Chad Hershock, CARNEGIE MELLON UNIVERSITY, hershock@cmu.edu
Laura Ochs Pottmeyer, CARNEGIE MELLON UNIVERSITY, lpottmey@andrew.cmu.edu
Jacqueline Pincus, CARNEGIE MELLON UNIVERSITY, jjpincus@cmu.edu
H. Elisabeth Ellington, CARNEGIE MELLON UNIVERSITY, hellingt@andrew.cmu.edu
Zach Mineroff, CARNEGIE MELLON UNIVERSITY, zminerof@andrew.cmu.edu

Unpacking SoTL "Failure": Lessons from a Fellowship Program Investigating Generative AI's Impacts

ABSTRACT

The experience of failure in designing, implementing, interpreting, and/or disseminating scholarship of teaching and learning (SoTL) research is not new to educational developers; it is, however, a relatively new focus for extended inquiry and exploration. The word failure may evoke complicated emotions or assumptions, but normalizing and learning from SoTL failure helps educational developers and instructors refine and elevate their practice. This paper contributes to the growing literature on SoTL failure by exploring project-level SoTL failure in the context of a fellowship program implemented by a center for teaching and learning (CTL). This fellowship is designed to support novice SoTL researchers from all disciplines as they evaluate the causal impact of technology-enhanced learning (TEL) interventions, specifically generative AI use, on student learning outcomes. Defined within the context of this fellowship, a failed project is one that did not collect at least one direct measure of student learning, did not analyze (or could not effectively interpret) the data regarding implications for future course design and delivery, or did not disseminate findings. We identify multiple sites for potential SoTL project failure within the processes of study design, implementation, and dissemination as well as consider how TEL interventions present further potential complications. Our goal is to share lessons learned through projects that failed or nearly failed and the preventative and mitigating strategies and scaffolding that we developed to address the most common causes of failure.

KEYWORDS

failure, educational development, generative AI use in the classroom, technology-enhanced learning, center for teaching and learning

INTRODUCTION

For instructors who are new to SoTL and the educational developers[1] who support them, failure can be a complicated, stigmatized concept. Chick, Cruz, Friberg, and Steiner (2023, 1) argue, "Ultimately, if we want to better understand 'unsuccessful' practices and other kinds of failure, we need to interrogate what we mean by 'success,' change scholarly conventions and institutional cultures, and embrace the value of failure with humility, curiosity, and grace." Their provocative essay reviews, categorizes, and invites deeper dialogue on the many ways "failure" manifests within SoTL. Their examples explore how data may reveal the extent to which students are failing to achieve learning objectives, why instructors may fail to engage or persist in practicing SoTL, and why

CC-BY-NC License 4.0 This is an Open Access article distributed under the terms of the Creative Commons – Attribution License 4.0 International which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly attributed.

individual projects may fail to be initiated, completed, and/or disseminated. Our essay continues to unpack this last category: SoTL failure at the project-level.

Specifically, in the context of fellowship programs designed to support novice SoTL researchers (e.g., Colby, Cruz, Cordaro, and Cruz 2022; Wright, Finelli, Meizlish, and Bergom 2011), we share observations regarding: (1) the ways in which projects can unexpectedly fail to implement high fidelity teaching interventions, yield interpretable data, and/or disseminate findings; (2) how technology-enhanced learning (TEL) interventions create additional places for such failures to occur; and (3) preventative strategies that fellowship program managers can adopt. We are not aware of any previous literature describing failures common within SoTL fellowship programs (SFPs) or regarding TEL, especially when the program's explicit intention is to evaluate causal impacts of teaching interventions on student outcomes.

Below, we introduce the context for our findings and define project-level success and failure within SFPs. Next, we report our findings regarding how SoTL fellowship and TEL projects can fail in these contexts. For each finding, we describe how we refined our SFP to minimize the frequency of such failures. For brevity and to report trends across projects, we report our findings and conclusions in the aggregate. We hope these hard-won lessons will help both SoTL practitioners and educational developers as well as inspire others to examine their own perceptions of "failure."

In what context did we observe SoTL failure in fellowship and TEL projects?

We are educational developers at a large center for teaching and learning (CTL) at a mid-sized, research-intensive university in the midwestern United States. Regardless of discipline, our goal is to cultivate evidence-based teaching by lowering the barriers to collect direct measures of student outcomes as a standard practice in order to inform course design and delivery. Our modus operandi is a systems-thinking educational development approach (Lovett, Hershock, and Brooks 2023). We leverage an ecosystem of complementary programs and services to motivate and educate instructor-scholars, facilitate SoTL project implementation, and disseminate results and implications for teaching and learning.

SFPs are an important part of our CTL's toolkit. At their best, SFPs foster diverse teaching innovations, provide an entry point to both SoTL and CTLs for instructors from all disciplines, provide actionable data to instructors, disseminate transferable lessons learned, and enhance teaching and learning in other courses (Colby et al. 2022; Wright et al. 2011; Wright, Lohe, Pinder-Grover, and Ortquist-Ahrens 2018). Typically, we select cohorts of three to ten applicants and offer fellows financial incentives, approximately $5,000 per course selected. Across 12–18 months, our CTL staff also provide substantial, in-kind effort to help fellows scope, design, implement, analyze, and disseminate projects. Additionally, CTL staff organize monthly meetings for fellows that facilitate dialogue as a SoTL community of practice.

In response to generative artificial intelligence's (GenAI) disruption to higher education (Bowen and Watson 2024; Magrill and Magrill 2024; Mollick 2024), we created a SoTL fellowship program as part of our campus-wide Generative Artificial Intelligence Teaching as Research (GAITAR) Initiative. The GAITAR Fellowship Program approaches this SoTL challenge as a set of testable, empirical questions (Eberly Center 2023). Does students' use of GenAI increase, decrease, or not change student performance on learning outcomes? For whom (i.e., are impacts equitable)? And, why? To date, three cohorts of GAITAR Fellows representing seven schools and colleges have implemented projects across 32 courses. Reporting on our 32 GAITAR Fellowship projects' results is beyond the scope of this essay. However, preliminary project results are reported elsewhere (Eberly Center 2025a).

While our SFP focused exclusively on GenAI's impacts, our lessons learned are generalizable to SoTL projects that investigate the impacts of TEL, and most also apply to studies without a TEL focus. Thus, except where indicated by use of the word "TEL" or "GenAI" in the heading, these findings are relevant to SoTL projects in general. The development process of many TEL tools creates additional avenues for the potential for failure. Furthermore, GenAI tools (a subset of TEL) have unique characteristics (e.g., more non-deterministic outputs, rapid updates/versions, often less user familiarity, and more open-ended functionality than most other TEL) that present added challenges to research studies.

What defines success and failure in SoTL fellowship programs?

Broadly speaking, we consider our SFPs successful when participating instructors have a positive professional development experience, feel supported as they innovate their teaching, are exposed to evidence-based teaching strategies, start to develop a toolkit and self-efficacy for SoTL, practice data-driven course design and delivery, and engage in an inclusive community of practice. Strictly speaking, we view individual fellowship projects as a "SoTL success" when they collect at least one direct measure of student performance or learning in a course, analyze and interpret the data regarding implications for future course design and delivery, and disseminate findings. Disseminating findings may take many forms, including publication in a peer reviewed journal, a conference presentation, poster session, etc. We acknowledge that this definition of SoTL success includes at least one direct measure of student learning or performance. Frequently, we also collect and analyze student perceptions and/or experience data, which is highly informative; however, student perception/experience data does not necessarily follow the same patterns as direct measures of their learning outcomes (Kruger and Dunning 1999). Given the time, money, and human resources we invest, and because we cannot fund every proposal, when a project does not achieve our minimum criteria for "SoTL success," it can feel like a "SoTL failure."

SFPs like the GAITAR Fellowship Program are explicitly designed to foster studies of the impacts of teaching interventions on student outcomes via tests of null hypotheses. Null hypotheses predict that student outcomes do NOT differ among a study's comparison groups. When data analyses suggest otherwise, we reject the null hypothesis, inferring that whatever differs among comparison groups caused the observed differences in student outcomes. We refer to these projects as "what works," "which is better," or "what causes" (hereafter WBC), depending on what pedagogical conditions are compared within or between study subjects (see Pincus and Hershock 2024). The WBC approach is not necessarily the only or best way to conduct SoTL (Chick 2014), but if one wishes to infer causality, an intervention and appropriate comparison groups are prerequisites. Therefore, WBC studies have been a conspicuous area of our practice, more so after launching the GAITAR Initiative. After analyzing data from a WBC project, if it remains difficult to draw causal inferences from a null hypothesis test, then feelings of SoTL failure or low return on investment may be particularly acute.

Please note that we did not say, "when a hypothesis test fails to refute a null hypothesis." Our discussion of SoTL failure is independent of whether or not the data collected via quantitative, qualitative, or mixed methods refutes or supports a null hypothesis. Null results from excellent SoTL study designs are informative and valuable.

Instead, we are highlighting project-level SoTL failure arising from the challenges of interpreting any result (including null findings) because we failed to negotiate a more rigorous WBC study design (as defined in Pincus et al. 2024) with an instructor and/or an instructor's implementation fidelity is unexpectedly lacking. In either case, failure to confidently interpret and

disseminate the data's implications for teaching and learning may occur because alternative, highly plausible, causal explanations surface.

Below, we describe the specific ways in which WBC projects on the impacts of GenAI could unexpectedly fail to achieve our criteria for minimum SoTL success. Our observations are based on the subset of GAITAR Fellowship projects that either narrowly avoided or resulted in SoTL failure. For the latter group, we sometimes could convince instructors to tweak and repeat their studies, but mostly they chose otherwise. Consequently, we made substantial refinements to our request for proposals, selection criteria, and consulting service model to cultivate more SoTL successes. Below, we share our lessons learned.

FINDINGS

The headings below represent lessons learned. For each, the first paragraph explains the challenges we observed and how they threatened SoTL success. Subsequent paragraphs describe how we adjusted our SFP to minimize the probability of failure. Once again, SOTL failure refers to the inability to draw meaningful conclusions from a project due to one or more aspects of implementation (e.g. lack of control group, implementation fidelity, etc.). A methodologically sound project with null results is incredibly valuable and offers meaningful conclusions, for all of the aforementioned reasons; if we are confident in those results, we would define that project as a SOTL success.

Proposals are an early indicator of (un)successful projects

Proposal writers benefit from SoTL scaffolding

Our CTL supports instructors of all ranks, postdocs, graduate students, and staff. Any instructor from any discipline may apply to our SFPs, which means that while they are experts in their field, their experience with educational research and WBC studies varies. Therefore, proposals differ in: (1) how well they articulate a research question and/or study design, and (2) how viable their articulated research question and study design are for classroom-based research. However, because we do not expect our instructors to be SoTL experts, we have accepted proposals historically that have study design "potential," i.e., our CTL colleagues can envision which data sources could be leveraged and how a comparison group can be formed. While this potential is sometimes realized with a willing instructor, we have run into resistance in other cases when suggesting the use of a comparison group or tweaking a particular assignment to be a more solid data source.

We have since provided multiple scaffolds to prepare applicants to consider and propose viable study designs for their projects. Our RFP explicitly requires applicants to include a comparison group, and we attached a resource that describes comparison group options. This one change alone has dramatically improved the quality of proposals. Additionally, instructors have many ways to engage with us to learn more about SoTL and study designs. First, we encourage applicants to consult with the CTL through "office hours" that they can drop into or via a one-on-one consultation. We also regularly offer a four-day institute to introduce instructors to SoTL and classroom-based research methods.

Provisional interviews mitigate risk-taking in the selection process

Despite increased scaffolding and support, another area of ambiguity is the level of detail provided by applicants. Although we include a suggested word limit per question for the sake of transparency and equity, we encounter applicants who go well beyond the limit and those who are severely under. Regardless, applicants may omit key information about their research question, teaching intervention, comparison groups, or data sources. In the past, we often took risks by funding

such proposals, especially if (as above) we can see the potential, but their SoTL success/failure has been variable.

Beyond increasing the scaffolding for our RFPs, we refined our list of selection criteria to more explicitly raise potential barriers to SoTL success and weighted certain criteria over others. The review process includes individual scoring of de-identified proposals as well as a full group conversation to reconcile different reviewer scores, account for info gaps in proposals, and identify potential problems. We have become more stringent in rejecting proposals with lower scores on prioritized criteria. Additionally, we added a step in the review process: holding provisional interviews for applicants with information gaps or problems. These interviews enable us to gather missing information, seek clarity from the applicant, and gauge their flexibility and willingness to engage. Consequently, we can make more informed decisions about a proposal's "potential" for SoTL success.

TEL interventions create their own SoTL implementation challenges

Limit SFP TEL projects to fully developed and tested tools

We often receive SFP applications from instructors proposing to build a new TEL tool and then test drive it in their course. Compared to using an off-the-shelf tool, building and user testing new TEL tools before deployment in a SoTL study is non-trivial. Whether the tool is created with GenAI or not, project timelines usually necessitate tool development in parallel with SoTL study and course (re)design. However, until fellows complete tool development and user testing, they routinely struggle to focus on anything else. Consequently, if tool development is unsuccessful, challenging, or slow, an instructor's divided attention often results in hasty, underdeveloped study designs and pedagogical implementation plans. Through two cohorts of GAITAR Fellows, we observed this outcome in some projects developing a home-grown GenAI TEL tool. For example, in one such project, the study design changed from using four sections of a course to a single section, using a pilot version of the tool because tool development took significantly longer than anticipated.

Given our observed high-risk of SoTL failure, we altered the SFP's RFP and selection criteria. For a project to be eligible, a fellow must develop the proposed GenAI tool fully and user test it by the proposal submission deadline. Because prospective fellows might view our SFP primarily as an opportunity to develop their idea for a GenAI tool, rather than conduct SoTL, we redirect them to a separate seed grant program at our institution designed specifically for TEL tool development. Once an instructor's tool is fully developed, they may reapply for our SFP.

Off-the-shelf TEL tools, including GenAI tools, may require repeated pilot testing

With GenAI tools, like Copilot and ChatGPT, one cannot assume that particular prompts will generate appropriate or reliable outputs (Eberly Center 2025b), and students will prompt GenAI tools in unexpected ways. Therefore, one should not assume that a student's interaction with a GenAI tool will unfold as an instructor intends. Additionally, the interfaces and underlying LLMs of off-the-shelf GenAI tools may evolve rapidly, causing tool performance and user experiences to change significantly, sometimes week to week. Even when tools perform as expected, poor instructions may unintentionally confuse students or lead to errors. All of the above can result in SoTL failure regarding implementation fidelity and/or because outcomes data become more difficult to interpret regarding hypothesis testing. The latter two scenarios are also relevant to projects investigating impacts of TEL tools in general.

If a SoTL intervention involves student-facing TEL, we recommend that instructors thoroughly test instructions and user interactions with the tool prior to implementation. Furthermore, if an instructor is using an off-the-shelf GenAI tool, they should be prepared to encounter changes in the

user experiences or tool performance before, during, or after their study (especially across semesters). Single pilot tests may not be sufficient. Instructors must test usability and performance frequently and in close proximity to study implementation, adjusting instructions for students and study designs accordingly. During dissemination, we advise scholars to report the specific versions of the TEL tool used (e.g., Microsoft Copilot web chat with enterprise account access in March 2024), so that others may extrapolate results and their implications accordingly, given the constantly shifting landscape of TEL in higher education.

Some data sources tend towards intractability for null hypothesis testing

Measuring individuals' learning requires individual deliverables

Many course syllabi have learning objectives framed as, "After completing this course, students should be able to . . . [insert student-centered, measurable learning objective here]." Implicitly, these statements assume each individual student. Likewise, many SoTL projects are motivated by hypotheses about individual students' learning. However, frequently, the only extant course assessment aligned with a research question or learning objective is a team project with group, rather than individual deliverables. As a data source, group deliverables demonstrate what at least one team member can do, not what each team member can do. If the research question or objective focuses on what groups can achieve, a group deliverable is fine, assuming a sufficient sample to draw a causal inference. Otherwise, instructors need to consider adding or substituting an individual deliverable to effectively test a hypothesis about individual student outcomes. Otherwise, SoTL failure is likely regarding the interpretation of a hypothesis test. Unfortunately, instructors often resist this type of pivot because they have prioritized team work or do not wish to increase student and/or instructor workload.

Consequently, our RFP now explicitly asks about which course assessments could be used as a data source to test proposed research questions. This allows us to be more selective on student assessment type when funding SFP proposals. Applicants must identify each assessment as an individual or team deliverable. When necessary, we conduct provisional interviews with applicants to explore opportunities for incorporating individual assessments. If instructors are unwilling to adjust, we respect their choice but do not fund the proposal.

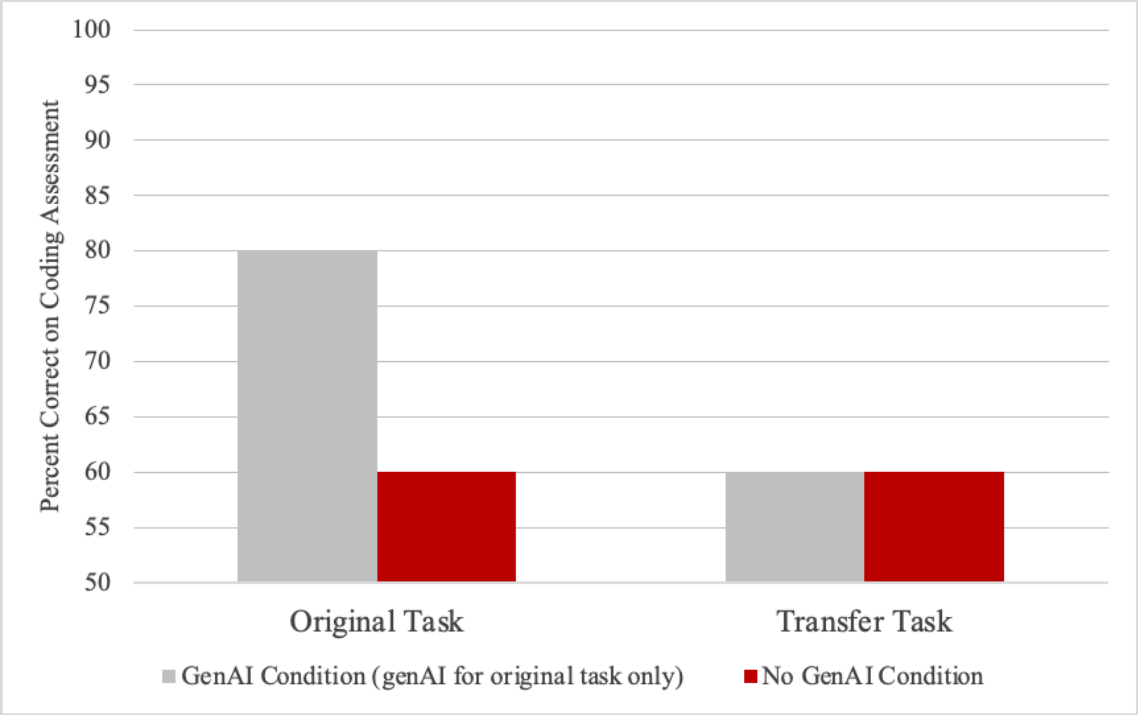
GenAI can confound performance and impact

Imagine a between-subjects study comparing two conditions: students authoring computer code with versus without access to GenAI. In the former condition, students are using GenAI to create part or all of the deliverable used as the study's data source. What if, on average, the deliverables from the GenAI condition are significantly higher quality? Do the students in the GenAI condition have stronger coding skills? Did the GenAI outperform the humans at coding? Or, some of both? When GenAI is both the intervention and used to generate the data source, researchers cannot interpret observed differences (or lack thereof) in performance as impacts of GenAI on student learning (or lack thereof). Some early studies on the impacts of GenAI suffer from this confound (e.g., Noy and Zhang 2023; Sun, Boudouaia, Zhu, and Li 2024), as did several projects in our SFP.

To isolate the impacts of GenAI on learning, we recommend that all study subjects complete a transfer task, i.e., an isomorphic version of the original task completed independent of the study conditions after a delay (e.g., Bastani, Bastanai, Sungu, Ge, Kabakcı, and Mariman 2024; Fan et al. 2024; Pankiewicz and Baker 2023; Pardos and Bhandari 2024). For the scenario above, after a delay, imagine all students are required to author computer code to solve a similar, new coding problem requiring the same knowledge and skills, but without access to GenAI. By comparing performance on

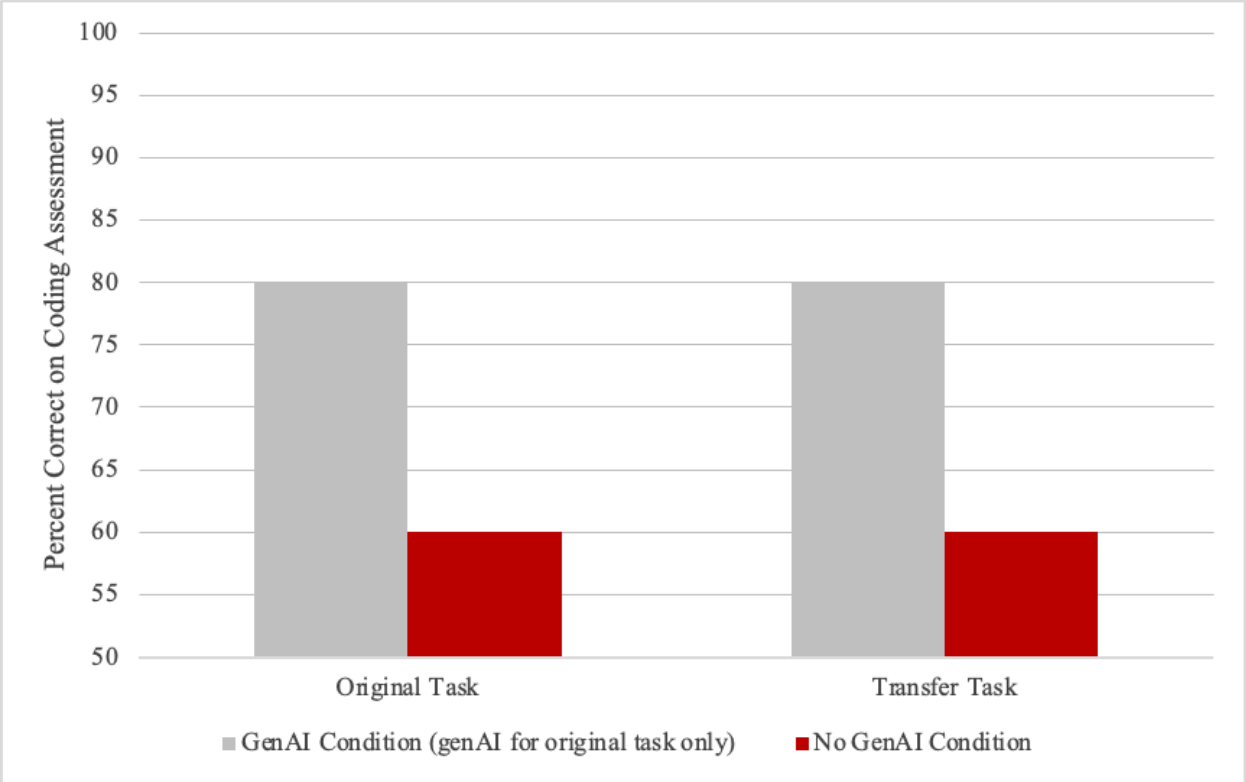
the transfer task across conditions and to performance on the previous task(s), we can better isolate the impacts of GenAI (or lack thereof) on student learning. We have expanded the scenario above and included hypothetical figures (for illustrative purposes only) to further convey the importance of a transfer task. For this scenario, what if students in the GenAI condition systematically performed lower on the transfer task than on the original task, and their transfer task performance did not differ from students' performance on the original task in the condition without GenAI? This result would suggest that GenAI is better at coding than the sample of students in the non-GenAI condition, not that students' coding skills improved due to GenAI access (Figure 1a). However, if student performance did not differ in the GenAI condition between the original and transfer task, this result would suggest that access to GenAI improved students' coding skills when not using the tool (Figure 1b). Regardless, without the transfer task, one would not be able to effectively interpret why performance on the original task differed between conditions.

Figure 1a. Hypothetical results of student performance on two coding assignments



Notes: Hypothetical results of student performance on two coding assignments (original task and transfer task) for students in two conditions. In both conditions, students completed the transfer task without GenAI. Results suggest that it is the tool rather than student skill or knowledge that is driving the larger original task score for the GenAI condition.

Figure 1b. Alternative hypothetical results on two coding assignments



Notes: Hypothetical results of student performance on two coding assignments (original task and transfer task) for students in two conditions. In both conditions, students completed the transfer task without GenAI. Results suggest that the higher student performance seen in the original task with the GenAI tool persists even when students are not using GenAI indicating that the tool has impacted student learning.

GenAI transcripts can be challenging data sources

Many GAITAR Fellows express interest in analyzing students' transcripts from interventions leveraging GenAI. However, sometimes this data source does not align with testing a hypothesis. For instance, projects comparing conditions with and without GenAI lack transcripts or comparable process documentation for the non-GenAI condition. Thus, there is nothing to compare between conditions to test a WBC[2] hypothesis. Other times, students' GenAI transcripts might be interesting data sources to compare across conditions implementing GenAI in different ways. However, SoTL failures can unexpectedly occur because instructors don't know how to conduct the required qualitative analyses or, like our CTL staff, lack the time and resources to complete them. We have not yet encountered a project in which the GenAI transcripts on their own are sufficient to test a null hypothesis about the impacts of GenAI on student learning.

When instructors show enthusiasm for GenAI transcripts as a data source, we proactively discuss the above issues to manage expectations and prevent SoTL failure. How will that work directly test a null hypothesis? How will the analysis directly support the interpretation of findings from another data source used to test a null hypothesis? Who will analyze the transcripts, when, and by what method? If the answers are not compelling, we encourage instructors to focus on other data sources.

WBC projects must have a viable comparison group

Without a viable comparison group, SoTL researchers cannot test a WBC null hypothesis effectively. We encountered a variety of surprising ways that SFPs might fail related to comparison groups.

Student self-selection into treatment conditions can be problematic

While this finding is applicable to non-GenAI studies, we found this particularly problematic in our GenAI-focused projects. For example, in several of the funded projects, instructors who did not want to impose GenAI use on their students offered them the choice, with the intention of comparing student outcomes between those who opted in versus out. Not only does this design introduce selection bias that would confound interpretation of the data (e.g., compared to random assignment of students), but also every time in our SFP, by happenstance, it resulted in severely unbalanced sample sizes across conditions. In some courses, almost everyone opted to use GenAI. In others, almost nobody did. In either case, SoTL failure resulted.

Consequently, we no longer select SFP proposals with comparison groups formed by student self-selection. Instead, prior to acceptance, if all other aspects of the proposal are promising, we conduct provisional interviews with instructors to explore their willingness to adopt another method of assigning students to conditions.

Beware of order effects in within-subjects designs

Within-subjects designs are potentially powerful strategies for testing SoTL null hypotheses, especially in smaller courses (e.g., Melville, Hershock, and Reeja-Jayan 2024). With these studies, each student experiences all conditions and thus serves as their own "control" (Pincus et al. 2024). Within-subjects designs in which all students experience the same condition synchronously and cycle through conditions are particularly appealing to SFP instructors because they minimize disruptions to course delivery. However, such designs are potentially confounded by order effects, especially when instructors have a small number of comparable assignments to leverage as data sources (e.g., the instructor assigned using GenAI to the first major assignment and not using GenAI to the second major assignment). Which condition did students experience first? Is the performance in the second condition truly independent of the first or impacted by it? If performance is greater in the second condition, is it due to the difference in conditions or because students improved from practice and feedback on the first assignment?

We recommend the SFPs implementing this type of within-subjects design ensure students experience multiple cycles of conditions (i.e., courses need more than two assignments as data sources). Within each cycle of conditions, assignments should be comparable in difficulty. Across cycles, instructors should randomize or alternate systematically the order of conditions. If courses are limited by having a number of assignments equal to the number of conditions, then we recommend considering collecting a second semester of data in which the order of conditions is altered. Or, within a semester, consider a within-subjects design in which all conditions are occurring for each assignment, and equal numbers of students are randomly assigned to each condition. Depending on the conditions, the latter option may be difficult to implement logistically because students may realize they are having a different experience than their peers. When students in the same section experience different conditions, we work closely with the instructor on student messaging and equity.

New courses focused on the intervention may lack comparison groups

We observed edge cases in which comparison groups for a within- or between-subjects design could not be identified in a single semester, precluding a WBC study. While this finding is applicable to non-GenAI studies, the following example references GenAI tools. Imagine an instructor creating a new course about using GenAI to compose music and generate artistic visualizations of the score. To fill a gap in the curriculum, they aim to provide hands-on creative opportunities for students to explore disciplinary applications for a variety of GenAI tools. SFP managers might see tremendous potential for WBC studies in this course. However, if the instructor is unwilling to have students do coursework without GenAI or to limit which GenAI tools may be used in which ways, viable WBC comparison groups would be difficult to identify. Because the instructor taught this course for the first time, a historical comparison group would not be possible either.

We no longer accept proposals that do not clearly identify comparison groups. When an applicant has not identified a historical or contemporaneous comparison group, we conduct provisional interviews to explore instructor flexibility. If consensus cannot be reached, we do not fund the proposal but will still offer in-kind staff support to the instructor for collecting "what happens" data, i.e., a SoTL study without a null hypothesis test. Additionally, we explore what we call the "SoTL long view." Sometimes, data from a "what happens" study inspires an alteration of course design or delivery. If the instructor decides to change something in the next iteration of the course, we offer to measure the same "what happens" outcomes again. By comparing the first and second iteration of the course, we now have a WBC study with comparison groups from different semesters (for affordances and limitations of such designs, see Pincus et al. 2024).

Threats to implementation fidelity happen

In our SFPs, we work closely with instructors to maximize implementation fidelity of comparison groups and interventions, checking for consensus and understanding. Despite the best of intentions, mistakes and/or miscommunications may occur regarding WBC study implementation. Sometimes instructors implement (or attempt to) in unexpected ways that compromise the fidelity of the original study design, risking SoTL failure. For example, an instructor might accidentally implement one condition for all students in a course, rather than randomly assign it to half of the students. No comparison group exists in these scenarios, and the WBC null hypothesis cannot be tested. Alternatively, an instructor might modify the teaching intervention itself upon instinct or because they forget a detail, introducing a confounding factor. We don't pass judgment, as we make mistakes and have misunderstandings in our own research projects too. The instructor can still collect valuable "what happens" pre/post data, but an opportunity to test the intended hypothesis is missed.

Other than trying again in a subsequent semester, we don't have a guaranteed solution for this type of SoTL failure. During consultations, we emphasize the importance of implementation fidelity for interpreting results before the instructor implements the study design. We also include a program milestone during which instructors sign off on a written summary of the study design as part of institutional review board compliance. Additionally, we attempt to schedule regular one-on-one check-in meetings, increasing the probability of surfacing a miscommunication or avoiding implementation infidelity. Debrief meetings scheduled after the semester may surface any implementation fidelity issues, which is particularly useful if the person analyzing the majority of the data is not the same person teaching the course.

Null results are NOT failures, but may impede dissemination

Many causes of SoTL failure described above result in uninterpretable results. Null results are not uninterpretable or uninformative by default. Nevertheless, some of our SoTL fellows confess to viewing null results as failures. In fact, believing that only statistically significant results warrant publication, they may consign conclusive findings to the file drawer. Rather than proceed to the paper drafting stage, many fellows return to their data for further scrutiny, hoping that significant differences will somehow come to light if they just keep digging. Faculty investment in a statistically significant result (and the hope of large effect sizes) is understandable; instructors not only want students to benefit from their SoTL intervention, they have invested considerable time and effort in study design, data collection, data analysis, and often in writing as well. If the SoTL intervention does not improve student learning, instructors may well wonder what value further dissemination of the research would possibly have to a wider audience. Additionally, fellows who developed a TEL tool specifically for this SoTL intervention find it particularly hard to accept a null result, because they sought validation of their tool or innovation through their research.

Because the effects of GenAI on student learning are still emerging, it is especially important for fellows to understand that null results are worthy of dissemination; they are possible, valid, and useful to other instructors. Consultation teams now prepare fellows for the possibility and potential meanings of null results as one outcome of educational research early in the process of study design. We further normalize null results and address publication bias by sharing examples of peer-reviewed published studies of null results that may serve as models for data analysis and interpretation.

Additionally, the center incentivizes the dissemination of null results through the stipend structure of the fellowship. Historically, we paid half of SFP awards up front, in case instructors required funds to implement their project, and half upon project completion, defined as receiving a brief, informal report on outcomes. Recently, we've experimented with dispersing awards incrementally upon completion of key project milestones with clear criteria, including sharing data with their CTL colleagues, disseminating at a campus event, or submitting a manuscript to a peer-reviewed outlet. Early compliance patterns are promising, especially for projects experiencing bottlenecks. Even if the instructor chooses not to pursue publication, the incentive to share data with CTL colleagues ensures that preliminary findings are made public on our website (Eberly Center 2025b).

We have found that mentorship and writing support help bridge the gap between data analysis and dissemination. As such, fellows receive a robust template for writing SoTL articles in the Intro-Methods-Results-Discussion (IMRD) format. We also curate a repository of different types of SoTL articles, including IMRD, for fellows. Weekly writing accountability groups provide an opportunity for fellows to discuss their results and receive feedback as they shape their findings into a compelling research story (e.g., Marquis, Healey, and Vine 2014; Proffitt, Boone, Janes, Hall, Lemoins, and Dunn 2023; Skarupski and Foucher 2018). Writing accountability groups also address other known barriers to dissemination, including low motivation, high workload, and low self-efficacy from unfamiliarity with the norms of educational research (e.g., Healey, Matthews, and Cook-Sather 2019; Moore, Felten, and Strickland 2013). Fellows receive support in analyzing and interpreting data from a CTL data science research associate, who also provides feedback on their manuscripts. Finally, fellows participate in a panel discussion Q and A with SoTL journal editors who are invited to discuss the value of null results and answer questions related to writing and publishing SoTL.

Perceived power dynamics may prevent negotiation of more rigorous SoTL designs

Who makes the decisions and whose voice has more influence on SFP projects, particularly as it relates to study design? Is it the instructor-scholars or the educational developers who support them? In terms of institutional ranking, CTL staff are frequently "outranked" by the faculty instructors they support. This may lead to the perception that CTL staff must defer to the instructors regarding implementation decisions, potentially increasing the likelihood of the SoTL failures outlined above.

Perhaps unsurprisingly, we have found that projects (and morale) are stronger when we leverage the strengths of all individuals involved, and CTL staff and the instructor take on equal co-creation roles. Power dynamics and decision making are now explicitly discussed as CTL staff professional development. In these professional development sessions, CTL staff practice navigating these situations in consulting scenario case studies and role plays. CTL staff are also explicitly empowered (and supported) by their supervisors to gently push back and/or advocate for study design changes when necessary.

CONCLUSIONS

In our SFPs, we define SoTL failure in WBC studies as: (1) not collecting data adequate for null hypothesis testing; (2) inability to interpret implications of data for teaching and learning; and/or (3) the absence of dissemination of findings. Who knew that such projects could (nearly) fail in so many ways? Because we are experienced at SoTL and provide significant support to SFP participants, one might assume that SoTL failure should be preventable. Vexingly, failures still occur in the predictable and unpredictable ways described above.

The good news is that educational developers can learn a lot from SoTL failure and near failure and parlay these lessons toward reducing its probability in WBC-focused SFPs. To preemptively minimize potential SoTL failure, we identified relatively small changes to the design and implementation of our SFP model regarding transparency, setting expectations, accountability, and scaffolding. Time will tell if and how these changes impact the rates of SoTL successes and (near) failures, though our earliest experiences are encouraging.

We also learned not to lower our standards. Our SFP goals continue to include fostering pedagogical innovation and providing entry points to SoTL, but historically, we accepted projects with conspicuous SoTL shortcomings outright on the optimistic assumption that we could negotiate a more rigorous pathway forward. We no longer do that; funding more projects is not necessarily better for SoTL success. This point will hopefully resonate with CTLs who have fewer or different levels of resources at their disposal. Our first two cohorts included 16 and 10 courses, and although we are well-resourced, this project load stretched our CTL too thin; both the educational developers and the projects suffered. For our third cohort, we accepted five fellows whose projects passed our more stringent criteria and provisional interviews. We continue to offer support to the authors of declined proposals, hoping that they will reapply with stronger proposals.

Many of our findings and lessons learned are applicable at the application selection or study design stages (e.g., importance of a control group, not using GenAI transcripts as data sources, etc.), and thus they may be transferable to CTLs of varying sizes and resource levels. Reducing the number of funded projects and/or the level of support on various aspects of the research cycle are a few ways in which CTLs can adapt these strategies to align with their resources. For example, if a CTL can fund a certain number of projects but does not have the capacity to take on the data analysis, they can still be stringent about the RFP requirements and what qualifies for funding. The RFP can also state that

the data analysis must be done by the applicant (or another unit that may be able to offer support). Or perhaps the CTL can fund two to three projects, including data analysis, but cannot support the dissemination directly.

Above all, we learned the value of normalizing and learning from SoTL failure for instructors and the educational developers supporting SFPs. Despite the best of intentions and efforts, SFP projects sometimes fail. At our CTL, we frequently discuss the concepts of productive failure and iterative refinement, even when programs are successful by most or all metrics. It is during our frequent CTL staff meetings, in which we discussed each of the projects in-depth and summarized the project findings, that we identified these "lessons learned." This manuscript is a product of those numerous discussions among educational developers leveraging knowledge from the education research literature rather than a systematic research evaluation of the program. Additionally, those meetings set the expectation that we will always be looking for ways to improve as a standard procedure. By cultivating a judgement-free process for discussing challenges and mitigating strategies, our CTL leveraged failure in order to refine our SFP program and elevate our SoTL practice.

ACKNOWLEDGMENTS

The SoTL in Process article by Nancy Chick and colleagues (2023) both motivated this contribution and provided a framework to share our experiences regarding failure. We appreciate the leadership of Carnegie Mellon University's provost Jim Garrett, who funded the GAITAR Initiative. We also thank the instructor-scholars who collaborated with and inspired Eberly Center colleagues. Finally, we are grateful to our Eberly Center colleagues who supported and challenged us (and fellows) as we designed, implemented, and iteratively refined the GAITAR Fellowship program, sometimes on short and stressful timelines. Members of the Eberly Center's staff writing accountability group provided feedback that improved early drafts of this paper.

NOTES

1. Defined by the POD Network as professionals who use learning science and theories to help educators improve their teaching (2025). Additionally, we include student learning, inclusion, and equity as a focus of the work of educational developers.
2. Refers to "what works," "which is better," or "what causes" studies

Chad Hershock, CARNEGIE MELLON UNIVERSITY, hershock@cmu.edu
Laura Ochs Pottmeyer, CARNEGIE MELLON UNIVERSITY, lpottmey@andrew.cmu.edu
Jacqueline Pincus, CARNEGIE MELLON UNIVERSITY, jjpincus@cmu.edu
H. Elisabeth Ellington, CARNEGIE MELLON UNIVERSITY, hellingt@andrew.cmu.edu
Zach Mineroff, CARNEGIE MELLON UNIVERSITY, zminerof@andrew.cmu.edu

AUTHOR BIOGRAPHIES

Chad Hershock (United States) is the executive director of the Eberly Center for Teaching Excellence and Educational Innovation at Carnegie Mellon University.

Laura Ochs Pottmeyer (United States) is the senior data science research associate at the Eberly Center for Teaching Excellence and Educational Innovation at Carnegie Mellon University.

Jacqueline Pincus (United States) is the associate director of graduate student and postdoc teaching initiatives at the Eberly Center for Teaching Excellence and Educational Innovation at Carnegie Mellon University.

H. Elisabeth Ellington (United States) is the senior teaching consultant at the Eberly Center for Teaching Excellence and Educational Innovation at Carnegie Mellon University.

Zach Mineroff (United States) is the assistant director of learning engineering at the Eberly Center for Teaching Excellence and Educational Innovation at Carnegie Mellon University.

DISCLOSURES

We did not use generative AI at any stage of the writing and preparation of this article.

ETHICS

The Institutional Review Board at Carnegie Mellon University approved this research.

REFERENCES

Bastani, Hamsa, Osbert Bastani, Alp Sungu, Haosin Ge, Özge Kabakcı, and Rei Mariman. 2024. "Generative AI Can Harm Learning." *The Wharton School Research Paper*. https://doi.org/10.2139/ssrn.4895486.

Bowen, Jose A. and Charles E. Watson. 2024. *Teaching with AI: A Practical Guide to a New Era of Human Learning*. Johns Hopkins University Press and American Association of Colleges and Universities.

Chick, Nancy L. 2014. "Methodologically Sound' Under the 'Big Tent': An Ongoing Conversation." *International Journal for the Scholarship of Teaching and Learning* 8 (2): 1. https://doi.org/10.20429/ijsotl.2014.080201.

Chick, Nancy L., Laura Cruz, Jennifer C. Friberg, and Hillary H. Steiner. 2023. "Making Space for Failure in the Scholarship of Teaching and Learning: A Blueprint." *Teaching & Learning Inquiry* 11. https://doi.org/10.20343/teachlearninqu.11.36.

Colby, Susan A., Laura Cruz, Danielle Cordaro, and Clare Cruz. 2022. "Fellow Travelers: Taking Stock of Faculty Fellows Programs in the Age of Organizational Development." *To Improve the Academy: A Journal of Educational Development* 41 (2): 2. https://doi.org/10.3998/tia.844.

Eberly Center. 2023. "Generative AI Teaching as Research (GAITAR) Initiative." Effective October 3. https://www.cmu.edu/teaching/gaitar/index.html.

Eberly Center. 2025a. "GAITAR Fellows Project Descriptions." Effective December 19. https://www.cmu.edu/teaching/gaitar/project-descriptions/index.html.

Eberly Center. 2025b. "Generative AI Tools FAQ." Effective May 19. https://www.cmu.edu/teaching/technology/aitools/index.html.

Fan, Yizhou, Luzhen Tang, Huixiao Le, Kejie Shen, Shufang Tan, Yueying Zhao, Yuan Shen, Xinyu Li, and Dragan Gašević. 2025. "Beware of Metacognitive Laziness: Effects of Generative Artificial Intelligence on Learning Motivation, Processes, and Performance." *British Journal of Educational Technology* 56 (2): 489–530. https://doi.org/10.1111/bjet.13544.

Healey, Mick, Kelly E. Matthews, and Alison Cook-Sather. 2019. "Writing Scholarship of Teaching and Learning Articles for Peer-Reviewed Journals." *Teaching & Learning Inquiry* 7 (2). https://doi.org/10.20343/teachlearninqu.7.2.3.

Kruger, J. and D. Dunning. 1999. "Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments." *Journal of Personality and Social Psychology* 77 (6): 1121.

Lovett, Marsha, Chad Hershock, and Judy Brooks. 2023. "Cultivating and Sustaining a Faculty Culture of Data-Informed Teaching: How Centers for Teaching and Learning Can Help." In *In Their Own Words: What Scholars and Teachers Want You to Know About Why and How to Apply the Science of Learning in your Academic Setting*, edited by Catherine. E. Overson, Christopher M. Hakala, Lauren L. Kordonowy, and Victor A. Benassi. Society for the Teaching of Psychology.

Magrill, Jamie, and Barry Magrill. 2024. "Preparing Educators and Students at Higher Education Institutions for an AI-Driven World." *Teaching & Learning Inquiry* 12: 1–9. https://doi.org/10.20343/teachlearninqu.12.16.

Marquis, Elizabeth, Mick Healey, and Michelle Vine. 2014. "Building Capacity for the Scholarship of Teaching and Learning (SoTL) Using International Collaborative Writing Groups." *International Journal for the Scholarship of Teaching and Learning* 8 (1): 12. https://doi.org/10.20429/ijsotl.2014.080112.

Melville, Michael C., Chad Hershock, and B. Reeja-Jayan. 2024. "Virtual Engineering in Minecraft: Helping Students Visualize and Manipulate the Properties of Materials." *Advances in Engineering Education* 13 (1): 11–26. https://doi.org//10.18260/3-1-1153-36059.

Mollick, Ethan. 2024. *Co-Intelligence: Living and Working with AI*. Penguin.

Moore, Jessie L., Peter Felten, and Michael Strickland. 2013. "Supporting a Culture of Writing: Faculty Writing Residencies as a WAC Initiative." In *Working with Faculty Writers,* edited by Ann E. Geller and Michele. A. Eodice. Utah State University Press.

Noy, Shakked, and Whitney Zhang. 2023. "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence." *Science* 381 (6654): 187–92. https://doi.org/10.1126/science.adh258.

Pankiewicz, Maciej, and Ryan S. Baker. 2023. "Large Language Models (GPT) for automating feedback on programming assignments." *International Conference on Computers in Education*. https://doi.org/10.58459/icce.2023.950.

Pardos, Zachary A., and Shreya Bhandari. 2024. "ChatGPT-Generated Help Produces Learning Gains Equivalent to Human Tutor-Authored Help on Mathematics Skills." *Plos One* 19 (5). https://doi.org/10.1371/journal.pone.0304013.

Pincus, Jacqueline and Chad Hershock. 2024. "A Narrative Review of Interdisciplinary Teaching and Learning Scholarship: Closing the Gap Between Theory and Practice. *International Journal for Academic Development* 30 (4): 535–52. https://doi.org/10.1080/1360144X.2024.2389446.

POD Network. 2025. *What is Educational Development.* https://podnetwork.org/about/what-is-educational-development/.

Proffitt, Rachel, Anna E. Boone, William E. Janes, Jamie B. Hall, Samantha Shea Lemoins, and Winnie Dunn. 2023. "Supporting Faculty Scholarship through a Peer Writing Group: A Model and Guide for Success." *International Journal for Academic Development* 30 (2): 201–15. https://doi.org/10.1080/1360144X.2023.2293864.

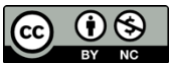
Skarupski, Kimberly A., and Kharma C. Foucher. 2018. "Writing Accountability Groups (WAGs): A Tool to Help Junior Faculty Members Build Sustainable Writing Habits." *Journal of Faculty Development* 32 (3): 47–54.

Sun, Dan, Azzeddine Boudouaia, Chengcong Zhu, and Yan Li. 2024. "Would ChatGPT-Facilitated Programming Mode Impact College Students' Programming Behaviors, Performances, and

Perceptions? An Empirical Study." *International Journal of Educational Technology in Higher Education* 21 (14). https://doi.org/10.1186/s41239-024-00446-5.

Wright, Mary C., Cynthia J. Finelli, Deborah Meizlish, and Inger Bergom. 2011. "Facilitating the Scholarship of Teaching and Learning at a Research University." *Change: The Magazine of Higher Learning* 43 (2): 50–6. https://doi.org/10.1080/00091383.2011.550255.

Wright, Mary C., Deborah R. Lohe, Tershia Pinder-Grover, and Leslie Ortquist-Ahrens. 2018. "The Four Rs: Guiding CTLs with Responsiveness, Relationships, Resources, and Research." *To Improve the Academy* 37 (2): 271–86. https://doi.org/10.1002/tia2.20084.



Copyright for the content of articles published in *Teaching & Learning Inquiry* resides with the authors, and copyright for the publication layout resides with the journal. These copyright holders have agreed that this article should be available on open access under a Creative Commons Attribution License 4.0 International (https://creativecommons.org/licenses/by-nc/4.0/). The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited, and to cite *Teaching & Learning Inquiry* as the original place of publication. Readers are free to share these materials—as long as appropriate credit is given, a link to the license is provided, and any changes are indicated.