CBE—Life Sciences Education
Vol. 13, 593–601, Winter 2014

Research Methods

Using Small-Scale Randomized Controlled Trials to Evaluate the Efficacy of New Curricular Materials

Dina Drits-Esser,* Kristin M. Bass,† and Louisa A. Stark*

*Genetic Science Learning Center, University of Utah, Salt Lake City, UT 84108; †Rockman et al, San Francisco, CA 94108

Submitted August 21, 2013; Revised July 14, 2014; Accepted July 29, 2014
Monitoring Editor: Nancy Pelaez

How can researchers in K–12 contexts stay true to the principles of rigorous evaluation designs within the constraints of classroom settings and limited funding? This paper explores this question by presenting a small-scale randomized controlled trial (RCT) designed to test the efficacy of curricular supplemental materials on epigenetics. The researchers asked whether the curricular materials improved students' understanding of the content more than an alternative set of activities. The field test was conducted in a diverse public high school setting with 145 students who were randomly assigned to a treatment or comparison condition. Findings indicate that students in the treatment condition scored significantly higher on the posttest than did students in the comparison group (effect size: Cohen's $d = 0.40$). The paper discusses the strengths and limitations of the RCT, the contextual factors that influenced its enactment, and recommendations for others wishing to conduct small-scale rigorous evaluations in educational settings. Our intention is for this paper to serve as a case study for university science faculty members who wish to employ scientifically rigorous evaluations in K–12 settings while limiting the scope and budget of their work.

INTRODUCTION

While the What Works Clearinghouse (WWC) has stated that randomized controlled trial (RCT) studies "can receive the highest WWC rating of *Meets WWC Group Design Standards without Reservations*" (U.S. Department of Education, 2011, p. 9), these studies can be logistically and technically difficult to implement in the classroom. RCT studies are distinguished from other study designs in that participants are randomly assigned to a treatment or control condition before the intervention to be studied begins. All other variables (including observable and unobservable characteristics) average out among treatments through the use of random distributions across the population. This design makes causal claims possible (National Research Council [NRC], 2002; U.S. Department of Education, 2011).

RCT designs are typically seen in large-scale rather than small-scale research studies in the K–12 science education literature (for a discussion of trade-offs researchers must consider in conducting large-scale RCTs, see Taylor *et al.*, 2013). One major challenge to conducting RCTs in K–12 classrooms involves the logistics required to carry them out. RCTs in educational settings must be sensitive to teacher and student effects and therefore require hierarchical designs. Ideally, these designs require at least 30 teachers and classrooms (and access to appropriate covariates) to have enough statistical power to detect the small effects typical of educational interventions (Raudenbush *et al.*, 2007). However, these sample sizes are impractical for the evaluation budgets of many projects with modest levels of funding. Developing partnerships with willing districts and schools and acquiring parental consent for students requires a substantial amount of time, which is beyond the reach of most grants.

Many researchers, therefore, seek to conduct comparison studies on a smaller scale. Small-scale comparison studies conducted in high school biology classrooms typically use matched-sample designs (e.g., Zahide *et al.*, 2001; Marbach-Ad *et al.*, 2008). Alternatively, studies are described as using random assignment but make little mention of process

DOI: 10.1187/cbe.13-08-0164
Address correspondence to: Dina Drits-Esser (Dina.Drits@utah.edu).

issues such as randomizing participants, measuring fidelity of implementation, identifying a meaningful comparison condition, selecting school sites, or obtaining human subjects approval.

In this article, we describe our process and the issues involved in conducting a small-scale RCT study in high school biology classrooms to test the efficacy of curricular materials on epigenetics. This is followed by a discussion of the critical considerations that must be made when planning and carrying out these studies. We intend for our research to be a case study for university science faculty members who wish to employ scientifically rigorous evaluations in K–12 settings while limiting the scope and budget for their work.

Background to the Evaluation: The Curricular Materials

Few high school–level curricular materials have been developed on epigenetics, which is an emerging area of importance in biology and genetics. Epigenetics is the study of changes in gene expression due to factors other than changes in a gene's DNA nucleotide sequence. The development and maintenance of an organism is orchestrated by chemical tags that bind to DNA or histones, switching parts of the genome off and on at strategic times and locations. Epigenetics researchers study these interactions and seek to understand the factors that influence them.

To bring this new science to classrooms, the Genetic Science Learning Center (GSLC) at the University of Utah developed a free curricular supplemental module on epigenetics in collaboration with high school biology teachers and scientists (GSLC, 2012a,b). The Epigenetics module was developed based on a theory of change that posits students will better understand epigenetics when curricular materials build on disciplinary core ideas that students have already learned, including the relationships between DNA, genes, and proteins. The module involves: 1) interactive multimedia learning experiences that allow students to manipulate processes and see their effects at the molecular level; 2) a hands-on model that reinforces key concepts; 3) movies that utilize creative approaches to presenting information, thereby engaging students' attention, which is the essential first step in learning; and 4) a take-home activity for students to apply what they have learned about factors that may affect their own epigenome. Our theory of change principles are consistent with the larger research base on science learning and learning through multimedia (e.g., NRC, 2000, 2011).

The Research Questions

To determine the effectiveness of the epigenetics curricular module on student learning, we formulated the following research questions:

- Do students exposed to the GSLC curricular supplemental materials (i.e., the treatment condition) improve their understanding of epigenetics more than students who are taught the same content using alternative materials (i.e., the comparison condition)?
- Does 2-wk retention of knowledge vary according to instructional condition?

These research questions were intended to determine whether the GSLC module *caused* learning in the treatment group and whether it did so to a greater degree than an alternative treatment (i.e., the comparison group). Because we asked a causal question, we determined that an experimental (or randomized) or quasi-experimental design was most appropriate. We considered the following statement from the NRC (2002) Committee on Scientific Principles for Education Research:

> Randomized field trials are an ideal method when entities being examined can be randomly assigned to groups. Experiments are especially well-suited to situations in which the causal hypothesis is relatively simple…. [Quasi-experimental studies are appropriate in] situations in which randomized field trials are not feasible or desirable. (pp. 109–110)

The GSLC partnered with Rockman et al, an independent, external research and evaluation firm, to review the research questions and conduct the study. The team determined that it was feasible and desirable to conduct an experimental study. This type of study would provide us with data that were the least likely to be subject to the vulnerabilities with confounding variables that are inherent in quasi-experimental studies, such as classroom culture effects. The research methods we used in our study are described next.

METHODS

This paper serves to inform university science faculty members who wish to conduct scientifically rigorous evaluations in K–12 settings while limiting their scope and budget. Therefore, we describe our deliberations and choices in more detail than is usually included in a methods section. Our goal is to be transparent about our thought processes and how they informed our design rationale and choices. Figure 1 provides an overview of our process of conducting an RCT study in K–12 classrooms.

Participants and School Context

Participant Sample Size. Statistical power is defined as the probability that a test will find a significant difference when this difference actually exists. Effect size quantifies the magnitude of the difference between two groups. It is often used to compare the relative effectiveness of two educational interventions. Statistical packages report many different estimates of effect size, as seen in the educational research literature. Among the most common are Cohen's *d* and partial eta-squared (for conversion tables, see Ferguson, 2009). For our study, we wanted to look for a "moderate" effect size, as defined relative to other comparable science education intervention studies (Lipsey *et al.*, 2012). Therefore, if we found a difference between the two conditions, we could claim we had found an effect that was both statistically and practically significant.

We used G*Power (Faul *et al.*, 2009) to calculate the sample size needed to detect the minimum desired effect. Having found no RCT studies of supplemental high school biology curricular materials that most closely matched our intervention, we relied on a meta-analysis of K–12 educational

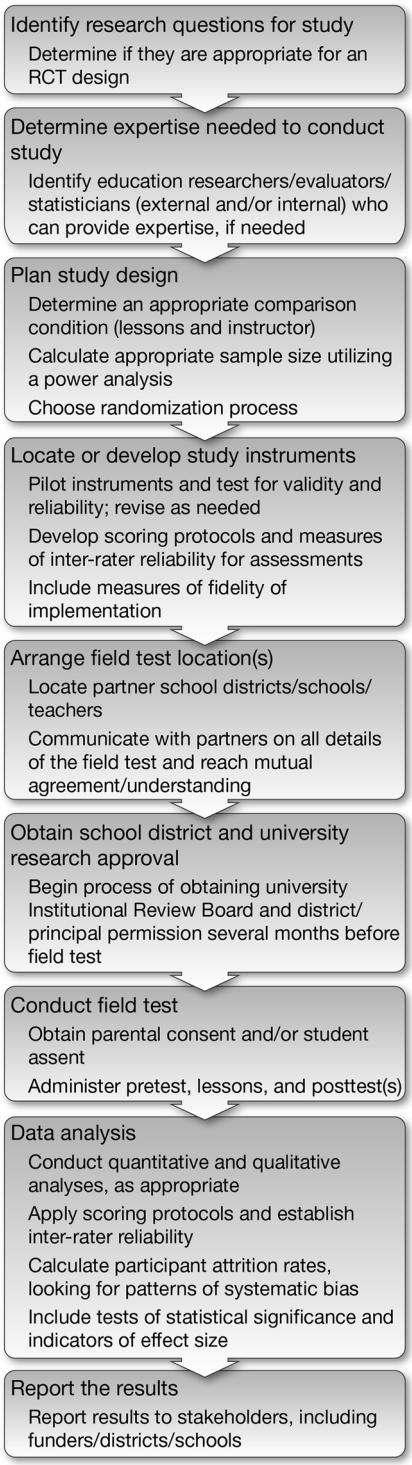

Figure 1. Process for conducting a curriculum efficacy RCT study in K–12 classrooms.

interventions to establish an effect size benchmark (Hill *et al.*, 2008). This study indicated that the average effect size for high school intervention studies using researcher-developed tests was 0.34 SDs. We decided that we should seek a minimum group difference of 0.4–0.5 SDs, as measured by Cohen's *d*, to ensure practical significance. Our power analysis subsequently determined that we needed 102–156 students to detect an effect size in our desired range.

Participant Selection. The field test was conducted with 153 students who were enrolled in a grade 9–12 biology course. Five classes from an ethnically and linguistically diverse public high school in the Intermountain West (41% Hispanic, 41% Caucasian, 8% Asian, 5% black, 4% Pacific Islander, 1% American Indian/Alaska Native, 55% free/reduced lunch, 32% English Language Learner), all taught by the same teacher, participated. We chose this school because of its diverse student population, which allows greater generalizability of the findings.

The school district's research committee determined that the epigenetics curriculum was an extension of students' existing biology curriculum and therefore parental consent was not required and all students in the test classrooms would be included in the curricular instruction. However, only students who read and approved the assent forms were required to complete the content knowledge tests and to turn in worksheets. The process of school site selection is described in the *Discussion* section.

Analytic Sample. Although we started with 153 students (75 treatment, 78 comparison), we ended with 121 students (63 treatment, 58 comparison) in the final sample after dropping students who lacked matching code numbers. We analyzed attrition data for two key variables: group assignment and baseline test scores. First, the attrition rates show that we lost 20.9% of the total students from our original sample, or 16% of the treatment students and 25.6% of the comparison students (a difference of 9.6% between the two groups). These values fell within acceptable WWC parameters for attrition in science interventions (WWC, 2012). A chi-square test also found no significant association between group assignment and attrition, $\chi^2(1) = 2.15$, $p = 0.143$. Second, a t test did not reveal significant differences in pretest scores between the students we dropped from the study (M = 1.43, SD = 0.87) and those we retained (M = 1.85, SD = 1.20), $t(151) = -1.89$, $p = 0.061$.[1] We concluded that systematic, differential attrition was generally not a concern, though the slightly lower pretest scores of the eliminated students could potentially limit the generalizability of our findings.

Instruments

Content Knowledge Test. The research team developed pre/post content knowledge tests using a systematic, iterative process of construct identification, question creation, and instrument review or validation (Wilson, 2005). The epigenetics test measured two overarching constructs: 1) the relationships between the epigenome, the genome, and the environment; and 2) the relationships between DNA, genes, and proteins. The former was the key learning objective for the unit, while the latter was included to assess whether the materials reinforced students' understanding of basic genetic principles. Each test had six multiple-choice and two

[1]Although the sample sizes are uneven, Levene's test indicated that the population variances were equal ($F(1151) = 2.24$, $p = 0.136$), thereby fulfilling one of the assumptions of a t test.

open-ended items. A full description of the development and validation process is described in another manuscript, currently in preparation (unpublished data). For detailed information on this process and for copies of the test and scoring rubrics, please contact the corresponding author.

Fidelity of Implementation. To describe the implementation of both sets of lessons and to compare the planned curriculum with the enacted one, we collected two kinds of data—observation field notes and the instructor's daily lesson reflection notes. First, the internal evaluator (D.D.-E.) observed each lesson in the field test and took field notes on the implementation for both conditions, including the content and affective aspects of the lessons. Because the field test was relatively short (2 d), it was feasible for the evaluator to observe each lesson. The field note forms were developed by the evaluator and were divided into sections that included time, scripting each activity and dialogue, evaluator reflective notes, and evaluator rating.

For the content aspects, she recorded the specific content and lesson being taught and the length of time for each activity within the lessons. For the affective aspects, she noted whether overall for that class period the instructor's level of enthusiasm was low, medium, or high. She also counted how often the instructor questioned students about what they were learning. Finally, she noted the quality of the instructor's responses to students by recording whether a response was low, medium, or high quality.

Second, the field test classroom instructor recorded lesson reflection notes at the end of each day to provide the researchers (D.D.-E. and K.M.B.) with an understanding of her experiences and any modifications made from the original lesson plans. Combined, the two data sources served as a quality check and a means of ascertaining potential threats to internal validity (e.g., imitation of treatment, compensatory equalization of treatment).

It would have been optimal to videotape these lessons for implementation fidelity scoring by an observer who was not familiar with the hypotheses being tested. However, we determined that receiving Institutional Review Board (IRB) approval would have been impossible without parent permission for videotaping, and the district research committee informed us that this would be very difficult to obtain from parents at this school.

Design and Procedure

Randomization. To maximize the rigor of our RCT design, we used a single-level, randomized block design in which students were randomly assigned to treatment and comparison conditions within classrooms. Randomization within classrooms, rather than between classrooms, enabled us to obtain a more accurate estimate of the treatment effect by controlling for classroom context. We could then pool those estimates across groups to produce a less-biased appraisal of the treatment's value (Trochim, 2006). Kirk (1995) noted that randomization within classrooms reduces variability and addresses issues of statistical power. Variability within a relatively *homogeneous* block is less than the variability of the entire sample, resulting in a more efficient estimate than without blocking. Further, if the classroom is the unit of analysis, a nested design is desirable, which requires large numbers of classrooms (e.g., 50–60 classrooms). If randomization is made by the individual, fewer individual participants are required, making an RCT design more feasible.

We recognized that randomizing within classrooms in a single school (as opposed to a hierarchical study with randomization between schools) opened the study to several threats to internal validity, namely diffusion or compensatory equalization of treatment by the instructor, or compensatory rivalry or resentful demoralization of the comparison group. Nevertheless, we believed that the control of variance afforded by a within-class block design and the logistical cost advantage of conducting a study in just one school outweighed other potential limitations. Furthermore, to monitor for the presence of internal validity threats, we observed all lessons and documented instructor behaviors and student reactions.

While there are many software programs available to guide the randomization process, these programs are actually considered pseudo–random number generators because they start with a specified set of seed numbers (Peterson, 1998). We instead chose to assign students to groups using a deck of cards that had been shuffled at least seven times to ensure a fully random distribution of values (Hedges, 2009).

Procedure. To control for teacher effects, the GSLC's senior education specialist designed and taught both the treatment and comparison lessons. It could be argued that it was also necessary to control for classroom effects even within the same teacher. Because this was a small-scale proof-of-concept study, however, the sample was not considered large enough to be sensitive to classroom-level effects.

The lessons lasted ~100 min over 2 d. Because students within a single classroom were randomly assigned into a condition, the regular classroom teacher took half of each class while the senior education specialist took the other half. The classroom teacher taught in her regular classroom, conducting a worksheet-based review of genetics; the topics covered were not directly related to epigenetics. Meanwhile, the education specialist taught either the treatment or comparison lessons in the computer laboratory across the hallway. The lessons for the treatment groups were taught first, followed by the lessons for the comparison groups.

The classroom teacher administered the pretest during class 1–2 d before the field test began, depending on class schedules. The education specialist administered the first posttest at the end of the lessons she taught. Approximately 2 wk after the first posttest, the classroom teacher administered a follow-up posttest to measure student knowledge retention.

Curricular Materials. One of the challenges in conducting an RCT is identifying an appropriate comparison curriculum. For our study, we used the only other widely available epigenetics materials appropriate for the high school level. Produced by NOVA scienceNOW (2007a,b,c,d), the curriculum included several videos designed for public audiences and accompanying lesson plans.

Our hypothesis was that the GSLC Epigenetics module would be more effective than the NOVA materials, due to several research-based design features. Interactive multimedia materials (e.g., the GSLC module) more strongly engage learners' attention—a fundamental prerequisite to learning—than multimedia materials that involve only passive viewing, such as videos (NRC, 2011). When using

interactive materials, learners can manipulate science processes and observe interactions that are too small to be seen or not available in the classroom. These materials facilitate inquiry, engaging learners in asking questions such as "What happens if I do ____?" "If I do ____ what will happen with ____?" Such questions, expressed overtly or subliminally, facilitate attention, engagement, and learning. Thus, interactive multimedia materials better support learning than passive watching (NRC, 2011). Noninteractive multimedia materials can better engage learners' attention if they include visualizations that are out of the ordinary or unexpected (e.g., the GSLC module) rather than more commonly seen visuals. Stimulating interest through unique materials can subsequently lead to greater learning (Hidi and Renninger, 2006). Furthermore, students can better learn new concepts if there are more connections to concepts they have previously learned (e.g., the GSLC module; NRC, 2000).

Both the treatment and comparison conditions began with students watching a 13-min video on epigenetics from NOVA scienceNOW (2007b). After this, the two conditions varied, although both included multimedia materials and a hands-on model. The education specialist selected materials from the GSLC's module that addressed the same topics as the NOVA materials (see Table 1 for a description of the treatment and control curricula). She did not use all of the materials from the GSLC module so that both sets of lessons were the same length and addressed the same topics.

Table 1. Summary of epigenetics curricula for the treatment and comparison conditions

Topic	Treatment condition	Comparison condition
Introduction to epigenetics	Two related lessons: 1. NOVA scienceNow: *Epigenetics*[a] (video) 2. *The Epigenome at a Glance*[b] (video)	NOVA scienceNow: *Epigenetics*[a] (video)
Environmental influence on epigenetics	Three related lessons: 1. *The Epigenetics of Identical Twins*[b] (video) 2. *Lick Your Rats*[b] (interactive animation) 3. Your Environment, Your Epigenome[d] (print-based activity)	*A Tale of Two Mice*[c] (video)
Relationship of epigenome to genome	*The Epigenome at a Glance*[b] (video)	Analogies and similes[e] (brainstorming activity, creating analogies and similes)
Gene control and regulation	Two related lessons: 1. *Gene Control*[b] (interactive animation) 2. DNA and histone model[d] (hands-on model)	Chromatin model[f] (hands-on model)

[a]Available at http://video.pbs.org/video/1525107473.
[b]Available at http://learn.genetics.utah.edu/content/epigenetics.
[c]Available at www.pbs.org/wgbh/nova/body/epigenetic-mice.html.
[d]Available at http://teach.genetics.utah.edu/content/epigenetics.
[e]Available at www.pbs.org/wgbh/nova/education/viewing/3411_02_nsn.html.
[f]Available at www.pbs.org/wgbh/nova/education/activities/pdf/3411_02_nsn.pdf.

Data Analysis

Test Data. Students could earn 12 points total on the eight-item content knowledge test; three points for each open-ended item (two items) and one point for each multiple-choice item (six items).

The research team, along with a GSLC content specialist, developed rubrics to score the two open-ended items. The rubrics were created using a top-down, bottom-up process in which we began with a priori expectations for different response levels and modified and consolidated those levels based on students' actual answers (Chi, 1997). We revised and refined the rubrics by scoring the items on a scale of 0–3. We gave three points to thorough answers that made correct and appropriate connections between concepts. We assigned one and two points to answers that fit in between these guidelines. We did not award points for answers that revealed significant misconceptions or contained all incorrect details. Using the final rubrics, the external evaluator (K.M.B.) served as the primary scorer for one of the open-ended items, and the internal evaluator (D.D.-E.) served as the reliability check; we did the reverse for the other item. For each item, the primary scorer rated all of the student responses, while the secondary scorer reviewed a random 25% of the answers. We then calculated the intraclass correlations (ICCs), an index of interrater reliability, to measure the consistency and agreement of the primary and secondary scorers' ratings. This approach was a justifiable practice, given the resources available and the low-stakes nature of this exploratory study (Stemler and Tsai, 2008). The ICCs were at least 0.80 for each item, which is considered sufficient for high reliability.

We performed analyses of covariance (ANCOVAs) to determine whether posttest scores were significantly different between the treatment and comparison groups, once the prior scores were held constant. ANCOVAs are often recommended over repeated-measures analysis of variance (ANOVA) for two reasons. First, the gain scores used in repeated-measures ANOVAs have questionable reliability. Second, ANCOVA controls for the relationship between the pretest and the outcome measures, thereby reducing error variance and increasing statistical power (Dimitrov and Rumrill, 2003; Knapp and Schafer, 2009). This is especially important in small-scale studies.

We conducted two ANCOVAs, one for each of the posttests. In the first analysis, we used the end-of-unit posttest as the dependent variable and the pretest as the covariate. Group assignment (i.e., treatment or comparison) was a fixed factor. To evaluate group differences in the long-term retention of epigenetics knowledge, we repeated the ANCOVAs for the 2-wk follow-up tests, using the second posttest as the dependent variable and the pretest as the covariate.

Implementation Data. We compared the observation notes and education specialist reflections with the lesson plans for evidence that they had been implemented with fidelity.

We also looked for evidence of program differentiation, or "whether critical features that distinguish the program from the comparison condition are present or absent during implementation" (O'Donnell, 2008, p. 34), by comparing observation notes with the planned lessons. Further, we compared notes from each condition on the education specialist's enthusiasm by comparing the frequency of low, medium, or high scores. We compared how often she questioned students across groups. Similarly, we compared the frequency of low-, medium-, or high-quality responses from the education specialist across the two conditions.

We looked for evidence of any "treatment infidelity" by assessing whether the education specialist truncated activities in one group and not in another, whether she allotted more time to either group for any defining features of the lessons, and whether she treated either group preferentially in affective terms. We recorded any differences between groups.

IRB Approval

This study was approved by the University of Utah IRB (IRB_00047518) and by the school district in which the study was conducted. The university's IRB approval process included submitting agreement letters from the district science specialist, the district's external review committee, and the school's principal. A student assent form, which students read and approved before the field test, was submitted to the IRB as part of the approval process. The final content knowledge pre- and posttests were also required by the IRB. We were not required by the school district (and thus the IRB) to consent parents or guardians, because requiring this would prevent completion of the field test; teachers in this diverse school typically receive < 20% of signed parent forms. Finally, the IRB approval process included completing an online application detailing the study, including such information as research purpose, methods, locations, researchers' background, data confidentiality issues, justification for sample sizes, and justification for conducing research on populations that are considered vulnerable (e.g., K–12 students).

RESULTS

Analysis of Knowledge Growth from Pretest to Posttests 1 and 2

On average across conditions, student test scores increased from the pretest to the first posttest by 2.49 points. We conducted a one-way ANCOVA to determine whether posttest 1 scores differed significantly by study condition when holding pretest scores constant. A preliminary analysis evaluating the homogeneity-of-slopes assumption indicated that the relationship between the covariate and the dependent variable did not differ significantly across the two levels of the independent variable, $F(1, 117) = 2.49$, mean squared error (MSE) = 8.77, $p = 0.118$. The ANCOVA for the field test indicated that there was a significant difference between the treatment and comparison conditions on the posttest after controlling for pretest scores, $F(1, 118) = 4.88$, MSE = 17.44 $p = 0.029$, Cohen's $d = 0.40$. The observed effect size is comparable with results from RCTs of high school interventions using researcher-developed tests (mean effect size = 0.34; Lipsey *et al.*, 2012) and small-scale RCTs of short-term college biology curricular supplements (e.g., Cohen's $d = 0.56$; Gutierrez, 2014).

A second ANCOVA examined whether knowledge retention on posttest 2 varied by treatment condition, again while holding pretest scores constant. As with the previous analysis, a preliminary test of the homogeneity-of-slopes assumption indicated that the relationship between the covariate and the dependent variable did not differ significantly across the two levels of the independent variable, $F(1, 117) = 0.63$, MSE = 2.64, $p = 0.431$. Proceeding with the ANCOVA, researchers found no differences in long-term change attributable to group assignment after controlling for the pretest, $F(1, 118) = 0.84$, MSE = 3.53, $p = 0.362$.

Fidelity of Implementation

Observations of the treatment and comparison lessons, combined with daily reflections from the instructor, suggested that all lessons were carried out as designed in the time allotted to them. Further, the lessons were comparable in the affective qualities of teaching, including level of enthusiasm, student questioning, and responsiveness to students. The data also indicated that the education specialist implemented all of the lessons as planned—she did not exclude or adapt any of the planned activities. These findings contributed to the internal validity of the study.

The observations and reflections also supported the assertion that the treatment groups differed in their essence. As expected, the critical difference between the treatment and comparison lessons was that students in the treatment condition used interactive, online materials to learn content that students in the comparison condition learned through videos and a brainstorming activity.

DISCUSSION

The focus of this paper is the feasibility and effectiveness of small-scale RCTs in K–12 education research, using our RCT study of epigenetics curricular materials as an example. The findings revealed that small-scale RCT studies can be conducted effectively in high school science classrooms. In this section, we describe critical considerations for researchers planning to conduct their own classroom-based RCT studies.

Deciding on a Fair Comparison for the Curricula

One of the greatest challenges we faced in our study was deciding on what constituted a fair comparison between the two conditions. We considered several options: 1) students would engage in lessons that were not related to epigenetics; 2) we would identify other curricular materials on epigenetics that were appropriate for high school students; or 3) teachers would use their usual materials on this topic ("business as usual"). The last option was not feasible, because epigenetics was too new a topic to have been widely included in high school curricula. We eliminated option 1 because we wanted to have a meaningful comparison and not just a claim that the GSLC's epigenetics materials were better than no epigenetics instruction at all. We decided that option 2 would give students equal opportunities to learn

the content, while varying the mechanisms by which the content was delivered (i.e., the presence or absence of online, scientifically accurate, interactive animations).

Having made this decision, we had to identify and select an appropriate, conceptually meaningful set of activities to serve as the comparison condition. This took more effort than implementing a more traditional business as usual program, and we had to plan accordingly. Furthermore, in order to best match the content covered in the two curricula, we had to adjust some of the GSLC materials. Although we included videos, two interactive animations, a hands-on model, and a print-based activity, we eliminated the module's Learn More topics on gene expression and cell signals, inheritance, nutrition, and the brain. We also truncated some of the GSLC lessons because we wanted to expose students to each topic for the same amount of time in both the treatment and comparison conditions. For example, this meant teaching three activities from the GSLC module in the time needed to teach one activity from the comparison curriculum.

Further, our implementation choices highlight the trade-offs between internal and external validity in narrowly controlled studies. On the one hand, our efforts to equalize content exposure eliminated a confounding factor in the study design and improved our ability to make claims about the effectiveness of the GSLC materials in one particular context. On the other hand, our curricular adaptations limit our ability to generalize this work to real-world classrooms. Had students been able to explore the GSLC materials freely, for example, they might have chosen to spend more time on them and learned the concepts in greater depth.

Deciding on the Appropriate School District to Approach for Study Participation

Having a supportive, flexible school district and school site was key to our successful implementation of the RCT. We found that it is possible to find a school that meets researchers' needs and has the resources to participate in a randomized controlled study. We utilized our strong partnership with a large local school district's science specialist and our strong relationship with several teachers in local districts. Because of these pre-established relationships, it was not challenging to locate interested teachers and obtain support from the science specialist.

For the epigenetics field test, we first decided that we wanted to conduct the study in a school with a diverse student population, for eventual generalizability of the findings. We identified teachers working in this school district with whom we had a relationship and who we thought might be amenable to working with us. After contacting a teacher and determining that she was interested, we obtained a letter of support from the district science specialist and applied to the district's external research board for approval to conduct the study. The external research approval from the district included completing forms that detailed study purposes, locations, methods, and content knowledge tests.

After approval had been received, we contacted the school's principal by mail, email, and eventually by telephone for permission to conduct the study. Receiving the principal's permission included determining that the school had the proper resources to accommodate the field test (namely, computer rooms and classroom space). The teacher we had recruited also voiced her support as part of the process of receiving principal permission. The letters of approval from the district and the principal were submitted to the university's IRB, which required these letters before giving final approval for the study. It took 4 mo to secure all of the approvals.

The lessons we learned while obtaining the approvals included the following: 1) Begin the school district and university IRB process early, as both of these take many months to clear; 2) if possible, have an internal (or local) evaluator or project manager obtain local school district and university IRB approval, manage the logistics of finding appropriate schools for field testing, and establish or utilize pre-established relationships with schools and teachers; 3) work with teachers who have buy-in and who are willing to assist with logistics such as discussing the benefits and feasibility of field testing in the school with the principal; 4) allot enough time for pilot testing and revising the assessment items before the field test is scheduled; 5) know the curricular sequence and approximate times for teaching units on each topic in the school testing site; and 6) work the field test into teachers' schedules and take into account the curriculum the students' will have been exposed to at the time of the field test.

Defining and Measuring Fidelity of Implementation as a Quality Control Check

Monitoring for and assuring fidelity of implementation is an important step in conducting studies in classroom settings, and one that is potentially overlooked in RCT studies. As such, in our research study, additional data provided by an observer on fidelity of implementation across the conditions allowed us to have greater confidence that the findings had internal validity. In other words, it increased our confidence in attributing changes to the treatment rather than to an external variable.

Securing the Necessary Evaluation Expertise from the Beginning

The partnership between the internal and external evaluators was a critical component to the study process, providing the opportunity to maximize what we could learn from the field test. The GSLC internal evaluator (D.D.-E.) knew the curricular material and had the necessary district and school connections. The external evaluator (K.M.B.) brought assessment development and research design experience and lent a more objective perspective to the study's results and implications. This collaboration required additional time and funding from the GSLC. Further, the external evaluator's resources and expertise improved the GSLC's capacity to conduct and critique research on its materials. In some cases, it is simpler to work exclusively with one evaluator, but in cases such as conducting the RCT study, it was advantageous to fund and support both evaluators. While the GSLC hired a private evaluation and consulting company for this evaluation, this is just one model of an effective partnership. If an evaluator external to the institution is not feasible, an education or other social science faculty member could conceivably play the same role.

Aligning Evaluation Expectations with Time and Budget

The project that funded the evaluation was primarily focused on developing several curricular supplemental modules over a 5-yr period. Within that time frame and budget, it was only feasible to conduct a small-scale RCT to test the effectiveness of the materials. Conducting this RCT, including reporting the results, involved ~400 h of staff time. With the proof-of-concept evidence from the current study completed, it then becomes feasible to conduct efficacy studies with a larger sample under more varied implementation conditions.

CONCLUSIONS AND IMPLICATIONS

While our study showed that conducting an RCT in a public school to test educational materials can be reasonably done, the process revealed critical steps that must be taken for successful completion and valid results. Several conclusions for researchers and evaluators can be drawn from our process and results. These are: 1) Small-scale RCTs can be conducted to evaluate student learning from curricula. This can be done in a time- and cost-effective manner. 2) RCTs take time to design and execute. This includes developing assessment items, obtaining approval from all entities involved, planning and conducting the field test. and analyzing the data. 3) It can be challenging to find appropriate comparison lessons for RCTs. 4) Small-scale designs have trade-offs. For instance, the feasibility of working with a single school is counterbalanced by internal validity threats related to the quality of teaching between conditions. This makes assessing fidelity of implementation especially important.

Our results show how one grant recipient responded to a call for rigorous evaluation and what it took to do this. With any type of rigorous design (experimental, quasi-experimental, or qualitative designs), there must be a logical consistency between research design, instrumentation and data collection, and conclusions and implications. This is necessary even in small-scale proof-of-concept evaluations, such as this one. Further, any evaluation should describe both the conceptual model underlying the intervention and provide data on how that intervention was carried out. By requiring program developers to gather data on implementation in the course of conducting rigorous research, funders and policy makers will not only establish which of their programs work but can begin to identify replicable, sustainable best practices.

ACKNOWLEDGMENTS

We thank a) the talented high school biology teachers from across the United States who participated in the Beyond the Central Dogma Summer Institute, during which they developed the learning goals and drafted ideas for the activities in the Epigenetics curriculum module; b) the Genetic Science Learning Center (GSLC) team, who further developed and produced the module; and c) Molly Malone, the GSLC's Senior Education Specialist, who planned and taught the treatment and control lessons for this study. The project described was supported by award number R25RR023288 from the National Center for Research Resources. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center for Research Resources or the National Institutes of Health.

REFERENCES

Chi M (1997). Quantifying qualitative analyses of verbal data: a practical guide. J Learn Sci *6*, 271–315.

Dimitrov DM, Rumrill P (2003). Pretest-posttest designs in rehabilitation research. WORK *20*, 159–165.

Faul F, Erdfelder E, Buchner A, Lang AG (2009). Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Meth *41*, 1149–1160.

Ferguson CJ (2009). An effect size primer: a guide for clinicians and researchers. Prof Psychol Res Pract *40*, 532–538.

Genetic Science Learning Center (2012a). Learn. Genetics: Epigenetics. http://learn.genetics.utah.edu/content/epigenetics (accessed 14 April 2014).

Genetic Science Learning Center (2012b). Teach. Genetics: Epigenetics Supplemental Materials. http://teach.genetics.utah.edu/content/epigenetics (accessed 14 April 2014).

Gutierrez AF (2014). Development and effectiveness of an educational card game as supplementary material in understanding selected topics in biology. CBE Life Sci Educ *13*, 76–82.

Hedges LV (2009). Basic experimental design. Presentation at the Institute for Education Sciences Summer Research Training Institute, held June 21 to July 3, 2009, in Nashville, TN.

Hidi S, Renninger KA (2006). The four phase model of interest development. Educ Psychol *41*, 111–127.

Hill CJ, Bloom HS, Black AR, Lipsey MW (2008). Empirical benchmarks for interpreting effect sizes in research. Child Dev Perspect *2*, 172–177.

Kirk RE (1995). Experimental Design: Procedures for the Behavioral Sciences, 3rd ed., Pacific Grove, CA: Brooks Cole.

Knapp TR, Schafer WD (2009). From gain score t to ANCOVA F (and vice versa). Pract Assess Res Eval 14(6). http://pareonline.net/getvn.asp?v=14&n=6 (accessed 20 August 2013).

Lipsey MW, Puzio K, Yun C, Hebert MA, Steinka-Fry K, Cole MW, Roberts M, Anthony KS, Busick MD (2012). Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms, NCSER 2013-3000, Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Marbach-Ad G, Rotbain Y, Stavy R (2008). Using computer animation and illustration activities to improve high school students' achievement in molecular genetics. J Res Sci Teach *45*, 273–292.

National Research Council (NRC) (2000). How People Learn: Brain, Mind, Experience, and School, expanded ed., Washington, DC: National Academies Press.

NRC (2002). Scientific Research in Education, Washington, DC: National Academies Press.

NRC (2011). Learning Science Through Computer Games and Simulations, Washington, DC: National Academies Press.

NOVA scienceNOW (2007a). A Tale of Two Mice. www.pbs.org/wgbh/nova/body/epigenetic-mice.html (accessed 14 April 2014).

NOVA scienceNOW (2007b). Epigenetics. http://video.pbs.org/video/1525107473 (accessed 14 April 2014).

NOVA scienceNOW (2007c). Epigenetics Teacher Guide. www.pbs.org/wgbh/nova/education/activities/pdf/3411_02_nsn.pdf (accessed 14 April 2014).

NOVA scienceNOW (2007d). Epigenetics Viewing Ideas. www.pbs.org/wgbh/nova/education/viewing/3411_02_nsn.html (accessed 14 April 2014).

O'Donnell CL (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. Rev Ed Res *78*, 33–84.

Peterson I (1998). The Jungles of Randomness: A Mathematical Safari, New York: Wiley.

Raudenbush SW, Martinez A, Spybrook J (2007). Strategies for improving precision in group randomized experiments. Educ Eval Policy Anal *29*, 15–29.

Stemler SM, Tsai J (2008). Best practices in interrater reliability. In: Best Practices in Quantitative Methods, ed. J Osbourne, Thousand Oaks, CA: Sage.

Taylor J, Kowalski S, Wilson C, Getty S, Carlson J (2013). Conducting causal effects studies in science education: considering methodological trade-offs in the context of policies affecting research in schools. J Res Sci Teach *50*, 1127–1141.

Trochim W (2006). Randomized Block Designs. www.socialresearchmethods.net/kb/expblock.php (accessed 20 August 2013).

U.S. Department of Education (2011). What Works Clearinghouse: Procedures and Standards Handbook, version 2.1. www.whatworks.ed.gov (accessed 20 August 2013).

What Works Clearinghouse (2012). WWC Evidence Review Protocol for Science Interventions, version 2.0. www.whatworks.ed.gov (accessed 20 August 2013).

Wilson MR (2005). Constructing Measures: An Item Response Modeling Approach, Hillsdale, NJ: Erlbaum.

Zahide Y, Ozden M, Aksu M (2001). Comparison of hypermedia learning and traditional instructing on knowledge acquisition and retention. J Educ Res *94*, 207–214.