Abstract Title Page

Title: How Malleable is a Teacher's Classroom Practice?: A Meta-Analysis of Randomized Field Studies with Classroom Observations

Authors and Affiliations:
Rachel Garrett, (presenter) American Institutes for Research
Martyna Citkowicz, American Institutes for Research
Ryan Williams, American Institutes for Research

Abstract Body

Context:
Policymakers, practitioners and researchers have collectively recognized the salience of having highly effective teachers in all of our nation's classrooms. Research has clearly demonstrated the importance of teacher quality for student achievement, beyond other school-level characteristics (Aaronson, Barrow, & Sander, 2007; Goldhaber, 2002; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004). Federal education policy has had a long-standing interest in teacher effectiveness, but has put a particular focus on effective teachers over the last 15 years. This began with the requirements for highly qualified teachers in every school, as mandated by the No Child Left Behind Act in 2002. This was followed by the federal Race to the Top (RTTT) competition and the No Child Left Behind (NCLB) waivers, which both continued to place significant emphasis on ensuring effective instruction, primarily through an emphasis on more robust teacher evaluation systems. More recently, federal accountability requirements for teacher evaluation have been loosened in the Every Student Succeeds Act reauthorization, but states and districts continue to place a strong emphasis on teaching quality, and to allocate significant resources towards professional development for teachers (Jacob & McGovern, 2015).

Alongside the increased focus on teacher effectiveness in federal policy, the field of education research has witnessed a substantial increase in rigorous education research in instruction. As shown in Appendix Figure 1, a search of the literature for randomized studies with a focus on developing classroom practice demonstrates this uptick. The research likely has responded to both the policy impetus and the related need from the field to understand how to support and promote effective instruction through professional learning programs for teachers. Yet questions remain about how much classroom practice is improved through intervention, and what approaches yield the most success.

Purpose:
Our purpose is to examine the question of malleability by conducting a meta-analysis of randomized experiments of interventions directed at classroom practice. We consider how changes in classroom practice may vary by specific aspects of practice, for example, classroom management or questioning and discussion techniques. We also examine how heterogeneity of effects may relate to features of the interventions, and seek to identify if there are particularly effective approaches to teacher professional learning. Through this work we address the following set of questions:

1) How malleable is a teacher's classroom practice?
2) Are specific aspects of classroom practice more or less malleable?
3) Are particular intervention features associated with improvements in classroom practice outcomes?

Approach:
In the first phase of the study, a systematic literature search was completed to find randomized studies that included both K–12 in-service teachers and an analysis of impacts on classroom practice measured through classroom observation. We searched for studies in multiple databases and manually examined the reference lists in 16 meta-analyses. We identified 40 studies for

inclusion. The studies were coded for all relevant effect sizes for impacts on classroom practice, in addition to coding for a wide-ranging set of intervention, observation measure, and contextual factors. Because some studies included multiple samples with multiple outcomes, the number of effects extracted from the 40 studies was 321 effects.

In the second phase of the study, we conducted a meta-analysis on the coded and computed effect sizes (Cook, Cooper, Cordray, Hartmann, Hedges, & Light, 1992; Stuart, Cole, Bradshaw, & Leaf, 2001; Weiss, Bloom, & Brock, 2014). We applied Hedges' (1981) small sample correction to adjust for small sample sizes and adjusted all effect sizes for nesting (e.g., teachers are typically nested within schools and districts; Hedges 2007, 2011; Hedges & Citkowicz, 2015). Using robust variance estimation to account for dependent effect sizes (i.e., effects calculated from the same studies; Fisher & Tipton, 2015; Hedges, Tipton, & Johnson, 2010; Tipton 2014), we first conducted a random-effects meta-analysis of all 321 effects. The meta-analysis produces a weighted mean effect of all effect sizes, without accounting for any additional study or sample characteristics. We also conducted separate meta-analyses for each outcome domain (classroom environment, instruction, overall effectiveness) and constructs within the domains (e.g., classroom culture, instructional format, math content, student engagement). We examined the heterogeneity, or variation, in effects by conducting separate meta-regressions for each moderator of interest, including school level, intervention dosage (based on intervention length measured in weeks and hours), timing of observational measurement (e.g., mid-stream, directly after, after time has passed), content taught, teacher sample size, and teachers' years of experience.

Findings:

We find that, on average, the randomized field trials targeting classroom practice yielded a positive, statistically significant effect of 0.42 standard deviations based on classroom observations, as presented in Table 2 in the Appendix. While this is promising, our results also give caution: we observed a substantial amount of heterogeneity across effect sizes. Our estimated I^2 indicates that 73% of the total variation in our estimates of the treatment effect is due to heterogeneity between effects rather than within effects (e.g., sampling error). Moreover, our absolute effect sizes range from -0.9 to 2.8 standard deviations, with a prediction interval of -0.38 to 1.22). This implies that while classroom practice is malleable, there is substantial variation in how much classroom practice changes across different interventions. Additionally, in our subgroup analyses that focus on specific aspects of instruction, the findings consistently demonstrate positive and mostly significant results across the different constructs. Again though the findings further indicate substantial variability in effects across studies (with the exception of mathematics content) as observed by the wide confidence intervals and the high I^2s.

Further results provide insights for intervention design and features. For example, our results indicate that longer, more intensive interventions are not more effective than shorter ones, and that interventions may boost changes to classroom practice when they include opportunities for active teacher learning, instructional materials, or training on data use. Additional results that examine heterogeneity of results related to teacher and study characteristics will be presented.

Conclusions:

Our results indicate that interventions directly targeting classroom practice through professional learning can bring about meaningful shifts in practice as measured through classroom observation. We find that the effects of these interventions are on average positive, and the magnitude of the changes in classroom practice are substantively notable. Moreover, these findings generalize across multiple domains and sub-constructs of observed classroom practice. This indicates that the malleability of classroom practice is not tied to a specific aspect of observable practice, and likewise that interventions can support teachers in developing different types of classroom practice skills. From this, we conclude that teachers stand to benefit from professional learning opportunities designed to promote their classroom practice development across a range of areas.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago Public High Schools. *Journal of Labor Economics*, *25*(1), 95–135.

Cook, T. D., Cooper, H., Cordray, D. S., Hartmann, H., Hedges, L. V., & Light, R. J. (1992). *Meta-analysis for explanation: A casebook*. New York, NY: Russell Sage Foundation.

Fisher, Z., & Tipton, E. (2015). robumeta: Robust variance meta-regression. R package version 1.6. http://CRAN.R-project.org/package=robumeta

Goldhaber, D. D. (2002). The mystery of good teaching. *Education Next*, *2*(1), 50-55.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107-128. doi: *10.2307/1164588*

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341-370. doi: 10.3102/1076998606298043

Hedges, L. V. (2011). Effect sizes in three-level cluster-randomized experiments. *Journal of Educational and Behavioral Statistics*, *36*(3), 346-380. doi: 10.3102/1076998610376617

Hedges, L. V., & Citkowicz, M. (2015). Estimating effect size when there is clustering in one treatment group. *Behavior Research Methods, 47*(4), 1295-1308. doi: 10.3758/s13428-014-0538-z

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools and academic achievement. *Econometrica*, *73*(2), 417-458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247–252.

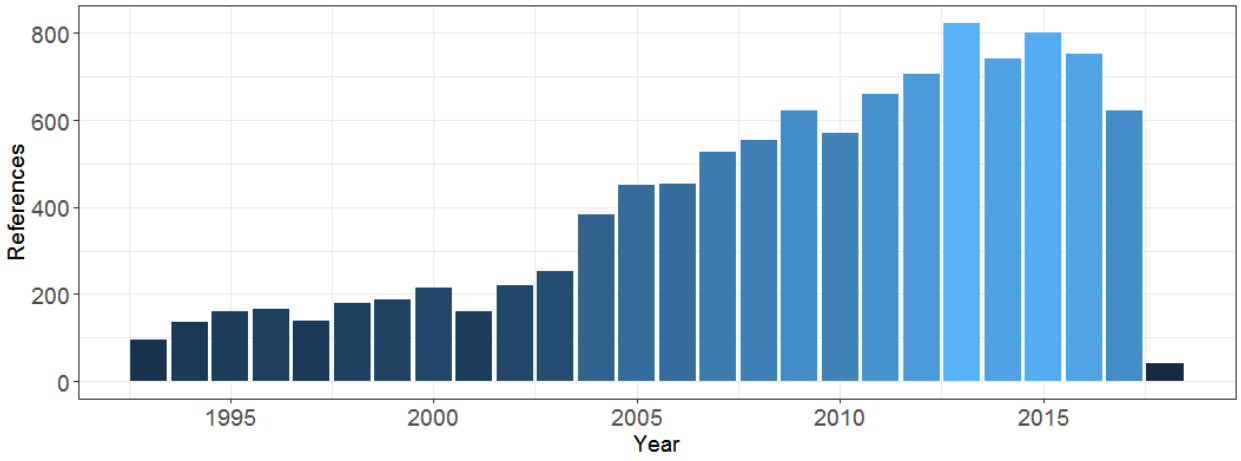
Stuart, E. A., Cole, S. R., Bradshaw, C. P., & Leaf, P. J. (2001). The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *174*(2), 369–386.

Tipton, E. (2014). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods, Advance online publication*. doi: dx.doi.org/10.1037/met0000011

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, *33*(3), 778–808.

Appendix

Figure 1. Number of references per year of randomized field studies that target classroom practice



Note. Figure shows the citation returns from an EBSCO Host search of the previous 25 years, by year, for the following search string: ("teacher practice" OR instruction OR "instructional practice" OR "classroom practice" OR "teacher effectiveness") AND (intervention OR strateg* OR program OR treatment) AND (experiment OR "randomized experiment" OR "randomized trial" OR "randomized control").

Table 1. Overall Mean Effects and Mean Effects by Classroom Practice Constructs

Construct	Mean	SE	Confidence Interval Lower	Confidence Interval Upper	*N*	*k*	I^2	τ^2
All Effects	0.42	0.07	0.28	0.56	40	321	72.60	0.16
Broad Domain								
Classroom Environment	0.27	0.08	0.10	0.44	17	112	51.76	0.07
Instruction	0.46	0.08	0.29	0.63	32	198	76.69	0.18
Overall Effectiveness	0.49	0.29	-0.20	1.18	8	11	84.90	0.44
Constructs within Broad Domains								
Classroom Environment								
Aggregate environment	0.32	0.13	-0.03	0.67	6	22	52.37	0.07
Classroom culture	0.16	0.06	0.01	0.31	9	32	41.02	0.03
Classroom management	0.43	0.24	-0.17	1.02	7	58	57.02	0.24
Instruction								
ELA content	0.55	0.16	0.13	0.96	6	49	84.70	0.30
Math content	0.31	0.04	0.20	0.41	7	21	0.00	0.00
Instructional format	0.21	0.08	0.03	0.40	10	22	16.07	0.01
Discourse	0.28	0.10	0.05	0.50	11	43	53.07	0.08
Student engagement	0.46	0.12	0.14	0.79	6	11	69.12	0.11
General / aggregate instruction	0.64	0.20	0.20	1.07	15	35	85.78	0.34

Note. Rho was set to 0.53, based on the overall mean effect across all studies. N = number of studies, k = number of effect sizes. I^2 = proportion of the total variance due to between effect size heterogeneity. τ^2 = total effect size heterogeneity.

"Aggregate environment" includes measures that aggregated across two or more items capturing classroom environment.

"Discourse" construct includes questioning, discussion and formative assessment.

"General/aggregate instruction" includes measures that aggregated across two or more items capturing instruction, or overall instructional measures.