What works in large-scale education reform

The single most robust finding across decades of education research is uncomfortable: most reforms fail, and the ones that succeed take far longer than politicians promise. Of the hundreds of interventions studied, only a handful show strong causal evidence of improving student outcomes — and even those depend heavily on context, implementation quality, and sustained political commitment. The evidence base is thinner than the confidence of reformers suggests, with much of what passes for knowledge resting on correlational studies, survivorship bias, and meta-analyses of questionable methodological rigor. This report synthesizes findings from McKinsey, Hattie's Visible Learning, OECD PISA, the World Bank, and peer-reviewed research to map what we actually know, what we don't, and where the field has been misled.

The McKinsey reports shaped global policy on shaky evidence

McKinsey & Company's two landmark education reports — *How the World's Best-Performing School Systems Come Out on Top* (2007) and *How the World's Most Improved School Systems Keep Getting Better* (2010) — have been among the most influential education policy documents of the 21st century. Their conclusions have shaped reforms from Abu Dhabi to Ontario. (Taylor & Francis Online) Yet both rest on a methodology that would not survive peer review.

The 2007 report, authored by Michael Barber and Mona Mourshed, (WorldCat) studied **25 school systems** (including 10 top performers such as Finland, Singapore, South Korea, and Hong Kong) and concluded that three things matter most: **getting the right people to become teachers, developing them into effective instructors, and ensuring every child receives quality instruction.** (Joe Kirby +2) The report found that Singapore accepts only 1 in 8 applicants for teacher training, (joe-kirby) (Joe Kirby) Finland draws from the top 20% of graduates, (Joe Kirby) (joe-kirby) and South Korea recruits primary teachers from the top 5% of the academic cohort. (joe-kirby) Its central claim — that "the quality of an education system cannot exceed the quality of its teachers" (ResearchGate) — became an axiom of global education policy.

The 2010 follow-up examined **20 school systems** (McKinsey & Company) and proposed a four-stage improvement model. Systems moving from **poor to fair** need scripted lesson plans and centralized control. (IATED Library) Those going from **fair to good** require data systems and accountability structures. The **good-to-great** transition demands increased teacher autonomy and peer learning. The **great-to-excellent** stage calls for distributed leadership and innovation. (IATED Library) (mckinsey) McKinsey identified six interventions common to all stages: building instructional skills, assessing students, improving data systems, facilitating policy frameworks, revising curriculum, and ensuring appropriate teacher compensation. (McKinsey & Company) (mckinsey) A key finding was that **significant gains can be achieved in six years or less**, (Jordan Tinney) (McKinsey & Company) with median leader tenure in successful systems at 6–7 years. (McKinsey & Company +2)

Why the reports don't hold up to scrutiny

The criticisms are devastating and well-documented. Frank Coffield's 2012 *Journal of Education Policy* analysis identified **10 fundamental deficiencies**, Taylor & Francis Online including conflation of correlation and causation, a narrow definition of education quality limited to test scores, and an assumption that governments have "complete control of education policy, so that all 'actors' will do as the Ministry of Education decides." service Henry Braun of Boston College noted the reports were "written in a rather impressionistic style" and suffered from **survivorship bias** Springer — examining only successful systems without investigating whether failures had implemented similar strategies. Springer Robin Alexander called the framework "methodologically reductionist, elevating simple statistical correlation over the exploration of culture." University of York ResearchGate

Both reports were published as grey literature — **never peer-reviewed** — by a consulting firm with financial interests in education advisory services. Barber himself was Tony Blair's former Chief Adviser on Delivery before joining McKinsey. Robinalexander Taylor & Francis Online The reports established McKinsey as a major player in global education consulting, opening contracts worldwide. ResearchGate Despite these limitations, their stage model remains the dominant framework for thinking about system improvement trajectories.

What the evidence actually shows about six popular interventions

Increasing teacher pay produces surprisingly weak direct effects

The strongest causal evidence comes from **De Ree, Muralidharan, Pradhan, and Rogers (2018)** in the *Quarterly Journal of Economics* — the only large-scale randomized controlled trial on unconditional salary increases. A permanent **doubling of teacher base salaries** in Indonesia improved teacher satisfaction and reduced outside employment, but produced **zero improvement in student learning outcomes** after 2–3 years. The estimate was precise enough to rule out even modest positive effects. Oxford Academic Vrije Universiteit Amsterdam

Cross-country correlational studies tell a different story. **Dolton and Marcenaro-Gutiérrez (2011)** found a positive association between relative teacher pay and student performance Sage Journals in PISA/TIMSS data, and García and Han (2022) found higher base salaries correlated with reduced achievement gaps in US districts. ResearchGate The reconciliation lies in the **mechanism**: higher pay may improve outcomes through recruiting better candidates into teaching, ERIC not through motivating incumbent teachers. The EEF Toolkit rates performance pay at just **+1 month of additional progress** with very low evidence security. Education Endowment Fou... Hattie's effect size for teacher performance pay is a mere **d = 0.05**, ranking 227th of 252 influences. visible-learning

The critical evidence gap is that **no RCT has tested whether higher initial salaries attract better candidates and improve outcomes over the long term** — the recruitment channel that cross-country evidence suggests matters most.

Class size reduction works under narrow conditions at enormous cost

The **Tennessee STAR experiment (1985–1989)** remains the gold standard — a fully randomized trial across 79 schools assigning ~6,500 students to small classes (13–17 pupils) or regular classes (22–25). (Ucsd) Effect sizes ranged from **d = 0.17 to 0.34**, (ERIC) with effects **roughly double for minority students**. (PubMed) Long-term follow-up found increased likelihood of taking college entrance exams. (NBER) This is genuine causal evidence.

Yet quasi-experimental studies consistently find null results. **Hoxby (2000)** used natural population variation in 649 Connecticut schools over 20 years and found **zero effect** of class size reductions from a base of 15–30 students, with estimates precise enough to rule out improvements of even 0.03 SD. (NBER) **Hanushek** has argued STAR results are biased upward (Stanford) by non-random attrition and Hawthorne effects. California's 1996 class size reduction program required hiring 25,000 new teachers quickly, many uncertified — the quality dilution offset any size benefit. (EdNC)

The EEF rates class size reduction at **+1 month** of additional progress overall, noting benefits appear only with reductions to **fewer than 20 (or even 15) students** and only if teaching practices actually change. (Education Endowment Fou…) Hattie's effect size is **d = 0.21** (rank 186/252). (visible-learning) The emerging consensus is that class size reduction is among the **least cost-effective** interventions available, (RS Network) with the Tennessee experiment's benefits achievable through far cheaper approaches.

School autonomy helps rich countries and can harm poor ones

The most important finding on school autonomy comes from **Hanushek, Link, and Woessmann (2013)** (Stanford) in the *Journal of Development Economics*: autonomy has a **significant positive effect on student achievement in developed countries** but a **negative effect in developing countries**. The key moderator is institutional quality — autonomy requires strong accountability, qualified teachers, and capable school leaders to function.

OECD PISA 2022 data reinforces this conditionality, identifying the three mechanisms that make autonomy work: teacher mentoring, monitoring through inspectors, and systematic recording of test results. (OECD) However, **Gunnarsson et al. (2009)** found that across 10 Latin American countries, positive effects of school autonomy disappeared entirely after correcting for endogeneity (IDEAS/RePEc) — suggesting much of the observed relationship is selection bias (better schools are granted more autonomy, not the reverse). Hattie does not provide a single effect size for school autonomy as a discrete category. The evidence quality is **moderate but heavily context-dependent**, with almost no true causal studies.

School choice evidence has turned sharply negative in recent years

The school choice evidence has undergone a remarkable reversal. Early studies of the **Milwaukee voucher program** and meta-analyses like **Shakeel, Anderson, and Wolf (2021)** — covering 21 RCTs — found modest positive effects: **+0.27 SD in reading and +0.15 SD in math**, (Jay P. Greene's Blog) driven primarily by developing country programs. US-only effects were small: roughly +0.05 SD in reading. (Wiley Online Library)

However, the most recent large-scale US voucher programs have produced **large negative effects**. The **Louisiana Scholarship Program** RCT found substantially lower test scores in math, English, and science after four years. **Indiana's Choice Scholarship Program** showed persistent negative math effects lasting four years. **Ohio's EdChoice** produced "overwhelming evidence" of substantial negative test score effects. (Chalkbeat) As Brookings scholar Mark Dynarski summarized in 2017: "Four recent rigorous studies reached the same result: on average, students that use vouchers to attend private schools do less well on tests." (Brookings)

Charter schools show more nuanced results. **CREDO's 2023 national study** of 1.8 million charter students found modest gains of **+6 days in math and +16 days in reading** nationally, (Hoover Institution) but **urban charters performed substantially better** (+40 days math, +28 days reading). (Gatesfoundation) Effect sizes remain small in standardized terms (Chalkbeat) (d ≈ 0.01–0.08). Hattie rates school choice at **d = 0.12** (rank 208/252) and charter schools at **d = 0.09** (rank 217/252). (visible-learning)

One consistent bright spot: voucher programs produce more positive effects on **educational attainment** (graduation and college enrollment) than on test scores, with 4–21 percentage point gains in high school graduation across multiple studies. (Fordham Institute) The competitive effects on nearby public schools are also consistently positive, (Fordham Institute) though small (d ≈ 0.02–0.03). (Cascadepolicy)

Accountability systems boost math but narrow the curriculum

The definitive NCLB study by **Dee and Jacob (2011)** in the *Journal of Policy Analysis and Management* used a comparative interrupted time series design and found statistically significant increases in 4th-grade math performance — an effect size of **d = 0.23 by 2007**. Improvements appeared at both lower and top percentiles and extended to 8th-grade math among low-achieving groups. However, NCLB produced **no detectable effect on reading achievement** in any grade.

The costs were significant. **Neal and Schanzenbach (2010)** found accountability disproportionately helped "bubble kids" near the proficiency threshold while neglecting those well above or below it. **Koretz (2017)** documented that high-stakes test score gains often did not transfer to independent measures — a phenomenon known as score inflation. Instructional time shifted toward tested subjects at the expense of science, social studies, and arts. (Umich) Jacob (2005) found high-stakes testing in Chicago led to strategic increases in special education placements. Hattie rates external accountability at **d = 0.31** (rank 148/252), below his hinge point of meaningful impact. (visible-learning)

School leadership is the underrated intervention

Branch, Hanushek, and Rivkin (2013) used Texas longitudinal data to estimate that highly effective principals raise achievement by **d = 0.13 in math and d = 0.09 in reading**, (Stanford) adding **2–7 months of additional learning per year** in high-poverty schools. (Uark) The **Wallace Foundation's landmark 2004 synthesis** (Leithwood et al.) concluded that "leadership is second only to classroom instruction among all school-related factors." (Fordham Institute)

Robinson's (2008) meta-analysis of 26 studies identified five leadership dimensions, with the most impactful being **"promoting and participating in teacher learning and development" at d = 0.84** — roughly twice the effect of the other four dimensions. (Winginstitute) Hattie rates school leadership at **d = 0.32** (rank 144/252). (visible-learning) The key limitation is that most evidence measures variation in principal quality, not whether specific training programs improve principal effectiveness. The effects are **indirect** — principals influence outcomes through teacher quality, school culture, and resource allocation — making causal identification inherently difficult. (Taylor & Francis Online)

Hattie's rankings are widely cited but deeply problematic

John Hattie's *Visible Learning* project is the largest synthesis of education research ever attempted. (Routledge) The original 2009 edition synthesized **800+ meta-analyses** covering 52,637 studies and approximately 80 million students (Pcmac) across 138 influences. (visible-learning) The 2023 sequel expanded to **2,100+ meta-analyses**, 130,000+ studies, 300–400 million students, and 320+ influences. (Amazon) (Amazon UK) Hattie set a "hinge point" of **d = 0.40** — the average effect across all interventions — below which interventions are not considered to have "desired effects." (visible-learning) (Inspirasifoundation)

The top-ranked influences as of the 2017 update include **collective teacher efficacy (d = 1.57)**, self-reported grades (d = 1.33), teacher estimates of achievement (d = 1.29), response to intervention (d = 1.29), and Piagetian programs (d = 1.28). (visible-learning) (Getting Better) Among the interventions most relevant to policy, feedback ranks high at **d = 0.70** (rank 32), direct instruction at **d = 0.60** (rank 48), metacognitive strategies at **d = 0.60** (rank 46), teacher-student relationships at **d = 0.52** (rank 75), and professional development at **d = 0.41** (rank 117). (visible-learning) The structural and policy interventions discussed above — teacher pay (d = 0.05), class size (d = 0.21), school choice (d = 0.12), school finances (d = 0.21) (visible-learning) — all rank well below the hinge point.

Influence	Cohen's d	Rank (/252)	Above hinge?
Collective teacher efficacy	1.57	1	Yes
Feedback	0.70	32	Yes
Direct instruction	0.60	48	Yes
Metacognitive strategies	0.60	46	Yes
Teacher-student relationships	0.52	75	Yes
Professional development	0.41	117	Borderline
School leadership	0.32	144	No
External accountability	0.31	148	No
Homework	0.29	159	No
Class size	0.21	186	No
School finances	0.21	185	No
School choice	0.12	208	No
Charter schools	0.09	217	No
Teacher performance pay	0.05	227	No

The criticisms undermine the entire framework

The methodological critiques are severe enough to question whether Hattie's rankings should inform policy at all. **Snook et al. (2009)** argued that Hattie excludes study quality entirely: "Any meta-analysis that does not exclude poor or inadequate studies is misleading." ResearchGate **Bergeron and Rivard (2017)** in the *McGill Journal of Education* characterized the methodology as "pseudoscience," ResearchGate demonstrating that Cohen's d from the same experiment can range from 0 to infinity depending on calculation method. Blogger **Slavin (2018)** argued Hattie is "merely shoveling meta-analyses containing massive bias into meta-meta-analyses that reflect the same biases." Wordpress

The "apples and oranges" problem is concrete, not abstract. Under "feedback," Hattie combines Standley (1996) — about **background music on production lines** — with Rummel and Feinberg (1988) — about **giving students money and sweets as rewards** — as if both were classroom feedback interventions. He also misreported one study's effect size as +0.60 when it was actually −0.60. Blogger The homework finding of d = 0.29 masks a dramatic split: **d = 0.15 for elementary students versus d = 0.64 for secondary students.** Blogger Teacherhead "Self-reported grades" (d = 1.33) measures the correlation between student predictions and actual performance — it does

not mean asking students to predict grades improves outcomes. **Kraft (2020)** and Dylan Wiliam have argued that effect size benchmarks should vary by student age and study duration, making cross-category comparison fundamentally invalid. Blogger

In the 2023 sequel, Hattie himself acknowledged these problems, admitting the ranking was added "at the last minute" as an appendix and that "the greatest misinterpretation was the misuse of the ranking." Substack ASCD He now emphasizes asking "What works best, for whom, and under what conditions?" Routledge — a question his methodology cannot answer.

Structural reforms show results in 3–6 years; cultural transformation takes decades

The timeline evidence reveals a sharp distinction between **structural reforms** that affect entire cohorts immediately and **capacity-building reforms** that transform teaching quality over generations.

Poland provides the cleanest natural experiment. The 1999 reform replaced early vocational tracking with a comprehensive gymnasium system, delaying academic sorting by one year. PISA 2000 captured pre-reform students; PISA 2003 captured the first reform cohort. The gains were dramatic: **+11 points in reading and +20 points in math within just 3 years**. By 2006, Poland had gained 29 reading points and 25 math points — moving from below the OECD average to 9th globally. World Bank researchers attributed the gains primarily to **delayed tracking**, which gave likely vocational students approximately 100 PISA points (~1 full standard deviation) of additional learning. By 2012, Poland had gained 48 points in math from its 2000 baseline. TheGlobalEconomy.com However, after the 2017 reversal of reforms, **all gains were erased by 2022**. OECD

Ontario, Canada demonstrated that focused system reform can show results within **2–3 years**. Premier McGuinty's 2003 reforms ResearchGate — creating the Literacy and Numeracy Secretariat, setting focused goals (75% of students meeting standard; 85% graduation), and emphasizing capacity building over punitive accountability Springer — produced Michaelfullan a rise from **54% to 72%** of students meeting provincial standards and graduation rates climbing from **68% to 84%** over 11 years. Michael Fullan, who served as Special Advisor, EdCan Network concluded that "the entire system should show positive, measurable results within two or three years." EdWeek

Finland's story is fundamentally different — a 30+ year cultural transformation. The comprehensive school reform enacted in 1968 Centre for Public Impact was implemented incrementally from 1972 to 1977. Apsce Teacher education was elevated to master's-degree level in 1979. Streaming was abolished in the mid-1980s. Finland's PISA dominance in the early 2000s Springer reflected a **lag effect of three decades of systematic reform**. Its subsequent decline has been equally dramatic Statista — reading dropped from 547 to 490 University of Jyväskylä and math from 548 to 484 between 2006 and 2022, University of Jyväskylä the largest decline among top-performing countries. Springer

Singapore's trajectory spans nearly 60 years across five distinct phases, from survival-driven (1959–1978) through efficiency-driven (1979–1996) to ability-driven (1997–2011) and beyond. Brookings CommonWealth Magazine Its 2022 math score of **575 is the highest any country has achieved in any PISA domain.** OECD Data Pandas Even Singapore's bottom socioeconomic quartile scores above the overall OECD average — no other country can claim this.

System	Reform start	First measurable gains	Peak performance	Time to peak
Poland	1999	3 years (2003)	13 years (2012)	13 years
Ontario	2003	2–3 years (2005–06)	~11 years (2014)	11 years
Finland	1968/1972	1980s–1990s (domestic)	2000–2006 (PISA)	~34 years
Singapore	1959/1965	Incremental throughout	2015–2022	~50–57 years

McKinsey's benchmark of "six years or less" for significant gains appears well-supported for structural reforms starting from low baselines. McKinsey & Company mckinsey But reaching and sustaining the highest performance levels takes considerably longer and requires fundamentally different interventions.

Five empirically identified barriers explain why most reforms fail

Political time horizons are fatally mismatched with reform timelines

McKinsey found that the **median tenure of successful reform leaders is 6 years for strategic leaders and 7 years for political leaders.** In contrast, US urban superintendents average **3 years**, English education secretaries average **2 years**, and French education ministers average **2 years**. McKinsey & Company mckinsey This mismatch is perhaps the single greatest structural barrier to sustained improvement. McKinsey's 2024 *Spark & Sustain* report identifies "leadership discontinuity" as a primary failure mode, noting that "rapid electoral cycles and short tenures for ministers of education can lead to a whipsaw of priorities." McKinsey & Company

Teacher unions can block or enable reform depending on approach

Terry Moe's *Special Interest* (2011) documents how US teachers' unions engage in "the politics of blocking" — systematically vetoing reforms that threaten job security or autonomy. Hoover Institution Academia.edu The NEA grew from a politically inactive professional association in the 1950s to a billion-dollar organization and the largest delegate-sending organization to Democratic conventions by 1990. EdWeek Mexico's SNTE (1.4+ million members) Wikipedia exemplified union capture at its worst: in Oaxaca, **36% of teachers directly inherited their positions** from family members, Dissent and reform attempts triggered the largest teacher mobilizations in Mexican history.

Yet the counter-evidence is equally compelling. **Finland's Teachers' Union** was deeply involved in designing the 2014 national curriculum. Ontario's success depended on McGuinty's deliberate strategy of rebuilding trust with teachers after the previous adversarial administration. (Springer) The World Bank's 2019 analysis found that "countries most effective in introducing and sustaining reforms considered the needs of various stakeholders. Those that failed to get the buy-in of a key group at the outset faced difficulties in implementing reforms." (worldbank) Unions are neither inherently obstructionist nor inherently progressive — their role depends on whether they are treated as partners or adversaries. (ERIC)

Implementation fidelity is the critical bottleneck

A **RAND Corporation study (2006)** of Comprehensive School Reform found that **no school had fully implemented all core components** of its chosen reform model. (RAND) Teachers received about half the recommended initial training and one-quarter of recommended ongoing professional development. (County Health Rankings) At observed implementation levels, CSR "can be expected to have little effect on student achievement" (rand) (RAND) — despite over $2 billion in federal investment across 8,000+ schools. (rand) (RAND)

Fullan's research identifies the **"implementation dip"** — an expected decline in performance and confidence when encountering innovations requiring new skills — as normal (Learning Forward) but frequently misinterpreted as evidence of failure. (EdCan Network) His broader estimate is that factors collectively contribute to the **"demise of 75% of reform attempts."** (Northeastern University Li...) UNESCO's 2025 analysis identifies three recurring failure dynamics: reforms guided only by technical design without attention to change management, capacity gaps at the "middle tier" that prevent national policies from reaching classrooms, and absent feedback loops. (UNESCO)

Funding matters less than how it is spent

The World Bank's 2019 analysis found that "changes in government spending on education was not strongly correlated with long-term learning trends." (worldbank) McKinsey's 2007 report documented that Singapore, one of the world's top performers, **spends less on primary education than 27 of 30 OECD countries.** (McKinsey & Company +2) The problem is not typically total spending but allocation — recurrent costs (especially teacher salaries) dominate budgets, leaving little for the capacity building that reform requires.

Cultural context makes policy borrowing treacherous

Singapore's success rests on a "unique configuration of historical experience, institutional arrangements and cultural beliefs" that renders its portability limited (The Conversation) — 360 schools, (Improving Teaching) political stability under the same party since 1965, (Brookings) and deep cultural emphasis on education. (Springer) Finland operates on professional trust with no word for "accountability" in Finnish. (World Economic Forum) Research consistently shows that borrowed policies become "almost unrecognisable from the policy lender's perspective" due to contextual factors. (The Hechinger Report +2) Hwa (2022) demonstrated that teacher accountability functions through entirely different psychological mechanisms in Finland (professional autonomy and

trust) versus Singapore (structured career pathways and performance management). Both work — but neither can be transplanted. Taylor & Francis Online

---

Neither system-wide nor targeted reform has a monopoly on evidence

The evidence does not resolve the debate between comprehensive system reform and targeted interventions — instead, it suggests each approach suits different conditions.

The case for **system-wide reform** rests primarily on observational evidence from Finland, Singapore, and Ontario, supplemented by Fullan's theoretical framework. All three achieved sustained improvement through a common pattern: few focused priorities, capacity building over punitive accountability, collaborative relationships with teachers, and leadership stability. Fullan's six fundamentals — developing the entire teaching profession, selecting a small number of priorities per school, balancing instruction and assessment, developing leadership, using nonpunitive intervention strategies, and integrating all components OECD — describe the Ontario approach that produced measurable gains within 2–3 years. EdWeek

The case for **targeted interventions** rests on stronger causal evidence. **Borman et al.'s (2003) meta-analysis** of 29 Comprehensive School Reform models found promising overall effects, with schools implementing CSR for 5+ years showing "particularly strong effects." ResearchGate **Success for All** has multiple RCTs showing positive reading effects. County Health Rankings A 2016 meta-analysis found that CSR schools "substantially to completely eliminated" Black-white achievement gaps in elementary and middle schools. PubMed Central However, RAND's finding that no school fully implemented its reform model raises questions about whether targeted programs can work at scale. rand RAND

McKinsey's stage model offers the most useful synthesis: the appropriate approach depends on the system's starting point. Jordan Tinney McKinsey & Company Systems at the **poor-to-fair** stage benefit from prescriptive, targeted approaches — scripted curricula, basic teacher training, direct accountability. McKinsey & Company At the **fair-to-good** stage, a mixture is needed. The **good-to-great** transition requires more systemic approaches — professional learning communities, increased autonomy, peer learning. And the **great-to-excellent** stage demands fully distributed, system-wide innovation. mckinsey The critical insight is that **interventions appropriate at one stage can be counterproductive at another** — centralized control that lifts a poor system will stifle an already-good one.

The evidence demands humility

Three conclusions emerge with reasonable confidence from this synthesis. First, **teaching quality is the most important in-school factor**, but we lack proven, scalable methods for systematically improving it — the Indonesia RCT shows that simply paying more doesn't work, SSRN SSRN and no country has found a way to replicate Finland's or Singapore's teacher recruitment without decades of cultural transformation. Second, **structural reforms that affect**

whole cohorts (like Poland's delayed tracking) can produce rapid, measurable gains, while capacity-building reforms take a generation or more. Third, **implementation, not design, is where reforms die** McKinsey & Company — the gap between policy intent and classroom reality swallows billions of dollars and decades of effort. UNESCO

What the evidence does not support is the confident prescriptivism of either the McKinsey reports or Hattie's rankings. The McKinsey model relies on survivorship bias and untested causal claims. Springer ResearchGate Hattie's effect sizes combine studies of such wildly different quality and type that, Site Title as Bergeron argued, they cross the line into pseudoscience. Wordpress The most honest assessment of education reform evidence is that **we know far less than we claim to know**, The Hechinger Report and the interventions with the strongest causal evidence (class size reduction through STAR, NCLB's math effects, voucher program failures) tend to be the most politically inconvenient for all sides. The field would benefit from fewer grand theories and more humility about the limits of transferring what works in one context to another. ERIC