



Growth Mindset Interventions and Academic Achievement: A Meta-Analysis of Implications for Educational Management

Endin Nasrudin¹, Dadang Sahroni², Eki Agustin³

¹ UIN Sunan Gunung Djati Bandung and endin.nasrudin@uinsgd.ac.id

² UIN Sunan Gunung Djati Bandung and dadang.sahroni@uinsgd.ac.id

³ Universitas Madani Nusantara and eqyagustine@gmail.com

Corresponding Author: endin.nasrudin@uinsgd.ac.id

ARTICLE HISTORY

Received Dec, 2025

Revised Dec, 2025

Accepted Dec, 2025

ABSTRACT

Growth mindset interventions are widely adopted in schools on the premise that teaching students that intellectual abilities can be developed will raise academic achievement, yet published syntheses disagree about how large, and how dependable, the effect is. This study conducted a systematic review and second-order meta-analysis of the strongest available evidence in order to clarify what the pooled effects imply for educational management. Following a PRISMA 2020-adapted protocol, we identified three large meta-analytic syntheses and four large randomized field experiments conducted in the United States, England, and Norway. Effect estimates were extracted, standard errors were reconstructed from confidence intervals, and estimates were compared through contrast tests, heterogeneity statistics, prediction-interval-based probabilities, and practical-significance translations. Pooled effects on academic achievement ranged from $d = 0.02$ (highest-quality studies) through $d = 0.05$ (all studies) to $d = 0.14$ (targeted subsamples with high implementation fidelity). The difference between the quality-restricted and the targeted estimate was statistically detectable ($z = 2.08$, $p = .038$), and between-estimate heterogeneity was high ($I^2 = 76.9\%$). Under a prediction interval of -0.08 to 0.35 , an approximately 89% chance of a positive effect in a new, comparable implementation is implied, but only about 63% of an effect of at least 0.10. We conclude that growth mindset interventions are best treated as a low-cost, context-dependent supplement rather than a system-wide strategy. Managers should target at-risk students, secure fidelity, align school structures with the message, and evaluate locally.

Keywords: Academic Achievement; Educational Management; Evidence-Based Decision Making; Growth Mindset; Meta-Analysis

INTRODUCTION

Few ideas from educational psychology have travelled as far, or as fast, as growth mindset. In its scholarly form, the idea derives from implicit theories of intelligence: people who believe that intellectual ability is a fixed entity tend to interpret difficulty as evidence of limited talent, whereas people who believe that ability can be developed tend to interpret difficulty as information about strategy and effort [1]. In its popular form, the idea was carried into staff rooms, school posters, parenting books, and commercial training programs by Dweck's [2] widely read volume on mindset, and it has since become one of the most visible psychological concepts in school improvement planning in many countries.

The scholarly case for intervening on mindsets rests on a plausible causal story. Students who hold a growth mindset are expected to seek challenge, persist after setbacks, and use feedback, and these behaviors are expected to accumulate into higher achievement over time [3]. Early field evidence appeared to confirm the story. Blackwell, Trzesniewski, and Dweck [4] reported that a

belief in malleable intelligence predicted a more favorable trajectory in mathematics grades across the transition to junior high school and that a short workshop on the malleability of the brain helped to reverse a decline in grades. Subsequent studies suggested that brief, online treatments could improve the performance of at-risk secondary students at very low marginal cost [5], especially after the materials were redesigned with adolescent input [6]. Because such treatments take little instructional time and almost no equipment, they became attractive to school leaders operating under tight budgets.

From the perspective of educational management, however, the central question is not whether a psychological mechanism is plausible but whether a specific decision, such as purchasing a program, scheduling a workshop series, or training teachers, is likely to repay the time, money, and attention it consumes. Research on school leadership consistently shows that the influence of leaders on student learning is exerted largely indirectly, through the conditions they create for teaching and learning and through the priorities they set for professional development and resource allocation [7], [8]. A school leader who commits scarce professional development days to an intervention with a trivial average effect forgoes alternative uses of that time. A leader who dismisses an intervention that works well under identifiable conditions forgoes a cheap opportunity. The quality, and the proper interpretation, of the evidence is therefore a managerial matter and not only a scientific one.

At present that evidence is contested. In the first large synthesis, Sisk, Burgoyne, Sun, Butler, and Macnamara [9] examined 43 intervention studies involving 57,155 participants and concluded that the overall effect on academic achievement was weak ($d = 0.08$), although students from low-income families and students at academic risk appeared to benefit more. Macnamara and Burgoyne [10] later reviewed 63 studies ($N = 97,672$), found a small overall effect ($d = 0.05$, 95% CI [0.02, 0.09]) that was no longer significant after correction for publication bias, and reported effects close to zero when only studies that changed mindsets as intended ($d = 0.04$) or only the highest-quality studies ($d = 0.02$) were considered. In the same issue of the same journal, Burnette et al. [11] analyzed 53 independent samples and reported that effects on academic achievement reached $d = 0.14$ (95% CI [0.06, 0.22]) in subsamples expected to benefit most when implementation fidelity was high, while emphasizing that the 95% prediction interval, from -0.08 to 0.35, indicates wide variation in what future interventions may deliver. Commentators have argued that the two sets of authors answered different questions and that average effects of interventions should be interpreted through the lens of heterogeneity [12], [13].

Large field experiments reinforce the ambiguity. The National Study of Learning Mindsets in the United States found that a brief online intervention improved the core-course grades of lower-achieving ninth graders by 0.10 grade points, with larger gains in schools whose peer norms supported challenging work [14]. A similar experiment in Norway replicated effects on challenge-seeking and advanced mathematics enrollment [15]. Yet a cluster randomized trial of a teacher-delivered program for 5,018 primary pupils in 101 English schools found no additional progress in reading, grammar, or mathematics [16].

Three gaps follow from this picture. First, the published syntheses are usually read one at a time, so practitioners encounter either an optimistic or a skeptical headline rather than a comparison of what each estimate means. Second, the estimates are rarely translated into quantities that managers can use, such as the number of students expected to be affected, the probability that a new

implementation will work, or the conditions that make a positive result more or less likely. Third, the implications for management functions such as targeting, fidelity assurance, procurement, school climate, and local evaluation have not been drawn together in one place.

This article addresses those gaps through a systematic review and second-order meta-analysis. It pursues four research questions. RQ1: What is the magnitude of the effect of growth mindset interventions on academic achievement according to the best available syntheses? RQ2: To what extent do estimates change when the evidence is restricted by study quality, by verified mindset change, or by participant subgroup and implementation fidelity? RQ3: How heterogeneous are the effects, and what does this imply for the chance that a new implementation will succeed? RQ4: What do the answers imply for educational management decisions? The contribution is threefold: a transparent side-by-side quantification of competing estimates, a set of practical-significance translations, and an evidence-based decision framework for school and system leaders. The findings are of immediate use to principals, foundation and ministry planners, and the providers of teacher professional development.

RESEARCH METHOD

Design and reporting framework

The study is a systematic review with a second-order (meta-synthesis) meta-analysis, that is, a quantitative comparison of effect estimates that were themselves produced by earlier meta-analyses, supplemented by a structured review of large primary experiments. Reporting was guided by the PRISMA 2020 statement [17] and adapted to the second-order design. This design was chosen deliberately. Re-pooling the original primary studies would require access to the full study-level data of each earlier meta-analysis and would double count the same trials across the syntheses. A second-order design instead asks how a decision maker should weigh the three most influential syntheses against one another and against the strongest experiments.

It must be stated plainly what the design does and does not do. It does not constitute a new exhaustive search and extraction of every primary study of growth mindset interventions. It does not produce a single new pooled estimate that could be treated as independent of the earlier ones, because the syntheses draw on overlapping primary studies. Its contribution lies in the transparent comparison of estimates and in their translation into managerial terms.

Eligibility criteria

Eligible sources met the following criteria, specified in a PICOS format. Population: school-age students, from primary school through upper secondary school and including early tertiary transitions as covered by the sources. Intervention: programs designed to teach or foster a growth mindset (the belief that intellectual abilities can be developed), delivered directly to students or through teachers, whether online, in workshops, or in classroom routines. Comparison: control or business-as-usual conditions. Outcome: academic achievement, such as grades, grade point average, test scores, or course progression. Study design: for syntheses, systematic reviews with meta-analysis in peer-reviewed journals that reported a pooled effect on academic achievement from interventions; for primary studies, randomized experiments with at least several hundred participants, reporting achievement or course-taking outcomes. Sources that analyzed only the correlation between mindset and achievement were excluded from the quantitative comparison, although the correlational literature is used in the discussion.

Information sources, search, and selection

Searches were conducted in October 2026 through web-based retrieval of indexed bibliographic records, publisher abstracts, institutional repositories, and trial-evaluation reports (including journal pages, PubMed records, university repositories, and the evaluation archive of a national education research foundation). Search strings combined the terms growth mindset, mindset intervention, implicit theories, academic achievement, meta-analysis, randomized controlled trial, and effectiveness trial. Additional large trials were identified from the retrieved pages that discussed or cited the syntheses. Titles and abstracts were screened against the criteria above. This purposive search yielded three syntheses and four primary experiments that met the criteria and were judged to be the most influential and the most methodologically informative. Because the search was purposive and not exhaustive, the possibility that relevant smaller or non-English-language studies were missed cannot be excluded (see Limitations).

Data extraction

For each included source the following were recorded: design, setting, number of studies (k) and participants (N), outcome type, the reported effect size with its 95% confidence interval or prediction interval, and any stated moderator findings or conditions of effectiveness. Extracted values were verified against the abstract or results text of the source. Where a source did not report a confidence interval (the overall estimate in Sisk et al. [9]), the point estimate is reported descriptively and is excluded from inferential computations. Primary experiments that reported effects in original units (for example grade points) are reported in those units and are not converted to standardized effect sizes, in order to avoid introducing errors through assumed standard deviations.

Quality and risk-of-bias considerations

The syntheses differ mainly in how they treat study quality. Macnamara and Burgoyne [10] coded adherence to best practices essential for causal inference and reported that authors with a financial incentive to report positive findings published significantly larger effects than authors without such an incentive. Burnette et al. [11] restricted the sample to randomized designs. For this review, quality was appraised at the level of the estimate, asking whether the estimate (a) was based on randomized designs, (b) was corrected for publication bias, (c) was conditioned on verified mindset change, and (d) was restricted to a prespecified subgroup or fidelity condition. These attributes determine the inferential status of each estimate and are used to interpret divergence (Table 1).

Statistical analysis

All analyses used standardized mean differences (Cohen's d). Standard errors were reconstructed from reported 95% confidence intervals as $SE = (\text{upper limit} - \text{lower limit}) / (2 \times 1.96)$, and z statistics as d / SE , with two-sided p values from the standard normal distribution. The difference between two estimates was tested as $(d1 - d2) / \sqrt{SE1^2 + SE2^2}$. Because the samples of the syntheses overlap, the sampling errors of the estimates are positively correlated; ignoring the covariance makes the variance of the difference larger than it truly is, so these tests are, if anything, conservative. They are nonetheless presented as descriptive contrasts and not as confirmatory tests, since the estimates also differ in the estimand they target.

Heterogeneity between the quality-restricted estimate and the targeted-subsample estimate was summarized with Cochran's Q and the I² statistic [18], computed under an inverse-variance

weighting and labeled illustrative because only two estimates were compared and because they are not independent. To express how much future implementations may vary, the 95% prediction interval reported by Burnette et al. [11] was converted to an approximate normal distribution (mean equal to the midpoint, standard deviation equal to the interval width divided by 2×1.96) and used to compute the probability that a new, comparable implementation would exceed given thresholds (0, 0.05, and 0.10). This follows the logic that a prediction interval, unlike a confidence interval, describes the spread of true effects across settings [19]. The probabilities are conditional on the targeted, high-fidelity condition and rest on a normality assumption.

Practical significance was assessed in three ways. First, effects were classified using the benchmarks proposed by Kraft [20] for causal studies with standardized achievement outcomes (below 0.05 small, 0.05 to below 0.20 medium, 0.20 or above large). Second, d was translated into the percentile standing of the average treated student within the control distribution, assuming normality, and into the binomial effect size display correlation, $r = d / \sqrt{d^2 + 4}$ [21]. Third, the expected number of additional students per 500 who would score above the control median was computed as $500 \times (\Phi(d) - 0.5)$, where Φ is the standard normal cumulative distribution function. Computations were performed in Python using standard numerical libraries; all inputs are reported in Tables 1 and 2 so that the analysis can be reproduced by hand.

RESULTS AND DISCUSSION

Characteristics of the evidence base

Seven sources met the criteria: three meta-analytic syntheses (reported as five estimates because the synthesis by Macnamara and Burgoyne yielded three nested estimates) and four large randomized experiments. Table 1 summarizes their design, scale, and principal findings. Together the syntheses cover tens of thousands of students across studies conducted in several countries, whereas the four experiments are concentrated in three high-income, mostly European or North American settings: the United States, England, and Norway. This geographic concentration matters for transferability and is revisited below.

Table 1 Characteristics of the Included Syntheses and Field Experiments

Source	Design and setting	Scale	Principal finding on achievement
Sisk et al. [9]	Meta-analysis of mindset interventions	$k = 43$; $N = 57,155$	$d = 0.08$ (overall); larger for low-income and at-risk students
Macnamara & Burgoyne [10], all studies	Meta-analysis, quality-coded; bias-corrected analyses	$k = 63$; $N = 97,672$	$d = 0.05$, 95% CI [0.02, 0.09]; not significant after publication-bias correction
Macnamara & Burgoyne [10], mindset changed	Subset with verified mindset change	$k = 13$; $N = 18,355$	$d = 0.04$, 95% CI [-0.01, 0.10]
Macnamara & Burgoyne [10], highest quality	Subset with the strongest design	$k = 6$; $N = 13,571$	$d = 0.02$, 95% CI [-0.06, 0.10]

Burnette et al. [11]	Meta-analysis of randomized interventions; 53 independent samples (all outcomes)	Targeted subsamples, high fidelity	$d = 0.14$, 95% CI [0.06, 0.22]; 95% PI [-0.08, 0.35]
Yeager et al. [14]	Preregistered national RCT, United States	$N = 12,490$ (65 schools in analysis)	Lower achievers +0.10 grade points; higher achievers $SMD = 0.01$; +3 percentage points in advanced math enrollment
Foliano et al. [16]	Cluster RCT, teacher-delivered program, England	101 schools; 5,018 pupils	No additional progress in reading, grammar, or mathematics
Rege et al. [15]	RCT (U-say), Norway	$N = 6,541$	Replicated effects on challenge-seeking and advanced math enrollment
Bettinger et al. [22]	RCT, one upper secondary school, Norway	$N = 385$	Mindset index +0.55 SD; benefits concentrated among fixed-mindset and lower-achieving students

Note. k = number of studies; N = participants; PI = prediction interval; SMD = standardized mean difference. Values as reported by the sources.

Two features of Table 1 shape everything that follows. The first is that the estimates are not competing measurements of the same quantity. The overall estimate of Macnamara and Burgoyne [10] describes the average of whatever interventions researchers have run, in whatever populations, with whatever fidelity. The estimate of Burnette et al. [11] describes an average for a deliberately selected condition, namely subsamples expected to benefit most and interventions implemented with high fidelity. The first answers the question "what has the field produced on average?", and the second answers "what can be expected when the intervention is well targeted and well delivered?" The second is the more relevant question for a manager who is able to choose targets and control delivery, but it is also the more vulnerable to selective reporting of subgroups, which is precisely the concern raised by critics.

The second feature is that the primary experiments differ in what they treat as the intervention. The American and Norwegian experiments tested short, standardized online modules delivered to students, whereas the English trial tested a teacher-delivered package that combined a one-day teacher training course with weekly lessons over eight weeks, and expected teachers to embed mindset messages in their daily practice [16]. Modality, intensity, and the identity of the deliverer therefore vary together with the setting, and this confounding limits the strength of any inference about which design feature explains the divergence in results.

Magnitude of pooled effects (RQ1)

Table 2 presents the inferential statistics reconstructed for the four estimates that reported confidence intervals, and Figure 1 displays them graphically alongside the Kraft [20] thresholds. The estimate across all 63 studies in Macnamara and Burgoyne [10] is small but statistically distinguishable from zero ($d = 0.05$, $SE = 0.018$, $z = 2.80$, $p = .005$). Restricting the evidence to studies that demonstrably changed mindsets reduces the point estimate slightly and removes statistical

significance ($d = 0.04$, $z = 1.43$, $p = .154$), and restricting it further to the six highest-quality studies reduces the point estimate to $d = 0.02$ with a confidence interval that comfortably includes values of either sign ($z = 0.49$, $p = .624$). By contrast, the estimate of Burnette et al. [11] for targeted, high-fidelity conditions is nearly three times the all-studies estimate and is clearly distinguishable from zero ($d = 0.14$, $SE = 0.041$, $z = 3.43$, $p < .001$). The earlier synthesis by Sisk et al. [9] lies between these values ($d = 0.08$), although, as noted in the Method, the absence of a reported confidence interval in the abstract prevents inferential comparison.

Table 2 Reconstructed Inferential Statistics for Pooled Estimates

Estimate	d	95% CI	SE	z	p
Macnamara & Burgoyne, all studies	0.05	[0.02, 0.09]	0.018	2.80	.005
Macnamara & Burgoyne, mindset changed	0.04	[-0.01, 0.10]	0.028	1.43	.154
Macnamara & Burgoyne, highest quality	0.02	[-0.06, 0.10]	0.041	0.49	.624
Burnette et al., targeted and high fidelity	0.14	[0.06, 0.22]	0.041	3.43	< .001

Note. $SE = (\text{upper limit} - \text{lower limit}) / (2 \times 1.96)$; $z = d / SE$; two-sided p . Sisk et al. [9] is excluded because no confidence interval was available.

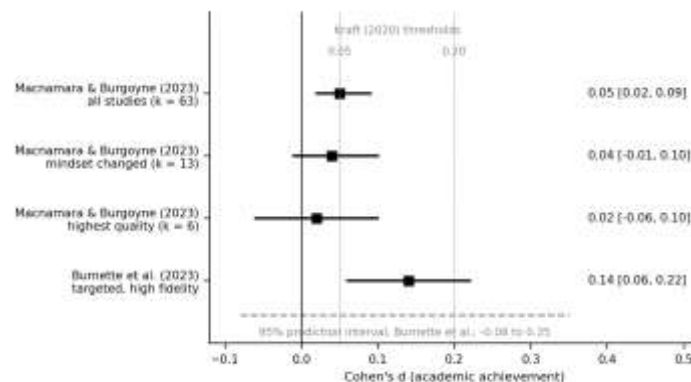


Figure 1 Pooled Estimates (Cohen's d, 95% CI) of Growth Mindset Interventions on Academic Achievement

Three observations follow. First, every confidence interval for an all-studies or quality-restricted estimate has an upper limit of 0.10 or lower. If the skeptical syntheses are right, the largest plausible average benefit of a typical intervention is on the order of one-tenth of a standard deviation, and the most credible central values are two to five hundredths. Second, the verified-mindset-change subset is smaller (13 studies, $N = 18,355$) than the full sample and yields a smaller, not larger, point estimate. This is theoretically awkward for the strong form of mindset theory, which predicts that interventions that succeed in changing the mediating belief should show larger achievement effects. Macnamara and Burgoyne [10] interpret the pattern as evidence that observed effects are not clearly produced through the proposed mechanism. A more charitable reading, offered by the proponents, is that small, noisy, and heterogeneous effects are statistically difficult to detect in subsets of this size [13]. Both readings remain compatible with the data.

Third, the point estimates should not be read as a ranking of the credibility of the authors but as a reflection of the eligibility rules and analytic frameworks. The sensitivity of the pooled value to such choices is itself a finding. Pooled effects in psychological intervention research are known to be sensitive to the inclusion of small studies, to multiple outcomes per sample, and to the treatment of subgroup analyses, which is one of the arguments for evaluating heterogeneity explicitly [12].

Why the estimates diverge, and how heterogeneous the effects are (RQ2 and RQ3)

Table 3 contrasts the targeted-condition estimate with each of the three nested estimates from Macnamara and Burgoyne [10]. The targeted estimate exceeds the all-studies estimate by 0.09 standard deviations ($z = 2.02$, $p = .043$), the mindset-change estimate by 0.10 ($z = 2.02$, $p = .044$), and the highest-quality estimate by 0.12 ($z = 2.08$, $p = .038$). Given the positive correlation between the sampling errors of overlapping syntheses, these tests are conservative, and the differences are therefore unlikely to be artefacts of sampling variance alone. They are better understood as the consequence of the different estimands and eligibility rules just described. Between the quality-restricted and the targeted estimate the heterogeneity statistics are $Q = 4.32$ ($df = 1$) and $I^2 = 76.9\%$. Although computed on only two non-independent estimates and therefore only illustrative, this value indicates that most of the observed dispersion between the two estimates exceeds what sampling error alone would be expected to produce.

Table 3 Contrasts Between the Targeted Estimate and the Macnamara and Burgoyne Estimates

Contrast (Burnette minus ...)	Difference	SE	z	p
All studies	0.09	0.045	2.02	.043
Mindset changed	0.10	0.050	2.02	.044
Highest quality	0.12	0.058	2.08	.038
Heterogeneity, highest quality vs. targeted	$Q = 4.32$	$df = 1$	$I^2 = 76.9\%$	-

Note. Contrasts are descriptive; overlapping samples make the tests conservative. Q and I^2 are illustrative for two non-independent estimates.

Four mechanisms jointly explain the divergence, and each has a managerial counterpart. The first is subgroup targeting. Both Sisk et al. [9], who reported larger benefits for students from low-income families and for students at academic risk, and Burnette et al. [11], who focused on subsamples expected to benefit most, converge on the proposition that average effects obscure larger effects for specific groups. The National Study of Learning Mindsets provides the cleanest confirmatory evidence: the effect on grades was concentrated among lower-achieving students, while the effect for higher-achieving students was approximately zero (standardized difference of 0.01) [14]. Applied to management, this suggests that a universal rollout dilutes whatever benefit exists and that a targeted rollout is the more efficient use of limited time.

The second mechanism is implementation fidelity. A program that is delivered in a way that departs from its design, for example by teaching the concept as a slogan without changing feedback practices, cannot be expected to produce effects that a faithful implementation might. Reviews of implementation science show that the degree of implementation is a robust predictor of outcomes across prevention and educational programs [23]. Dweck [24] herself cautioned against a "false growth mindset" in which effort is praised regardless of learning, and Yeager and Dweck [25] argue that the intervention should be judged against the standard of what was intended. For managers,

fidelity is not a research nicety but a controllable variable that can be audited through observation, teacher logs, and student measures.

The third mechanism is context and affordance. In the National Study, the gain among lower-achieving students was larger, 0.15 grade points in core courses and 0.17 grade points in STEM courses, in medium- and lower-performing schools in which the peer climate supported the pursuit of challenging work [14]. The Norwegian experiment of Rege et al. [15] produced a similar insight from a structural direction: it provided evidence that school policies governing when and how students can opt in to advanced mathematics determine whether students are able to profit from a mindset intervention. A message that "you can grow" has limited effect if the organization does not offer the opportunity to attempt more challenging work. These findings position the school as a system and the principal as the person responsible for the affordances that allow a psychological message to become behavior.

The fourth mechanism is bias in the production of the evidence. Macnamara and Burgoyne [10] found that authors with a financial incentive to report positive findings published significantly larger effects than other authors, and that the all-studies estimate lost significance after correcting for publication bias. This finding does not imply misconduct; it shows that the incentive structure around commercially distributed programs can distort the literature, and it implies that managers who procure programs should weigh independent evaluations more heavily than vendor-sponsored ones. The case for caution is strengthened by the English effectiveness trial, which was independently evaluated and found no effect.

The English result deserves a further comment because it illustrates that a null result does not by itself falsify the theory. The evaluators noted that most teachers in the comparison schools were familiar with growth mindset and that more than a third had received training days on it, so the trial compared a structured package with approaches rooted in the same theory and not with a counterfactual of no mindset activity at all [16]. The trial can therefore be understood as showing that a structured package added no measurable progress beyond the diffuse mindset messages already circulating in schools, a conclusion that is of direct practical interest to managers, who face the same crowded environment. It underlines that the marginal return to any additional mindset activity in a school that already works on mindsets may be especially low.

Turning to RQ3, the prediction interval of Burnette et al. [11] for achievement, from -0.08 to 0.35, conveys that even under favorable conditions future implementations may produce small negative effects or moderately large positive ones. Under the normal approximation described in the Method, the interval implies a mean of 0.135 and a standard deviation of 0.110. The probability that a new, comparable implementation yields a positive effect is then approximately .89, the probability that it exceeds 0.05 is approximately .78, and the probability that it reaches at least 0.10 is approximately .63. Put differently, roughly one new implementation in nine would be expected to show no benefit or a harm, and more than a third would fall short of 0.10, even in the favorable, high-fidelity condition. These probabilities do not apply to implementations that are untargeted or poorly delivered, where the lower estimates of Macnamara and Burgoyne [10] are the better guide, and they rest on a distributional assumption that the original authors did not make. They are nonetheless useful because they convert an abstract interval into the language of risk, which is how managers think about decisions.

Practical significance (RQ2)

Whether an effect is worth pursuing depends on its size relative to cost and alternatives, and not on statistical significance alone. Kraft [20] argues that the common benchmark of 0.20 or 0.40 standard deviations, often attributed to the synthesis of Hattie [26], is unrealistic for rigorously evaluated educational interventions with standardized achievement outcomes, and proposes that effects below 0.05 be regarded as small, from 0.05 to below 0.20 as medium, and 0.20 or above as large. By this standard, the quality-restricted estimate ($d = 0.02$) is small, whereas the all-studies, Sisk, and targeted estimates (0.05, 0.08, and 0.14) are medium, although the first sits at the lower boundary of that category. The skeptical and the optimistic syntheses thus place the effect on opposite sides of the line that separates "small" from "medium" in a framework that is already adjusted for the modesty of typical educational effects.

Table 4 Practical Translation of Effect Sizes

d	Kraft [20] category	Percentile of average treated student	r (BESD)	Variance explained	Extra students above control median per 500
0.02	Small	50.8	.010	0.01%	4.0
0.05	Medium (lower bound)	52.0	.025	0.06%	10.0
0.08	Medium	53.2	.040	0.16%	15.9
0.14	Medium	55.6	.070	0.49%	27.8

Note. Percentile = $100 \times \Phi(d)$; $r = d / \sqrt{d^2 + 4}$; extra students = $500 \times (\Phi(d) - 0.5)$. Assumes normality.

Table 4 translates each estimate into more tangible quantities. At $d = 0.14$, the average treated student would stand at about the 56th percentile of the control distribution, and a school with 500 treated students would expect about 28 additional students to score above the control median. At $d = 0.05$, the corresponding numbers are the 52nd percentile and about 10 students, and at $d = 0.02$, they are the 51st percentile and about 4 students. In terms of variance explained, even the targeted estimate corresponds to less than one-half of one percent. These magnitudes are small in absolute terms, and they would be negligible if the intervention required substantial resources. But they would be economically attractive if the incremental cost per student were very low, as is the case for brief online modules, and if the effect were concentrated among students whose outcomes matter most for school completion, such as those at risk of failing core courses. In the National Study, the intervention also raised the proportion of students enrolled in more advanced mathematics, by 3 percentage points, among both lower- and higher-achieving students [14]. Course-taking is a plausible pathway through which small immediate effects become larger long-run outcomes, a possibility that cannot be settled with the short-term data currently available.

A cautionary point is needed on interpretation. The conversion in Table 4 assumes that the standardized effects are comparable across the underlying outcome measures and that the distribution is approximately normal. The effect of the American experiment was reported in grade points, and the comparison with standardized values is therefore qualitative. The numbers should be read as orders of magnitude for planning, not as forecasts for a given school.

For planning purposes, three scenarios can be distinguished. In a pessimistic scenario, the implementation resembles the average of the existing literature or its highest-quality studies (d of about 0.02 to 0.05), the expected gain is between roughly 4 and 10 additional students per 500 above the control median, and the program is justified only if it costs almost nothing. In a central scenario, in which the program is aimed at lower-achieving students and delivered with checked fidelity (d of about 0.08 to 0.10, between the earliest synthesis and the lower end of the targeted estimate), the expected gain is roughly 16 to 20 students per 500. In an optimistic scenario, which resembles the conditions of the best-supported experiments (d of about 0.14 or more), the gain is about 28 students per 500 or more. A prudent manager would budget on the pessimistic scenario, plan delivery on the central scenario, and treat the optimistic scenario as an upside that must be demonstrated locally before it is claimed.

Evidence from large field experiments

The four experiments provide a complementary lens because they are less affected by the selection and reporting problems that complicate meta-analysis. The National Study of Learning Mindsets [14], a preregistered experiment with a nationally representative sample of United States schools, reported that lower-achieving students gained 0.10 grade points, with larger gains in supportive peer climates, and that effects were negligible for higher-achieving students. The Norwegian U-say experiment, conducted in nearly all schools in two counties ($N = 6,541$), replicated the effects on a behavioral measure of challenge-seeking and on advanced mathematics enrollment [15], and an earlier randomized experiment with 385 first-year students in one Norwegian upper secondary school found that the program raised a growth mindset index by 0.55 standard deviations and that benefits were concentrated among students with fixed mindsets or lower prior achievement [22]. The English trial, in contrast, found no impact on national reading, grammar, or mathematics tests, including for pupils eligible for free school meals [16]. Independent replication attempts of related laboratory and classroom effects have likewise reported null results [27].

Taken together, these experiments suggest a pattern that is more informative than any single mean effect. Positive findings come from experiments that used brief, student-facing, carefully designed online modules, that targeted adolescents transitioning into more demanding school settings, and that measured effects among students with room to improve or on outcomes such as challenge-seeking and course-taking. The null finding comes from a program delivered to younger pupils through teachers, whose outcome was a high-stakes national test, and whose comparison schools were already exposed to mindset ideas. The mapping is not a controlled comparison, and it would be a mistake to conclude that age, delivery mode, or outcome type individually explains the divergence. It is nevertheless sufficient to caution managers against extrapolating from one setting, particularly a high-income, Western one, to a different school system.

Whether mindset interventions work in Southeast Asian school systems is an open question. Among the sources examined here, we did not identify large randomized trials conducted in Indonesia or neighboring countries. The cultural meaning of effort, the structure of examinations, the organization of the school day, and teacher-student relations all differ from the settings of the main trials. Because the notion that ability grows with effort may be more culturally fluent in some societies than in others, transfer to other cultural contexts cannot be taken for granted. For ministries and school networks in such contexts, the sensible implication is that local piloting precedes scaling.

Implications for educational management (RQ4)

The evidence reviewed supports a posture that is neither enthusiastic adoption nor wholesale rejection. We distill it into seven management implications, summarized in Table 5 and elaborated below.

Table 5 Evidence-Based Management Implications

Management domain	Evidence signal	Recommended action
Investment sizing	Credible central estimates range from 0.02 to 0.14	Fund as a supplement to instruction; keep cost per student low
Targeting	Effects concentrated among lower-achieving and at-risk students	Pre-specify the target group and flag eligible students
Fidelity	Most favorable evidence comes from high-fidelity delivery	Written protocol, short training, delivery logs, check of mindset change
Structures and norms	Larger gains where peer norms and course policies afford challenge	Review placement, reassessment, and feedback policies
Teacher practice	Adult responses to failure matter; the English trial was null	Pair content with observable changes in feedback and task design
Procurement	Authors with financial incentives reported larger effects	Require independent, preregistered evaluations; discount post hoc subgroups
Local evaluation	Prediction interval spans zero (-0.08 to 0.35)	Staggered rollout with a pre-agreed decision rule

Prioritize and size the investment realistically. When the credible central estimates lie between 0.02 and 0.14, a growth mindset intervention is not a substitute for instructional improvement, curriculum quality, or the teaching of specific learning strategies. Leadership research indicates that the strongest school-level influence on student outcomes arises from promoting and participating in teacher learning and from ensuring a stable orderly environment [8]. Managers should regard a mindset program as a possible supplement to those core activities, funded in proportion to the modest expected effect. A rule of thumb that follows from Table 4 is that the program is worth considering when its full cost per student, including staff time, is low enough to be justified by a gain on the order of one to three percentile points in the average case.

Target instead of blanketing. The most robust moderator across the sources is the characteristics of the student. Lower-achieving and at-risk students benefited most in the National Study and were the group highlighted by Sisk et al. [9] and Burnette et al. [11]. A management information system that flags students whose grades fall below the school median, or who are at risk in core courses, allows leaders to concentrate the intervention where the expected return is highest. Universal delivery may remain defensible for reasons of equity and simplicity, but it should not be justified by the expectation of a school-wide gain in average achievement.

Treat fidelity as a managerial control variable. Because the high-fidelity estimate is the basis of the most favorable evidence, leaders who adopt an intervention without securing fidelity cannot expect to obtain that evidence's results. Practical safeguards include a written delivery protocol, brief training for those who deliver the sessions, observation or logs to record whether sessions were

completed as designed, and a short check of whether students understood the core message. The latter corresponds to the manipulation check that Macnamara and Burgoyne [10] found in only a subset of studies, and it allows a school to see whether the mediating belief changed at all before expecting achievement to follow.

Align structures and norms with the message. The evidence from the National Study and the Norwegian experiment shows that effects depend on whether the school gives students the opportunity and the social permission to act on a growth mindset. Principals can influence several affordances: course-placement and opt-in policies for advanced classes, grading and reassessment rules that allow students to improve work after feedback, the norms that colleagues and peers convey about taking on challenge, and the responses of teachers to errors. Such actions are within the control of management and can be pursued whether or not a formal program is purchased. Bryk and Schneider [28] show that relational trust in schools is a core resource for improvement; trust is also plausibly a condition under which students believe a message that ability can grow.

Invest in teachers as carriers of the message, with realism about the evidence. Haimovitz and Dweck [29] argue that children's mindsets are shaped by how adults respond to failure, which implies that teacher and parent practices should be consistent with the message students receive. Yet the English trial shows that training teachers and providing materials is not sufficient to improve test performance in a context in which mindset ideas were already widespread. Managers should therefore combine any mindset content with specific, observable changes in feedback and task design, and avoid the slogan-level practices that Dweck [24] warned about.

Govern procurement and evidence use. The association between financial incentive and larger reported effects [10] justifies a procurement standard: request independent, preregistered evaluations with randomized or credibly quasi-experimental designs, check whether the evaluation measured mindset change and achievement, and discount claims based on vendor-sponsored studies or on post hoc subgroup analyses. The same standard should apply to internal advocacy. A school that wishes to be evidence-informed needs a culture in which the question "what is the evidence, and who produced it?" can be asked without being interpreted as hostility.

Evaluate locally and plan for heterogeneity. Because the prediction interval spans zero, no synthesis can guarantee that an intervention will work in a particular school. Managers should therefore build evaluation into rollout. A staggered or lottery-based rollout, in which some classes or schools start earlier than others, creates a comparison group at little additional cost and gives leaders credible local evidence. Pre-specified outcomes, a pre-specified target group, and a decision rule agreed in advance, for example continuing only if the local estimate for the target group exceeds a threshold, protect against the temptation to find a subgroup that "worked" after the fact. This is the logic of the heterogeneity revolution advocated by Bryan et al. [12]: effects are not constants of the intervention but properties of the interaction between intervention, population, and context.

Beyond achievement, managers may weigh outcomes such as challenge-seeking, course enrollment, and engagement, for which the experiments reported more consistent effects than for test scores. The syntheses reviewed here also report stronger effects on psychological outcomes such as mental health in the work of Burnette et al. [11], who found a pooled effect of $d = 0.32$ in targeted, high-fidelity subsamples. If student well-being is part of a school's mission, the case for brief mindset programs may rest partly on these outcomes, though the present article, focusing on achievement, does not evaluate them further.

Limitations and directions for future research

Several limitations qualify the conclusions. First, the second-order design relies on summary estimates and cannot correct for biases that exist within the underlying primary studies, nor can it model the dependence among the overlapping syntheses. The contrast tests and heterogeneity statistics are therefore descriptive. Second, the search was purposive and conducted through web-based retrieval; it was not an exhaustive multi-database search with dual independent screening, and relevant smaller or non-English studies may have been missed. Third, one synthesis [9] did not provide a confidence interval in the source text available to us and is therefore reported descriptively. Fourth, the probabilities derived from the prediction interval rest on a normality assumption and are conditional on the high-fidelity, targeted condition. Fifth, the primary experiments report outcomes on different scales, and the American result is expressed in grade points, which restricts quantitative comparison. Sixth, the evidence is concentrated in high-income countries, and its generalization to Indonesia and other Southeast Asian systems is untested.

Future research should, in our view, pursue five directions. Large preregistered trials in Southeast Asian school systems would address the generalization gap. Studies that measure and report mindset change and implementation fidelity for every treated unit would allow a better test of the mediating mechanism. Research on the structural affordances of schools, such as placement policy and assessment practices, would clarify how managers can create the conditions under which the intervention works. Long-term follow-up would test whether small immediate effects on grades and course-taking accumulate into graduation and later outcomes. Finally, a full individual-participant-data meta-analysis, with all primary studies coded by independent teams and with random-effects modeling of moderators, would settle several of the questions that a second-order synthesis can only frame.

CONCLUSION

This article compared the most influential syntheses and the largest field experiments on growth mindset interventions and academic achievement, and translated their implications into the language of educational management. The pooled effect on achievement ranges from approximately 0.02 for the highest-quality studies, through 0.05 for all studies and 0.08 in the earliest synthesis, to 0.14 for targeted subsamples with high implementation fidelity. The differences among these estimates are not random noise but reflect different questions, different eligibility rules, and real heterogeneity in effects, with a 95% prediction interval from -0.08 to 0.35 for the favorable condition. Even there, roughly one new implementation in nine would be expected not to help, and more than one in three would fall below 0.10.

Three conclusions follow for practice. First, growth mindset interventions should be regarded as a low-cost, context-dependent supplement and not as a system-wide strategy for raising achievement; neither the claim of large gains for all students nor the claim that the idea is worthless is supported. Second, expected benefits are greatest when the intervention is directed at lower-achieving or at-risk students, delivered with demonstrable fidelity, and embedded in schools whose policies and peer norms give students opportunities to take on challenge. Third, managers cannot rely on a synthesis to tell them what will work in their own school, and they should therefore build local evaluation, independent evidence standards, and decision rules into every adoption.

Suggestions follow from these conclusions. School leaders should pilot interventions with pre-specified target groups and outcomes before scaling, invest equal effort in aligning school structures with the message, and avoid presenting mindset slogans as a substitute for instructional improvement. Policymakers and funders should support independent, preregistered trials in underrepresented regions, including Southeast Asia, and require reporting of mindset change and fidelity. Researchers should report heterogeneity and prediction intervals as a matter of routine, so that the decisions of managers rest on what is known about the variation of effects and not only on their average.

REFERENCES

- [1] C. S. Dweck and E. L. Leggett, "A social-cognitive approach to motivation and personality," *Psychological Review*, vol. 95, no. 2, pp. 256–273, 1988, doi: 10.1037/0033-295X.95.2.256.
- [2] C. S. Dweck, *Mindset: The new psychology of success*. New York, NY: Random House, 2006.
- [3] D. S. Yeager and C. S. Dweck, "Mindsets that promote resilience: When students believe that personal characteristics can be developed," *Educational Psychologist*, vol. 47, no. 4, pp. 302–314, 2012, doi: 10.1080/00461520.2012.722805.
- [4] L. S. Blackwell, K. H. Trzesniewski, and C. S. Dweck, "Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention," *Child Development*, vol. 78, no. 1, pp. 246–263, 2007, doi: 10.1111/j.1467-8624.2007.00995.x.
- [5] D. Paunesku, G. M. Walton, C. Romero, E. N. Smith, D. S. Yeager, and C. S. Dweck, "Mind-set interventions are a scalable treatment for academic underachievement," *Psychological Science*, vol. 26, no. 6, pp. 784–793, 2015, doi: 10.1177/0956797615571017.
- [6] D. S. Yeager et al., "Using design thinking to improve psychological interventions: The case of the growth mindset during the transition to high school," *Journal of Educational Psychology*, vol. 108, no. 3, pp. 374–391, 2016, doi: 10.1037/edu0000098.
- [7] K. Leithwood, A. Harris, and D. Hopkins, "Seven strong claims about successful school leadership," *School Leadership & Management*, vol. 28, no. 1, pp. 27–42, 2008, doi: 10.1080/13632430701800060.
- [8] V. M. J. Robinson, C. A. Lloyd, and K. J. Rowe, "The impact of leadership on student outcomes: An analysis of the differential effects of leadership types," *Educational Administration Quarterly*, vol. 44, no. 5, pp. 635–674, 2008, doi: 10.1177/0013161X08321509.
- [9] V. F. Sisk, A. P. Burgoyne, J. Sun, J. L. Butler, and B. N. Macnamara, "To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses," *Psychological Science*, vol. 29, no. 4, pp. 549–571, 2018, doi: 10.1177/0956797617739704.
- [10] B. N. Macnamara and A. P. Burgoyne, "Do growth mindset interventions impact students' academic achievement? A systematic review and meta-analysis with recommendations for best practices," *Psychological Bulletin*, vol. 149, no. 3-4, pp. 133–173, 2023, doi: 10.1037/bul0000352.
- [11] J. L. Burnette et al., "A systematic review and meta-analysis of growth mindset interventions: For whom, how, and why might such interventions work?," *Psychological Bulletin*, vol. 149, no. 3-4, pp. 174–205, 2023, doi: 10.1037/bul0000368.
- [12] C. J. Bryan, E. Tipton, and D. S. Yeager, "Behavioural science is unlikely to change the world without a heterogeneity revolution," *Nature Human Behaviour*, vol. 5, no. 8, pp. 980–989, 2021, doi: 10.1038/s41562-021-01143-3.
- [13] E. Tipton, C. Bryan, J. Murray, M. A. McDaniel, B. Schneider, and D. S. Yeager, "Why meta-analyses of growth mindset and other interventions should follow best practices for examining heterogeneity: Commentary on Macnamara and Burgoyne (2023) and Burnette et al. (2023)," *Psychological Bulletin*, vol. 149, no. 3-4, pp. 229–241, 2023, doi: 10.1037/bul0000384.
- [14] D. S. Yeager et al., "A national experiment reveals where a growth mindset improves achievement," *Nature*, vol. 573, no. 7774, pp. 364–369, 2019, doi: 10.1038/s41586-019-1466-y.
- [15] M. Rege et al., "How can we inspire nations of learners? An investigation of growth mindset and challenge-seeking in two countries," *American Psychologist*, vol. 76, no. 5, pp. 755–767, 2021, doi: 10.1037/amp0000647.
- [16] F. Foliano, H. Rolfe, J. Buzzeo, J. Runge, and D. Wilkinson, *Changing mindsets: Effectiveness trial*. London, UK: Education Endowment Foundation, 2019.
- [17] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, Art. no. n71, 2021, doi: 10.1136/bmj.n71.

- [18] J. P. T. Higgins and S. G. Thompson, "Quantifying heterogeneity in a meta-analysis," *Statistics in Medicine*, vol. 21, no. 11, pp. 1539–1558, 2002, doi: 10.1002/sim.1186.
- [19] J. Int'Hout, J. P. A. Ioannidis, M. M. Rovers, and J. J. Goeman, "Plea for routinely presenting prediction intervals in meta-analysis," *BMJ Open*, vol. 6, no. 7, Art. no. e010247, 2016, doi: 10.1136/bmjopen-2015-010247.
- [20] M. A. Kraft, "Interpreting effect sizes of education interventions," *Educational Researcher*, vol. 49, no. 4, pp. 241–253, 2020, doi: 10.3102/0013189X20912798.
- [21] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein, *Introduction to meta-analysis*. Chichester, UK: Wiley, 2009, doi: 10.1002/9780470743386.
- [22] E. Bettinger, S. Ludvigsen, M. Rege, I. F. Solli, and D. Yeager, "Increasing perseverance in math: Evidence from a field experiment in Norway," *Journal of Economic Behavior & Organization*, vol. 146, pp. 1–15, 2018, doi: 10.1016/j.jebo.2017.11.032.
- [23] J. A. Durlak and E. P. DuPre, "Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation," *American Journal of Community Psychology*, vol. 41, no. 3–4, pp. 327–350, 2008, doi: 10.1007/s10464-008-9165-0.
- [24] C. S. Dweck, "Carol Dweck revisits the "growth mindset"," *Education Week*, vol. 35, no. 5, pp. 20–24, 2015.
- [25] D. S. Yeager and C. S. Dweck, "What can be learned from growth mindset controversies?," *American Psychologist*, vol. 75, no. 9, pp. 1269–1284, 2020, doi: 10.1037/amp0000794.
- [26] J. Hattie, *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London, UK: Routledge, 2009, doi: 10.4324/9780203887332.
- [27] Y. Li and T. C. Bates, "You can't change your basic ability, but you work at things, and that's how we get hard things done: Testing the role of growth mindset on response to setbacks, educational attainment, and cognitive ability," *Journal of Experimental Psychology: General*, vol. 148, no. 9, pp. 1640–1655, 2019, doi: 10.1037/xge0000669.
- [28] A. S. Bryk and B. Schneider, *Trust in schools: A core resource for improvement*. New York, NY: Russell Sage Foundation, 2002.
- [29] K. Haimovitz and C. S. Dweck, "The origins of children's growth and fixed mindsets: New research and a new proposal," *Child Development*, vol. 88, no. 6, pp. 1849–1859, 2017, doi: 10.1111/cdev.12955.