consent to data collection activities that involved direct interactions, including the survey. Wake Tech staff de-identified all administrative data prior to any analysis by the research team.

Additional data collected as part of this first phase of the project included quantitative and qualitative survey measures of students' participation in the program, online log-in information, and implementation interviews with the staff. These measures are described in the sections that focus on how the Four C's Framework was applied.

The intent of this first phase of the project (and what was originally planned to be the full project) was to examine the impact of the project on student outcomes, including successful completion of a course (defined as passing the course). These outcomes were examined for the

Table 1. Impact of Project COMPASS on Successful Course Completion, by Course.

	Treatment group (T)		Control group (C)		Estimated effects	
Population	Adjusted *M*	*SD*	*M*	*SD*	Adjusted mean difference	*p*
All Students, *n* (T) = 912, *n* (C) = 1,031	56.3%	0.494	53.4%	0.499	2.8%	.343
Course A, *n* (T) = 434, *n* (C) = 598	55.1%	0.500	46.5%	0.499	8.6%**	.023
Course B, *n* (T) = 478, *n* (C) = 433	56.9%	0.486	63.0%	0.483	−6.1%	.170

Note. Asterisks indicate statistically significant difference in the impact estimates between courses.
*$p < .1$. **$p < .05$. ***$p < .01$.

full population of students and for certain subgroups, including demographic subgroups and by course. Outcome data came from administrative records collected by Wake Tech. The analyses were conducted as intent-to-treat analyses (Institute of Education Sciences, 2005) using a multilevel model with students clustered by section. More detail on the intervention and the study methodology can be found in Edmunds et al., 2021.

When we analyzed the impacts by course, we found something interesting. As Table 1 shows, there was an overall descriptively positive, but nonsignificant impact on successful course completion. When the results were broken out by course, however, we saw a significant impact for Course A and a nonsignificant, but negatively trending, impact for Course B. As Table 1 also shows, the difference in impacts between the two courses was statistically significant. Understanding the reasons behind these variations in impacts became an important question for our study.

These findings of variation in impact across courses led us to revise our study design to incorporate a second phase, representing a mixed methods sequential explanatory design (Creswell et al., 2003) that addressed the Four C's Framework. Under this phase, we reanalyzed data that we had already collected so that we could look at the results by course. We also collected additional data (observations, data on instructor characteristics, and follow-up interviews). Given that this second phase of the study was not preplanned, time and resource constraints meant that we were unable to capture all of the data that we might have liked. Figure 2, which has been adapted from a diagram presented by Plano Clark et al. (2013) presents an overview of the two phases of our study design and the data sources.

Applying the Four C's Framework to Our Study

In the second phase of our evaluation, we sought to understand why the project had a positive impact in Course A and not in Course B. We explored this variation in impact using the applicable categories from the adapted Four C's Framework (Weiss et al., 2014), focusing particularly on dimensions that might have provided insight into the potential mechanisms underlying variations in impact.

In our study, we looked at the first ''C'' or the contrast between students' experiences in the treatment and control groups in the two different courses. To explore this issue, we used qualitative observation data and qualitative survey data that were then quantitized. We also focused on understanding the impact of the program on specific mediators that are expected to lead to long-term impacts, using quantitative survey data.

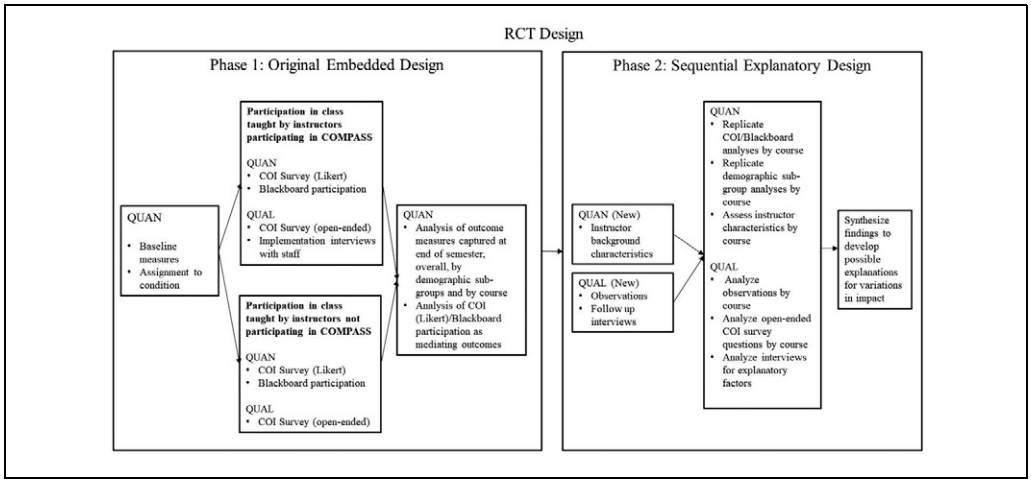

Figure 2. Two phases of project COMPASS evaluation.

We then explored the second ''C,'' or participant characteristics, looking at the differences in impacts for specific populations of students in each of the two classes. We conducted subgroup analyses using quantitative data.

Weiss et al.'s third ''C,'' or the broader program context, is important when thinking about the potential applicability of our findings to other education settings, such as 4-year institutions. The nature of the experiment we conducted did not allow us to address it in our study because both courses were taught at the same community college and during the same academic year. Nevertheless, we briefly discuss how a project could use mixed methods to explore the program context.

Finally, program characteristics (what we have called the fourth ''C'') were relevant, given that there might have been certain differences among the two courses that affected how the COMPASS intervention was implemented and thus might have affected the treatment-control contrast. We used both teacher background information and qualitative interview data to explore these ideas. Table 2 summarizes the data sources used in this article and connects them with each part of the Four C's framework. As we discuss in each of the following sections, there are places where we would recommend additional data collection and we include those recommendations in Table 2. We also include recommended data sources for exploring the third C, which was not relevant for our study but may be relevant for others.

The remainder of the article is organized by each of the Four C's: (1) Treatment-control **C**ontrast and mediators, (2) Participant **C**haracteristics, (3) Program Context, and (4) Program **C**haracteristics. In each section, we describe the specific research question, the methodology used to answer that question and our findings relative to the questions. As we mentioned earlier, our study evolved; therefore, under each ''C,'' we also identify places where we might have collected additional data if we had purposefully planned from the beginning for this study to examine variations in impact and reflect this framework.

The First "C": Treatment-Control Contrast and Mediators

Rationale and Research Questions

A key factor associated with the level of impact of the program is the extent to which the experiences students have in the treatment group are actually different than the experiences that

Table 2. Actual and Recommended Data Sources and Alignment With Four C's.

Relevant C	Data source	Purpose	Data collection	Analytic approach
N/A	Outcome administrative data from Wake Tech	Determine program impact	Wake Tech provided data on course completion and course grades for all students in the sample.	Analyses used a logit regression model examining difference between treatment and control.
First C: Treatment-Control Contrast	Structured classroom observations of each participating instructor	Examine the treatment-control contrast in implementation of targeted instructional practices	Research team member selected four recorded online sessions to review. The observer used a protocol and recorded the use and frequency of specific targeted instructional practices.	Instructional practices were coded according to levels of implementation set by the developer. These were summed to create an implementation score. Differences between treatment and control and between courses were descriptively compared.
First C: Treatment-Control Contrast	Blackboard log-ins	Determine treatment-control contrast in potential mediator	The Blackboard system provided information on number of times that students logged in.	Analyses used a linear regression model examining difference between treatment and control.
First C: Treatment-Control Contrast	Community of Inquiry Survey	Determine treatment-control contrast in potential mediator Open-ended question on contribution to student learning	Instructors asked students to complete a survey that measured implementation of the various components of the Community of Inquiry Framework.	Analyses used a linear regression model examining difference between treatment and control. Open-ended question was coded to identify themes.
Second C: Participant Characteristics	Demographic data from Wake Tech	Explore the connection between participant characteristics and outcomes	The administrative data from Wake Tech included background characteristics for students.	Outcome analyses were repeated for various subgroups of students.
Second C: Participant Characteristics	Recommended: Student interviews	Explore the connection between participant characteristics and outcomes	Recommend collecting detailed interview data from students of different backgrounds to understand how they respond to the intervention.	Interviews would be transcribed and then inductively coded to identify themes around response to the intervention.
Third C: Program Context	Recommended: Review of policy documents	Explore variations across context	Recommend collecting qualitative data around policies.	Code policies according to similarities and differences.
Third C: Program Context	Recommended: Site-level demographics	Explore variations across context	Recommend collecting demographic characteristics across sites.	Conduct subgroup analyses/use site-level factors as predictors in regression analyses.
Fourth C: Program Characteristics	Teacher baseline instructional quality	Explore the connection between program characteristics and outcomes	Wake Tech provided data on previous performance by participating teachers.	Conducted descriptive analyses by examining the baseline performance of Wake Tech instructors.
Fourth C: Program Characteristics	Teacher interviews	Explore the connection between program characteristics and outcomes	The evaluation team conducted interviews with treatment instructors across the two courses.	Interviews were recorded, transcribed, and summarized to track themes.

students have in the control group or the counterfactual (Hulleman & Cordray, 2009; Lemons et al., 2014) and then the extent to which those different experiences affect student outcomes. For example, a recent study of the variation in impact for early college high schools found that the impact was larger when the difference in quality was larger between the early college and the high school students would have otherwise attended (Miratrix et al., 2018). One possible explanation for the difference in impacts between the two courses could thus be that the treatment–control contrast was higher on practices that impact outcomes in Course A than in Course B. To test this idea, we asked two research questions:

Research Question 1: What is the impact of Project COMPASS, by course, on key mediators including: students' registration in Blackboard and their perceptions of the teaching, social, and cognitive presences in each course?
Research Question 2: What is the difference in instructional practices in the treatment and control groups for both courses?

Research Question 1

We used two different approaches to look at Research Question 1, which addresses potential mediators. As a measure of students' participation in the course, we used the number of times that a student logged in to Blackboard. We were able to use the full experimental sample of 1,943 students for these analyses and used the same methodology as the overall impact analyses.

To measure social, teaching, and cognitive presence, the evaluation team administered the Community of Inquiry Survey (Arbaugh et al., 2008) in the middle of each semester for a 2-week period. A total of 461 treatment students and 178 control students responded to the survey and provided a valid email address that allowed us to link to their administrative data. Excluding students who dropped or withdrew from the class prior to the date when the survey was administered, the overall response rates were 73% for treatment students and 28% for control students. Because the attrition rates were large and substantially different between the two groups, we followed the What Works Clearinghouse guidance (Institute of Education Sciences, 2017) and examined the baseline characteristics for the two groups. As shown in Table A 2 in the appendix, the groups were equivalent on key baseline characteristics.

The number of respondents was low overall and, when broken out by course, it was even lower. As such the course-level analyses should be considered suggestive and not conclusive. Additionally, although the treatment and control groups were comparable on observed characteristics, it is possible that there were differences in motivation between the two groups.

Table 3 presents the impact on the mediators by course. Overall, results show impacts that were generally consistent with the outcome findings. Results from the Blackboard analyses show that the treatment students in Course A had statistically significantly more log-ins as compared with control students. This is in contrast to the lack of a significant difference in log-ins between the treatment and control groups in Course B.

Looking at the other set of mediators, results from the Community of Inquiry Scales analyses show positive and statistically significant impacts for all three of the COI presences for Course A. In contrast, all three impacts were smaller for Course B and only one impact (cognitive presence) was statistically significant. Additionally, the differences in impact estimates for Teaching Presence and Social Presence were statistically significant at the 5% level.

Research Question 2

To answer Research Question 2, we conducted retrospective structured observations of each treatment and control instructor during the 2017-2018 academic year. This involved

Table 3. Impact of Project COMPASS on Mediators.

		Treatment group (T)		Control group (C)		Estimated effects		
Mediator	Variable	Adjusted *M*	*SD*	*M*	*SD*	Adjusted mean difference	Effect size	*p*
Course A								
Community of Inquiry Scales, *n* (T) = 224, N(C) = 104	Teaching presence	4.62	0.73	4.28	0.88	0.34***	0.44	.001
	Social presence	4.22	0.79	3.86	0.83	0.36***	0.45	.000
	Cognitive presence	4.31	0.76	3.92	0.85	0.39***	0.49	.000
Blackboard participation, *n* (T) = 434, *n* (C) = 598	Log-ins per student	60.45	43.91	46.78	37.78	13.67***	0.36	.000
Course B								
Community of Inquiry Scales, *n* (T) = 237, *n* (C) = 74	Teaching presence	4.42	0.68	4.39	0.79	0.03	0.04	.765
	Social presence	4.17	0.73	4.05	0.82	0.12	0.16	0.114
	Cognitive presence	4.33	0.66	4.14	0.7	0.19	0.28	.011
Blackboard participation, *n* (T) = 478, *n* (C) = 433	Log-ins per student	40.61	25.26	37.43	27.76	3.18	0.11	.355

Note. Asterisks indicate statistically significant difference in the impact estimates between courses.

*$p < .1$. **$p < .05$. ***$p < .001$.

Table 4. Average Weighted Scores of Observed Protocol Strategies.

Protocol strategies	Course A		Course B	
	Treatment	Control	Treatment	Control
Synchronous Events	2.25	0.00	.67	0.00
Announcements	2.00	1.29	2.22	1.13
Personalized Videos (Internal, Orientation, Getting Started)	1.75	0.00	1.83	0.00
Reducing Barriers	0.38	0.00	0.50	0.31
Threaded Discussions	2.25	2.63	2.00	2.25
Total Implementation Score	**1.73**	**0.78**	**1.44**	**0.74**

reviewing online course documentation for a sample of 4 weeks for one section for each instructor. We developed an observation tool that assessed the frequency of instructors' implementation of observable activities from the Project COMPASS protocol. The frequency of each behavior was noted, and each teacher was given an implementation score based on criteria set by the program staff. Scores for teachers across the two semesters were then averaged by intervention status and by course. We also added open-ended questions to the Community of Inquiry survey that captured specific activities that instructors used to develop each of the COI presences. These open-ended questions were coded to identify themes and were then summarized by course.

Results from the observations showed that there were differences in the observed instructional practices between treatment and control groups in both courses. Differences were particularly pronounced in synchronous events, texting, and personalized videos. Similar to what we might have expected based on the outcomes, the overall difference between instructional practices among the treatment and control groups was more pronounced in Course A, where the average implementation score for treatment instructors was over twice the average score for control instructors (Table 4).

Results from the analyses of open-ended questions showed that the primary difference in the treatment–control contrast and the primary difference between courses was in synchronous events. Approximately a quarter of treatment students in Course A reported that synchronous events helped the instructor develop a strong teaching presence and 18% indicated that it contributed to an increased social presence. No Course A control students mentioned this as important and no Course B students (either treatment or control) mentioned synchronous events either.

Our exploration of the treatment–control contrast so far provides generally consistent evidence about what could be causing the differences in impact. The observations indicate that there were differences in the targeted instructional practices between the treatment and control groups in both courses, with higher differences seen for Course A. Additionally, the Blackboard analyses and the Community of Inquiry analysis showed larger positive impacts for Course A, which are consistent with the developers' theory of change.

Interviews with the college faculty also noted that the Course B lead instructor, who was the head instructor for his subject area, introduced some of the COMPASS strategies into a course shell. For example, he incorporated videos that he had created as well as a core set of discussion prompts. The shell could be used in its entirety by any of the other instructors of Course B, including the control instructors. Despite this situation, many of the Project COMPASS practices were not included in the protocol.

Additional Data Collection

The data we collected above suggested patterns and frequency of the targeted instructional practices and the impact of the program on potential mediators. These data, however, did not allow us to explore whether some practices were more important than others. For example, Course A had higher implementation of synchronous events than Course B and it is possible that these strategies played the most important role in keeping students enrolled in the course. The open-ended survey questions suggested that the synchronous events were useful but this is an area where more in-depth follow-up interviews with students in each of the two courses could have explored the extent to which students identified specific instructional practices that were associated with them performing better in the course.

The Second "C": Participant Characteristics

Rationale and Research Question

A substantial body of research exists that connects students' background characteristics to their success in postsecondary education (Kao & Thompson, 2003). For example, students who are low-income or are members of certain minority groups are more likely to struggle in school (Ross et al., 2012) as are students who enter college with lower high school GPA. Therefore, it was important for us to examine whether differences in impacts between the two courses were connected in some way to students' background characteristics. We first explored whether there were significant differences in the demographic characteristics between the students in the two courses. We found statistically significant differences in gender and GPA; these results are shown in Table A 1 in the appendix. We then explored the following research question to see if students with different background characteristics responded differently to the intervention in the courses:

> **Research Question 2:** To what extent are the impacts for certain subgroups different between the two courses?

Research Question 3

To answer the question, we ran our core impact model separately for several subgroups within each course and then estimated average treatment effects for each subgroup, which are reported in Table 5. Specifically, we considered if the intervention had differential impact on minority students, low-income students, as measured by Pell Grant receipt, students with differing levels of academic performance and by gender. P values are included for an additional null hypothesis in Table 5, namely, whether the impact estimates differ across groups within the same course.

When the subgroup impacts were analyzed by course, we saw that the impacts for the more at-risk populations were higher in Course A than the impacts in Course B and also higher than the impacts for not at-risk populations in Course A, although not all of these differences were statistically significant. For example, as shown in Table 5, Course A had a statistically significant positive impact on successful course completion of 16 percentage points for minority students compared with a 2 percentage point increase for nonminority students in Course A and a 13 percentage point decrease for minority students in Course B, with the first and third of these estimates being statistically different from zero. In this case, the differences in the estimated effects between minority and nonminority students in Course A and between minority students in Course A and minority students in Course B were significant at the 5% level.

The analysis of participant characteristics suggests that the increased impacts in Course A were occurring primarily among the more at-risk populations of minority, Pell-eligible, and

Table 5. Impact Estimates by Subgroup and Course, Successful Course Completion.

Population	Course A		Course B	
	Impact estimate	*p* of impact estimate	Impact estimate	*p* of impact estimate
Minority students	15.88%***	.0003	−13.02%	.0157
Nonminority students	2.15%	.6130	−1.67%	.7218
p of difference between subgroups within course	.0117		.0137	
PELL eligible	14.41%***	.0060	−16.62%	.0006
Non-PELL eligible	4.08%	.2668	3.72%	.4566
p of difference between subgroups within course	.0534		.0004	
Incoming performance below median	12.46%***	.0076	−10.68%	.0193
Incoming performance above median	6.57%	.1067	−2.80%	.5944
p of difference between subgroups within course	.1699		.1604	
Female	8.19%**	.0035	−9.04%	.1119
Male	10.17%	.1046	−0.84%	.8738
p of difference between subgroups within course	.8418		.2027	

Note. Asterisks indicate statistically significant difference in the impact estimates between courses (*p* values not shown):
*$p < .1$. **$p < .05$. ***$p < .001$.

low-performing students. In Course B, on the other hand, outcomes for these more at-risk populations were statistically significantly worse for minority students and Pell-eligible students.

Additional Data Collection

A true shortcoming in our study is that we did not conduct interviews with students, at least partly because the online environment made it challenging. Interviews with members of populations showing higher benefits could have provided additional insights into why they may have been seeing higher impacts, particularly in Course A.

The Third "C": Program Context

The third ''C'' is the program context, which reflects contextual differences in settings that might influence the scope of the impact. As mentioned earlier, this was not relevant for our study because the two courses were implemented in the same institution during the same semester. In other cases, however, an evaluation might want to examine characteristics of the settings that might be associated with impact. For example, in cases where projects are implemented in different communities, evaluators would want to collect qualitative information about policies that are in place. As another example, evaluators may want to collect quantitative information about the demographic characteristics of the settings that might influence how a program is implemented and its impact. Additionally, researchers who are working in different cultural settings may want to collect specific information about the differences in cultural backgrounds between the program implementors and the program recipients.

Although we did not explore program context in our project, we did explore characteristics of program implementation, our fourth ''C.''

The Fourth "C": Program Characteristics

Rationale and Research Questions

The final ''C'' is program characteristics, which have also been associated with differing levels of impact. We explored two specific aspects of program implementation—the program instructors and differences in how the program was implemented in the two courses. Because this study randomly assigned students to instructors and because there were relatively few instructors (seven treatment and 14 control), we focused our efforts on understanding the quality of instructors and the extent to which that might have influenced the variations in impact between the two courses.

Our study was an Individual Randomized Group Treatment Trial (IRGT) in which students were randomly assigned to teachers but the intervention occurred at the teacher level; this is in contrast to a study where the teachers were randomly assigned to receive the intervention (Weiss, 2010). In IRGT studies, such as ours, there can be concerns that the impact of the program may have been confounded with the impact of the instructors if (1) there were variations in teacher effects (i.e., some teachers were more effective than others); and (2) the process of selecting teachers to participate in the intervention was nonrandom (Weiss, 2010).

We did not initially know whether there was variation in teacher effects, although previous research has shown that individual teachers can have widely varying impacts on student achievement (Goldhaber, 2008; Nye et al., 2004). We did know, however, that, in the Project COMPASS study, the process of selecting teachers was nonrandom for two reasons. First, the lead instructors who created the intervention and knew it very well were also two of the

Table 6. Baseline Instructor Performance.

	Course A		Course B	
	Treatment (n = 4)	Control (n = 7)	Treatment (n = 3)	Control (n = 7)
Average baseline performance	56.4%	58.11%	70.7%	58.4%
Range baseline performance	50.0% to 66.5%	55.0% to 66.0%	65.0% to 77.6%	50.4% to 70.0%
# New instructors	1	4	0	2

Note. The table reports the instructor-specific share of students passing the courses in the two semesters prior to the study period.

treatment instructors, one each in Course A and in Course B. In fact, the Course A lead instructor was the instructor who first conceptualized the intervention and proposed it for the grant. Second, the other treatment instructors had to voluntarily agree to participate so we can assume that they had some initial interest in using the intervention and/or comfort with using the practices associated with the intervention. This opened the door for the participating teachers to have been different in some way from the control instructors and for there to be potential bias in the results. To explore the role of teacher effects in the estimate of intervention impact, we asked the following research questions:

Research Question 4: To what extent did the baseline quality of teachers across the two courses differ?

Research Question 5: To what extent did teachers' descriptions of implementation differ across the two courses?

Research Question 4

To answer Research Question 4, we looked descriptively at the percentage of students successfully completing the course at baseline (in the semester prior to the start of the intervention) and during the 2017-2018 academic year, when the intervention was being implemented. We also conducted exploratory impact analyses that removed the lead instructors from the impact analyses as these individuals had more experience with the intervention and could have been considered to be more effective at implementing it.

The analyses of successful completion rates in prior semesters for instructors who taught the study courses preintervention showed that the difference in baseline instructor performance by treatment status was greater in Course B than in Course A (Table 6). This was opposite to what we might have expected, given the positive impacts in Course A. Given the baseline higher performance in Course B treatment instructors, we might have expected that Course B would have the larger impact.

We also attempted to explore the influence of the lead instructors by removing them from the analyses. The results showed a drop in both impact estimates. For Course A, the percentage point impact on completing the course successfully dropped from a statistically significant 8.6% to a nonsignificant 5.5%. Course B experienced a smaller change in impacts; the impact on the percentage of students completing the course with a C or better increased slightly from −6.1% to −5%. These findings indicate that the impact of the lead instructor was higher in Course A, suggesting that part of the increased impact might be due to the instructor. As we noted earlier, this instructor was the original developer of the intervention and was extremely enthusiastic about the model.

Research Question 5

Finally, we conducted interviews with treatment instructors across the two courses to understand more about instructor-level implementation factors that might also be accounting for differences across the two courses. Interviews with the instructors suggested that there were differences in implementation between the two courses. As noted earlier, the Course B lead instructor developed and used a course shell that included some of the COMPASS strategies and could be used by all Course B faculty, including both the treatment and control instructors. Additionally, while both the Course A and Course B instructors incorporated some synchronous events—primarily related to orientation—into their courses, the lead instructor for Course A is the only instructor who reported weekly synchronous events throughout the semester. The lead instructor for Course A also reported using the COMPASS strategies within the context of a redesigned course with a game-like atmosphere in which students completed some course activities using an online ''avatar'' to explore solutions to proposed challenges. Additionally, this redesigned course included an explicit focus on skills that students could use to be successful in both online and community college courses. Given that our earlier analyses showed a good portion of Course A's impacts being driven by the lead instructor and the fact that he had substantially different implementation than the other instructors, it is likely that a large part of the differences in impact may have been due to the gamification aspect of this redesigned course. As a result of these findings, the college began exploring the possibility of creating a course shell that would allow multiple instructors to embed this gamification approach into their instruction.

Synthesis and Conclusions

The past two decades have seen an explosion in RCTs in the field of education (Connolly et al., 2018). As the field of impact evaluations has matured, people have increasingly realized that an intervention's story is not necessarily only about whether it works on average but in understanding why it might work in some settings and not others. MMR is uniquely suited to helping researchers and evaluators understand differences in implementation and has historically been used to understand differences in individuals' experiences.

We believe that having a structured framework, such as the Four C's Framework we tested in this article, can help researchers more systematically use mixed methods to explore different potential causes of variation. If researchers plan their impact studies in advance to collect both quantitative and qualitative data around each of the ''C's,'' they can gain more in-depth understanding of the mechanisms by which the program may be working. In this article, we modelled how this might work as we explored the treatment–control contrast in both students' experiences and in mediators, students' characteristics, and program characteristics that might be associated with program implementation.

Conclusions From the RCT

The challenge with our study, as with many mixed methods studies, was synthesizing the various sources of information in a way that creates a cohesive understanding of the project. Our analyses relative to the treatment–control contrast showed us that the course with the higher impacts (Course A) also had higher impacts on mediating factors, including Community of Inquiry dimensions and Blackboard log-ins. This provided initial evidence for the downstream portion of the program's theory of change. The observations supported this conclusion with the larger treatment-control differences in Course A than Course B.

When we explored impacts by subgroup (participant characteristics), we saw that the overall higher impacts in Course A were being driven by higher impacts among minority students and students with at-risk characteristics. This indicated that there was something about the way the intervention was being implemented differently in Course A that was particularly effective for those populations, although it did not really help explain the differences in effects across the two courses. In other words, the strategies being used in Course A were more likely to reach these populations and it also suggests that there may be other activities that might differentiate the Course A treatment experience above and beyond those for which we conducted observations.

When we explored the program characteristics, we saw that the lead instructor in Course A was driving a large portion of the results. We also learned from interviews that this lead instructor had developed a gamification approach for his course that allowed students to select an online personality of different races or ethnicities. It is possible that it may have been these additional strategies, and not necessarily the COMPASS strategies, that may have been causing the increased impact for minority students. If this was in fact the case, our work in exploring variations in impact could have helped the project staff identify other strategies that might have an increased likelihood of attaining the impacts they are seeking.

It is important to note that, although we have explored some possible mechanisms for understanding the variations in impact, more than many studies do, we have not comprehensively assessed all possible mechanisms. For example, another possible explanation for variations in impact could be the alignment between the instructional strategies and the course content, which may have been stronger for one course. Nevertheless, the exploration that we have done has resulted in useful insights that the project staff can utilize as they seek to expand this work.

Contributions to the Field of Mixed Methods

In this article, we built on the points made by other researchers that MMR has tremendous potential to supplement the work of causal impact studies (de Salis et al., 2008; Donovan et al., 2002; Plano Clark et al., 2013), particularly to illuminate the mechanisms by which intervention might work and to understand the reasons that there might be variations in impact across individuals or settings (Johnson et al., 2019). A conceptual framework, such as the Four C's (Weiss et al., 2014) highlighted in this article, can provide a structured approach that will help researchers use mixed methods more systematically and purposefully to explore variations in impact.

We would hope that teams involved in causal research studies would include quantitative and qualitative researchers who would collaboratively develop a set of research questions housed within the Four C's framework. These questions should be posed in such a way that they can investigate the more particular or individual causal relationships that collectively contribute to the general causal study (Johnson et al., 2019). Sample questions might include the following:

- Treatment-Control Contrast: ''What does the treatment-control contrast look like across settings?''''How do individuals in different settings report experiencing the treatment or control conditions?''
- Client or Participant Characteristics: ''To what extent do results differ according to participants' characteristics?''''How do participants with different characteristics experience the intervention?''
- Program Context: ''How do policies and practices differ across contexts?''''What is the relationship between those policies and practices and program outcomes?''
- Program Characteristics: ''How do implementers' backgrounds differ across settings and what is the connection between those backgrounds and program implementation?''

We recognize that many studies will not be able to investigate all these questions due to resource constraints. Nevertheless, utilizing the Four C's Framework can help causal researchers be more purposeful, intentional, and proactive about understanding how an intervention works.

Appendix

Table A 1. Course Characteristics.

Characteristic	Course A (n = 1,032)	Course B (n = 911)
Student characteristics		
% Female	66.4***	53.1
% Hispanic	10.9	9.0
% Black	29.9	31.1
% White or Asian	53.2	52.8
% Identified as disabled	1.6	1.5
% Pell eligible	50.0	46.5
GPA at start of semester (if available)	2.61**	2.43
Availability of GPA, %	55.7*	51.8
Instructor characteristics		
Number of treatment instructors	4	3
Number of control instructors	7	7
Number of instructors new to the course	5	2
General course characteristics		
Number of treatment sections	16	8
Number of control sections	20	9
Academic division	Arts, humanities, and social sciences	Business and public services technologies
Course shell available for use by all instructors	No	Yes

Note. Asterisks indicate statistically significant difference in the student characteristics between courses.
*$p < .1$. **$p < .05$. ***$p < .001$.

Table A 2. Baseline Characteristics for Survey Sample.

	Overall		
Characteristic	Treatment M (n = 461)	Control M (n = 178)	Effect size (SD)
Female	61%	71%	−0.210 (0.480)
Hispanic	10%	10%	0.010 (0.305)
Black	22%	28%	−0.141 (0.493)
White or Asian	59%	57%	0.051 (0.493)
Identified as disabled	2%	0%	0.153 (0.111)
Pell eligible	46%	53%	−0.152 (0.500)
GPA at start of semester (if available)	2.71 (n = 284)	2.89 (n = 94)	−0.19 (0.873)
Availability of GPA	62%	53%	0.18 (0.492)

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the U.S. Department of Education's First in the World program, through Grant #P116F150082 to Wake Technical Community College. The opinions expressed are those of the authors and do not represent views of the U.S. Department of Education or other individuals within the SERVE Center, the University of North Carolina at Greensboro, or Wake Technical Community College.

ORCID iD

Julie A. Edmunds https://orcid.org/0000-0002-9439-8413

Notes

1. We use pseudonyms for the courses.
2. Note that we are discussing the differences here between the impact estimates for each course, not the differences between treatment and control groups (Bloom & Michalopoulos, 2010).

References

Arbaugh, J. B. (2007). An empirical verification of the Community of Inquiry framework. *Journal of Asynchronous Learning Networks*, *11*(1), 73-85. https://doi.org/10.24059/olj.v11i1.1738

Arbaugh, J. B., Cleveland-Innes, M., Diaz, S. R., Garrison, D. R., Ice, P., Richardson, J. C., & Swan, K. P. (2008). Developing a community of inquiry instrument: Testing a measure of the Community of Inquiry framework using a multi-institutional sample. *Internet and Higher Education*, *11*(3-4), 133-136. https://doi.org/10.1016/j.iheduc.2008.06.003

Bamberger, M., Rao, V., & Woolcock, M. (2010). *Using mixed methods in monitoring and evaluation: experiences from international development*. The World Bank Development Research Group, Poverty and Inequality Team http://documents.worldbank.org/curated/en/884171468156574032/pdf/WPS5245.pdf

Bitler, M. P., Gelbach, J. B., & Hoynes, H. W. (2006). What mean impacts miss: Distributional effects of welfare reform experiments. *American Economic Review*, *96*(4), 988-1012. https://doi.org/10.1257/aer.96.4.988

Bloom, H. S., & Michalopoulos, C. (2010). *When is the story in the subgroups? Strategies for interpreting and reporting intervention effects for subgroups*. MDRC. https://www.mdrc.org/publication/when-story-subgroups

Connolly, P., Keenan, C., & Urbanska, K. (2018). The trials of evidence-based practice in education: A Systematic review of randomised controlled trials in education research 1980-2016. *Educational Research*, *60*(3), 276-291. https://doi.org/10.1080/00131881.2018.1493353

Cragun, D., Pal, T., Vadaparampil, S. T., Baldwin, J., Hampel, H., & DeBate, R. D. (2016). Qualitative comparative analyses: A hybrid method for identifying factors associated with program effectiveness. *Journal of Mixed Methods Research*, *10*(3), 251-272. https://doi.org/10.1177/1558689815572023

Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advanced mixed methods research designs. In A. Tashakkori & C. Teddlie (Eds.), *Handbook of mixed methods in social and behavioral research* (pp. 209-240). Sage

de Salis, I., Tomlin, Z., Toerien, M., & Donovan, J. (2008). Using qualitative research methods to improve recruitment to randomized controlled trials: The Quartet study. *Journal of Health Services Research & Policy*, *13*(3), 92-96. https://doi.org/10.1258/jhsrp.2008.008028

Donovan, J., Mills, N., Brindle, L., Frankel, S., Smith, M., Jacoby, A., Peters, T., Frankel, S., Neal, D., & Hamdy, F. (2002). Improving design and conduct of randomised trials by embedding them in qualitative research: ProtecT (prostate testing for cancer and treatment) study. *British Medical Journal*, *325*(7367), 765-768. https://doi.org/10.1136/bmj.325.7367.766

Edmunds, J.A., Gicheva, D., Thrift, B., & Hull, M. (2021). High Tech, High Touch: The impact of an online course intervention on academic performance and persistence in higher education. *The Internet and Higher Education*, *49*, 100790. https://doi.org/10.1016/j.iheduc.2020.100790

Evens, T. M. S., & Handelman, D. (2006). *The Manchester School: Practice and ethnographic praxis in anthropology*. Bergahn Books.

Garrison, D. R., Anderson, T., & Archer, W. (2001). Critical thinking, cognitive presence and computer conferencing in distance education. *American Journal of Distance Education*, *15*(1), 7-23. https://doi.org/10.1080/08923640109527071

Gicheva, D., Edmunds, J. A., Thrift, B., Hull, M., & Bray, J. (2020). Conducting a randomized controlled trial in education: Experiences from an online postsecondary setting. *SAGE Research Methods Cases*. https://dx.doi.org/10.4135/9781529707533

Gleason, P., Clark, M., Tuttle, C. C., & Dwyer, E. (2010). *The evaluation of charter school impacts: Final report (NCEE 2010-4029)*. National Center for Educational Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. https://ies.ed.gov/ncee/pubs/20104029/pdf/20104030.pdf

Goldhaber, D. (2008). Teachers matter, but effective teacher quality policies are elusive. In H. F. Ladd & E. B. Fiske (Eds.), *Handbook of research in education finance and policy* (pp. 146-165). Routledge.

Gorard, D., & Taylor, D. (2004). *Combining methods in educational and social research*. Open University Press.

Hanley, P., Chambers, B., & Haslam, J. (2016). Reassessing RCTs as the ''gold standard'': Synergy not separatism in evaluation designs. *International Journal of Research & Method in Education*, *39*(3), 287-298. https://doi.org/10.1080/1743727X.2016.1138457

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: The role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness*, *2*(1), 88-110. https://doi.org/10.1080/19345740802539325

Institute of Education Sciences. (2005). *Key items to get right when conducting a randomized controlled trial in education*. https://files.eric.ed.gov/fulltext/ED531654.pdf

Institute of Education Sciences. (2017). *What works clearinghouse, procedures and standards handbook* (Version 4.0). https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

Johnson, R. B., Russo, F., & Schoonenboom, J. (2019). Causation in mixed methods research: The meeting of philosophy, science, and practice. *Journal of Mixed Methods Research*, *13*(2), 143-162. https://doi.dox.org/10.1177/1558689817719610

Kao, G., & Thompson, J. S. (2003). Racial and ethnic stratification in educational achievement and attainment. *Annual Review of Sociology*, *29*, 417-442. https://doi.org/10.1146/annurev.soc.29.010202.100019

Lemons, C. J., Fuchs, D., Gilbert, J. K., & Fuchs, L. S. (2014). Evidence-based practices in a changing world: Reconsidering the counterfactual in education research. *Educational Researcher*, *43*(5), 242-252. https://doi.org/10.3102/0013189X14539189

Mannell, J., Davis, K., Akter, K., Jennings, H., Morrison, J. Kuddus, A., Fottrell, E. (2021). Visual participatory analysis: A qualitative method for engaging participants in interpreting the results of randomized controlled trials of health interventions. *Journal of Mixed Methods Research*, *15*(1), 18-36.

McCormick, M. P., CapPella, E., O'Conner, E., Hill, J. L., & McClowery, S. (2016). Do effects of social-emotional learning programs vary by level of parent participation? Evidence from the randomized trials of INSIGHTS. *Journal of Research on Educational Effectiveness*, *9*(3), 364-394. https://doi.org/10.1080/19345747.2015.1105892

Miratrix, L., Furey, J., Feller, A., Grindal, T., & Page, L. C. (2018). Bounding, an accessible method for estimating principal causal effects, examined and explained. *Journal of Research on Educational Effectiveness*, *11*(1), 133-162. https://doi.org/10.1080/19345747.2017.1379576

Morrison, L. G., Hargood, C., LIn, S. X., Dennison, L., Joseph, J., Hughes, S., Michaelides, D. T., Johnston D., Johnston, M., Michie, S., P., Smith, P. W. F., Weal, M. J., & Yardley, L. (2014). Understanding usage of a hybrid website and smartphone app for weight management: A mixed methods study. *Journal of Medical Internet Research*, *16*(10), Article e201. https://doi.org/10.2196/jmir.3579

Nielsen, K., & Miraglia, M. (2016). What works for whom in which circumstances? On the need to move beyond the ''what works'' question in organizational intervention research. *Human Relations*, *70*(1), 40-62. https://doi.org/10.1177/0018726716670226

Nye, B., Konstantopolous, S., & Hedges, L. V. (2004). How large are teacher effects?*Educational Evaluation and Policy Analysis*, *26*(3), 237-257. https://doi.org/10.3102/01623737026003237

O'Cathain, A. (2018). *A practical guide to using qualitative research with randomized controlled trials*. Oxford University Press.

O'Cathain, A., Thomas, K. J., Drabble, S. J., Rudolph, A., Goode, J., & Hewison, J. (2014). Maximising the value of combining qualitative research and randomized controlled trials in health research: The QUAlitative Research in Trials (QUART) study—A mixed methods study. *Health Technology Assessment*, *18*, Article 38. https://doi.org/10.3310/hta18380

Plano Clark, V. L., Schumacher, K., West, C., Edrington, J., Dunn, L. B., Harzstark, A., Melisko, M., Rabow, M. W., Swift, P. S., & Miaskowski, C. (2013). Practices for embedding an interpretive qualitative approach within a randomized clinical trial. *Journal of Mixed Methods Research*, *7*(3), 219-242. https://doi.org/10.1177/1558689812474372

Porta, D. (2008). Comparative analysis: Case-oriented versus variable-oriented research. In D. Porta & M. Keating (Eds.), *Approaches and methodologies in the social sciences: A pluralist perspective* (pp. 198-222). Cambridge University Press.

Raudenbush, S. W., & Bloom, H. S. (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation*, *36*(4), 475-499. https://doi.org/10.1177/1098214015600515

Ross, T., Kena, G., Rathbun, A., KewalRamani, A., Zhang, J., Kristapovich, P., & Manning, E. (2012). *Higher education: Gaps in access and persistence study*. Department of Education, National Center for Education Statistics. https://nces.ed.gov/pubs2012/2012046.pdf

Sammons, P., Day, C., Kington, A., Gu, Q., Stobart, G., & Smees, R. (2007). Exploring variations in teachers' work, lives, and their effects on pupils: Key findings and implications from a longitudinal mixed-method study. *British Educational Research Journal*, *33*(5), 681-701. https://doi.org/10.1080/01411920701582264

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Company.

Van Scoy, L. J., Green, M. J., Creswell, J., Thiede, E., Wiegand, D., La, I.S., Lipnick, D., Johnson, R., Dimmock, A. E., Foy, A., & Lehman, E. (2020). Generating a new outcome variable using mixed methods in a randomized controlled trial: The Caregiver Study—An Advance Care Planning investigation. *Journal of Mixed Methods Research*. Advance online publication. https://doi.org/10.1177/1558689820970686

Weiss, M. J. (2010). The implications of teacher selection and the teacher effect in individually randomized group treatment trials. *Journal of Research on Educational Effectiveness*, *3*(4), 381-405. https://doi.org/10.1080/19345747.2010.504289

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, *33*(3), 778-808. https://doi.org/10.1002/pam.21760

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? Evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, *10*(4), 843-876. https://doi.org/10.1080/19345747.2017.1300719