The Role of Context in Educational RCT Findings: A Call to Redefine "Evidence-Based Practice"

Avi Kaplan[1], Jennifer Cromley[2], Tony Perez[3], Ting Dai[4], Kyle Mara[5], and Michael Balsai[1]

In this commentary, we complement other constructive critiques of educational randomized control trials (RCTs) by calling attention to the commonly ignored role of context in causal mechanisms undergirding educational phenomena. We argue that evidence for the central role of context in causal mechanisms challenges the assumption that RCT findings can be uncritically generalized across settings. Anchoring our argument with an example from our own multistudy RCT project, we argue that the scientific pursuit of causal explanation should involve the rich description of contextualized causal effects. We further call for incorporating the *evidence* of the integral role of context in causal mechanisms into the meaning of "evidence-based practice," with the implication that effective implementation of practice in a new setting must involve context-oriented, evidence-focused, design-based research that attends to the emergent, complex, and dynamic nature of educational contexts.

Keywords: classroom research; experimental research; learning environments; research utilization

In a critical meta-analysis of findings from 141 rigorous educational randomized control trials (RCTs), Lortie-Forgues and Inglis (2019) found only negligible effect sizes, wide confidence intervals, and low Bayes factors. The authors determined, "Such trials allow us to conclude neither that an intervention should be implemented at scale nor that this should be avoided to prevent the waste of public money" (p. 164). They proposed several possible reasons for this lamentable state of affairs in RCT educational research: that these RCTs were based on unreliable theory and research, were poorly designed or implemented, lacked power to detect actual effects, or some or all of the above. Here, we join Lortie-Forgues and Inglis in a constructive critique of educational RCTs. However, rather than poor theory or flawed design, we call attention to the long recognized yet commonly ignored role of context in educational phenomena. We argue that evidence for the central role of context in the causal mechanisms that give rise to educational phenomena challenges the prevalent epistemological assumption that RCT findings reflect, primarily, mechanisms that can be uncritically generalized across settings. Consequently, we call for incorporating the *evidence of the integral role of context in causal mechanisms* into the interpretation and scientific pursuit of "evidence-based practice." We anchor our argument with an example from our own multi-study RCT project.

For the past 5 years, we have engaged in an Institute of Educational Sciences (IES)–funded RCT intervention project that aimed to promote undergraduate students' academic success in introductory biology courses (Cromley et al., 2019). In the project, we tested the effects of different combinations of cognitive and motivational supports that were administered through the courses' online management systems to students who were randomly assigned to different conditions. In several iterative replications, we compared groups of students who accessed different combinations of cognitive and motivational supports with each other and with groups of students who accessed only cognitive, only motivational, or no supports at all. Altogether, we conducted 10 experiments, over four academic years, at three different institutions, in first- or second-semester biology courses for STEM (science, technology, engineering, and mathematics) majors, with a cumulative sample of 3,092 participants.

A meta-analysis of the intervention's effects on course grades across 50 comparisons found a statistically significant overall

[1]Temple University, Philadelphia, PA
[2]University of Illinois at Urbana-Champaign, Champaign, IL
[3]Old Dominion University, Norfolk, VA
[4]University of Illinois at Chicago, Chicago, IL
[5]University of Southern Indiana, Evansville, IN

Educational Researcher, Vol. 49 No. 4, pp. 285–288
DOI: 10.3102/0013189X20921862

effect of $g = .30$—a moderate and significant effect, with practical implications for some students. However, the effect sizes varied broadly across the different comparisons. Of the 50 effects, 41 were significant and positive, 3 were nonsignificant, and 6 were significant and negative. Moreover, the 41 positive effects also varied greatly (gs = .20 to .66). Univariate moderation tests by specific supports, fidelity of implementation, university, academic year, semester, students' biology background, course content, and timing of the study in our iterative development process were all significant. This suggested the crucial role of context in the way our almost identical intervention was "received" in the different settings, and in the same setting at different times.

We conducted post-hoc observations and documents analysis in an attempt to understand these contextual differences. These suggested that although our 10 contexts shared features such as student age, lecture-based instruction, introductory biology content, and exams as the main basis for grades, they also differed on quite substantive characteristics. For example, contexts differed in the size and reputation of the universities; the specific nature of reading materials and assignments; and, perhaps most important to our project, the motivational climate of the courses—some courses had a reputation for weeding out students while others did not, syllabi differed in emphasis on student support, and instructors had different reputations as teachers. Alas, our post-hoc data could not fully explain the effects variability.

Intertwined with these structural contextual differences, administering the intervention included unanticipated events that may have had impact in particular cases. For example, in one experiment, students "crammed" on the intervention; rather than accessing the supports in coordination with the course content, most students in this course accessed the entire set of supports during the last 2 weeks of the semester. Another example of an unanticipated event involved a changed schedule due to snow days that affected the timing of release of supports relative to the course content.

In summary, our project employed the "gold standard" method in testing for replication and for accumulating evidence of the benefit of our particular intervention. The findings from the meta-analysis suggest that, at the aggregate—across studies, contexts, and individuals within settings—students' access of and engagement in combinations of particular cognitive and motivational supports were beneficial to their grades relative to no access or to access of only cognitive or only motivational supports. However, our findings also suggest that these aggregated findings mask tremendous variability. The moderator analysis and unanticipated events suggest that each experiment in our multiexperiment study constituted a distinctive case with unique contextual features that framed the intervention's causal effects. This highlighted to us that our significant and positive aggregate finding does not provide justification to expect a similar effect in any other setting—even one that may seem very similar to the original study's in context and student characteristics.

Findings from other multisite experimental projects corroborate the crucial role of contextual and student individual differences in RCTs. For example, Borman et al. (2018) tested in high school contexts the effects of a self-affirmation intervention that was found to be effective in middle school. At the aggregate, their intervention was impressively effective (reduction of 50% in the growth of the racial achievement gap across the high school transition). Yet Borman et al. also emphasized the substantial variability across contexts, being careful to highlight the "potential" of the intervention "if implemented broadly" and to note that "these effects will depend on both contextual and individual factors" (p. 1773). Educational contexts are complex settings, characterized by unique and dynamic features, which involve nonlinear and serendipitous happenings that belie the expectation that a direct transfer of successful practices from one setting to another, similar as they may seem, would result in a simple replication.

Importantly, the understanding that an aggregate coefficient masks relevant variability pertains also to coefficients from a single RCT. Whereas randomization is assumed to create probabilistic equality in participants' characteristics across conditions, effects within conditions constitute aggregates across students with such different characteristics. Many, if not most, educational interventions aim at processes and outcomes conceptualized at the level of the individual student—learning, motivation, engagement, achievement. But RCTs produce average effects that reflect a mismatch between the conceptual and analytical units of analysis, raising a concern with the expectation that any effect would generalize to new individual students.

The implications are that effective transfer of RCT findings to new settings, groups, and individuals would only take place, not by transferring the same practices that were successful in one context to another, but by meticulous translation of robust *theoretical* understanding, including contextual causal mechanisms, to the particular characteristics of the new setting and students. Notably, RCT effects and tests of moderation by nominal-categorical variables (e.g., student age and demographics, school type, time of year) do not provide substantial insight into causal mechanisms by which context may affect an intervention. As the notable proponents of experimental design, Shadish et al. (2002) acknowledged, "experiments do less well in clarifying the mechanisms through which and the conditions under which [a] causal relationship holds"; that is, RCTs do not provide a "*causal explanation*" (p. 9). Instead, the data about ways by which contextual features and events in our studies may have affected the intervention would have been generated by approaching each experiment as a case study, with an intentional, rigorous, longitudinal ethnographic investigation that generates a "thick description" (Geertz, 1973) of the intervention's unfolding in the particular setting.

Over four decades ago, Lee Cronbach (1975) pointed to the inevitable limitation of experimental designs, which always concern the interaction of a treatment with the personal characteristics (aptitude) of the participant. Cronbach concluded that when Aptitude by Treatment Interactions (ATIs) are present, "a general statement about a treatment effect is misleading because the effect will come or go depending on the kind of person treated" (p. 119). Cronbach recognized the futility of drawing any firm conclusions about the effect of any intervention when there are moderators involved: "Once we attend to interactions, we enter a hall of mirrors that extends to infinity. However far we carry

our analysis—to third order or fifth order or any other—untested interactions of a still higher order can be envisioned" (p. 119).

The traditional understanding of RCT as the "gold standard" design that provides a basis for "evidence-based practice" fails to account for Cronbach's critique. The IES' What Works Clearinghouse (WWC) rates RCTs as meeting WWC standards "without reservations" (WWC, 2019a). In purporting to "focus on high-quality research to answer the question 'what works in education?'" (WWC, 2019b), this approach equates "evidence-based practice" with the aggregated effects of experiments (with the possible qualification of a moderator), thus ignoring the strong evidence that what has "worked" in one context, may not have actually "worked" for everyone in that context, and may very well not "work" the same way in any other context—similar as it may seem to the original study's setting.

Cronbach's (1975) solution was an epistemological change to the purported goal of RCTs. He emphatically recommended that instead of conducting experiments that aim at cross-contextual generalizations, researchers should investigate and theorize the causal processes around practice in the context in which it is implemented:

> Instead of making generalization the ruling consideration in our research, I suggest that we reverse our priorities. An observer collecting data in one particular situation is in a position to appraise a practice or proposition in that setting, observing effects in context. In trying to describe and account for what happened, he [*sic*] will give attention to whatever variables were controlled, but he [*sic*] will give equally careful attention to uncontrolled conditions, to personal characteristics, and to events that occurred during treatment and measurement. As he [*sic*] goes from situation to situation, his [*sic*] first task is to describe and interpret the effect anew in each locale, perhaps taking into account factors unique to that locale of series of events. (pp. 124–125)

In line with Cronbach's view, we contend that for RCTs to actually uphold their reputation for generating causal evidence, researchers would need to replace the goal of RCTs from generalizing findings to the population to generalizing findings to a robust theoretical understanding of the contextualized causal mechanisms underlying the phenomenon. Accepting the assumption that unique, dynamic, configurations of contextual features are inevitably intertwined with the causal mechanisms underlying the intervention's effects calls for collecting data, explicating, and conceptualizing these contextualized causal mechanisms (Maxwell, 2004). This would entail incorporating into the RCT systematic methods employed in case study research (Yin, 2018), guided by existing comprehensive frameworks of the myriad influences on interventions' causal effects (Weiss et al., 2013). As findings are incorporated from such RCTs in diverse contexts, the theory concerning the causal mechanisms underlying the intervention's effects would become more inclusive and robust. Still, translating these theoretical understandings to new settings is an inductive endeavor—any new setting is a unique context, and the effect can be anticipated to manifest differently from theoretical expectations. Hence, WWC's recommendations for implementing any practice should involve an emphasis on contextualizing it—employing a design-based process that maximizes the fit of the practice to the unique context while contributing further to theoretical understanding of the contextualized phenomenon.

In conclusion, we argue that the meaning of "evidence-based practice" should be modified to account for the *evidence* that context matters greatly; that impactful contextual features are inductive and unpredictable; that interventions unfold in complex ways; that findings about "what worked" in educational interventions should serve to build robust theories of contextual causal mechanisms; and that these, in turn, should provide a starting point for evidence-focused design-based implementation that attend to the emergent, complex, and dynamic nature of educational contexts (Kaplan et al., 2012; Penuel et al., 2011).

NOTE

The research reported herein was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140602 to Temple University. All statements reflect the authors' own opinions and do not reflect the policies of IES or the U.S. government.

ORCID ID

Avi Kaplan https://orcid.org/0000-0002-2898-0085

REFERENCES

Borman, G. D., Grigg, J., Rozek, C. S., Hanselman, P., & Dewey, N. A. (2018). Self-affirmation effects are produced by school context, student engagement with the intervention, and time: Lessons from a district-wide implementation. *Psychological Science*, *29*(11), 1773–1784.

Cromley, J. G., Perez, T., Kaplan, A., Dai, T., Mara, K., & Balsai, M. (2019, April). *Combined cognitive-motivational interventions with substantial benefits for undergraduate biology grades: A meta-analysis of 10 new experiments*. Paper presented at the annual meeting of the American Educational Research Association, Toronto, Canada.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*(2), 116–127.

Geertz, C. (1973). *The interpretation of cultures*. New York, NY: Basic.

Kaplan, A., Katz, I., & Flum, H. (2012). Motivation theory in educational practice: Knowledge claims, challenges, and future directions. In K. R. Harris, S. G. Graham, & T. Urdan (Eds.), *APA educational psychology handbook: Vol. 2. Individual differences, cultural considerations, and contextual factors in educational psychology* (chap. 7, pp. 165–194). American Psychological Association.

Lortie-Forgues, H., & Inglis, M. (2019). Rigorous large-scale educational RCTs are often uninformative: Should we be concerned? *Educational Researcher*, *48*(3), 158–166. doi:10.3102/0013189X19832850

Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher*, *33*(2), 3–11.

Penuel, W. R., Fishman, B. J., Cheng, B. H., & Sabelli, N. (2011). Organizing research and development at the intersection of learning, implementation, and design. *Educational Researcher*, *40*(7), 331–337. doi:10.3102/0013189X11421826

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

Weiss, M. J., Bloom, H. S., & Brock, T. (2013, June). *A conceptual framework for studying the sources of variation in program effects*. New York, NY: MDRC. Retrieved from https://www.mdrc.org/sites/default/files/a-conceptual_framework_for_studying_the_sources.pdf

What Works Clearinghouse (WWC). (2019a). *How the WWC rates a study: Rating group designs*. Retrieved March 8, 2019, from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_info_rates_061015.pdf

What Works Clearinghouse (WWC). (2019b). *What we do*. Retrieved March 8, 2019, from https://ies.ed.gov/ncee/wwc/WhatWeDo

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE.

AUTHORS

AVI KAPLAN, PhD, is a professor of educational psychology at the College of Education, Temple University, Philadelphia, PA 19122; *akaplan@temple.edu*. His research focuses on motivation and identity in educational settings.

JENNIFER CROMLEY, PhD, is a professor of educational psychology at the College of Education, University of Illinois Urbana-Champaign, Champaign, IL 61820; *jcromley@illinois.edu*. Her research focuses on reading comprehension of illustrated scientific text and on cognitive and motivational predictors of STEM students' achievement and retention.

TONY PEREZ, PhD, is an assistant professor of educational psychology at the Darden College of Education and Professional Studies, Old Dominion University, Norfolk, VA 23529; *acperez@odu.edu*. His research focuses on the relations among motivation, identity, and STEM achievement, and applying motivational principles to interventions designed to support STEM achievement and persistence.

TING DAI, PhD, is an assistant professor of educational psychology at the College of Education, University of Illinois at Chicago, Chicago, IL 60607; *tdai@uic.edu*. Her research focuses on measurement of student motivation, epistemic cognition, and achievement in STEM.

KYLE MARA, PhD, is an assistant professor of biology at the College of Science, Engineering and Education, University of Southern Indiana, Evansville, IN 47712; *kyle.mara@usi.edu*. His research focuses on vertebrate functional morphology and biomechanics, as well as in improving teaching and learning in STEM through educational interventions.

MICHAEL BALSAI, PhD, is an assistant professor of biology at the College of Science and Technology, Temple University, Philadelphia, PA 19122; *mjbalsai@temple.edu*. His scholarly interests focus on paleontology, the morphology of lizards, and the improvement of STEM instruction.

Manuscript received April 21, 2019
Revisions received August 29, 2019; October 26, 2019
Accepted December 27, 2019