RUNNING HEAD: Replication vs. Generalization

How to interpret discrepancies in empirical results from educational intervention studies

Daniel M. Oppenheimer[1] & Michelle R. Ellefson[2]

[1] Department of Social and Decision Sciences, Carnegie Mellon University

[2] Faculty of Education, University of Cambridge

This manuscript was accepted for publication in *Scholarship of Teaching and Learning in Psychology* on 22 July 2024. This is the accepted version. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite this preprint without authors' permission. The citation for the final, published article is:

Oppenheimer, D. M., & Ellefson, M. R. (2024). How to interpret discrepancies in empirical results from educational intervention studies. *Scholarship of Teaching and Learning in Psychology*. Advance online publication. https://doi.org/10.1037/stl0000427

Author Note

Daniel M. Oppenheimer https://orcid.org/0000-0002-2363-4220

Michelle R. Ellefson https://orcid.org/0000-0003-0407-9767

Daniel M. Oppenheimer played an equal role in writing–original draft. Michelle R. Ellefson played an equal role in writing–original draft.

Correspondence concerning this article should be addressed to Daniel M. Oppenheimer, Department of Social and Decision Sciences, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States. Email: oppenheimer@cmu.edu

Abstract

When Mueller and Oppenheimer documented the Longhand Advantage (2014) -- the finding that students learn better when they take notes using pen and paper rather than a laptop – the study went viral, influencing classroom laptop policies around the world. However, more recently Urry et al. (2019) followed up on the original finding but found no evidence for a longhand advantage. Similarly, the literature on whether interventions to improve executive functions lead to improved classroom performance is full of mixed and contradictory findings. Instructors who hope to use evidence-based interventions in the classroom are confronted with a dilemma when they encounter discrepant studies regarding whether or not an intervention is effective. One helpful approach to navigating this challenge is to consider the differences between direct replications and generalization studies. While direct replications recreate the conditions of the original study as closely as possible to determine whether the results are reliable and trustworthy, generalization studies investigate whether the results of the original study are robust to novel conditions, populations or environments. Because no study can perfectly replicate the conditions of another study, it is not always clear whether a study should be considered a replication or generalization. To resolve this, we describe a framework based in the intervention literature that proposes five principles for evaluating the match between original studies and their follow ups. We illustrate the utility of this framework using the discrepant findings from Urry et al. (2019) and Mueller and Oppenheimer (2014), and the mixed evidence from the executive control interventions literature as case studies. Broader implications and practicalities for interpreting findings from replication attempts are discussed.

How to interpret discrepancies in empirical results from educational intervention studies

"Every replication is different in innumerable ways from the original."

-Klein et al., 2014

The learning science literature is full of interventions that purportedly improve learning and warnings against activities that impede learning. Most of these have some empirical backing, but in the age of the replication crisis some educators have become skeptical of how well those findings will hold up in follow-up studies or in the classroom. When new studies yield different outcomes than the original studies, how are teachers and scholars to make sense of those discrepancies? The answer, of course, relies on the nature of those discrepancies.

Discussing the differences between direct replications, conceptual replications, and generalizations (among many other similar and related concepts) is difficult because the terms are used in different ways by different authors and in different academic disciplines. (For a particularly cogent discussion of terminology see Plesser, 2018). This is likely particularly true in interdisciplinary fields like the learning sciences, where scholars come from various backgrounds, and use discipline-specific norms, discipline-specific jargon. For the purposes of the present article, we would like to define a distinction between what we will call direct replication and generalization. The former recreates the conditions of the original study as closely as possible to determine whether the results are reliable and trustworthy. The latter investigates whether the results of the original study are robust to novel conditions, populations or environments. Both are valuable to scientific understanding, but because they allow for fundamentally different inferences, it is important not to conflate them.

All studies are embedded in cultural contexts which shift over time, and often studies are run in different locations and on different samples of the population. Moreover, even good faith attempts to engage in direct replications can run into logistical or practical challenges that necessitate changes to the original study design. When considering whether a study constitutes a direct replication, scholars typically attend to whether the replication uses the same materials as the original study, while neglecting whether those materials fulfill the same purpose for the populations being tested and given the conditions of testing. Scholars stress sample size to ensure adequate power but often overlook environmental factors that may reduce effect sizes. Given that all replications differ from the original study in some ways, how can we determine if those differences are meaningful enough to consider the study a generalization rather than a direct replication?

One possibility is to import frameworks from intervention science[1]. In intervention science, efficacy trials establish whether a program works with a tightly controlled population under tightly controlled circumstances. Effectiveness trials test whether the efficacious programs can be generalized to a wider population in a variety of settings. Singal et al. (2014) provided a useful framework for comparing efficacy and effectiveness on five dimensions: (1) *Research Question* – is the focus of the research on demonstrating a phenomenon in ideal circumstances vs. in real-world contexts; (2) *Population* - homogeneous vs. heterogeneous; (3) *Research Setting* - ideal with no distractions vs. more distractions (4) *Research Team* (which Singal et al. refer to as "providers") - highly experienced / fully-trained vs. less experience/training; and (5)

[1] We chose to use the term "intervention science" rather than the related, but subtly distinct "implementation science". Although there is not perfect agreement in the literature about how these terms are defined, we see intervention science as focusing on research design details while implementation science focuses more on the logistical details and implications of practitioners more widely adopting empirical findings.

Fidelity (which Singal et al. refer to as "intervention") - strict enforcement / standardized research protocols vs. flexible/unstandardized research protocols, in other words, the extent to which different researchers on the research team adhere to experimental or intervention protocols[2].

A good practice for understanding if a result is replicable aligns with establishing efficacy. Many efforts for wide-scale replication have focused almost exclusively on using identical stimuli and methodologies for stimulus presentation (e.g., Many Labs, 2014). This focus on the protocol fits well with Singal et al.'s (2014) first dimension - *Research Question*. Some replication projects have started to pay attention to whether the *Population* is the same (e.g. Manylabs 2; 2018). However, *Research Setting*, *Research Team* and *Fidelity* are often neglected.

For example, in a recent paper, Urry et al. (2021; henceforth "Urry") attempted to replicate Mueller and Oppenheimer's (2014; henceforth "Mueller") longhand advantage – the finding that students who take notes on laptops learn less than students who take notes using pen and paper. While participants in the original manuscript were run under distraction free, tightly controlled laboratory conditions, the replication was run as part of a classroom project, by undergraduates who had not been fully trained in the original experimental protocols. As a result, the data collection for the attempted replication occurred in "sessions on campus at various times of day outside of class. Many noted that sessions were subject to distractions and errors; sessions also varied in formality and equipment (i.e., laptops and headphones)." Moreover, the population tested in Urry had different baseline rates of laptop usage.

---

[2] We changed some of the terminology to allow it to more effectively apply to a wider research context than Singal et al. had originally referred to, and to avoid confusing readers about which concepts were directly imported from Singal et al., and which are extensions of that conceptual framework for application to the present questions.

While Urry described their study as a direct replication, the *Population, Research Setting*, *Research Team* and *Fidelity* all fundamentally differ from Mueller, and thus would be better described as an attempt at generalization. To understand why these factors are important in this case study, pay attention to *mechanism* – not just *if* a particular intervention is effective but *why* that intervention is purportedly effective. This is especially relevant for papers using lab rather than field studies, because it is possible to fully control and manipulate the lab environment to better test causal and mechanistic claims. In the case of Urry and Mueller, the differences across these dimensions are theoretically meaningful in that they have direct relevance to the purported mechanism of the longhand advantage.

In the case of the longhand advantage there are several mechanisms that have been proposed for what is likely a multiply determined effect. The first is being off task; students using computers typically have access to social media, email, etc., and when students focus on those uses of a laptop instead of educational content, it can undermine student learning (e.g., Glass & Kang 2019). Indeed, some studies suggest that students on laptops spend only about a third of class time actually on task (c.f. Ragan et al., 2014), and this can hurt performance. This mechanism was not under investigation in either the Mueller or the Urry study, because participants in both studies were being monitored by the research team and kept from using the laptop to browse the internet.

The second proposed mechanism, and the one advocated by Mueller, suggests that because people type faster than they can handwrite, laptop note-takers can take down lectures verbatim, while longhand note-takers cannot, and thus are forced to take notes in their own words. This process of summarizing content in their own words requires depth of processing, which improves

learning (Craik & Tulving, 1975). In other words, longhand note-takers are forced to process more deeply, which serves as a form of desirable difficulty (Bjork & Bjork, 2011).

Some difficulties are desirable in that they create so-called "germane load," which supports learning. These include most interventions that push students toward deeper processing, such as requiring students to generate (rather than review) answers, and to interleave (rather than chunk) examples (see Bjork & Bjork, 2011 for examples and further discussion). However, some distractions create so-called "extraneous load," which interferes with students' processing (Sweller, 2010). Splitting attention to non-germane distractions can make otherwise desirable difficulties harmful (for more on Cognitive Load Theory, see Chandler & Sweller, 1991). Cognitive load affects cognitively effortful processes more than relatively effortless processes (Thomson & Oppenheimer, 2022). In the case of notetaking, cognitive load should disrupt the effortful processing of the longhand note takers (summarizing content in their own words) while having little effect on the relatively effortless processing of laptop note taking (verbatim transcription). In other words, Urry introduced features such as distraction which creates extraneously load; if Mueller's mechanistic account is correct, these features would be expected to eliminate the longhand advantage, which is exactly what happened.[3]

In many ways, the conditions in Urry have more ecological validity than the artificial lab conditions of Mueller. After all, real world classroom environments are highly varied and rife with distractions. Given that Mueller has received significant media attention and has been used to justify laptop bans in classrooms, identifying that the finding may not generalize beyond

[3] In addition to the Urry paper, other scholars, most notably Morehead et al. (2019), have conducted replication attempts of the Mueller paper, with complex and nuanced results. The full details of that nuance and complexity are beyond the scope of the present manuscript, but we direct interested readers to the online supplement which describes them in more detail.

distraction-free environments is valuable. Nonetheless, the Urry paper cannot speak to the reliability of the original finding, because the study was not designed to demonstrate the efficacy of the intervention.

The efficacy/effectiveness distinction (e.g., Singal et al., 2014) also has important implications for inclusion criteria for meta-analyses. For example, while Urry's meta-analysis shows null results, another recent meta-analysis of 14 studies and over 3,000 participants (Allen et al., 2020) found reliable longhand advantages (~25% improvement). Many of the studies in the Allen et al. differed from Mueller, but largely involved highly controlled research settings, research teams, and high fidelity. (See also an even more recent meta-analysis of 24 studies with over 3000 participants with similarly robust results for the longhand advantage; Flanigan et al., 2024).

Urry are to be commended for their commitment to open science and for reporting and acknowledging the weaknesses of their study, specifically the distracting and variable environments, and errors in experimental protocols. It is precisely that level of honest self-scrutiny that allows science to reconcile differences across findings and enables progress. In this case, it helps in the interpretation of the null results, suggesting that their paper may be better classified as a failure to generalize rather than a failure to replicate. To improve our science and allow practitioners to understand how to apply the insights of our studies to the classroom, learning scientists need to ensure that it is standard to attend to, and explicitly report details of, research settings, research team training, and fidelity (see Ellefson & Oppenheimer, 2022, for a discussion of the statistical effects of fidelity violations).

It is important to note that notetaking and laptop use is hardly unique in the learning sciences when it comes to mixed and confusing evidence about which findings are robust and when they will generalize. Interventions to improve executive functions are another example of heterogeneous generalization. Executive functions are the higher order cognitive skills needed for goal-directed behaviors (Diamond, 2013). Many studies find a reliable link between executive functions and academic outcomes, indicating that interventions should both improve executive function skills and transfer to academic achievement (e.g., Ellefson et al., 2019). However, studies on the impact of executive functions training on educational outcomes find inconsistent results.

For example, The CogMed intervention (Klingberg et al., 2005) targets working memory in a computer game. Although working memory often improves, different studies find different results about whether these improvements transfer to academic attainment (e.g., Bharadwaj et al., 2022 - no evidence of transfer; Holmes & Gathercole, 2014 - mixed evidence of transfer). Similar inconsistencies appear in studies exploring whether training other executive functions leads to gains in the classroom (e.g., Inhibition: Ganesan et al., 2024; Wilkinson et al., 2020, Cognitive flexibility: Buttleman & Karbach, 2017; van Bers et al. 2020). Interventions targeting multiple executive functions seem to have more robust transfer effects, but they are more complicated and require more research team expertise and training. For example, Tools of the Mind seems to improve executive functions and academic skills in young children (Diamond et al., 2007; 2019), but the intervention design is complex and does not always generalize to other educational contexts (Baron et al., 2017). The new ONE program (Orchestrating Numeracy and the Executive) embeds executive functions training in early numeracy teaching. There is great heterogeneity across studies regarding whether or not transfer to educational settings is observed

(Scerif et al., 2023). Although most interventions target children, there are a growing number focused on adolescents and adults (for a review and discussion see Diamond & Ling, 2016 and Hillman et al., 2019).

Again, exploring *mechanism* provides insights about what could be causing these mixed results. Some of the discrepancies in findings could be due to the differently trained research teams, different methods or different populations studied. For example, for the ONE program quantitative fidelity metrics indicate that the observed heterogeneity in results across studies is greatly influenced by the extent to which teachers adhere to the program (Scerif et al. 2023). Interpretation of findings should account for assessment methodology, especially with regard to implementation fidelity (for more detailed discussion see Smith et al., 2019)

So, what is a teacher to do? Does this inconsistency in findings mean that executive functions interventions should be abandoned? Does the mixed evidence on laptop use and note taking mean that practitioners should adopt or abandon laptop bans in the classroom?

Unfortunately, the answer is not simple. Teachers should consider the evidence knowing that any new study cannot account for all of the complexities in their classrooms. At the end of the day, part of the reasons that some findings are not robust across studies is because there are different materials, and different contexts, individual differences between students, and variations in the effectiveness of instruction and/or implementation fidelity of the interventions. Only by understanding the purported mechanisms by which the intervention is presumed to work, and whether that mechanism is relevant to the particulars of their classroom, can teachers make informed decisions about adopting specific interventions.

This is analogous to how the child development research can help us understand much about cognitive development but little about how to respond to a specific child in a given moment. Parents' response to a child in a specific situation needs to incorporate their knowledge of that child, knowledge of the child/parent relationship, and knowledge of the context, not just the general principles derived from the developmental science literature. Similarly, teachers should consider how results from studies in the learning sciences can be applied to their classrooms by using their expertise in successful pedagogy and considering the needs and abilities of their specific students, school and local community.

For example, in a quiet, distraction-free classroom with focused students, note-taking in one's own words might facilitate deeper processing and better learning, However, in a noisier, more distracting classroom with less focused students, pen and paper may provide less benefit. In a classroom where educators have the time and bandwidth to learn and adhere to the fairly complex requirements of the ONE program, one might see transfer to educational outcomes. However, in classrooms where teachers are less able to adhere to the precise details of the program, less transfer may be observed. Indeed, teachers need to honestly evaluate their own skill sets, and whether or not they have the ability, training, bandwidth, and interest to adopt various interventions.

Teachers also need to think about the content, goals, and structure of specific lesson plans and assessment instruments, and whether specific mechanisms facilitate or impede those goals. Verbatim note taking may inhibit deeper processing of materials but may also create a more complete record of a lecture. The latter may be a more important goal if a student is taking notes for an absent friend, to create an archive of classroom activities, or where the literal verbatim content is otherwise important (e.g. maintaining the integrity of quotes for journalistic or legal

purposes, mastering APA style, instruction about the difference between plagiarism and paraphrasing, etc.).

Finally, teachers can assess the effectiveness of teaching methods in their own courses. If an intervention seems like it might be appropriate for a particular class, one can try it out. If the intervention improves outcomes in the class, then the teacher can continue to use it in future iterations. To assist with this, teachers should refer to the growing literature in implementation science (e.g. Soicher et al., 2020), which explores the tools that can be used to effectively translate work from the academic literature into practice, and which also engages with efficacy/effectiveness distinctions when considering the conditions under which findings will successfully generalize to classroom environments.

As the replication movement grows across psychological science, attending to these multiple dimensions of how studies may differ can help replicators design better studies, statisticians design better meta-analyses, and psychological scientists more generally interpret discrepant results. Importantly, it can also help educators understand how to make sense of inconsistent findings in the literature, and best decide whether a particular finding is relevant to their specific classrooms.

References

Allen, M., LeFebvre, L., LeFebvre, L. & Bourhis, J. (2020). Is the pencil mightier than the keyboard? A meta-analysis comparing the method of notetaking outcomes, *Southern Communication Journal*, *85*(3), 143-154. https://doi.org/10.1080/1041794X.2020.1764613

Baron, A., Evangelou, M., Malmberg, L.-E., Melendez-Tores, G.-J. (2017). The Tools of the Mind curriculum for improving self-regulation in early childhood: A systematic review. *Campbell Systematic Reviews*, *13*, 1-77. https://doi.org/10.4073/csr.2017.10

Bharadwaj, S. V., Yeatts, P., & Headley, J. (2022). Efficacy of CogMed working memory training program in improving working memory in school-age children with and without neurological insults or disorders: A meta-analysis. *Applied Neuropsychology: Child*, *11*(4), 891–903. https://doi.org/10.1080/21622965.2021.1920943

Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R. Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

Buttelmann, F., & Karbach, J. (2017). Development and plasticity of cognitive flexibility in early and middle childhood. *Frontiers in Psychology*, *8*, 1040. https://doi.org/10.3389/fpsyg.2017.01040

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, *8*(4), 293-332. https://doi.org/10.1207/s1532690xci0804_2

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268-294. https://doi.org/10.1037/0096-3445.104.3.268

Desselle, S., & Shane, P. (2018). Laptop versus longhand note taking in a professional doctorate course: Student performance, attitudes, and behaviors. *Innovations in Pharmacy*, 9(3), Article 15. https://doi.org/10.24926/iip.v9i3.1392

Diamond A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Diamond, A., Barnett, W. S., Thomas, J., & Munro, S. (2007). Preschool program improves cognitive control. *Science*, *318*(5855), 1387–1388. https://doi.org/10.1126/science.1151148

Diamond, A., Lee, C., Senften, P., Lam, A., & Abbott, D. (2019). Randomized control trial of Tools of the Mind: Marked benefits to kindergarten children and their teachers. *PloS One*, *14*(9), e0222447. https://doi.org/10.1371/journal.pone.0222447

Diamond, A., & Ling, D. (2016). Conclusions about interventions, programs, and approaches for improving executive functions that appear justified and those that, despite much hype, do not. *Developmental Cognitive Neuroscience*, *18*, 34-48. https://doi.org/10.1016/j.dcn.2015.11.005

Dumontheil, I., Wilkinson, H. R., Farran, E. K., Smid, C., Modhvadia, R., Mareschal, D., & the UnLocke team (2023). How do executive functions influence children's reasoning about

counterintuitive concepts in mathematics and science? *Journal of Cognitive Enhancement*, *7*, 257–275 (2023). https://doi.org/10.1007/s41465-023-00271-0

Ellefson, M. R., Baker, S. T., & Gibson, J. L. (2019). Lessons for successful cognitive developmental science in educational settings: The case of executive functions. *Journal of Cognition and Development*, *20*(2), 253–277. https://doi.org/10.1080/15248372.2018.1551219

Ellefson, M. R., & Oppenheimer, D. M. (2023). Is replication possible without fidelity? *Psychological Methods*, *28*(6), 1446–1455. https://doi.org/10.1037/met0000473

Glass, A. L., & Kang, M. (2019). Dividing attention in the classroom reduces exam performance. *Educational Psychology*, *39*(3), 395-408. https://doi.org/10.1080/01443410.2018.1489046

Ganesan, K., Thompson, A., Smid, C. R., Cañigueral, R., Li, Y., Revill, G., Puetz, V., Bernhardt, B. C., Dosenbach, N. U. F., Kievit, R. & Steinbeis, N. (2024). Cognitive control training with domain-general response inhibition does not change children's brains or behavior. *Nature Neuroscience*, Advance Online Article. https://doi.org/10.1038/s41593-024-01672-w

Hillman, C. H., McAuley, E., Erickson, K. I., Liu-Ambrose, T., & Kramer, A. F. (2019). On mindful and mindless physical activity and executive function: A response to Diamond and Ling (2016). *Developmental Cognitive Neuroscience*, *37*, Article 100529. https://doi.org/10.1016/j.dcn.2018.01.006

Holmes, J., & Gathercole, S. E. (2014). Taking working memory training from the laboratory into schools. *Educational Psychology*, *34*(4), 440–450. https://doi.org/10.1080/01443410.2013.797338

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition*, 15(4), 332-340. https://doi.org/10.3758/BF03197035

Horbury, S. R., & Edmonds, C. J. (2021). Taking class notes by hand compared to typing: Effects on children's recall and understanding. *Journal of Research in Childhood Education, 35*(1), 55–67. https://doi.org/10.1080/02568543.2020.1781307

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., Rédei, A. C., Cai, H., Cambier, F., Cantarero, K., Carmichael, C. L., Ceric, F., Chandler, J., Chang, J.-H., Chatard, A., Chen, E. E. … & Nosek, B. A. (2018). Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science, 1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., Gillberg, C. G., Forssberg, H., & Westerberg, H. (2005). Computerized training of working memory in children with ADHD--a randomized, controlled trial. *Journal of the American Academy of Child and Adolescent Psychiatry*, *44*(2), 177–186. https://doi.org/10.1097/00004583-200502000-00010

Mueller, P. A., & Oppenheimer, D. M. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science*, *25*(6), 1159-1168. https://doi.org/10.1177/0956797614524581

Plesser, H. E. (2018). Reproducibility vs. replicability: A brief History of a confused terminology. *Frontiers in Neuroinformatics*, *11*, Article 76. https://doi.org/10.3389/fninf.2017.00076

Ragan, E. D., Jennings, S. R., Massey, J. D., & Doolittle, P. E. (2014). Unregulated use of laptops over time in large lecture classes. *Computers & Education*, *78*, 78-86. https://doi.org/10.1016/j.compedu.2014.05.002

Gür, T. (2021). The effect of verbatim and generative notes taken by hand and keyboard at university level on success and persistence. *Education Quarterly Reviews*, *4*(3). https://doi.org/10.31014/aior.1993.04.03.325

Scerif, G., Gattas, S., Hawes, Z., Howard, S., Merkley, R., & O'Connor, R. (2023). *Orchestrating numeracy and the executive: The ONE programme*. PsyArXiv. https://doi.org/10.31234/osf.io/2gxzv

Singal, A. G., Higgins, P. D., & Waljee, A. K. (2014). A primer on effectiveness and efficacy trials. *Clinical and Translational Gastroenterology*, *5*(1), e45. https://doi.org/10.1038/ctg.2013.13

Smith, K., Finney, S., & Fulcher, K. (2019). Connecting assessment practices with curricula and pedagogy via implementation fidelity data. *Assessment & Evaluation in Higher Education,* 44(2), 263-282.

Soicher, R. N., Becker-Blease, K. A., & Bostwick, K. C. P. (2020). Implementation of the utility value intervention: An adaptation of implementation science frameworks for higher education teaching and learning research. *Cognitive Research: Principles and Implications*. *5*(1), Article 54. https://doi.org/10.1186/s41235-020-00255-0

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22*(2), 123-138. https://doi.org/10.1007/s10648-010-9128-5

Thomson, K. S., & Oppenheimer, D. M. (2022). The "effort elephant" in the room: What is effort, anyway? *Perspectives on Psychological Science*, *17*(6), 1633-1652. https://doi.org/10.1177/174569162110648

Urry, H. L., Crittle, C. S., Floerke, V. A., Leonard, M. Z., Perry, C. S., 3rd, Akdilek, N., Albert, E. R., Block, A. J., Bollinger, C. A., Bowers, E. M., Brody, R. S., Burk, K. C., Burnstein, A., Chan, A. K., Chan, P. C., Chang, L. J., Chen, E., Chiarawongse, C. P., Chin, G., Chin, K., … Zarrow, J. E. (2021). Don't ditch the laptop just yet: A direct replication of Mueller and Oppenheimer's (2014) Study 1 plus mini meta-analyses across similar studies. *Psychological Science*, *32*(3), 326–339. https://doi.org/10.1177/0956797620965541

van Bers, B. M. C. W., van Schijndel, T. J. P., Visser, I., & Raijmakers, M. E. J. (2020). Cognitive flexibility training has direct and near transfer effects, but no far transfer effects, in preschoolers. *Journal of Experimental Child Psychology*, *193*, 104809. https://doi.org/10.1016/j.jecp.2020.104809

Wilkinson, H. R., Smid, C., Morris, S., Farran, E. K., Dumontheil, I., Mayer, S., Tolmie, A., Bell, D., Porayska-Pomsta, K., Holmes, W., Mareschal, D., Thomas, M. S. C., & the UnLoke Team. (2020). Domain-specific inhibitory control training to improve children's learning of counterintuitive concepts in mathematics and science. *Journal of Cognitive Enhancement*, *4*, 296–314. https://doi.org/10.1007/s41465-019-00161-4

Online Supplement 1

Morehead et al. (2019), also attempted a replication that more faithfully emulates the conditions of Mueller's studies (sterile, controlled, lab environment). The results from the Morehead et al. are nuanced and mixed – they found a significant longhand advantage for factual material, and a directional, but not significant longhand advantage for conceptual material for immediate recall. But they found no evidence for the longhand advantage after introducing a delay (while Mueller had found a longhand advantage after a delay; Study 3).

Perhaps more crucially, Morehead et al. (2019) looked at the relationship between the amount of verbatim note taking and subsequent recall; while they found that longhand users took significantly fewer verbatim notes (replicating Mueller) they found no link between the amount of verbatim transcription of the lecture and subsequent recall (failing to replicate Mueller). That said, a subsequent study by Gur (2021) randomly assigned participants to take notes verbatim vs. in their own words (rather than relying on naturalistic variation) and found that verbatim notes led to lower performance, suggesting that all else being equal verbatim note taking does hurt learning, but that all else is not always equal, depending on the conditions of testing.

It is also worth noting that in addition to the Urry and Morehead papers, which do not find consistent evidence of the longhand advantage, there are other papers that do, such as Horbury and Edmonds (2020) who generalize the effect to elementary school classrooms, and Desselle and Shane (2018) who generalize the effect to graduate student classrooms.

This muddled pattern of results highlights the need for the framework that we discuss in the main manuscript. The findings of Mueller seem to generalize well across different populations, but not very well across different conditions of testing. It thus becomes essential to

pay close attention both to the context of the studies that explore a phenomenon as well as specific classroom environments to which one is trying to adopt an intervention, and determine the extent to which they match