The GET Curriculum and Student Well-Being: A Quasi-Experimental Evaluation

David Matischek

University of Graz

Author Note

This study was registered (open-ended registration, without preregistered hypotheses or an analysis plan) on the Open Science Framework (https://osf.io/pm6uv/, linked to project https://osf.io/hb79r/). The analysis script, the anonymized dataset, and the Quarto source of this manuscript are available as a self-contained reproducible package in the study's Open Science Framework project (https://osf.io/hb79r/). Generative AI tools (Claude, Anthropic) were used to assist with code development, data visualization, and language editing; all statistical analyses, results, and interpretations were verified by the author. The author declares no conflicts of interest.

Correspondence concerning this article should be addressed to David Matischek, Email: david.matischek@uni-graz.at

Abstract

This study evaluated the effects of the *Global Educational Transformation* (GET; https://www.get-ngo.com) curriculum—a timetabled school subject grounded in positive psychology—on psychosocial outcomes in secondary-school students. Using a quasi-experimental pre-post design across 13 schools and 59 classes in Austria and Germany (N = 528 after attention-check exclusion), we assessed 10 outcomes including self-esteem, flourishing, empathy, growth mindset, self-efficacy, classroom climate, locus of control, and learning orientation. The study was registered without hypotheses or an analysis plan, so all inferential tests are exploratory in their specification. Multilevel models revealed no intervention effect on any of the 10 outcomes, either uncorrected or after correction, and the largest group difference was d = 0.18. Equivalence tests placed three constructs within d = ± 0.20 bounds, and Bayesian model comparison favoured the null for all 10 outcomes (BF_{01} = 7.1–21.7; strong for nine, moderate for empathy). Classroom climate is a class-level rather than a student-level construct and was additionally analysed with the class as the unit (ICC(1) = 0.30, 44 classes); the estimate again showed no effect, but with an interval wide enough that no conclusion about this outcome is warranted. Classes that delivered the curriculum intensively did not differ from classes that never delivered it, so the null result cannot be attributed to weak implementation alone. Restricting the comparison to the schools that ran both intervention and control classes, which removes all between-school confounding, likewise produced no significant effect on any outcome. Six of 30 Baseline × Group interaction tests survived FDR correction, with students who began lower changing more favourably than those who began higher. The pattern also appears on constructs the curriculum does not address, and the design cannot separate a genuine moderation from differential regression to the mean, so it is reported as exploratory. The findings are discussed alongside recent large-scale null results in school-based mental health promotion.

Keywords: mental health promotion, school-based intervention, quasi-experimental design, social-emotional learning, positive psychology, positive education, null result, implementation fidelity, equivalence testing, baseline moderation

The GET Curriculum and Student Well-Being: A Quasi-Experimental Evaluation

Introduction

Mental health problems among children and adolescents represent one of the most pressing public health challenges of the 21st century. Epidemiological estimates suggest that approximately 13% of young people worldwide experience clinically significant mental health difficulties (Polanczyk et al., 2015), with many cases remaining undetected and untreated (Kieling et al., 2011). The World Health Organization has identified adolescent mental health as a global priority, noting that half of all mental health conditions emerge before the age of 14, yet the majority of affected young people receive no adequate intervention (World Health Organization, 2022). These concerns have intensified in recent years, with longitudinal data indicating secular increases in internalizing problems—including anxiety and depression—across high-income countries. The COVID-19 pandemic further exacerbated these trends (Racine et al., 2021; Ravens-Sieberer, Kaman, et al., 2023), reinforcing the urgency of preventive approaches that can reach young people at scale (World Health Organization, 2022).

The conceptualization of mental health has itself undergone a shift that bears directly on how interventions are designed and evaluated. The traditional disease model, in which mental health is equated with the absence of psychopathology, has been increasingly supplanted by dual-factor and complete state models that treat well-being and ill-being as partially independent dimensions (Keyes, 2005; Suldo & Shaffer, 2008). Keyes (2005) demonstrated empirically that the presence of positive mental health—characterized by emotional well-being, psychological well-being, and social well-being—is not merely the inverse of mental illness but constitutes a distinct dimension that independently predicts important life outcomes including productivity, physical health, and healthcare utilization. Suldo and Shaffer (2008) extended this framework to adolescent populations and identified four mental health groups among school-aged youth: those who were flourishing without psychopathology, those who were symptomatic yet still functioning well, those who were languishing despite the absence of diagnosable conditions, and those who

were both symptomatic and languishing. The languishing group—adolescents without clinical diagnoses but with low well-being—exhibited academic and social functioning comparable to that of the symptomatic group, suggesting that the absence of disorder is insufficient as a marker of healthy development. This dual-factor perspective provides the theoretical rationale for interventions that target well-being promotion rather than symptom reduction, and it anticipates the possibility that an intervention may successfully enhance flourishing without producing detectable changes in clinical symptom measures, or vice versa. The GET evaluation draws on this framework by measuring positive constructs—self-esteem, self-efficacy, flourishing, empathy, and growth mindset—rather than clinical symptom inventories, thereby evaluating the curriculum on the dimension it was designed to address.

Given that children spend a substantial proportion of their waking hours in educational settings, schools have been increasingly recognized as strategic environments for delivering mental health promotion and prevention programs (Weare & Nind, 2011). School-based approaches offer several distinct advantages over clinic-based models: they provide universal access regardless of socioeconomic background, reduce the stigma associated with help-seeking, and can be integrated into existing institutional structures without requiring families to navigate external services (Domitrovich et al., 2017). Furthermore, schools enable the sustained delivery of curricula over extended periods, which may be necessary for meaningful psychological change to occur and consolidate. In their systematic review, Weare and Nind (2011) concluded that the most effective school-based programs are those that are comprehensive, embedded in the school culture, and implemented with fidelity over time. Domitrovich et al. (2017) further emphasized that social-emotional competence developed in school contexts serves as a protective factor against both internalizing and externalizing problems, positioning schools not merely as convenient delivery sites but as developmentally appropriate contexts for psychosocial intervention (Greenberg et al., 2017).

The developmental significance of the school context becomes particularly salient when considered alongside what is known about normative psychological trajectories during early and

middle adolescence. Eccles and Roeser (2011) characterized schools as one of the primary developmental contexts of adolescence, arguing that the fit—or misfit—between adolescent developmental needs and the school environment critically shapes trajectories of motivation, belonging, and psychological adjustment. This stage-environment fit perspective suggests that interventions delivered during this period must contend with powerful normative developmental processes that may either amplify or attenuate program effects. Self-esteem, for instance, follows a well-documented curvilinear trajectory across the lifespan: it is relatively high in late childhood, declines during adolescence as young people encounter new social comparison processes and pubertal changes, and then gradually recovers through late adolescence and into adulthood (Orth & Robins, 2014). An intervention targeting self-esteem in students aged 10 to 16 is therefore operating against a normative developmental headwind during precisely the period when self-evaluations are most vulnerable to decline. Similarly, perceptions of classroom climate tend to become more negative across the school year and across grade levels, as the novelty of a new school environment wears off and as increasing academic demands erode students' sense of belonging and support (Eccles & Roeser, 2011). These normative developmental patterns have two important implications for the present study. First, they suggest that maintaining stable levels of well-being during early adolescence may itself represent a meaningful intervention effect, even if it does not manifest as a pre-to-post increase. Second, they imply that effect sizes observed in this age group may be systematically smaller than those observed in younger children or older adolescents, a consideration that should inform the interpretation of null or small effects.

The most extensively evaluated framework for school-based mental health promotion is Social and Emotional Learning [SEL; Collaborative for Academic, Social, and Emotional Learning (2020)]. In a landmark meta-analysis of 213 universal school-based SEL programs, Durlak et al. (2011) reported significant improvements across multiple domains, with effect sizes ranging from $d = 0.22$ for conduct problems to $d = 0.57$ for social-emotional skills. These findings generated considerable optimism about the potential of SEL programs to produce broad, reliable benefits across diverse populations and settings. However, more recent evidence has substantially

tempered this optimism. Cipriano et al. (2023) conducted an updated meta-analysis covering SEL interventions published between 2008 and 2020 and reported considerably smaller average effect sizes than those found by Durlak et al. (2011), ranging from $d = 0.13$ to $d = 0.23$ depending on the outcome domain. These smaller estimates raise important questions about the role of selection bias and publication bias in the earlier literature. Taylor et al. (2017) examined follow-up effects of SEL programs and found that academic-performance benefits persisted at an average of 3.75 years post-intervention, though the magnitude of follow-up effects was smaller than immediate post-test effects, suggesting some attenuation over time.

Parallel to the SEL movement, positive psychology has offered a complementary framework for school-based intervention. Seligman et al. (2009) argued that well-being can and should be taught in schools, proposing that evidence-based exercises targeting character strengths (Peterson & Seligman, 2004; Proctor et al., 2011), positive emotions, engagement, and meaning can be integrated into regular curricula alongside traditional academic content (Seligman, 2011). The positive education movement has since spawned numerous programs, including the Penn Resiliency Program (Brunwasser et al., 2009) and the Geelong Grammar School model, among others (Seligman et al., 2005; Waters & Loton, 2019; White & Murray, 2015). In a comprehensive review of school-based positive psychology interventions, Waters (2011) identified elements that recur across the programs with the strongest evidence: explicit instruction in well-being skills, integration of those skills with academic content, teachers who are trained in the content and model it themselves, and—in the most far-reaching models—change at the level of institutional culture, policy, and everyday school practice (Norrish et al., 2013; White & Murray, 2015). The same review notes that the evidence base remains heavily skewed toward programs developed and evaluated in Anglo-American contexts.

Programs differ considerably in how many of these elements they realise, and the difference is a plausible source of heterogeneity in reported effects. The curriculum evaluated here implements several of them: it is a timetabled subject with explicit instruction, it is delivered by teachers who complete a dedicated training program, and it is supported by a

purpose-developed textbook. What it does not attempt is the whole-school transformation that the most comprehensive models pursue, in which well-being is embedded in institutional culture and staff practice beyond the subject itself. Whether a curriculum-centred model of this kind produces measurable effects without that broader institutional change is an open empirical question, and one the present study is positioned to address. All these programs share an emphasis on building psychological resources rather than merely preventing deficits, aligning with the broader shift in clinical and developmental psychology toward promoting flourishing alongside reducing psychopathology (Norrish et al., 2013; Seligman et al., 2005, 2009; Suldo & Shaffer, 2008).

Despite the generally favorable meta-analytic evidence, several high-profile studies have produced null results, challenging the assumption that school-based positive psychology and SEL programs are uniformly effective. The MYRIAD trial (Kuyken et al., 2022), the largest school-based mindfulness study to date with over 8,000 students across 85 schools in the United Kingdom, found no evidence that mindfulness training improved mental health or well-being in adolescents compared to teaching as usual. This null result was particularly striking because the trial was conducted with exceptional methodological rigor—cluster-randomized, adequately powered, and pre-registered—and because it contradicted a body of smaller, less rigorous studies that had reported positive effects. Similarly, Galla et al. (2024) reanalyzed data from 22 RCTs involving over 16,000 adolescents and concluded that universal school-based mindfulness interventions produced no significant effects across eight outcome categories. The growth mindset literature has become similarly emblematic of the replication challenges facing positive psychology interventions in schools. While Dweck (1999) originally proposed that teaching students about the malleability of intelligence could improve motivation and achievement—and Blackwell et al. (2007) provided early experimental evidence that a growth mindset intervention could improve academic trajectories across an adolescent transition—subsequent large-scale replications have yielded much smaller effects than initially reported (Yeager et al., 2019), and recent meta-analytic work has estimated the average effect of growth mindset interventions on academic achievement at approximately $d = 0.05$, a magnitude that is difficult to distinguish from

zero in any single study. A comprehensive meta-analysis by Macnamara and Burgoyne (2023) covering 63 studies with nearly 100,000 students reported a negligible overall effect of growth mindset interventions (d = 0.05) after correcting for publication bias. Systematic reviews of universal resilience-focused interventions have similarly found limited evidence of effectiveness (Dray et al., 2017; Lösel & Beelmann, 2003). These developments suggest that the field may have systematically overestimated the effects of universal school-based interventions, particularly when programs are delivered under real-world conditions rather than the carefully controlled circumstances of efficacy trials (Cipriano et al., 2023).

Programs of this kind are rarely evaluated against a single outcome, and there is a theoretical reason for that. They are not assumed to act on well-being directly. They rest on a chain in which instruction first changes what students believe about themselves and their capacity to influence outcomes, those beliefs then shape how students approach learning and one another, and only downstream does subjective well-being shift (Durlak et al., 2011; Seligman, 2011). The ten outcomes assessed here were chosen to cover successive links in that chain rather than to sample the affective domain broadly. A first group captures *agency beliefs*—generalized self-efficacy (Beierlein et al., 2012), school-related self-efficacy (Jerusalem & Satow, 1999), and internal versus external locus of control (Kovaleva et al., 2012)—the social-cognitive constructs through which, on Bandura's account, instruction is expected to translate into behavior (Bandura, 1997). A second group captures *motivational orientation*—implicit theories of intelligence (Blackwell et al., 2007; Dweck, 1999) and learning goal orientation (Spinath et al., 2002)—which specify how those beliefs are brought to bear on academic tasks. A third group is *social and relational*: empathy (Vossen et al., 2015) and perceived classroom climate (Rauer & Schuck, 2003), the interpersonal dimension that a timetabled subject taught in an intact class group is particularly positioned to affect and which corresponds to the relationships component of PERMA (Seligman, 2011). A fourth group contains the *distal well-being outcomes* by which such programs are ultimately justified—flourishing (Diener et al., 2010) and global self-esteem (Rosenberg, 1965)—alongside satisfaction with school performance as a subjective attainment indicator.

Assessing the chain rather than only its endpoint has a consequence for how a null result can be read. A curriculum that left distal well-being unchanged but shifted agency beliefs or motivational orientation would warrant a different conclusion than one in which nothing moved anywhere: the first pattern would suggest a mechanism engaging but not yet reaching the outcomes it is meant to serve, the second that the assumed mechanism did not engage at all. Reporting all ten outcomes together also serves a second purpose here. Because no primary outcome was designated in advance, presenting the complete set—rather than selecting one after inspecting the results—is what prevents the largest effect from being promoted to the study's headline finding after the fact.

Among these constructs, empathy merits closer attention because of its multidimensional structure and its distinctive developmental trajectory during adolescence. Davis (1983) proposed an influential model distinguishing cognitive empathy—the capacity to adopt another person's psychological perspective—from affective empathy—the tendency to experience emotional reactions congruent with another's emotional state. These two components are served by partially distinct neural and psychological processes and follow different developmental trajectories across adolescence (Eisenberg et al., 2006). Cognitive empathy, which depends on the maturation of perspective-taking abilities and executive function, tends to increase gradually throughout adolescence as prefrontal cortical regions continue to develop (Eisenberg et al., 2006). Affective empathy, by contrast, is relatively stable across the adolescent period, as the basic capacity for emotional resonance is established earlier in development. Gender differences in empathy are among the most robust findings in developmental psychology, with girls consistently scoring higher than boys on both self-report and behavioral measures of empathic responding, a difference that tends to widen during adolescence (Davis, 1983; Eisenberg et al., 2006). These developmental and gender-related patterns matter for interpreting any school-based evaluation that includes empathy: because perspective-taking capacity changes normatively across this age range, a comparison group is itself a moving baseline, so an observed group difference can reflect differing rates of normative development rather than an intervention effect in either direction.

Empathy was assessed here as one of ten outcomes, using an instrument whose items permit separate examination of the cognitive, affective, and sympathy components. The subscale analyses reported below were conducted after the composite result was known and are therefore exploratory; the developmental literature summarized here is offered as an interpretive frame for that result, not as a prediction that preceded it.

In the German-speaking countries (the DACH region: Germany, Austria, Switzerland), adolescent mental health has received increasing attention, with the longitudinal COPSY study documenting elevated rates of psychological difficulties among German-speaking youth (Ravens-Sieberer et al., 2021, 2022; Ravens-Sieberer, Devine, et al., 2023) and epidemiological work in Austria revealing substantial prevalence of mental health problems in adolescents (Wagner et al., 2017). Against this backdrop, the concept of teaching well-being as a dedicated school subject has gained traction through the *Schulfach Glueck* ("Happiness as a School Subject") initiative (Fritz-Schubert, 2008), developed by Ernst Fritz-Schubert and implemented in various forms across Austrian and German schools. Despite growing adoption, rigorous evaluation of this approach has been limited. The scarcity of rigorous evaluations in the DACH region mirrors a broader pattern observed internationally: large-scale national evaluations of school-based social-emotional programs have frequently produced disappointing results. The most prominent example is the national evaluation of the Social and Emotional Aspects of Learning (SEAL) program in the United Kingdom, in which Humphrey et al. (2010) assessed the program across approximately 40 secondary schools (22 intervention and 19 comparison schools) and found no significant effects on students' social-emotional skills, mental health difficulties, or school engagement compared to matched comparison schools. The SEAL null finding was attributed in part to the wide variation in implementation quality across schools and the difficulty of achieving meaningful whole-school culture change through a program that was often delivered superficially or inconsistently—a pattern of implementation heterogeneity that bears striking resemblance to the present study's context. An internal pilot evaluation of the GET curriculum, reported in an unpublished thesis (Hemmelmayr, 2025) (N = 173), found a significant effect only

for empathy (d = 0.30), with no significant changes in self-esteem, flourishing, self-efficacy, or growth mindset. Because this pilot was an internal, unpublished evaluation conducted within the GET project rather than an independent study, its figures could not be externally verified; together with its small sample and the lack of a randomized control condition at the school level, this precludes strong causal inference.

The *Global Educational Transformation* (GET; https://www.get-ngo.com) curriculum represents an attempt to embed positive psychology content into the regular school timetable through a dedicated subject called *Self-Development* (*Persoenlichkeitsentwicklung*). The curriculum covers topics including self-awareness, emotion regulation, growth mindset, empathy, self-efficacy, goal-setting, and cooperative learning skills, drawing on the PERMA framework of well-being (Seligman, 2011), whose operationalization and measurement in student populations has been described by Kern et al. (2015). Crucially, unlike supplementary SEL programs that exist alongside the regular curriculum as occasional workshops or add-on activities, GET is designed to be timetabled as a weekly lesson in participating schools, occupying a formal curricular slot equivalent to other academic subjects. This structural integration is important because Durlak and DuPre (2008) have demonstrated that implementation intensity and institutional embedding are key moderators of program effectiveness, with programs delivered as integral parts of the school day producing effects two to three times larger than those delivered as peripheral supplements. Whether this weekly implementation ideal can be realized consistently across diverse school contexts remains an open question, as implementation science research has consistently demonstrated that program fidelity varies substantially in real-world educational settings (Dane & Schneider, 1998; Fixsen et al., 2005).

Implementation variability is not merely a methodological nuisance to be controlled away; it can serve as a valuable window into the conditions under which programs succeed or fail (Durlak & DuPre, 2008; Fixsen et al., 2005). When programs are delivered with varying intensity across sites, researchers can examine whether outcomes track with implementation fidelity, providing quasi-experimental evidence about program mechanisms. We capitalize on this

naturally occurring variation in implementation intensity to test the dose-response proposition, using teacher-reported implementation data to construct a dose variable that captures the actual exposure each class received. This approach acknowledges the ecological reality that school-based interventions are rarely delivered with uniform fidelity and seeks to extract information from that variability rather than treating it solely as error.

A final theoretical consideration concerns the question of who benefits from universal interventions. Meta-analytic evidence from prevention science has consistently shown that intervention effects tend to be larger for participants who enter with higher risk or lower baseline functioning—a phenomenon variously termed the compensation hypothesis, the risk-moderation effect, or the aptitude-treatment interaction (Horowitz & Garber, 2006; Stice et al., 2009). Stice et al. (2009), in a meta-analysis of depression prevention programs, found that selective programs targeting high-risk youth produced substantially larger effects than universal programs, and that within universal programs, participants with elevated baseline symptoms showed the greatest improvement. Similarly, Yeager et al. (2019) found that the growth mindset intervention in the National Study of Learning Mindsets produced significant effects primarily among lower-achieving students, with school achievement level operating as a further moderator, suggesting that baseline context moderates intervention efficacy. If this pattern extends to school-based positive education curricula, an aggregate group comparison could in principle obscure benefits confined to students who enter with lower psychosocial functioning, while students who are already functioning adequately show no change. This literature provided the rationale for the baseline-moderation analyses reported below. Because the registration contained no analysis plan, those analyses were specified after data collection and are reported as exploratory; they are not presented as tests of an a priori hypothesis. A moderation pattern of this kind is also difficult to interpret on its own, because regression to the mean produces the same pattern in the absence of any differential treatment effect (Barnett et al., 2005; Nesselroade et al., 1980).

The present study evaluated the GET curriculum using a quasi-experimental pre-post

design across schools in Austria and Germany. The study was registered on the OSF as an Open-Ended Registration (pm6uv), which documents the design, the target constructs, and the procedures. It contains **no** directional hypotheses, no designation of primary outcomes, and no statistical analysis plan. We therefore state research questions rather than hypotheses, and we make the resulting limitation explicit at the outset: every inferential test reported here was specified after the data were collected. No result below should be read as confirming a prediction made in advance, and the *p*-values carry the corresponding inflation of error rates for post hoc testing.

Within that constraint, the study addressed four questions. **RQ1:** Does participation in the GET curriculum change the ten assessed psychosocial outcomes—spanning well-being, general and school-related self-efficacy, self-esteem, empathy, classroom climate, locus of control, growth mindset, learning orientation, and satisfaction with school performance—relative to non-participating classes, once pre-test differences and the nesting of students in classes and schools are accounted for? **RQ2:** If no difference is detected, is the evidence merely inconclusive, or does it positively support the absence of a practically relevant effect? This question calls for equivalence testing and Bayesian model comparison rather than a null-hypothesis test, because a non-significant *p*-value alone cannot distinguish the two cases. **RQ3:** Given the substantial variation in how much of the curriculum was actually delivered, does outcome change track the delivered dose? **RQ4:** Does any effect depend on where students start, that is, do students with lower baseline scores change differently from those with higher scores? RQ4 is the most exploratory of the four and, as noted above, is confounded with regression to the mean by construction; the analyses reported below therefore include a direct test of that alternative explanation rather than treating a moderation pattern as evidence of differential efficacy.

The registration additionally named the identification of differential effectiveness by socioeconomic background as a study goal. The published dataset does not contain a direct indicator of socioeconomic status; parental country of birth was collected and is used here as a proxy for migration background, which is related to but not interchangeable with socioeconomic position. That subgroup analysis is reported among the exploratory results, and the substitution is

listed among the deviations from the registration in Appendix E. Two further sets of analyses—individual-level reliable change and class-level outcome profiles—are reported as descriptive supplements; they address neither of the four questions directly but characterize the heterogeneity underlying the aggregate results.

Method

Participants

The final analysis sample comprised 528 students from 13 schools and 59 classes in Austria and Germany. Schools were located in both urban and rural settings, representing a range of school types common in the Austrian and German secondary education systems. Students ranged in age from 9 to 16 years (M = 12.2, SD = 1.5), with 53.8% identifying as female and two students identifying as diverse (coded as missing for gender-based analyses). Group membership is defined in two ways, and the distinction matters for reading the tables that follow. The *broad* coding assigns every student by enrolment: the class was either enrolled in the GET curriculum or it was not, giving IG n = 304 and CG n = 224. This is the coding shown in Figure 1 and Table 1, and under it no student is unassigned. The *strict* coding counts a class as intervention only if the curriculum was actually delivered; the 53 students in classes that were enrolled but reported no or minimal delivery are therefore set aside rather than counted as treated. The primary model M1 uses the strict coding and consequently rests on N = 468 to 472 per construct rather than on the full 528; Model M3 estimates the same contrast under the broad coding, and the two are reported side by side throughout. Table 1 summarizes the key demographic characteristics of the analysis sample.

Table 1
Sample Demographics

Characteristic	Value
Total N (analysis sample)	528
Schools	13
Classes	59
Age M (SD)	12.2 (1.5)
Age range	9–16
Female n (%)	284 (53.8%)
Male n (%)	242 (45.8%)
Intervention group (IG) n	304
Control group (CG) n	224

Note. IG = intervention group (classes enrolled in the GET curriculum); CG = control group. Percentages are column percentages within each group. Counts refer to the analysis sample after attention-check exclusion.

Sample Flow

Figure 1 presents the participant flow from initial data collection to the final analysis sample. Of 798 students assessed at pre-test, 19 were excluded because their identification code was missing, unusable, or a test entry. A further 184 could not be linked to a post-assessment, either because the student did not take part at post-test or because no post-test record could be assigned to them. Of the 595 linked pairs, 67 failed the attention check at one or both occasions (fewer than 2 of 3 directed-response checks correct), leaving 528 students in the analysis sample. Every one of these could be assigned to a binary study group; a class-level implementation level was determinable for all but 58 of them, whose classes returned no usable implementation report (see Table 2).

The analysis sample covers 13 of the 14 participating schools. One school is absent: its single participating class was a control class of eight students, none of whom completed the post-test, so no pair from that school could be formed. The remaining 13 schools and 59 classes constitute the final analytic sample, which is used unchanged for every analysis reported below.

Demographics by Implementation Level

Implementation levels varied substantially across schools and classes, as shown in Table 2. The largest group was *none/minimal* (0–2 reported teaching hours over the academic year), comprising all control classes plus intervention classes in which teachers reported minimal or no delivery despite initial enrollment.

The two tables partition the same students differently, which is worth stating explicitly. The 224 control-group students all fall into *none/minimal*, whereas the 304 intervention-group students distribute across *none/minimal* (53), *low* (29), *moderate* (9), *intensive* (155) and *unknown* (58). The 277 students in the *none/minimal* row of Table 2 are therefore the 224 control students plus the 53 intervention students whose classes never delivered the subject; the binary totals in Figure 1 and the implementation totals in Table 2 describe the same 528 students under two different partitions. This distribution underscores the implementation heterogeneity that characterizes the present study and motivates the distinction between binary group comparisons and dose-response analyses.

Table 2
Implementation Level by Study Condition

Implementation Level	IG	CG	Total	% of sample
None/Minimal	53	224	277	52.5
Low	29	0	29	5.5
Moderate	9	0	9	1.7
Intensive	155	0	155	29.4

Table 2
Implementation Level by Study Condition

Implementation Level	IG	CG	Total	% of sample
Unknown	58	0	58	11.0
Total	304	224	528	100.0

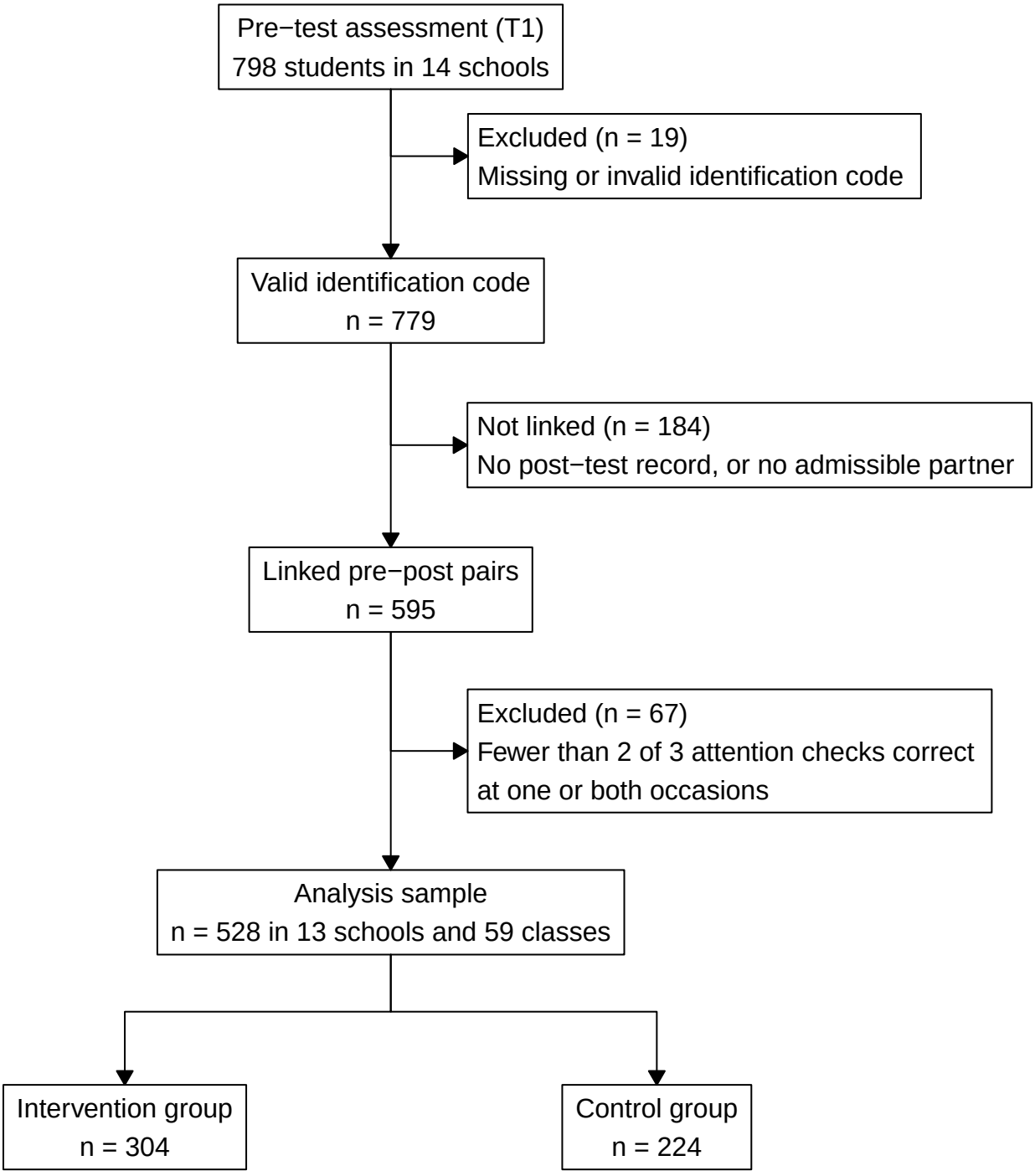
Note. Implementation level is a class-level rating of how much of the curriculum was actually delivered over the academic year: none/minimal = 0-2 reported teaching hours, low = sporadic delivery (approximately 3-4 sessions), moderate = regular but not weekly, intensive = approximately weekly throughout the year. Unknown = classes for which no implementation report was available. By definition every control class falls into none/minimal, since control classes received no curriculum; the none/minimal row therefore combines all control students with the intervention students whose classes did not deliver the subject. IG = intervention group; CG = control group.

Measures

Ten outcome variables were assessed at both pre-test and post-test. Where an authorized German version of an instrument existed, it was used. Where none existed, the items were translated into German for this study and, where the original wording was too demanding for students aged 9 to 16, simplified in vocabulary and sentence structure while preserving the content of the item. Simplification of this kind can shift an item's psychometric properties, so the reliabilities reported below are given for this sample rather than assumed from the source publications, and the full German wording of every item is reproduced in Appendix F so that the adaptations can be inspected. Unless otherwise noted, response scales were Likert-type. All reliability coefficients reported below are Cronbach's alpha computed separately for the pre-test and post-test administrations.

Figure 1

Participant flow from pre-test assessment to the final analysis sample, following CONSORT reporting conventions. Boxes on the right give the reason and number of exclusions at each stage. Group counts refer to the students assignable to a school and class.



School-Related Self-Efficacy (SWB). This five-item scale measures students' confidence in handling academic demands, drawing on social-cognitive theory (Bandura, 1997; Schwarzer & Jerusalem, 1995), and was rated on a 4-point scale ranging from 1 (*does not apply at all*) to 4 (*fully applies*). The scale was developed by Jerusalem and Satow (1999) and has been widely used in German-speaking educational research. Internal consistency was acceptable at both time points (α_{pre} = 0.79, α_{post} = 0.79), comparable to the range of .71 to .89 reported across student samples by Jerusalem and Satow (1999).

Flourishing Scale (SWS). Eight items capture subjective well-being and psychological flourishing [conceptually related to subjective happiness; Lyubomirsky and Lepper (1999)] on a 7-point scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*), adapted from Diener et al. (2010) using the authorized German translation (Esch et al., 2013). The scale showed good internal consistency (α_{pre} = 0.86, α_{post} = 0.88), closely matching the α = .87 reported by both Diener et al. (2010) and the German validation (Esch et al., 2013). Notably, the distribution was substantially left-skewed, with approximately 31 to 33 percent of students scoring at or near the scale maximum (skewness = -1.07 at pre-test), indicating a ceiling effect that limits the capacity to detect positive change.

Adolescent Measure of Empathy and Sympathy (AMES). Twelve items assess cognitive empathy, affective empathy, and sympathy in adolescents on a 5-point scale ranging from 1 (*never*) to 5 (*always*), based on Vossen et al. (2015). Internal consistency was good (α_{pre} = 0.82, α_{post} = 0.82), comparable to the subscale alphas of .75 to .86 reported by Vossen et al. (2015); a direct total-scale comparison is not possible because the original validation reports subscale rather than composite reliabilities. The scale showed an approximately normal distribution (skewness = -0.28). The three-factor structure of the AMES permits subscale analyses examining cognitive empathy, affective empathy, and sympathy separately.

General Self-Efficacy Short Scale (ASKU). Three items measure generalized self-efficacy beliefs on a 5-point scale ranging from 1 (*does not apply at all*) to 5 (*fully applies*), based on Beierlein et al. (2012; see also Beierlein et al., 2013 for the English-language validation).

A separate nonsense control item was embedded within the ASKU item block to serve as an attention check (see Appendix F). Internal consistency for the three substantive items was α_{pre} = 0.69, α_{post} = 0.73, somewhat lower than the ω = .81–.86 reported in the adult norm samples (Beierlein et al., 2013), as expected given the younger sample and the inherent limitation of three-item scales. The pre-test distribution contained 18 impossible values (likely data entry errors coded as sentinel values), which were set to missing prior to analysis.

Classroom Climate (CC). Six self-developed items measuring perceived cooperative classroom atmosphere on a 4-point scale ranging from 1 (*not true*) to 4 (*completely true*), conceptually based on the FEESS battery (Rauer & Schuck, 2003) (see also Stiglbauer et al., 2013 for the role of classroom climate in Austrian adolescent well-being). As a self-developed measure, the psychometric properties of this scale have not been independently validated. Internal consistency was acceptable (α_{pre} = 0.79, α_{post} = 0.88). This scale showed the largest pre-post decline across the full sample, a trend observed in both groups and likely reflecting normative developmental changes across the academic year.

Locus of Control (IE). Six self-developed items measuring internal versus external control beliefs were administered on a 5-point scale ranging from 1 (*strongly agree*) to 5 (*strongly disagree*), conceptually based on the IE-4 short scale [3 internal, 3 external items; Kovaleva et al. (2012); see Nießen et al. (2022) for the English-language IE-4 validation]. Due to the LimeSurvey response scale direction, higher composite scores indicate stronger *external* control beliefs; all IE results should be interpreted accordingly. Internal consistency was notably low (α_{pre} = 0.51, α_{post} = 0.54), though this is comparable to the ω = .53–.71 reported for the original IE-4 subscales (Kovaleva et al., 2012; Nießen et al., 2022), reflecting a known limitation of ultra-short scales rather than a sample-specific problem. All results involving the IE scale should be interpreted with caution given this limited reliability.

Implicit Theories of Intelligence (ITIS). Eight items measure students' beliefs about whether intelligence is fixed or malleable on a 6-point scale ranging from 1 (*strongly disagree*) to 6 (*strongly agree*), based on Dweck (1999; German factorial validation reported by Gnambs et al.,

2020). Internal consistency was acceptable (α_{pre} = 0.73, α_{post} = 0.76), lower than the α > .90 reported for adult samples (Dweck, 1999; Gnambs et al., 2020) but matching the α = .78 found in adolescent samples (e.g., Blackwell et al., 2007). One reverse-coded ITIS item additionally served as an attention check.

Self-Esteem (SES). Ten items from the Rosenberg Self-Esteem Scale (Rosenberg, 1965), using the German adaptation by Collani and Herzberg (2003), were rated on a 4-point scale ranging from 1 (*strongly disagree*) to 4 (*strongly agree*). Internal consistency was good (α_{pre} = 0.85, α_{post} = 0.88), exceeding the α = .72–.84 reported for the German adaptation (Collani & Herzberg, 2003) and consistent with the meta-analytic range of .77–.88 for adolescent samples. One reverse-coded SES item was embedded as an attention check.

Learning Goal Orientation (learning). Five self-developed items measuring intrinsic learning motivation, conceptually based on the SELLMO battery [Scales for the Assessment of Learning and Achievement Motivation; Spinath et al. (2002)], on a 4-point scale ranging from 1 (*not true*) to 4 (*completely true*), based on Spinath et al. (2002; 2nd ed., Spinath et al., 2012). One item was reverse-coded. Internal consistency was marginal (α_{pre} = 0.62, α_{post} = 0.62), below the α = .75 reported for the original SELLMO Lernziele subscale (Spinath et al., 2002) though within its published range of .58–.86; this represents a limitation for this outcome.

Subjective School Performance Satisfaction (subperform). A single item assessing students' self-rated satisfaction with their school performance was rated on a 4-point scale. As a single-item measure, no internal consistency coefficient can be computed; however, single-item scales can provide valid assessments of subjective states (Abdel-Khalek, 2006).

Attention and Comprehension Checks

Four control items (CNTRL) were embedded across the survey within the ASKU, SES, ITIS, and SWB scales. Three of these—within the ASKU, SES, and ITIS scales—were directed-response attention check items with unambiguously correct answers (e.g., "An elephant fits in my schoolbag," "Snow is mostly white"; see Appendix F for all item texts). These items

were scored as correct or incorrect. The fourth control item, embedded in the SWB scale ("I have never lied"), assessed social desirability rather than attention and was not included in the exclusion criterion. Participants were required to answer at least two of the three attention check items correctly to be retained in the analysis, a criterion that led to the exclusion of 67 participants (11.3% of the matched sample). Additionally, a comprehension check item assessed whether students understood the survey instructions on a 4-point scale.

Procedure

The study was approved by the Ethics Committee of the University of Graz (approval reference no. GZ. 39/224/63 ex 2024/25). Written informed consent was obtained from parents or legal guardians of all participating students, and written assent was obtained from the students themselves. Teacher data were collected as part of the broader GET evaluation but are not included in the present analysis, which focuses exclusively on secondary-school student outcomes. All procedures were conducted in accordance with the Declaration of Helsinki and applicable data protection regulations.

Data were collected at two time points: pre-test (September through November 2025) and post-test (May through June 2026), spanning approximately one academic year. Surveys were administered digitally via LimeSurvey at most schools, supplemented by paper-and-pencil versions at selected sites where digital administration was not feasible. All responses were pseudonymized using individually pre-assigned five-character alphanumeric codes that were distributed to students on laminated cards at each time point.

Implementation levels were coded from structured research protocols. At the end of the school year, the research assistants asked the teacher of each intervention class how much of the curriculum had actually been delivered, and recorded the answer in the class protocol. Four levels were distinguished: *none/minimal* (0–2 reported teaching hours per year), *low* (sporadic use, approximately 3 to 4 sessions across the year), *moderate* (regular but not weekly), and *intensive* (approximately one weekly lesson throughout the school year). These were recoded into a

numeric dose variable (0 to 3) for multilevel models. Control group classes were uniformly coded as dose = 0.

Pre- and post-test records were linked by a five-character identification code; where the recorded code was unavailable, records were linked on demographic and response information.

Data Analysis

All analyses were conducted in R (R Core Team, 2024) using the tidyverse ecosystem for data manipulation (Wickham et al., 2019), knitr for reproducible reporting (Xie, 2015), and ggplot2 for visualization (Wickham, 2016). Multilevel linear models were estimated using the `nlme` package (Pinheiro et al., 2023) to account for the nested data structure [students within classes within schools; Snijders and Bosker (2012)]. All models specified random intercepts for school and class. Random slopes for the group effect were not included because the treatment was assigned at the class level and the number of classes per school was insufficient for stable random slope estimation; this represents a limitation that may have produced slightly anti-conservative standard errors for the group coefficient. The class level carried the bulk of the clustering (see Table 3), whereas school-level ICCs were negligible for most constructs (e.g., 3.5% for school-related self-efficacy, 1.9% for flourishing, 0.3% for locus of control, and 1.2% for learning orientation; only empathy reached a non-trivial 0.2%). The school random intercept therefore contributed little variance and was retained for consistency of the nesting structure rather than because of substantial between-school heterogeneity; the near-zero school variance for several constructs is also the most likely source of the occasional convergence difficulties. As a robustness check, the primary (M1) models were refit without the school level. For the 9 constructs whose reduced model converged, every group-effect conclusion was unchanged (maximum shift in Cohen's d = 0.069; no construct crossed the p = .05 significance threshold in either direction). Missing data were handled by listwise deletion (the default in `nlme`), which is valid under the assumption of missing completely at random (MCAR). The plausibility of this assumption was supported by the observation that the pattern of missingness was similar across

groups, though formal tests of differential attrition were not conducted.

A sensitivity analysis was conducted to determine the minimum detectable effect size (Shadish et al., 2002). Accounting for the multilevel design effect [DEFF = 1.6, based on an average cluster size of 7 and a mean ICC of .10; see Hedges and Hedberg (2007) for ICC benchmarks in educational research], the effective sample size was approximately 316 for the main group comparison. Given this effective sample size, the minimum detectable effect size at 80% power was $d = 0.25$ at $\alpha = .05$. This means the study was adequately powered to detect effects at or above $d = 0.25$ but may have been underpowered for smaller effects. This limitation must be considered when interpreting the null main effects.

Model Specifications

The modelling strategy was not a stepwise search. An intercept-only model was fitted first for every outcome to quantify how much variance lies between classes and schools; those results are reported as intraclass correlations in Table 3 and establish that the nesting must be modelled. From there, each substantive model was specified in full and applied unchanged to all ten outcomes, rather than adding predictors sequentially and retaining those that improved fit. This choice is deliberate: with ten outcomes, stepwise selection would make the reported specification a function of the data and inflate the error rate in a way no correction accounts for. The trade-off is that model fit is not optimised per outcome, and the covariate set is the same everywhere whether or not it contributes.

Seven model variants were tested for each of the ten outcome constructs. Models M1 and M2 used a strict binary group coding (IG vs. CG only, excluding ambiguous cases). Models M3 and M4 used a broad binary coding that included all participants with any implementation information. Models M5 and M6 used the ordinal dose variable (0 to 3). Model M7 used categorical implementation levels. All models with covariates included age (centered) and gender as control variables.

Model M1 (Strict binary main effect with covariates). The primary model tested

whether assignment to the experimental group predicted post-test scores after controlling for pre-test scores, age, and gender:

$$\text{post}_i = \beta_0 + \beta_1 \cdot \text{pre}_i + \beta_2 \cdot \text{IG}_i + \beta_3 \cdot \text{age}_i + \beta_4 \cdot \text{gender}_i + u_{0,\text{school}} + u_{0,\text{class}} + \varepsilon_i$$

Model M2 (Strict binary Baseline × Group interaction with covariates). To test whether baseline levels moderated the group effect:

$$\text{post}_i = \beta_0 + \beta_1 \cdot \text{pre}_i + \beta_2 \cdot \text{IG}_i + \beta_3 \cdot (\text{pre}_i \times \text{IG}_i) + \beta_4 \cdot \text{age}_i + \beta_5 \cdot \text{gender}_i + u_{0,\text{school}} + u_{0,\text{class}} + \varepsilon_i$$

Models M3–M4 (Broad binary coding). Parallel to M1–M2 using the broader group classification to test robustness.

Model M5 (Dose-response with covariates). Using the ordinal dose variable (0 to 3):

$$\text{post}_i = \beta_0 + \beta_1 \cdot \text{pre}_i + \beta_2 \cdot \text{dose}_i + \beta_3 \cdot \text{age}_i + \beta_4 \cdot \text{gender}_i + u_{0,\text{school}} + u_{0,\text{class}} + \varepsilon_i$$

Model M6 (Baseline x Dose interaction with covariates). To test whether baseline levels moderated the dose effect.

Model M7 (Categorical implementation). Using implementation level as a factor with *none/minimal* as the reference category, to test whether specific levels differed from the reference. M5 and M7 answer the dose question under different assumptions: M5 treats the four implementation levels as equally spaced steps on a single scale, whereas M7 imposes no spacing assumption and estimates each level separately. Because implementation was rated at the class level, the dose effect in both models is identified between classes; its effective sample size is therefore the number of classes rather than the number of students, which the class random

intercept accounts for in the standard errors.

Assumption checks. The multilevel models assume normally distributed level-1 residuals, homogeneous residual variance across fitted values and groups, a linear relation between pre-test and post-test scores, and approximately normal random intercepts. Each assumption was checked for the primary model of every outcome. Residual normality was assessed by quantile-quantile plots together with skewness and excess kurtosis rather than by a formal normality test, because at this sample size such tests reject trivial departures that leave the estimates unaffected. Homoscedasticity was assessed by a Breusch-Pagan test of the squared residuals on fitted values and, separately, by the ratio of residual standard deviations between the intervention and control groups. Linearity was assessed by a likelihood-ratio test comparing the primary model against one containing a quadratic pre-test term. The distribution of the estimated class-level intercepts was summarized by its skewness and excess kurtosis. Full results are reported in Appendix J; departures are noted there and, where relevant, in the interpretation of the corresponding outcome.

Multiplicity across the primary outcomes. No primary outcome was designated in the registration. All outcomes therefore constitute a single family of tests, and the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) applied to the interaction tests was applied to the primary group contrasts as well. Both uncorrected p-values and corrected q-values are reported, so that the effect of the correction is visible rather than implicit.

Specification robustness (ANCOVA versus change score). Models M1 through M7 regress post-test on pre-test scores. In randomized designs this ANCOVA specification is more efficient than analyzing change scores, but in non-randomized designs the two estimate different quantities and can diverge—the phenomenon known as Lord's paradox (Miller & Chapman, 2001; Van Breukelen, 2006). Because the present groups were formed by self-selection and differ at baseline on several constructs, the group contrast was additionally estimated on the change score, with the same covariates and the same nesting structure. Reporting both makes visible any conclusion that rests on the modeling choice rather than on the data.

Pre-test scores in interaction models were centered at their grand mean to facilitate

interpretation. The significance level was set at $\alpha = .05$. Effect sizes are reported as Cohen's d (Cohen, 1988), computed as the model-estimated regression coefficient divided by the residual standard deviation from the multilevel model.

Interaction Probing

For constructs showing significant Baseline × Group interactions, the Johnson-Neyman technique was approximated by re-estimating the interaction model with the pre-test score re-centered at specific percentiles (10th, 25th, 50th, 75th, and 90th). At each percentile, the estimated group effect and its significance were obtained, allowing identification of the baseline range at which the intervention effect is statistically significant.

Equivalence Testing and Bayesian Analysis

To move beyond the absence of a significant effect toward positive evidence for the null hypothesis, equivalence tests using the Two One-Sided Tests (TOST) procedure (Lakens, 2017) and BIC-based Bayes Factors (Wagenmakers, 2007) were computed. Equivalence bounds were set at Cohen's $d = \pm 0.20$, corresponding to the lower end of the effect size distribution reported by Cipriano et al. (2023). The TOST procedure was computed using the group coefficient and its standard error from the primary multilevel model (M1), with Cohen's d derived as $d = \hat{\beta}_{\text{group}}/\sigma_{\text{residual}}$ and the 90% confidence interval constructed from the model-based standard error, thereby correctly accounting for the nested data structure. The BIC-based Bayes Factors were computed by comparing the full model (including baseline, group, age, and gender) against a reduced model omitting the group predictor, both estimated with maximum likelihood (ML) to ensure valid likelihood-ratio-based comparisons across models differing in their fixed-effect structure (Pinheiro et al., 2023).

Reliable Change Index

Reliable Change Index (RCI) values were computed following Jacobson and Truax (1991) to classify each individual's change as reliably improved, unchanged, or reliably deteriorated. Proportions of reliably improved participants were compared between IG and CG using chi-square tests. The index is defined for the individual case and has no established multilevel counterpart: its standard error comes from the test–retest reliability of the instrument, not from a variance decomposition, so there is no accepted way to let it absorb the clustering of students within classes. The classification is therefore reported as descriptive, and the inferential question—whether the intervention changed the *proportion* of students who improve reliably—is answered instead by a generalized linear mixed model on the binary improved/not-improved outcome with the same random-intercept structure as the main models, which does handle the clustering. This analysis is presented as supplementary throughout.

Sensitivity Analyses

To assess whether the findings depend on decisions about which participants to retain, both the primary group contrast and the Baseline × Group interaction were recomputed under six progressively stricter inclusion criteria, ranging from the standard attention-check rule to a combined filter on attention checks and pre-post response agreement. Appendix B lists the criteria and reports the full set of re-estimated models.

Level of Analysis for Classroom Climate

Nine of the ten outcomes describe the individual student. Classroom climate does not: it describes the class. Every student in a class rates the same object, so their ratings are not independent observations of separate phenomena but multiple indicators of one shared property—a referent-shift consensus construct in the terms of Chan (1998). Analysing it exactly like the other nine, one row per student with the class entered as a nuisance random effect, answers a question about students when the construct is a property of classes, and it counts each

class as many times as it has respondents.

Classroom climate was therefore additionally examined at the level at which it is defined. Two indices support that step (Bliese, 2000; LeBreton & Senter, 2008). ICC(1) is the proportion of variance lying between classes, that is, how much of a student's rating is determined by which class they are in. ICC(2) is the reliability of the *class mean*, the quantity a class-level analysis actually uses; because averaging over roughly nine students cancels much individual noise, a construct can have a moderate ICC(1) and still yield a highly reliable class mean, and values at or above .70 are conventionally treated as adequate for aggregation. Both were computed for all ten constructs so that classroom climate can be read in context. The group contrast for classroom climate was then re-estimated with one row per class: class mean post-test score regressed on class mean pre-test score and condition, with school as a random intercept, restricted to classes with at least five respondents. This model has far fewer observations than the student-level one and is reported as a check on the direction and the interval, not as a replacement.

Sociodemographic Moderators

Country of birth (student, mother, father) and language spoken at home were used to test whether the intervention worked differently across sociodemographic groups. Migration background was graded rather than dichotomised—no migration background, one parent born abroad, both parents born abroad, and student born abroad—because these situations differ in what they imply for a student, and a binary indicator collapses that distinction. Home language was treated separately, since a family can have migrated and speak German at home or not have migrated and speak another language. Group × moderator interactions were estimated for these two variables and for gender across all ten outcomes, with pre-test score and age as covariates and the same nesting structure as the main models. Because the questions were asked at different occasions, migration background is available for 440 of the 528 students and home language for all of them.

Preregistration and Deviations

The study was registered on the Open Science Framework (OSF pm6uv, linked to project hb79r; registered 07.09.2025, before T1 data collection). The registration is an Open-Ended Registration that describes the design, target constructs, sample, and procedures, but does not contain preregistered hypotheses, a primary outcome designation, or a statistical analysis plan. Consequently, the inferential strategy employed in this report—including the multilevel modeling approach, equivalence testing, Bayesian model comparison, false discovery rate (FDR) correction, and interaction probing—was developed after data collection and should be considered exploratory in its analytical specification, even though the target constructs and moderator interest (socioeconomic background) were registered as study goals. The study departs from the registration in five respects, each itemised with its severity in Appendix E. Two of them bear directly on how the results should be read and are therefore stated here. The registration described a single "newly developed, comprehensive scale"; before data collection this was replaced by a battery of established, validated instruments, which we regard as an improvement but which means the measurement level was not preregistered. And although differential effectiveness by socioeconomic background was a registered study goal, no direct indicator of socioeconomic status was collected, so that question can be addressed only indirectly through a proxy. The remaining three concern the classroom-climate instrument, the source of the achievement measure, and the grade range analysed. The qualitative component (teacher and student interviews) was conducted but is reported separately.

Results

Preliminary Analyses

Intraclass Correlations

Table 3
Intraclass Correlations (ICC) for Pre-Test Scores at School and Class Level

Scale	ICC School (%)	ICC Class (%)	ICC Total (%)
SWB	3.5	2.9	6.4
SWS	2.9	5.0	7.8
AMES	2.0	6.2	8.2
ASKU	2.9	3.6	6.5
CC	8.9	25.7	34.6
IE	0.1	4.9	5.0
ITIS	5.0	4.1	9.1
SES	3.3	5.8	9.1
learning	0.6	2.6	3.2

Note. ICC = proportion of total pre-test variance lying between clusters, estimated from an intercept-only multilevel model with classes nested in schools. School ICC = variance between schools; class ICC = additional variance between classes within schools. Values near zero indicate negligible clustering at that level. Scale abbreviations are defined in Appendix A.

Table 3 presents the intraclass correlations for pre-test scores. Meaningful clustering emerged at the *class* level for several scales—most strongly for classroom climate (ICC_{class} = 25.7%), followed by flourishing (5%), self-esteem (5.8%), and growth mindset (4.1%). School-level ICCs, by contrast, were small for nearly all constructs (see Table 3 and the Model Specifications section). This class-level clustering justifies the use of multilevel modeling to account for the nested data structure (Donner & Klar, 2000) and indicates that ignoring the clustering would have produced biased standard errors and inflated Type I error rates. Classroom climate is a special case rather than merely the largest value in the column: it is the only outcome that describes the class rather than the student, and its clustering is of a different order from the rest. It is examined separately at the class level below.

Descriptive Statistics and Baseline Equivalence

Table 4
Descriptive Statistics for All Outcome Scales: Full Sample

Scale	N	Pre M	Pre SD	Post M	Post SD	Change M	Change SD	Alpha Pre	Alpha Post
SWB	524	2.998	0.524	3.048	0.553	0.050	0.529	0.786	0.791
SWS	525	5.850	0.890	5.872	0.918	0.022	0.797	0.858	0.880
AMES	525	3.406	0.597	3.476	0.614	0.071	0.564	0.820	0.825
ASKU	524	3.489	0.719	3.609	0.755	0.121	0.760	0.694	0.725
CC	526	2.847	0.729	2.736	0.702	-0.111	0.730	0.787	0.881
IE	525	2.546	0.542	2.493	0.572	-0.053	0.599	0.508	0.537
ITIS	527	4.478	0.724	4.546	0.813	0.069	0.746	0.734	0.764
SES	528	2.879	0.569	2.917	0.591	0.039	0.482	0.851	0.880
learning	527	3.008	0.459	3.023	0.447	0.015	0.481	0.623	0.622
subperform	524	3.204	0.707	3.172	0.703	-0.032	0.778	NA	NA

Note. Full analysis sample, both groups combined. Change = post-test minus pre-test. Alpha Pre and Alpha Post = Cronbach's alpha at each measurement occasion, computed from item-level data with reverse-coded items recoded; alpha is not defined for subjective school performance (subperform), which is a single item. Scale abbreviations and response formats are given in Appendix A.

Table 4 presents descriptive statistics for all outcome scales across the full analysis sample. Mean scores were generally stable from pre-test to post-test, with change scores close to zero for most measures. Classroom climate showed the largest mean decline (M_{change} = -0.11), a trend observed in both groups and likely reflecting normative developmental changes across the academic year. Empathy showed a small mean increase (M_{change} = 0.07). Table 5 presents descriptive statistics separated by group assignment together with the baseline equivalence tests,

confirming that raw mean changes were similar between IG and CG across all scales.

Table 5
Pre-Post Descriptive Statistics and Baseline Equivalence by Group

Scale	Group	n	Pre M (SD)	Post M (SD)	Change M (SD)	t	p	d
SWB	IG	301	3.01 (0.54)	3.05 (0.56)	0.04 (0.58)	0.81	.419	0.07
	CG	223	2.98 (0.51)	3.04 (0.54)	0.06 (0.45)	–	–	–
SWS	IG	302	5.83 (0.93)	5.88 (0.95)	0.05 (0.80)	-0.71	.479	-0.06
	CG	223	5.88 (0.84)	5.87 (0.88)	-0.01 (0.79)	–	–	–
AMES	IG	301	3.36 (0.61)	3.41 (0.60)	0.06 (0.58)	-2.22	.027	-0.20
	CG	224	3.47 (0.58)	3.56 (0.62)	0.09 (0.54)	–	–	–
ASKU	IG	301	3.43 (0.73)	3.57 (0.76)	0.14 (0.81)	-2.32	.021	-0.20
	CG	223	3.57 (0.69)	3.66 (0.75)	0.09 (0.69)	–	–	–
CC	IG	302	2.93 (0.70)	2.75 (0.68)	-0.18 (0.72)	3.17	.002	0.28
	CG	224	2.73 (0.75)	2.71 (0.73)	-0.01 (0.74)	–	–	–
IE	IG	302	2.55 (0.55)	2.50 (0.56)	-0.06 (0.64)	0.36	.717	0.03
	CG	223	2.54 (0.54)	2.49 (0.59)	-0.05 (0.55)	–	–	–
ITIS	IG	304	4.44 (0.72)	4.49 (0.80)	0.05 (0.74)	-1.56	.119	-0.14
	CG	223	4.54 (0.72)	4.62 (0.83)	0.09 (0.75)	–	–	–
SES	IG	304	2.87 (0.58)	2.92 (0.58)	0.06 (0.49)	-0.64	.524	-0.06
	CG	224	2.90 (0.56)	2.91 (0.61)	0.01 (0.47)	–	–	–
learning	IG	303	3.01 (0.47)	3.01 (0.44)	-0.00 (0.51)	0.12	.906	0.01
	CG	224	3.00 (0.45)	3.05 (0.45)	0.04 (0.44)	–	–	–
subperform	IG	301	3.19 (0.72)	3.19 (0.72)	0.01 (0.85)	–	–	–
	CG	223	3.23 (0.69)	3.14 (0.68)	-0.08 (0.67)	–	–	–

Note. IG = intervention group; CG = control group. *M* and *SD* are raw (unadjusted) values; Change = post-test minus pre-test. The three rightmost columns report Welch's *t* test comparing

IG and CG on pre-test scores and are given once per scale, on the IG row. Cohen's d = (M_{IG} - M_{CG}) / pooled SD, so positive values indicate higher IG baseline scores. Cell n reflects complete pre- and post-test cases per scale. Subjective school performance (subperform) is a single item and was not included in the baseline-equivalence tests. Scale abbreviations are defined in Appendix A. Inferential group comparisons are the baseline-adjusted multilevel models of Model M1, reported under Primary Analyses.

The baseline columns of Table 5 report the equivalence tests between the intervention and control groups. The groups were comparable on most pre-test measures, but not on all of them: three constructs differ at $p < .05$, namely classroom climate (CC, $d = 0.28$, $p = .002$), general self-efficacy (ASKU, $d = -0.20$, $p = .021$), and empathy (AMES, $d = -0.20$, $p = .027$). These imbalances must be borne in mind when interpreting the corresponding outcomes—particularly classroom climate, where the intervention group starts higher, which also shows the highest class-level clustering (see Table 3) and features centrally in the implementation analyses. That the groups differ on several baseline measures is expected in a design where schools and classes selected themselves into the conditions; it tempers the assumption of quasi-experimental comparability and does not rule out unmeasured confounders that differ between participating and non-participating schools.

Figure 2 visualizes the mean unadjusted pre-post changes by group. Raw change scores were small for most constructs; the baseline-adjusted group contrasts—which are the basis for inference—are reported in Table 6.

Primary Analyses

Overall Intervention Effects (Model M1)

Figure 2

Mean unadjusted *pre-post change scores by group (IG vs. CG) for the nine multi-item outcome scales; error bars represent standard errors. Satisfaction with school performance is a single item and is therefore shown separately in Figure 3 rather than here. This figure is descriptive: the inferential tests are baseline-adjusted multilevel models (see Table 6 and Figure 3). In those adjusted models, no outcome showed a significant group difference.*

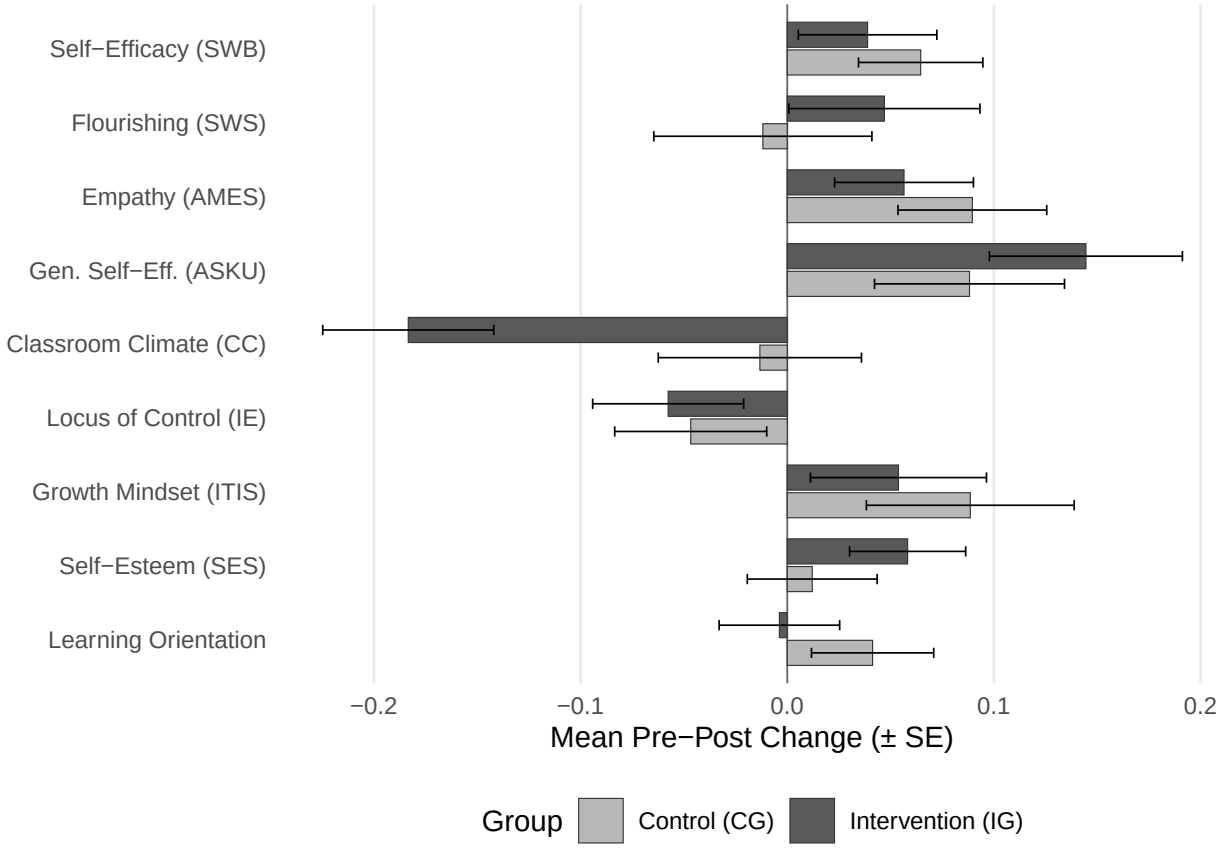


Table 6

Model M1 Results: Effect of Group Assignment (IG vs. CG) on Post-Test Scores With Covariates

Construct	N	Beta	SE	t	p	q	d	95% CI for d	Sig
SWB	469	0.014	0.052	0.272	.786	.939	.031	[-.190, .252]	
SWS	470	0.016	0.073	0.216	.829	.939	.022	[-.178, .222]	
AMES	470	-0.085	0.058	-1.475	.141	.939	-.179	[-.417, .059]	
ASKU	468	0.007	0.067	0.105	.916	.939	.011	[-.197, .219]	
CC	471	0.007	0.092	0.077	.939	.939	.013	[-.316, .342]	

Table 6

Model M1 Results: Effect of Group Assignment (IG vs. CG) on Post-Test Scores With Covariates

Construct	N	Beta	SE	t	p	q	d	95% CI for d	Sig
IE	469	-0.021	0.054	-0.385	.700	.939	-.042	[-.253, .170]	
ITIS	471	-0.065	0.077	-0.850	.396	.939	-.096	[-.318, .126]	
SES	472	0.005	0.048	0.100	.920	.939	.011	[-.206, .228]	
learning	471	-0.044	0.042	-1.056	.292	.939	-.114	[-.325, .097]	
subperform	469	0.051	0.076	0.679	.497	.939	.084	[-.159, .327]	

Note. IG = intervention group; CG = control group. Beta is the estimated group difference in post-test scores adjusting for pre-test score, age, gender, and the nesting of students in classes within schools. Effect size d = beta / residual SD, with its 95% confidence interval. q is the Benjamini-Hochberg adjusted p-value across the outcomes in this table; the same correction is applied to the interaction tests reported later, so that both families of tests are treated alike. Asterisks mark uncorrected $p < .05$; no effect remains significant after correction.

Table 6 presents the results from Model M1, the primary model testing the binary group effect (IG vs. CG) on post-test scores while controlling for pre-test scores, age, gender, and the nested data structure. Of the 10 outcomes estimable in this model, 0 reached significance before correction for multiple outcomes. The largest contrast was empathy (AMES; d = -0.18, p = .141), numerically favouring the control group; the same contrast under the broad binary coding gave a comparable estimate (Model M3: d = -0.17, p = .142). All other constructs showed effects of $|d| \leq 0.15$.

Because no primary outcome was designated in advance, all outcomes in this family constitute a single set of tests, and the Benjamini-Hochberg correction applied to the interaction tests below was applied here as well. After correction, 0 of the 10 effects remained significant.

The groups differed at baseline on three of the ten constructs (see Table 5). In non-randomized designs, an ANCOVA on post-test scores and a model of change scores estimate

different quantities and can diverge when groups are not equivalent at baseline (Miller & Chapman, 2001; Van Breukelen, 2006). The group contrast was therefore re-estimated on the change score with the same covariates and nesting structure. No construct changed its significance status under this choice; for empathy, the construct with the largest contrast, the two specifications agreed (ANCOVA: d = -0.18, p = .141; change score: d = -0.12, p = .324). Table 7 reports both specifications for every construct.

Table 7
Specification Robustness: Group Effect Estimated by ANCOVA and by Change Score

Construct	N	d ANC	p ANC	d chg	p chg	Sign flip	Sig. flip
SWB	472	.031	.786	.029	.766	no	no
SWS	473	.022	.829	.076	.481	no	no
AMES	473	-.179	.141	-.121	.324	no	no
ASKU	471	.011	.916	.117	.217	no	no
CC	474	.013	.939	-.134	.381	yes	no
IE	472	-.042	.700	-.041	.687	no	no
ITIS	474	-.096	.396	-.035	.767	no	no
SES	475	.011	.920	.082	.454	no	no
learning	474	-.114	.292	-.054	.597	no	no

Note. ANC = ANCOVA, Model M1, in which post-test score is regressed on pre-test score and the group indicator; chg = the change-score model, which regresses post-test minus pre-test on the group indicator. Both include age and gender as covariates and random intercepts for school and class. Negative values indicate higher scores in the control group. The two specifications answer subtly different questions and are known to diverge when groups differ at baseline; a difference between them is therefore informative rather than an error. Sign flip = the estimated direction differs between specifications; significance flip = the two specifications disagree about significance at $p < .05$. The table covers the nine multi-item scales; satisfaction with school

performance is a single item and is not included here. Scale abbreviations are defined in Appendix A.

Dose-Response Analyses (Model M5)

Table 8
Model M5 Results: Dose-Response Relationship (Implementation Dose 0–3) With Covariates

Construct	N	Beta	SE	t	p	d	Sig
SWB	516	0.001	0.017	0.058	0.954	0.002	
SWS	517	0.004	0.025	0.182	0.855	0.006	
AMES	517	-0.009	0.018	-0.471	0.638	-0.018	
ASKU	516	0.021	0.023	0.930	0.353	0.032	
CC	518	0.025	0.032	0.780	0.436	0.046	
IE	517	-0.020	0.019	-1.063	0.289	-0.039	
ITIS	519	-0.002	0.027	-0.081	0.935	-0.003	
SES	520	-0.006	0.016	-0.345	0.730	-0.013	
learning	519	0.003	0.014	0.201	0.841	0.007	
subperform	516	0.017	0.025	0.653	0.514	0.027	

Note. Dose coded as 0 (none/minimal), 1 (low), 2 (moderate), 3 (intensive). * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 8 presents dose-response results from Model M5, in which reported teaching hours enter as an ordinal predictor. No construct showed a significant dose coefficient, and every coefficient was smaller than $|d| = 0.05$, empathy included ($d = -0.02$, $p = .638$). Outcomes therefore did not track the amount of curriculum actually delivered in either direction. Because classes that never delivered the subject are indistinguishable from classes that delivered it weekly, the absence of a group effect cannot be attributed to insufficient dose alone.

Within-School Comparison

The comparison reported so far pools intervention and control classes across schools, so it is confounded with everything that differs between schools—catchment, staffing, and the self-selection that led a school to adopt the curriculum at all. Several schools ran both intervention and control classes. Restricting the contrast to those schools and entering the school as a fixed effect removes all between-school variation by construction, leaving a comparison between classes of the same school. This is the strongest quasi-experimental contrast the design supports, and it is reported here as a robustness check on the primary models. Table 9 gives the results for the 293 students in schools contributing at least five students to each arm.

The conclusion is unchanged: no outcome shows a significant group difference, and the largest effect size is $|d| = 0.21$. Removing school-level confounding therefore does not reveal an effect that the pooled analysis had masked. The estimates are less precise than the primary models—the widest confidence interval spans 0.75 units of d—because roughly half the sample contributes, so this analysis constrains large effects rather than small ones. Two further limits apply: the subsample is dominated by a small number of schools, and contamination works against this design in particular, since control classes in the same building are the most likely to encounter curriculum content informally.

Table 9
Within-School Group Effect: Intervention Versus Control Classes of the Same School

Scale	N	d	95% CI	p
SWB	291	.056	[-.256, .367]	.727
SWS	291	.212	[-.107, .532]	.194
AMES	292	-.102	[-.409, .205]	.514
ASKU	292	.100	[-.190, .391]	.499
CC	293	.172	[-.204, .548]	.371
IE	291	-.130	[-.437, .176]	.405

Table 9
Within-School Group Effect: Intervention Versus Control Classes of the Same School

Scale	N	d	95% CI	p
ITIS	292	.039	[-.283, .361]	.812
SES	293	.128	[-.182, .438]	.419
learning	293	-.029	[-.330, .272]	.850

Note. Group effect on post-test scores adjusting for pre-test score, with the school entered as a fixed effect and a random intercept for class, restricted to schools contributing at least five students to each arm. Only variation between classes of the same school contributes. Positive *d* favours the intervention group. Confidence intervals are wide because roughly half the sample contributes; the analysis constrains large effects, not small ones. Scale abbreviations are defined in Appendix A.

Equivalence Testing and Bayesian Analysis

Table 10
Equivalence Testing (TOST) and Bayesian Analysis Results for All Outcome Constructs

Construct	d	90% CI Lower	90% CI Upper	Equivalent	BF01	Interpretation
SWB	0.031	-0.155	0.216	Partial	20.8	Strong H0
SWS	0.022	-0.146	0.190	Yes	21.1	Strong H0
AMES	-0.179	-0.379	0.021	Partial	7.1	Moderate H0
ASKU	0.011	-0.164	0.186	Yes	21.4	Strong H0
CC	0.013	-0.263	0.288	Partial	20.1	Strong H0
IE	-0.042	-0.219	0.136	Partial	20.1	Strong H0
ITIS	-0.096	-0.282	0.090	Partial	14.7	Strong H0

Table 10

Equivalence Testing (TOST) and Bayesian Analysis Results for All Outcome Constructs

Construct	d	90% CI Lower	90% CI Upper	Equivalent	BF01	Interpretation
SES	0.011	-0.171	0.193	Yes	21.7	Strong H0
learning	-0.114	-0.291	0.064	Partial	12.5	Strong H0
subperform	0.084	-0.120	0.288	Partial	15.6	Strong H0

Note. Equivalence bounds set at $d = \pm 0.20$. BF_{01} = Bayes Factor in favor of the null hypothesis, computed from ML-estimated models (values > 10 indicate strong evidence for the null; 3–10 moderate evidence). TOST = Two One-Sided Tests procedure. All 10 outcome constructs are included.

The absence of significant *p*-values in the primary models indicates only that the data are insufficient to reject the null hypothesis; it does not constitute evidence that the null hypothesis is true. To provide positive evidence regarding the null, equivalence tests using the TOST procedure (Lakens, 2017) and BIC-based Bayes Factors (Wagenmakers, 2007) were computed, with equivalence bounds set at $d = \pm 0.20$, corresponding to the lower end of the effect size distribution reported by Cipriano et al. (2023) for SEL interventions.

Table 10 presents the combined results. Using multilevel model-based effect sizes and standard errors, 3 of the 10 tested constructs were formally equivalent to zero, meaning the 90% confidence interval for the effect size fell entirely within the equivalence bounds, so that any effect on them is statistically negligible. The remaining constructs showed partial equivalence, with one bound of the confidence interval crossing outside the equivalence region, which precludes a definitive equivalence conclusion from the TOST procedure alone. The confidence intervals are wider than a naive change-score analysis would give, because they take the clustered data structure into account and the effective sample is correspondingly smaller. Bayesian model comparison points the same way: BF_{01} ranged from 12.5 to 21.7, with empathy the least informative case

(BF_{01} = 7.1, moderate rather than strong evidence for the null). According to conventional benchmarks (Wagenmakers, 2007), BF_{01} > 10 constitutes strong evidence for the null and values between 3 and 10 moderate evidence; on that scale the data are at least seven times more likely under the null than under the alternative for every one of the 10 constructs. Formal equivalence for 3 constructs together with moderate-to-strong Bayesian evidence across all 10 supports the conclusion that the curriculum did not produce effects distinguishable from zero on the assessed outcomes. The qualification is one of precision rather than direction: with the present sample, effects of the magnitude meta-analyses report for universal school programs would not reliably have been detected, so the finding is best read as an absence of the moderate effects the curriculum was designed to produce rather than as proof of exactly zero. Figure 3 illustrates the effect sizes and their 90% confidence intervals, with the shaded equivalence region (d = ± 0.20) highlighting that most effects fall well within negligibility bounds.

Exploratory Analyses

Important methodological note. The analyses reported in this section are exploratory and were not prespecified in the study registration. They involve multiple comparisons across constructs, subgroups, and analytic approaches without correction for multiplicity. Accordingly, these results should be interpreted strictly as hypothesis-generating for future research.

Baseline × Group Interaction (Model M2)

Table 11
Model M2 Results: Baseline × Group Interaction Effects With Covariates

Construct	N	Beta (Pre x IG)	SE	t	p	Sig
SWB	469	-0.245	0.086	-2.861	0.004	**
SWS	470	0.025	0.077	0.321	0.748	
AMES	470	-0.084	0.075	-1.113	0.266	

Table 11
Model M2 Results: Baseline × Group Interaction Effects With Covariates

Construct	N	Beta (Pre x IG)	SE	t	p	Sig
ASKU	468	-0.154	0.083	-1.850	0.065	.
CC	471	0.100	0.079	1.264	0.207	
IE	469	-0.214	0.087	-2.451	0.015	*
ITIS	471	-0.033	0.088	-0.370	0.711	
SES	472	-0.084	0.074	-1.141	0.255	
learning	471	-0.204	0.081	-2.526	0.012	*
subperform	469	-0.171	0.084	-2.033	0.043	*

Note. Pre-test scores are grand-mean centered. IG = intervention group; CG = control group. * $p < .05$. ** $p < .01$. *** $p < .001$.

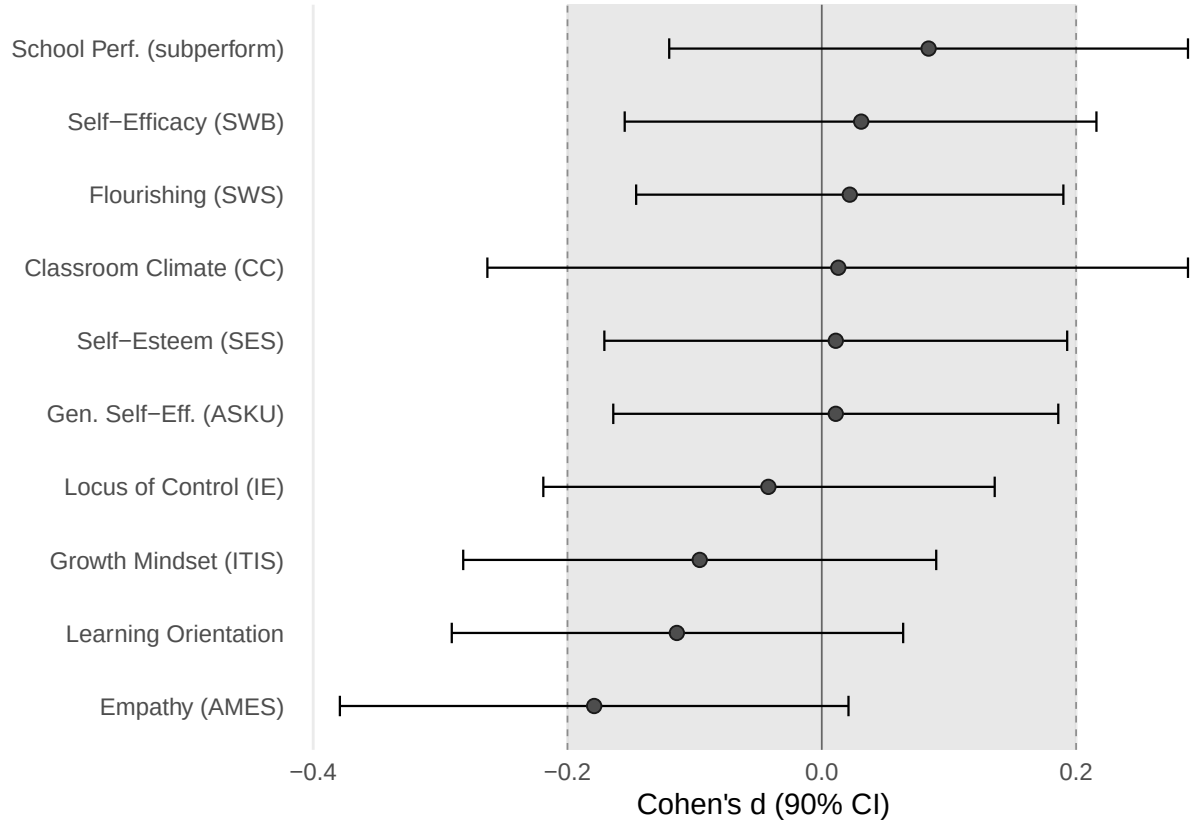
Model M2 shows a consistent pattern of negative Baseline × Group interaction coefficients, significant for 4 of the 10 constructs (Table 11). 8 of the 10 coefficients are negative, and those reaching conventional significance are SWB ($p = .004$), learning ($p = .012$), IE ($p = .015$), subperform ($p = .043$). The remaining constructs—SWS, AMES, ASKU, CC, ITIS, SES—did not (all $p > .05$; see Table 11 for the individual values). A negative coefficient means that the relationship between pre-test and post-test scores was weaker in the IG than in the CG: relative to the CG, students with lower baseline scores ended higher and students with higher baseline scores ended lower. The same pattern appeared under all three group codings (strict binary M2, broad binary M4, and dose-based M6), so it does not depend on the group classification used.

FDR Correction for Interaction Tests

To address the multiple comparisons concern inherent in testing Baseline × Group interactions across 10 constructs and 3 group codings (30 tests total), Benjamini-Hochberg FDR correction (Benjamini & Hochberg, 1995) was applied to all interaction p-values. Of 30 tests, 6

Figure 3

Forest plot of effect sizes (Cohen's d) with 90% confidence intervals for all outcome constructs. The shaded region represents the equivalence bounds ($d = \pm 0.20$), marked by dashed vertical lines. Three constructs (SWS, ASKU, SES) fall entirely within the equivalence region. Bayesian model comparison favours the null for all 10 constructs, most strongly for the self-evaluation scales and least so for empathy, where the evidence is moderate rather than strong.



survived FDR correction at $q < .05$ and 24 did not. The surviving tests are concentrated on a few constructs rather than spread across the outcome set, and they appear under more than one group coding, so they are not an artefact of a single specification. Because these tests were specified after the data were collected, FDR correction reduces but does not remove the inflation of error rates that post hoc testing entails. The full set of corrected *p*-values is reported in Table A23; whether a single common mechanism accounts for the pattern is examined in the next section and taken up in the Discussion.

Interaction Probing

To locate the baseline range in which the group effect changes sign, interaction probing was conducted by re-centering the pre-test variable at specific percentiles and re-estimating the group effect at each point. The pattern is illustrated here for self-esteem; probing results for all constructs are reported in Table A4. At the 10th percentile of baseline self-esteem (pre-test score = 2.1), the group coefficient was positive (β = 0.070, p = .349); it declined monotonically across the distribution, passing through approximately zero at the median (pre-test = 2.9; β = 0.003, p = .945) and reaching β = -0.055 (p = .441) at the 90th percentile.

This is the crossover the interaction term implies, and the probing makes its practical size visible: across all 25 probes of all constructs, 1 reached conventional significance. The interaction is therefore detectable as a trend across the baseline range without the group difference being reliably distinguishable from zero at any particular point on it. 6 of 30 interaction tests survive FDR correction.

Figure 4 illustrates the crossover interaction pattern for five selected constructs. The regression line for the IG is consistently flatter than for the CG, indicating that the pre-post relationship is attenuated in the intervention group. The two regression lines cross within the observed range of baseline scores, so the pattern is a crossover rather than a uniform attenuation. The group difference at any single point on that range is nevertheless small and not statistically reliable: at the 90th percentile of baseline self-esteem it is β = -0.055 (p = .441).

Regression to the Mean Considerations

Table 12

Pre-Post Correlations by Group With Fisher z-Transformation Test for Differential Stability

Scale	r (IG)	r (CG)	Diff (CG-IG)	z	p (Fisher)
SWB	0.437	0.637	0.200	-3.20	0.001
learning	0.381	0.515	0.135	-1.91	0.056

Table 12

Pre-Post Correlations by Group With Fisher z-Transformation Test for Differential Stability

Scale	r (IG)	r (CG)	Diff (CG-IG)	z	p (Fisher)
IE	0.342	0.529	0.188	-2.63	0.009
subperform	0.309	0.519	0.211	-2.89	0.004
ASKU	0.408	0.551	0.143	-2.10	0.036

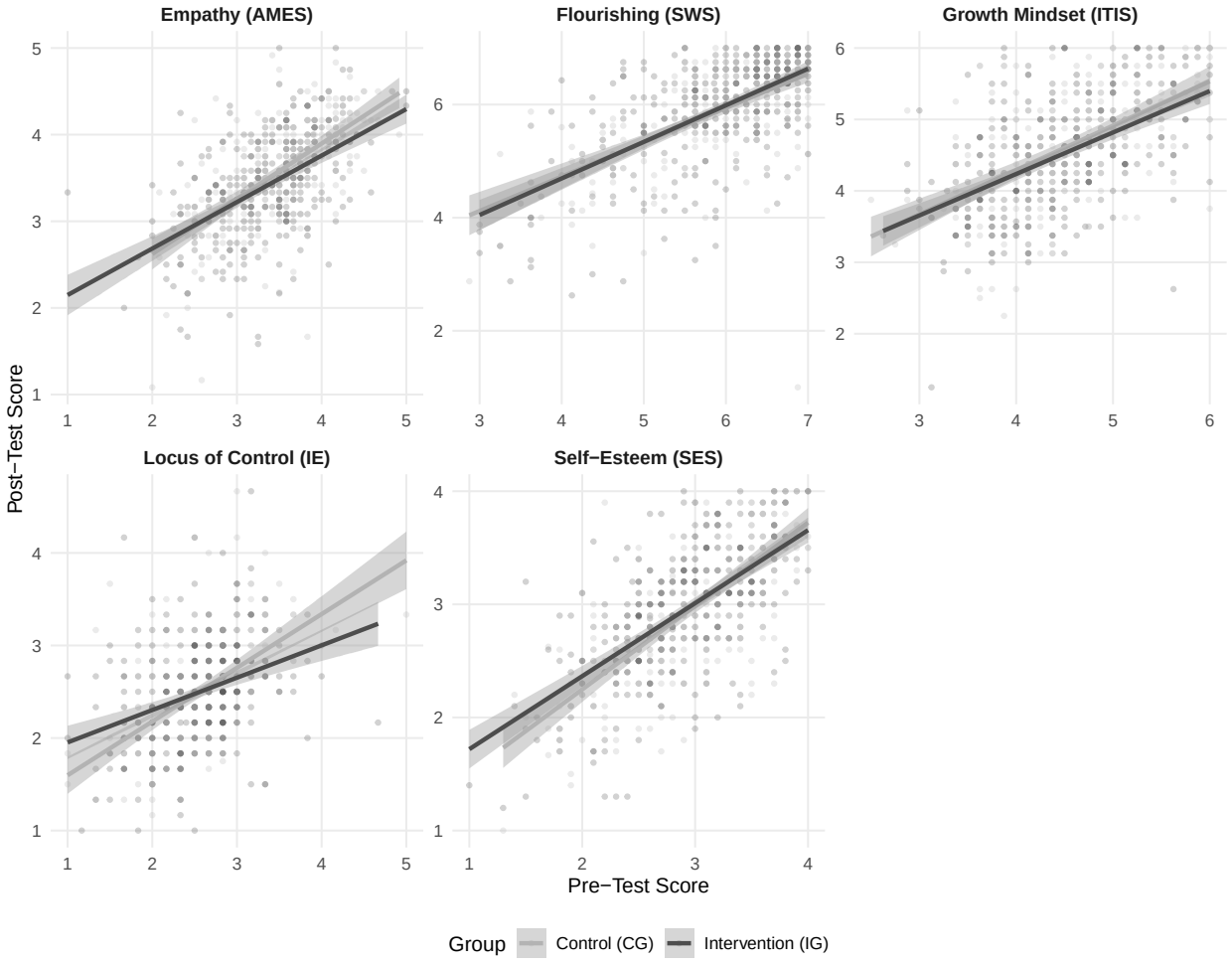
Note. Fisher z-test compares pre-post correlations between IG and CG. A significant difference indicates differential temporal stability, consistent with regression to the mean as a competing explanation for the observed interaction. IG = intervention group; CG = control group.

The interpretation of these interaction effects requires careful consideration of regression to the mean (RTM). Barnett et al. (2005) demonstrated that although covariate adjustment for the pre-test can address RTM in main-effect estimates, it does not protect interaction terms involving the pre-test variable from RTM confounds; moreover, in non-randomized designs such as the present one, adjusting for a pre-test that differs systematically between groups can itself introduce bias (Miller & Chapman, 2001; Van Breukelen, 2006). Table 12 presents pre-post correlations for the 5 constructs with the strongest interactions (SWB, learning, IE, subperform, ASKU), along with Fisher z-transformation tests (Nesselroade et al., 1980) comparing temporal stability between groups. Pre-post correlations were lower in the IG than in the CG for 5 of the 5 constructs, 4 of them significantly so. The residual diagnostics in Appendix J show the same asymmetry from a different angle: residual variability exceeded that of the CG for every outcome. What these two results jointly imply for the interpretation of the interaction effects is taken up in the Discussion.

The RTM account makes a further prediction that can be tested directly. If the interactions arise from measurement error in the pre-test, they must be systematically stronger for less reliable scales, because unreliability is what drives regression toward the mean. Across the 9 multi-item outcomes, the correlation between a scale's unreliability and the magnitude of its interaction effect was $r = .57$, $p = .109$, with a 95% confidence interval from -.15 to .90. The point estimate is

Figure 4

Baseline × Group Interaction: Pre-Test Scores (x-axis) vs. Post-Test Scores (y-axis) by Group. Regression lines illustrate the differential pre-post relationship in IG (dark grey) and CG (light grey). A flatter slope in the IG indicates that low-baseline IG students show more improvement and high-baseline IG students show less improvement relative to CG.



far below what the RTM account requires, but with only 9 scales the interval spans almost the entire range of possible values, so this test does not adjudicate between the explanations. It does establish that the RTM account has not been confirmed on its own central prediction.

Two limits of the Fisher z procedure should be stated here, because they bound what the preceding numbers can support. First, the Fisher z-test is a *diagnostic* for differential temporal stability, not a *correction* for it: it can show that an RTM explanation is tenable, but it cannot partial RTM out of the interaction estimates, which therefore remain confounded. Second, RTM in this setting originates in the unreliability of the pre-test, so the statistically appropriate remedy is to model the baseline as a latent variable with measurement error fixed from the observed reliabilities—a latent change score or errors-in-variables specification—rather than to enter the observed pre-test score as a manifest predictor. Such a model requires item-level responses, which are not part of the published dataset; the baseline-moderation results reported here should accordingly be read as descriptive of a pattern whose source has not been resolved. Future work with randomized allocation, or with a measurement-error-corrected baseline, is needed to determine whether any moderated effect survives once differential RTM is accounted for.

Factor-Level Analysis

To examine whether the observed patterns hold at the level of latent constructs, an exploratory factor analysis of the pre-test scores was conducted. Although the Kaiser criterion suggested three components, a three-factor solution was inadmissible (it produced a communality above unity), so the 2-factor solution reported here is the most differentiated one the data support. Multilevel models on the resulting factor scores reproduced the construct-level pattern: no significant main effects under either the binary or the dose coding, and a Baseline × Group interaction that reached nominal significance for Factor 1 (p = .044) but not for Factor 2 (p = .102). Full loadings and factor-level model results are reported in Appendix H (Table A19 and Table A20).

Class-Level Patterns

To characterize the heterogeneity of outcomes across implementation sites, class-level change profiles were constructed for all 59 class units in the dataset (including subgroups within classes for schools with individual-level group assignment). Each unit was classified as SUCCESSFUL (improvement on three or more of the 10 scales, defined by positive change exceeding 0.2 SD), UNSUCCESSFUL (decline on three or more scales), or MIXED (neither criterion met). Of these 59 units, 10 were classified as successful, 6 as unsuccessful, and 0 as mixed. This distribution indicates that positive outcomes were the exception rather than the rule, with the majority of classes showing an ambiguous pattern of small, inconsistent changes across constructs.

Mean pre-post change scores by implementation level and construct are displayed as a heatmap in Appendix I (Figure A1); it shows no consistent dose-response gradient and the counterintuitive negative changes at moderate implementation.

A critical finding from the class-level analysis is the absence of a dose-response relationship at this level of aggregation. The correlation between implementation intensity (coded 0 to 3) and composite change across classes was $r = -0.095$, which is essentially zero and slightly negative, providing no evidence that higher-dose classes achieved better outcomes. Furthermore, baseline-change correlations at the class level were uniformly and strongly negative (ranging from $r = -.30$ to $r = -.62$), consistent with regression to the mean at the aggregate level: classes that started higher tended to decline, while classes that started lower tended to improve, regardless of intervention status.

The Implementation Paradox

Table 13

Model M7 Results: Estimated Effects by Implementation Level (Reference: None/Minimal)

Construct	Level	Beta	SE	t	p	d	Sig
SWB	Low	-0.089	0.100	-0.890	.380	-.191	
SWB	Moderate	-0.174	0.163	-1.070	.292	-.373	
SWB	Intensive	0.040	0.053	0.754	.451	.085	
SWS	Low	-0.158	0.152	-1.040	.305	-.222	
SWS	Moderate	0.102	0.247	0.411	.683	.143	
SWS	Intensive	0.039	0.080	0.488	.626	.055	
AMES	Low	-0.101	0.112	-0.903	.372	-.208	
AMES	Moderate	0.074	0.170	0.434	.667	.152	
AMES	Intensive	-0.010	0.062	-0.165	.869	-.021	
ASKU	Low	-0.276	0.139	-1.989	.054	-.413	.
ASKU	Moderate	-0.082	0.227	-0.363	.719	-.123	
ASKU	Intensive	0.079	0.072	1.103	.271	.118	
CC	Low	-0.198	0.238	-0.835	.409	-.354	
CC	Moderate	-0.320	0.288	-1.109	.275	-.571	
CC	Intensive	0.042	0.112	0.371	.711	.074	
IE	Low	0.211	0.115	1.830	.076	.416	.
IE	Moderate	0.034	0.182	0.188	.852	.067	
IE	Intensive	-0.096	0.060	-1.590	.113	-.189	
ITIS	Low	-0.382	0.169	-2.268	.029	-.571	*
ITIS	Moderate	-0.047	0.250	-0.187	.853	-.070	
ITIS	Intensive	0.042	0.087	0.479	.632	.062	
SES	Low	-0.132	0.106	-1.238	.224	-.306	
SES	Moderate	0.057	0.160	0.354	.725	.132	
SES	Intensive	-0.027	0.054	-0.493	.622	-.062	

Table 13
Model M7 Results: Estimated Effects by Implementation Level (Reference: None/Minimal)

Construct	Level	Beta	SE	t	p	d	Sig
learning	Low	-0.096	0.092	-1.043	.304	-.240	
learning	Moderate	-0.106	0.144	-0.735	.467	-.265	
learning	Intensive	0.015	0.048	0.323	.747	.038	
subperform	Low	-0.237	0.148	-1.596	.119	-.370	
subperform	Moderate	-0.176	0.227	-0.778	.442	-.276	
subperform	Intensive	0.127	0.076	1.663	.097	.198	.

Note. Reference category is none/minimal. IG = intervention group; CG = control group. * $p <$ .05. ** $p <$ .01. *** $p <$ .001.

Table 13 presents the results from Model M7, which compared each implementation level to the *none/minimal* reference category.

The two partial implementation categories were associated with lower post-test scores than no implementation at all. For low implementation the coefficient was negative for 9 of 10 constructs (mean d = -0.25, 1 significant), and for moderate implementation for 6 of 10 constructs (mean d = -0.12, 0 significant). Intensive implementation did not differ from no implementation: coefficients were negative for 3 of 10 constructs, the mean effect was 0.04, and 0 reached significance.

These contrasts differ enormously in how much evidence stands behind them. Implementation level is a property of the class, so the effective sample size for each contrast is the number of classes at that level, not the number of students. The reference category comprises 33 classes and the intensive category 16, but the low category rests on 3 classes and the moderate category on 2. The significant coefficients in the low and moderate categories therefore cannot be separated from the individual teachers and classes concerned, and should not be read as effects of implementation intensity. Only the comparison between intensive and none/minimal

implementation rests on enough clusters to support inference, and that comparison is null.

Reliable Change Index

Table 14
Reliable Change Index: Proportions of Reliably Improved and Deteriorated Students by Group

Scale	IG Improved	CG Improved	IG Deteriorated	CG Deteriorated	Chi-squared	p
SWB	39 (13%)	20 (9%)	24 (8%)	8 (3.6%)	1.66	0.198
SWS	17 (5.6%)	3 (1.3%)	17 (5.6%)	11 (4.9%)	5.31	0.021
AMES	19 (6.3%)	12 (5.4%)	14 (4.7%)	14 (6.2%)	0.07	0.786
ASKU	23 (7.6%)	7 (3.1%)	13 (4.3%)	12 (5.4%)	4.01	0.045
CC	13 (4.3%)	15 (6.7%)	31 (10.3%)	15 (6.7%)	1.02	0.312
IE	28 (9.3%)	16 (7.2%)	29 (9.6%)	17 (7.6%)	0.49	0.485
ITIS	27 (8.9%)	17 (7.6%)	17 (5.6%)	16 (7.2%)	0.13	0.721
SES	14 (4.6%)	7 (3.1%)	12 (3.9%)	8 (3.6%)	0.40	0.525
learning	26 (8.6%)	18 (8%)	21 (6.9%)	12 (5.4%)	0.00	0.949

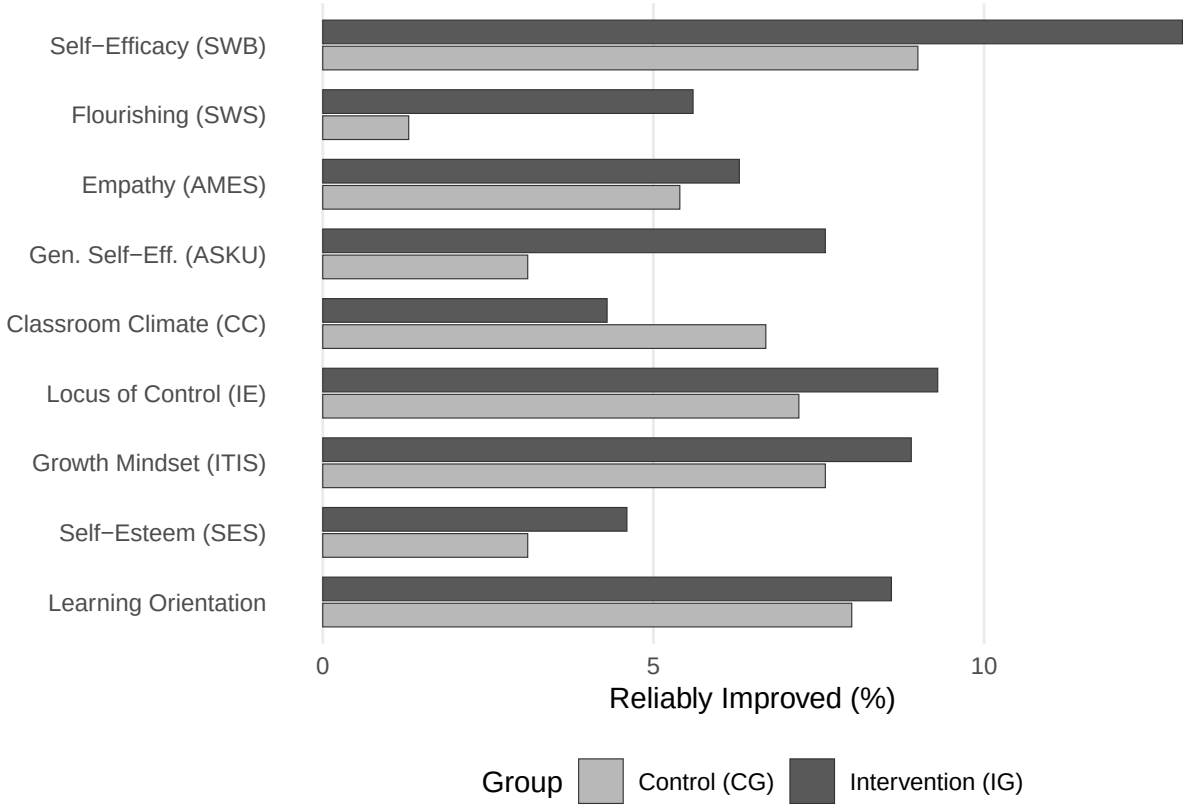
Note. RCI computed following Jacobson and Truax (1991). IG = intervention group; CG = control group. Chi-squared tests do not account for the nested data structure.

Table 14 presents the Reliable Change Index results. The RCI analysis does not account for the nested data structure, and the chi-square tests treating each student as an independent observation may produce anti-conservative *p*-values. To evaluate this concern, a cluster-adjusted generalized linear mixed model with a binary improved/not-improved outcome and school and class random intercepts was fitted for empathy, the construct with the largest raw difference in reliable improvement; the group effect was not significant once clustering was accounted for. Notably, the IG showed not only more reliably improved individuals but also more reliably deteriorated individuals compared to the CG, a pattern consistent with greater overall instability

in the intervention group rather than a unidirectional benefit. The RCI results should therefore be interpreted with substantial caution and are presented as supplementary to the multilevel models. Figure 5 visualizes the proportions of reliably improved students by group across all constructs.

Figure 5

Proportions of reliably improved students (RCI) by group for each outcome construct. Note that the chi-squared tests do not account for clustering; the AMES group difference was no longer significant after cluster adjustment (see text).



Subgroup Analyses

Table 15

Group Effects on AMES Subscales (Cognitive Empathy, Affective Empathy, Sympathy)

Subscale	N	Beta	SE	t	p	d	Sig
Cognitive Empathy	525	0.074	0.066	1.127	.261	.114	

Table 15
Group Effects on AMES Subscales (Cognitive Empathy, Affective Empathy, Sympathy)

Subscale	N	Beta	SE	t	p	d	Sig
Affective Empathy	525	-0.075	0.088	-0.854	.393	-.104	
Sympathy	525	0.010	0.065	0.156	.876	.015	

Note. IG = intervention group; CG = control group. * $p < .05$. ** $p < .01$. *** $p < .001$.

The composite AMES scale was broken into its three subscales to check whether a group difference is concealed by aggregation (Table 15). None of the three shows a significant effect: cognitive empathy $\beta = 0.074$ ($p = .261$, $d = 0.11$), affective empathy $\beta = -0.075$ ($p = .393$, $d = -0.10$), and sympathy $\beta = 0.010$ ($p = .876$, $d = 0.02$). The composite null result is therefore not an aggregation artefact: the three components behave alike, and none of them moves.

Splitting the IE composite into its internal and external locus of control subscales confirmed that neither subscale showed a significant group effect (internal: $\beta = 0.009$, $p = .888$, $d = 0.01$; external: $\beta = 0.029$, $p = .634$, $d = 0.04$; both $n = 525$), consistent with the null composite result and the scale's low internal consistency.

Age × Group interactions were tested for all 9 multi-item outcomes to examine whether the intervention operated differently for younger versus older students; none reached statistical significance (all $p \geq .011$, all $|d| \leq 0.18$). The largest effects were observed for implicit theories of intelligence (ITIS: $\beta = 0.007$, $p = .881$, $d = 0.01$) and self-esteem (SES: $\beta = 0.074$, $p = .011$, $d = 0.17$), both in the direction of slightly more favorable outcomes for older students in the intervention group, and both above the conventional significance threshold.

Table 16
Group Effects Within Students With Migration Background

Construct	N	Beta	SE	t	p	d	Sig
SWB	217	-0.050	0.067	-0.746	.457	-.109	

Table 16
Group Effects Within Students With Migration Background

Construct	N	Beta	SE	t	p	d	Sig
SWS	217	-0.025	0.109	-0.229	.819	-.033	
AMES	218	-0.093	0.063	-1.470	.144	-.201	
ASKU	216	-0.058	0.096	-0.606	.546	-.083	
CC	218	0.005	0.127	0.038	.969	.008	
IE	217	0.057	0.076	0.751	.454	.105	
ITIS	218	-0.125	0.101	-1.242	.216	-.185	
SES	218	0.097	0.079	1.226	.222	.220	
learning	218	-0.098	0.061	-1.597	.112	-.246	
subperform	216	0.191	0.083	2.299	.023	.316	*

Note. Subsample restricted to students reporting migration background. IG = intervention group; CG = control group. * $p < .05$. ** $p < .01$. *** $p < .001$.

Among students with a migration background (n = 216 to 218), one contrast reached conventional significance: satisfaction with school performance developed more favourably in the IG than in the CG ($\beta = 0.191$, $p = .023$, $d = 0.32$; Table 16). Self-esteem moved in the same direction without reaching significance ($\beta = 0.097$, $p = .222$, $d = 0.22$). No group difference emerged for any construct among students without a migration background. These contrasts were among many examined across constructs and subgroups and were not specified in advance.

Sensitivity Analyses

Across 60 re-estimations of the group main effect under 6 inclusion criteria (10 scales, n = 343–528), 1 reached significance at $p < .05$, with effect sizes between $d = -0.299$ and $d = 0.136$. The null main effect is not an artefact of the inclusion rules. Appendix B reports the full matrix.

Classroom Climate at the Class Level

Table 17
Reliability of the Class Mean and Effective Sample Size by Construct (Post-Test)

Construct	ICC(1)	ICC(2)	Design effect	Students	Effective n
SWB	0.066	0.388	1.53	527	345
SWS	0.058	0.353	1.46	527	362
AMES	0.073	0.411	1.57	526	334
ASKU	0.050	0.321	1.40	528	378
CC	0.301	0.793	3.38	527	156
IE	0.036	0.249	1.28	528	411
ITIS	0.053	0.332	1.42	527	371
SES	0.095	0.485	1.76	528	301
learning	0.039	0.267	1.31	527	402
subperform	0.090	0.469	1.71	526	307

Note. ICC(1) = proportion of variance lying between classes. ICC(2) = reliability of the class mean, computed as $k \cdot ICC(1)/(1 + (k - 1) \cdot ICC(1))$ with k the average class size; values at or above .70 are conventionally treated as adequate for aggregation to the class level. Design effect = $1 + (k - 1) \cdot ICC(1)$, the factor by which clustering inflates the variance of the group contrast. Effective n = the number of independent observations the clustered sample is equivalent to. Classroom climate is the only construct whose class mean is reliable, and the only one whose effective sample is a small fraction of the students measured. Scale abbreviations are defined in Appendix A.

Classroom climate behaves unlike the other nine outcomes, and Table 17 shows how far apart it sits. Its ICC(1) at post-test is 0.30 (pre-test 0.34), roughly 3.2 times the largest value among the remaining constructs (SES, 0.10); for most outcomes the figure lies between .03 and .10. A third of the variation in how students rate their classroom climate is therefore explained

simply by which class they are in—which is what one would expect of a property that belongs to the class rather than to the student, and is a point in favour of the measure rather than against it.

Two consequences follow. The first is that the class mean is a reliable quantity for this construct and for no other: ICC(2) = 0.79, against 0.48 or below elsewhere. Aggregating classroom climate to the class level is therefore defensible in a way that aggregating the other outcomes would not be. The second is a cost. The design effect for classroom climate is 3.38, so the 527 students who answered it carry the information of roughly 156 independent observations. Whatever this study can say about classroom climate rests on a far smaller effective sample than the headline *N* suggests, and that is true irrespective of how the model is specified.

Re-estimating the group contrast at the level the construct is defined on—one row per class, 44 classes with at least five respondents covering 467 students, class mean post-test on class mean pre-test and condition with school as a random intercept—leaves the conclusion unchanged: b = -0.073, SE = 0.116, $t(31)$ = -0.63, p = .534. Expressed on the student-level standard deviation so it can be compared with the main models, the estimate is d = -0.10, 95% CI [-0.44, 0.23]. The interval is wide, as it must be with 44 units, and it is the honest width for this question: it spans effects that would be worth having in both directions. The student-level and class-level analyses agree on the absence of a detectable effect, but only the class-level interval states correctly how much uncertainty remains.

Sociodemographic Composition and Moderation

Table 18

Migration Background and Home Language by Study Condition

Variable	Level	IG n	IG %	CG n	CG %
Migration background	no migration background	103	39.9	117	64.3
Migration background	one parent abroad	25	9.7	28	15.4
Migration background	both parents abroad	55	21.3	12	6.6

Table 18
Migration Background and Home Language by Study Condition

Variable	Level	IG n	IG %	CG n	CG %
Migration background	student born abroad	75	29.1	25	13.7
Language at home	German at home	145	47.7	171	76.3
Language at home	other language at home	159	52.3	53	23.7

Note. Migration background is graded from parental and student country of birth, asked at post-test: *one parent abroad* and *both parents abroad* refer to the parents' country of birth with the student born in the country; *student born abroad* takes precedence over the parental categories. Home language was asked at pre-test and is not redundant with country of birth—families can have migrated and speak German at home, or not have migrated and speak another language. Percentages are within condition and are based on the students for whom the variable is known; the totals therefore differ between the two blocks, because migration background was asked at post-test and home language at pre-test. IG = intervention group; CG = control group.

The two conditions differ substantially in sociodemographic composition (Table 18). 52.3% of intervention students speak a language other than German at home, against 23.7% of control students, $\chi^2(1) = 42.84$, $p < .001$. The graded migration variable shows the same split: 64.3% of the control group has no migration background against 39.9% of the intervention group, $\chi^2(3) = 41.78$, $p < .001$. This is a property of which schools took part in which condition, not of the curriculum, and it is the clearest single respect in which the two groups are not interchangeable.

Whether it changes the conclusion is a separate question, and the answer is that it does not. Group × moderator interactions were estimated for graded migration background, home language, and gender across all ten outcomes, with pre-test score and age as covariates and the same nesting structure as the main models (50 tests in total). Four were nominally significant at $p < .05$—about the two expected by chance at that rate—and none survived Benjamini-Hochberg correction

(smallest q = .106). The largest single coefficient ($|d|$ = 0.96) rests on one of the thin migration cells and should not be read as an effect. Nothing here indicates that the curriculum works for one sociodemographic group and not another; what the analysis does establish is that the null result is not concealing a subgroup effect along these lines.

Discussion

This evaluation followed 528 students in 59 classes at 13 schools in Austria and Germany across a full academic year. Its central finding is a null. No outcome showed an intervention effect, before or after correction for the number of outcomes examined, and the picture did not change when the group contrast was replaced by a dose-response analysis, by a model estimating each implementation level separately, or by a within-school comparison that removes between-school confounding altogether.

On the first research question—whether the GET curriculum changes psychosocial outcomes—the answer these data support is no. No contrast reached significance on any of the ten outcomes, before or after adjusting for the number of outcomes examined, and the largest of them was d = -0.18. The same holds under the broad group coding, under the dose-response coding, and under every inclusion rule examined in the sensitivity analyses. The defensible reading is that no effect of the curriculum on any assessed outcome was detected.

This is not the same as having learned nothing. Equivalence testing and Bayesian model comparison carry the conclusion past a mere failure to reject: for three constructs the data positively support the absence of a practically relevant difference (Lakens, 2017), and for the remainder they favour the null over the alternative (Wagenmakers, 2007). The dose-response analyses point the same way—outcome change did not track how much of the curriculum classes actually received—and this convergence matters, because a genuine but small effect diluted by weak implementation should have left a gradient across classes that received more.

The baseline-moderation analyses produced the one pattern that looked substantively interesting: students who began low appeared to gain, while students who began high appeared to

lose. It is a robust pattern in the statistical sense—it survives correction for the number of tests, appears under more than one group coding, and is of similar magnitude under every inclusion criterion examined (Appendix B). What the design cannot do is establish what produces it.

One feature invites caution. Where the pattern appears, it appears on constructs the curriculum addresses and on constructs it does not. A mechanism acting on self-esteem, growth mindset, and locus of control alike is easier to explain as a property of how the constructs were measured than of what was taught. That is an argument for caution rather than a demonstration: a general effect of the curriculum on how students appraise themselves would also spread across constructs, and the two possibilities make the same prediction here.

Two readings of these crossover interactions are available, and they are hard to tell apart with these data. Under the differential treatment effect interpretation, the GET curriculum genuinely benefits students who enter with lower psychosocial functioning, consistent with the well-documented compensation hypothesis in prevention science and resilience research (Masten, 2014; Rutter, 1979, 1987; Werner & Smith, 1992)—the finding that interventions tend to produce larger effects for higher-risk participants (Horowitz & Garber, 2006; Stice et al., 2009). If this interpretation were correct, it would have important practical implications: the curriculum could function as a targeted intervention delivered through a universal mechanism, reaching at-risk students without requiring screening or selective enrollment.

The second reading is less flattering to the curriculum but requires fewer assumptions. Under the regression-to-the-mean (RTM) interpretation, the observed interactions are statistical artifacts arising from differential measurement stability between groups (Barnett et al., 2005). The mechanism is worth stating plainly, because everything that follows turns on it: any measurement contains error, extreme scores contain more of it than central ones, and a student who scored unusually low at pre-test is therefore likely to score closer to the average at post-test whether or not anything happened in between. If that error is larger in one group, the drift toward the average is larger in that group too—which is exactly the crossover pattern, produced by nothing but noise. While ANCOVA controls for RTM in main effects (Miller & Chapman, 2001),

it does not protect interaction terms involving the pre-test variable from RTM confounds. If the IG exhibits greater measurement error or instability than the CG, more extreme baseline scores in the IG will regress more strongly toward the mean, producing the crossover pattern without any genuine treatment effect (Nesselroade et al., 1980). Two lines of evidence support this reading. Pre-post stability was lower in the intervention group for 5 of the 5 constructs with the strongest interactions (see Table 12), and residual variability was greater in the intervention group for every outcome (Appendix J). Measurements were, in this sense, noisier in the intervention group.

A third line of evidence, and the only one that could have settled the question, does not support it. If unreliability generates the interactions, the effect must be larger for scales that measure less reliably, since that is the mechanism by which regression to the mean operates. Across the outcomes, interaction magnitude showed no appreciable association with scale unreliability. With ten scales the test is weak—the confidence interval is wide enough to accommodate a substantial association—so this is not a refutation. But it means the RTM account remains unconfirmed on the one prediction that distinguishes it from a substantive explanation, and it should be presented as the leading candidate rather than as the resolution.

The alternative worth taking seriously is response shift (Schwartz & Sprangers, 1999). A curriculum built around self-reflection and emotional awareness may change not only where students stand on a construct but the internal standard against which they rate themselves. That would produce lower apparent stability in the intervention group—the same signature attributed above to measurement error—while being a genuine consequence of the intervention rather than noise. The present design cannot separate the two, because both predict the observed pattern. Distinguishing them requires modeling the baseline as a latent variable with measurement error fixed from the item responses, which the published dataset does not support. Until that is done, the honest position is that the intervention group's responses became less predictable from their own baseline, and that whether this reflects instrument noise or the curriculum doing something to how students appraise themselves is unresolved. That the pattern recurs on constructs the curriculum does not address is the strongest reason for caution, though not a decisive one: a

subject that changes how students appraise themselves in general could plausibly reach locus of control and learning orientation as well as self-esteem and flourishing. A common statistical mechanism—differential measurement instability in the intervention group—is the more economical explanation of the two, but economy is not evidence. Randomized designs are needed to determine whether any baseline-moderated effects survive when differential regression to the mean is experimentally controlled.

Concretely, a student who initially rated their well-being highly may, after a year of psychoeducational content about emotional awareness, recognise aspects of their functioning they had previously overlooked, and report a lower post-test score that reflects greater self-knowledge rather than deterioration. If that is what happened, the null main effects could be masking genuine improvement that recalibrated self-assessment offsets. Then-tests—retrospective pre-tests administered alongside the post-test—would separate this from measurement error and should be built into any future evaluation of a curriculum of this kind.

The absence of any dose-response relationship deserves equal attention. Neither the linear dose model nor the class-level analysis showed outcomes tracking how much of the curriculum was delivered. Two readings are available. The first is that the content does not produce the intended psychological changes at any delivery intensity, which would call the program's theory of change into question. The second, and in our view more likely, is that teacher-reported hours do not capture what actually matters about implementation—enthusiasm, responsiveness to the class, the depth of engagement a lesson achieves (Durlak & DuPre, 2008). A teacher reporting weekly delivery may be teaching it thinly; a teacher reporting occasional use may be teaching a handful of sessions that land.

The categorical analysis appears at first to sharpen this: classes with partial delivery scored below classes that received no instruction, and several of those contrasts were significant. That reading does not survive a look at how the classes are distributed. Because implementation is a property of the class rather than of the student, each contrast rests on the number of classes at that level, and the partial-delivery categories contain only a handful of classes between them.

Those coefficients are indistinguishable from the effects of the particular teachers involved, and we do not interpret them.

What remains interpretable is the comparison that does rest on a substantial number of classes on both sides: intensive delivery versus none. It is null. This is the more consequential result, because it removes the most natural defence of a program that shows no aggregate effect—that it was simply never delivered properly. In the classes where it was delivered as designed, outcomes were no better than in classes that never received it. Whether a threshold of fidelity exists beyond which the curriculum works therefore remains untested here, but the study provides no positive evidence for one.

These findings closely parallel the national evaluation of the Social and Emotional Aspects of Learning (SEAL) program in England (Humphrey et al., 2010), which similarly found no overall effects on student outcomes across a large-scale implementation characterized by high variability in delivery quality and intensity. As with SEAL, the GET curriculum was introduced into an ecologically complex school environment in which competing demands, varying levels of teacher buy-in, and inconsistent administrative support rendered the nominal treatment condition far more heterogeneous than the program design intended. The null dose-response finding is also consistent with the meta-analytic evidence from the Penn Resiliency Program (PRP), where Brunwasser et al. (2009) found that program effects tended to be larger in developer-led implementations than in those delivered by regular school staff, although this difference between delivery modes was not statistically significant. This apparent developer-delivery gap points to a core challenge for scalable school-based programs: the conditions under which a curriculum is developed and initially tested may be qualitatively different from the conditions under which it is subsequently disseminated, and the resulting gap between efficacy and effectiveness may be sufficient to eliminate detectable effects entirely. The GET evaluation, in which the curriculum was delivered by classroom teachers with varying degrees of preparation and commitment, exemplifies this translation problem.

Empathy deserves separate comment, because it is the construct on which the internal

pilot evaluation reported the one significant effect it found (Hemmelmayr, 2025). That effect is not reproduced here. The empathy contrast is d = -0.18 (p = .141), the largest of the ten but well short of significance, and none of the three subscales—cognitive empathy, affective empathy, or sympathy—shows a group difference either, so the composite result is not concealing a component effect. Empathy is also one of the constructs on which the groups differed at baseline, which is a further reason to read its estimate cautiously in either direction.

What remains is a genuine non-replication rather than a contradiction. The confidence interval around the present empathy estimate excludes an effect of the magnitude the pilot reported, so an effect that size can be ruled out in this sample; smaller effects cannot. Given that the pilot was internal, unpublished, and substantially smaller, and that the present estimate is the more precise of the two, the reasonable conclusion is that the empathy signal did not carry over to the full evaluation.

The exploratory subgroup analyses produced three nominally significant results, all of which require the same caution. Age moderated the group effect on self-esteem and on classroom climate, in both cases such that older students in the intervention group fared slightly better relative to their controls. Among students with a migration background, satisfaction with school performance developed more favourably in the intervention group than in the control group (see Table 16). It is tempting to read the last of these as the curriculum supplying psychosocial resources that matter most to students negotiating questions of belonging. That reading should be resisted. None of the three subgroups was specified in advance; the migration indicator is a proxy derived from parental country of birth rather than a measure of social position; three significant results out of 34 tests is close to what chance alone would produce at the conventional threshold; and none of the three would survive correction for the size of the exploratory set. They are hypotheses for a future study, not findings of this one.

These findings fit a broader pattern in the school-based intervention literature. The MYRIAD trial (Kuyken et al., 2022), the largest school-based mindfulness RCT to date, found no evidence that mindfulness training improved mental health or well-being in over 8,000

adolescents. Cipriano et al. (2023) updated the seminal Durlak et al. (2011) meta-analysis and reported substantially smaller effect sizes (d = 0.13 to 0.23 vs. the original d = 0.27 to 0.57), well below the benchmarks for educational interventions synthesized by Hattie (2009), with the most rigorous studies producing the smallest effects. The growth mindset literature, which directly overlaps with a core component of the GET curriculum, has shown a similar pattern of diminishing effects as methodological rigor increases (Dweck, 1999). Collectively, these findings raise the possibility that the field of school-based positive psychology has systematically overestimated the effects of universal interventions, particularly when programs are delivered under real-world conditions rather than the carefully controlled circumstances of efficacy trials. The GET evaluation, with its ecological validity, implementation heterogeneity, and transparent reporting of null results, contributes to a more balanced and realistic evidence base for school-based mental health promotion (Cipriano et al., 2023; Kuyken et al., 2022), a domain that has attracted considerable policy attention (OECD, 2015).

A key difference between the present study and the more optimistic earlier literature is the quality and intensity of implementation. Durlak and DuPre (2008) concluded that well-implemented programs produce effect sizes two to three times larger than poorly implemented ones. In the present study, a substantial proportion of the nominally experimental group received minimal or no actual GET instruction, severely diluting any potential treatment signal. This implementation failure is itself an important finding: it suggests that the GET curriculum, in its current form, does not generate sufficient institutional commitment or teacher engagement to ensure consistent delivery across sites. Whether this reflects inadequate teacher training, insufficient administrative support, competing curricular demands, or a fundamental limitation of the curriculum's design cannot be determined from the present data (Forman & Barakat, 2011), but it represents a critical barrier that must be addressed before any program effects could plausibly emerge at scale.

Practical Implications

For schools weighing whether to adopt the curriculum, the finding that carries the most weight is not the absent main effect but the absent dose-response gradient. A universal program that shows no aggregate benefit can usually be defended on the grounds that it was never delivered as intended; here it was, in a substantial number of classes, and those classes did not differ from classes that never received it. On present evidence the curriculum should not be adopted on the expectation of measurable psychosocial gains at the population level, and the timetable hours it requires have an opportunity cost that must be set against that. Equally, no evidence of harm emerged: on no outcome did the intervention group develop significantly less favourably than the control group.

Two things this study cannot support recommendations about should be named as such. It cannot advise on partial delivery, because the classes delivering the curriculum partially are too few to separate from the teachers who taught them. And it cannot yet justify redirecting the program towards students who start low: that possibility rests on the Baseline × Group interactions, and this design cannot separate a genuine moderation from regression to the mean. Should a randomized design confirm genuine moderation, reconceiving the curriculum as a selective rather than universal prevention strategy would follow the compensation hypothesis (Horowitz & Garber, 2006; Stice et al., 2009) and could be delivered through routine screening (Eccles & Roeser, 2011; Keyes, 2005)—but that is a recommendation for future evidence to make, not this one.

What the implementation data do support is a point about the conditions of delivery. A sizeable proportion of nominally intervention classes received minimal or no instruction despite teachers having completed the training and having the textbook available, which indicates that materials and training are necessary but not sufficient. Delivery also needs ongoing coaching and administrative structures that protect the timetabled hours from competing demands (Durlak & DuPre, 2008; Forman & Barakat, 2011)—adherence, dosage, quality, responsiveness, and differentiation are distinct dimensions of integrity, and only dosage was measured here (Dane &

Schneider, 1998). More generally, schools and authorities considering any positive education program should ask for evaluations conducted independently of the developer, under conditions resembling their own, with an active control condition and pre-registered analyses (Brunwasser et al., 2009).

One targeted recommendation does follow from the empathy result discussed above. Because the interval there still admits effects smaller than the pilot reported, the question worth asking next is not whether the package as a whole works but whether the perspective-taking and cooperative-learning components specifically do—which calls for dismantling studies that isolate individual curriculum elements rather than further evaluations of the whole.

Several limitations constrain the interpretation of both the null main effects and the exploratory interaction findings. The quasi-experimental design means that schools and classes self-selected into the program, introducing selection bias at both the school and class level: schools that adopted the GET curriculum may differ systematically from non-adopting schools in ways not captured by the measured covariates, and within schools, the decision about which classes received the curriculum was typically made by administrators or individual teachers rather than through random allocation. Pre-existing differences between IG and CG classes therefore cannot be ruled out even though baseline equivalence was generally supported, and one such difference is large and documented: the intervention group contains far more students with a migration background and far more who speak a language other than German at home than the control group (Table 18). Because that composition is a property of which schools volunteered for which condition, it cannot be separated from the treatment contrast by any analysis of these data. Adjusting for it is not a remedy either, since it stands in for whatever else distinguishes the participating schools—catchment area, resources, prior exposure to comparable programs. The moderation tests reported above give no indication that the curriculum works differently across these groups, so the imbalance is unlikely to be masking an effect; but it does mean the two conditions are not exchangeable populations, and a reader should treat the comparison as being between two sets of schools rather than between two otherwise equivalent groups of students.

Three further consequences of the design follow from the same root. Teacher effects are fully confounded with the treatment: each class had a single teacher who both delivered (or did not deliver) the curriculum and influenced outcomes through their general pedagogical approach, classroom management, and personal relationship with students. In schools where both IG and CG classes coexisted, contamination cannot be excluded—students from different classes may have shared curriculum content informally, and teachers who delivered the GET curriculum in some classes may have inadvertently applied its principles in their control classes as well. This bears directly on the within-school comparison reported above, which is the design least vulnerable to school-level selection and the most vulnerable to contamination; the two biases work in opposite directions and neither can be quantified here. Implementation heterogeneity, while informative, also means that many IG classes received minimal treatment, diluting any potential effect.

Power is a separate constraint, and it binds regardless of the design issues above. The study was underpowered for small effects. Once the clustering of students within classes is taken into account, the effective sample is considerably smaller than the raw number of students suggests, and the sensitivity analysis reported in the Method section shows that power to detect a small effect was well below conventional standards. The equivalence and Bayesian analyses partly compensate, since they provide positive evidence that effects are near zero rather than merely failing to reject the null. They do not compensate entirely: effects of the magnitude that meta-analyses report for real-world universal interventions (Cipriano et al., 2023) remain within the range this study could have missed.

Classroom climate warrants a caveat of its own, and it cuts deeper than the wording of any single result. It is the one outcome that is a property of the class rather than of the student, and it carries by far the strongest clustering of the ten (Table 17). The consequence is that this study is much less able to speak about classroom climate than the overall sample size implies: the design effect reduces 527 respondents to roughly 156 independent observations, and the class-level analysis rests on 44 classes. Classroom climate is also the construct on which the groups differ most at baseline and the one that carries the implementation analyses, so the weakest

measurement situation coincides with the heaviest inferential load. The class-level estimate agrees with the student-level one in finding no effect, but its confidence interval—d = -0.10, 95% CI [-0.44, 0.23]—is wide enough to contain effects worth having. For this outcome the appropriate conclusion is not that the curriculum had no effect but that the study could not determine whether it did. A design aimed at classroom climate would need many more classes rather than more students per class.

Ceiling effects represent a further constraint on the capacity to detect improvement. Several scales showed substantial proportions of students scoring at or near the scale maximum at baseline, with flourishing exhibiting the most pronounced ceiling compression, followed by classroom climate, school-related self-efficacy, and learning orientation (see Table 4). When a large proportion of students already score near the maximum, there is limited room for upward movement, and any positive intervention effect is mathematically constrained. This ceiling compression has particular implications for the interpretation of the differential baseline effects: the finding that low-baseline students appear to benefit while high-baseline students do not may partly reflect the asymmetric capacity for change across the baseline distribution rather than a genuine differential treatment mechanism. Notably, self-esteem shows a more symmetric distribution than flourishing, which may contribute to why the SES interaction survived both FDR correction and RTM testing while the SWS interaction did not.

The absence of an active control condition represents a design limitation that affects the interpretability of any effects that might have been detected. The control group received business-as-usual instruction, not a matched alternative activity. Had the intervention produced significant effects, it would have been impossible to determine whether these reflected specific curriculum content or non-specific factors such as attention, novelty, or the Hawthorne effect. While the null main effects render this concern less urgent for the primary findings, it remains relevant for the exploratory interaction and RCI results, where positive effects in the intervention group could partly reflect differential attention rather than curriculum-specific mechanisms.

The implementation assessment relied exclusively on retrospective, categorical teacher

self-reports that capture only the quantity of curriculum delivery, not its quality. Structured classroom observations, student engagement ratings, or qualitative interviews with teachers and students would have provided a richer picture of how the curriculum was actually delivered and experienced (Dane & Schneider, 1998; Durlak & DuPre, 2008). The crude categorical coding (none/minimal, low, moderate, intensive) may obscure meaningful variation in delivery quality within each category, and the absence of independent fidelity verification means that the implementation data are unvalidated. Future evaluations should integrate qualitative process evaluation alongside quantitative outcome assessment to distinguish between program failure and implementation failure.

All outcomes were assessed via self-report questionnaires, which are susceptible to social desirability bias and demand characteristics, particularly in adolescent populations where response behavior may be influenced by perceived expectations, peer norms, or the desire to present oneself favorably. Students in the intervention group, aware that they were receiving a dedicated well-being curriculum, may have adjusted their responses to align with the curriculum's aims—though the null main effects suggest that any such demand effect was insufficient to produce inflated scores. Conversely, younger adolescents may lack the metacognitive awareness to accurately report on constructs such as locus of control or growth mindset, introducing measurement noise that attenuates effect sizes. No behavioral outcomes—such as academic grades, attendance records, or disciplinary incidents—were collected, precluding any assessment of whether the curriculum influenced observable behavior even in the absence of self-reported attitudinal change.

Pre-post linkage carries residual uncertainty, because not every record carried a recorded identification code at both occasions (see Method). This bears more on the within-person analyses—the baseline interactions, regression to the mean, and the reliable change indices—than on the group main effects, which compare aggregates and are insensitive to which particular pre-test record is paired with which post-test record. The intervention period of approximately eight months may be insufficient for lasting psychological change, particularly for deeply rooted

constructs like self-esteem, which follows a normative developmental trajectory spanning years rather than months (Orth & Robins, 2014), and locus of control (Taylor et al., 2017). Measurement quality varied considerably across the ten outcomes, and this bears directly on how the null results should be read. Locus of control fell far below conventional internal-consistency standards, with learning orientation and general self-efficacy also below the usual threshold (see Table 4). Unreliability attenuates observed effects toward zero, so for these constructs a genuine effect of moderate size could have produced an observed effect too small to detect. The null results are correspondingly weaker evidence of absence for those three scales than for the well-measured ones, and locus of control is additionally the construct whose models failed to converge. Any future evaluation should replace or revise that scale before treating a null result on it as informative. Finally, no long-term follow-up was conducted, precluding conclusions about whether effects accumulate or emerge at later time points, as Taylor et al. (2017) demonstrated can occur with some SEL programs.

Set against these limitations are several features that make the null informative rather than merely inconclusive. The sample is large for an evaluation of a positive-education curriculum in the German-speaking region and is powered for effects of moderate size, though not for the smallest ones. The multilevel models account for the nesting of students within classes and schools, which analyses of school-based programs frequently neglect and which materially affects the standard errors of a class-level intervention. The group contrast was estimated under seven model specifications and checked against interaction probing, factor-level analysis, equivalence testing, Bayesian model comparison, reliable change indices, class-level profiling, and sensitivity to the inclusion rules; the answer did not depend on the choice. Reporting a null of this kind contributes to a more balanced evidence base in a field with a documented publication bias (Cipriano et al., 2023; Durlak, 2015).

The most important next step would be a randomized controlled trial with school-level randomization, in which entire schools are randomly assigned to intervention and control conditions prior to the start of the academic year. School-level randomization would eliminate the

self-selection confound that pervades the present quasi-experimental design, provide an unambiguous test of the baseline interaction (ruling out RTM as an alternative explanation), and allow for estimation of school-level intraclass correlations that could inform power analyses for future multi-site trials. Such a trial should include an active control condition—for example, a study skills or health education curriculum of equivalent duration and teacher contact—to isolate curriculum-specific effects from non-specific factors such as attention and novelty. Longer follow-up periods of two or more years would test whether effects accumulate with sustained exposure, and whether the developmental trajectory of self-esteem, which normatively dips during early adolescence before rising again (Orth & Robins, 2014), interacts with intervention timing in ways that short-term assessments cannot capture.

A clear gap is the absence of process data that could illuminate how the curriculum is actually experienced by students and delivered by teachers. Future investigations should incorporate structured lesson observation protocols, student engagement logs, and teacher self-reports of delivery quality—not merely dosage—to capture the multidimensional nature of implementation fidelity as conceptualized by Dane and Schneider (1998). Real-time fidelity monitoring, integrated from the outset rather than assessed retrospectively, would enable researchers to distinguish between program failure (the curriculum does not work even when delivered as intended) and implementation failure (the curriculum was not delivered as intended and therefore cannot be evaluated). Dismantling studies that isolate specific curriculum components—perspective-taking exercises, cooperative learning activities, emotional awareness training—would establish whether any single element does what the package as a whole did not. Such designs would have practical value for curriculum developers seeking to concentrate resources on the most effective components rather than delivering a broad curriculum with diffuse effects.

Finally, several broader research directions would strengthen the evidence base for positive education in the DACH region. Adaptive or targeted delivery models should be tested: if low-baseline students genuinely benefit, the curriculum could be offered as an elective or

delivered in small-group settings to students identified through routine well-being screening, rather than administered universally. Cost-effectiveness analyses should accompany future evaluations, as even small effects may justify the investment if the per-student cost is sufficiently low, whereas null effects with high costs would argue for reallocation of resources to alternative programs (Eccles & Roeser, 2011). Cross-cultural comparison studies could determine whether the curriculum functions differently in the Austrian and German school systems, which differ in structure, tracking practices, and pedagogical traditions, and whether country-level moderators explain some of the implementation heterogeneity observed in the present multi-site evaluation. Mixed-methods approaches incorporating qualitative data from students and teachers could illuminate the mechanisms through which the curriculum does or does not produce change, and whether the disruption-without-depth phenomenon observed at moderate implementation levels reflects a genuine barrier to program effectiveness or merely an artifact of the crude implementation measure available in the present data.

In conclusion, the GET *Self-Development* curriculum, as delivered in Austrian and German schools across one academic year, produced no detectable mean-level improvement in any of the psychosocial outcomes assessed. The single contrast that reached nominal significance favoured the control group, did not survive correction for the number of outcomes examined, and did not hold under an alternative model specification. Equivalence testing and Bayesian model comparison provide positive evidence that the effects lie near zero rather than merely leaving the null unrejected. The one pattern that looked substantively interesting—students starting low appearing to gain while students starting high appeared to lose—emerged across nearly every construct and every group coding. Responses in the intervention group were demonstrably less predictable from their own baseline, which is consistent with differential regression to the mean but equally consistent with the curriculum changing the standard against which students rate themselves. This study cannot separate the two, and that question, rather than the null main effect, is the result most worth pursuing. What this study cannot establish is whether the curriculum would work under better conditions, because in most participating classes it was not delivered as

designed. Answering that requires randomization at the school level, monitoring of what is actually taught rather than what was planned, and a follow-up long enough for effects on constructs that develop over years to become visible.

References

Abdel-Khalek, A. M. (2006). Measuring happiness with a single-item scale. *Social Behavior and Personality*, *34*(2), 139–150. https://doi.org/10.2224/sbp.2006.34.2.139

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. H. Freeman.

Barnett, A. G., Pols, J. C. van der, & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, *34*(1), 215–220. https://doi.org/10.1093/ije/dyh299

Beierlein, C., Kemper, C. J., Kovaleva, A., & Rammstedt, B. (2013). Short scale for measuring general self-efficacy beliefs (ASKU). *Methods, Data, Analyses*, *7*(2), 251–278. https://doi.org/10.12758/mda.2013.014

Beierlein, C., Kovaleva, A., Kemper, C. J., & Rammstedt, B. (2012). ASKU – Allgemeine Selbstwirksamkeit Kurzskala [ASKU – General Self-Efficacy Short Scale]. *GESIS Working Papers*, *2012/17*.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, *78*(1), 246–263. https://doi.org/10.1111/j.1467-8624.2007.00995.x

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349–381). Jossey-Bass.

Brunwasser, S. M., Gillham, J. E., & Kim, E. S. (2009). A meta-analytic review of the Penn Resiliency Program's effect on depressive symptoms. *Journal of Consulting and Clinical Psychology*, *77*(6), 1042–1054. https://doi.org/10.1037/a0017671

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, *83*(2), 234–246. https://doi.org/10.1037/0021-9010.83.2.234

Cipriano, C., Strambler, M. J., Naples, L. H., Ha, C., Kirk, M., Wood, M., Sehgal, K., Zieher, A. K., Eveleigh, A., McCarthy, M., et al. (2023). The state of evidence for social and emotional learning: A contemporary meta-analysis of universal school-based SEL interventions. *Child Development*, *94*(5), 1181–1208. https://doi.org/10.1111/cdev.13968

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Collaborative for Academic, Social, and Emotional Learning. (2020). *CASEL's SEL framework: What are the core competence areas and where are they promoted?* https://casel.org/casel-sel-framework-11-2020/

Collani, G. von, & Herzberg, P. Y. (2003). Eine revidierte Fassung der deutschsprachigen Skala zum Selbstwertgefühl von Rosenberg. *Zeitschrift für Differentielle Und Diagnostische Psychologie*, *24*(1), 3–7. https://doi.org/10.1024//0170-1789.24.1.3

Dane, A. V., & Schneider, B. H. (1998). Program integrity in primary and early secondary prevention: Are implementation effects out of control? *Clinical Psychology Review*, *18*(1), 23–45. https://doi.org/10.1016/S0272-7358(97)00043-3

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113–126. https://doi.org/10.1037/0022-3514.44.1.113

Diener, E., Wirtz, D., Tov, W., Kim-Prieto, C., Choi, D., Oishi, S., & Biswas-Diener, R. (2010). New well-being measures: Short scales to assess flourishing and positive and negative feelings. *Social Indicators Research*, *97*(2), 143–156. https://doi.org/10.1007/s11205-009-9493-y

Domitrovich, C. E., Durlak, J. A., Staley, K. C., & Weissberg, R. P. (2017). Social-emotional competence: An essential factor for promoting positive adjustment and reducing risk in

school children. *Child Development*, *88*(2), 408–416. https://doi.org/10.1111/cdev.12739

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. Arnold.

Dray, J., Bowman, J., Campbell, E., Freund, M., Wolfenden, L., Hodder, R. K., McElwaine, K., Tremain, D., Bartlem, K., Bailey, J., Small, T., Palazzi, K., Oldmeadow, C., & Wiggers, J. (2017). Systematic review of universal resilience-focused interventions targeting child and adolescent mental health in the school setting. *Journal of the American Academy of Child & Adolescent Psychiatry*, *56*(10), 813–824. https://doi.org/10.1016/j.jaac.2017.07.780

Durlak, J. A. (2015). *Handbook of social and emotional learning: Research and practice*.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology*, *41*, 327–350. https://doi.org/10.1007/s10464-008-9165-0

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, *82*(1), 405–432. https://doi.org/10.1111/j.1467-8624.2010.01564.x

Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. Psychology Press.

Eccles, J. S., & Roeser, R. W. (2011). Schools as developmental contexts during adolescence. *Journal of Research on Adolescence*, *21*(1), 225–241. https://doi.org/10.1111/j.1532-7795.2010.00725.x

Eisenberg, N., Fabes, R. A., & Spinrad, T. L. (2006). Prosocial development. In W. Damon, R. M. Lerner, & N. Eisenberg (Eds.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (6th ed., pp. 646–718). Wiley.

Esch, T., Jose, G., Gimpel, C., Scheidt, C. von, & Michalsen, A. (2013). Die Flourishing Scale (FS) von Diener et al. Liegt jetzt in einer autorisierten deutschen Fassung (FS-D) vor:

Einsatz bei einer Mind-Body-medizinischen Fragestellung. *Forschende Komplementärmedizin*, *20*(4), 267–275. https://doi.org/10.1159/000354414

Fixsen, D. L., Naoom, S. F., Blase, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature* (Nos. FMHI Publication #231). University of South Florida, Louis de la Parte Florida Mental Health Institute, National Implementation Research Network.

Forman, S. G., & Barakat, N. M. (2011). Cognitive-behavioral therapy in the schools: Bringing research to practice through effective implementation. *Psychology in the Schools*, *48*(3), 283–296. https://doi.org/10.1002/pits.20547

Fritz-Schubert, E. (2008). *Schulfach Glück: Wie ein neues Fach die Schule verändert*. Herder.

Galla, B. M. et al. (2024). Adolescents do not benefit from universal school-based mindfulness interventions. *Frontiers in Psychology*, *15*, 1384531.

Gnambs, T., Freund, M., & Großmann, N. (2020). The factorial structure and construct validity of a German translation of Dweck's Implicit Theories of Intelligence Scale under consideration of the wording effect. *Psychological Test and Assessment Modeling*, *64*(3), 293–313.

Greenberg, M. T., Domitrovich, C. E., Weissberg, R. P., & Durlak, J. A. (2017). Social and emotional learning as a public health approach to education. *The Future of Children*, *27*(1), 13–32.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. https://doi.org/10.3102/0162373707299706

Hemmelmayr, M. (2025). *Evaluation des Schulfachs Selbstentwicklung: Eine quasi-experimentelle Pilotstudie* [Master's thesis]. University of Graz.

Horowitz, J. L., & Garber, J. (2006). The prevention of depressive symptoms in children and

adolescents: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, *74*(3), 401–415. https://doi.org/10.1037/0022-006X.74.3.401

Humphrey, N., Lendrum, A., & Wigelsworth, M. (2010). Social and emotional aspects of learning (SEAL) programme in secondary schools: National evaluation. *Department for Education Research Report*, *DFE-RR049*.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12–19. https://doi.org/10.1037/0022-006X.59.1.12

Jerusalem, M., & Satow, L. (1999). Schulbezogene Selbstwirksamkeitserwartung [School-related self-efficacy expectation]. *Skalen Zur Erfassung von Lehrer- Und Schülermerkmalen*.

Kern, M. L., Waters, L. E., Adler, A., & White, M. A. (2015). A multidimensional approach to measuring well-being in students: Application of the PERMA framework. *The Journal of Positive Psychology*, *10*(3), 262–271. https://doi.org/10.1080/17439760.2014.936962

Keyes, C. L. M. (2005). Mental illness and/or mental health? Investigating axioms of the complete state model of health. *Journal of Consulting and Clinical Psychology*, *73*(3), 539–548. https://doi.org/10.1037/0022-006X.73.3.539

Kieling, C., Baker-Henningham, H., Belfer, M., Conti, G., Ertem, I., Omigbodun, O., Rohde, L. A., Srinath, S., Ulkuer, N., & Rahman, A. (2011). Child and adolescent mental health worldwide: Evidence for action. *The Lancet*, *378*(9801), 1515–1525. https://doi.org/10.1016/S0140-6736(11)60827-1

Kovaleva, A., Beierlein, C., Kemper, C. J., & Rammstedt, B. (2012). IE-4: Konstruktion und Validierung einer Kurzskala zur Messung von Kontrollüberzeugung [IE-4: Construction and validation of a short scale for the assessment of locus of control]. *GESIS Working Papers*, *2012/9*.

Kuyken, W., Ball, S., Crane, C., Ganguli, P., Jones, B., Sheridan-Sheridan, F., Taylor, L., Dalgleish, T., et al. (2022). Effectiveness and cost-effectiveness of universal school-based mindfulness training compared with normal school provision in reducing risk of mental

health problems and promoting well-being in adolescence: The MYRIAD cluster randomised controlled trial. *Evidence-Based Mental Health*, *25*(3), 99–109. https://doi.org/10.1136/ebmental-2021-300396

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods*, *11*(4), 815–852. https://doi.org/10.1177/1094428106296642

Lösel, F., & Beelmann, A. (2003). Effects of child skills training in preventing antisocial behavior: A systematic review of randomized evaluations. *The Annals of the American Academy of Political and Social Science*, *587*(1), 84–109. https://doi.org/10.1177/0002716202250793

Lyubomirsky, S., & Lepper, H. S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research*, *46*(2), 137–155. https://doi.org/10.1023/A:1006824100041

Macnamara, B. N., & Burgoyne, A. P. (2023). Do growth mindset interventions impact students' academic achievement? A systematic review and meta-analysis with recommendations for best practices. *Psychological Bulletin*, *149*(3-4), 133–173. https://doi.org/10.1037/bul0000352

Masten, A. S. (2014). *Ordinary magic: Resilience in development*. Guilford Press.

Miller, G. A., & Chapman, J. P. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology*, *110*(1), 40–48. https://doi.org/10.1037/0021-843X.110.1.40

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. *Psychological Bulletin*, *88*(3), 622–637. https://doi.org/10.1037/0033-2909.88.3.622

Nießen, D., Schmidt, I., Groskurth, K., Rammstedt, B., & Lechner, C. M. (2022). The Internal–External Locus of Control Short Scale–4 (IE-4): A comprehensive validation of

the English-language adaptation. *PLoS ONE*, *17*(7), e0271289. https://doi.org/10.1371/journal.pone.0271289

Norrish, J. M., Williams, P., O'Connor, M., & Robinson, J. (2013). An applied framework for Positive Education. *International Journal of Wellbeing*, *3*(2), 147–161. https://doi.org/10.5502/ijw.v3i2.2

OECD. (2015). *Skills for social progress: The power of social and emotional skills*. OECD Publishing. https://doi.org/10.1787/9789264226159-en

Orth, U., & Robins, R. W. (2014). The development of self-esteem. *Current Directions in Psychological Science*, *23*(5), 381–387. https://doi.org/10.1177/0963721414547414

Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues: A handbook and classification*. American Psychological Association; Oxford University Press.

Pinheiro, J., Bates, D., & Team, R. C. (2023). *nlme: Linear and nonlinear mixed effects models*. https://CRAN.R-project.org/package=nlme

Polanczyk, G. V., Salum, G. A., Sugaya, L. S., Caye, A., & Rohde, L. A. (2015). Annual research review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. *Journal of Child Psychology and Psychiatry*, *56*(3), 345–365. https://doi.org/10.1111/jcpp.12381

Proctor, C., Maltby, J., & Linley, P. A. (2011). Strengths use as a predictor of well-being and health-related quality of life. *Journal of Happiness Studies*, *12*(1), 153–169. https://doi.org/10.1007/s10902-009-9181-2

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Racine, N., McArthur, B. A., Cooke, J. E., Eirich, R., Zhu, J., & Madigan, S. (2021). Global prevalence of depressive and anxiety symptoms in children and adolescents during COVID-19: A meta-analysis. *JAMA Pediatrics*, *175*(11), 1142–1150. https://doi.org/10.1001/jamapediatrics.2021.2482

Rauer, W., & Schuck, K. D. (2003). *FEESS 3-4: Fragebogen zur Erfassung emotionaler und*

sozialer Schulerfahrungen von Grundschülkindern dritter und vierter Klassen. Hogrefe.

Ravens-Sieberer, U., Devine, J., Napp, A.-K., Kaman, A., Saftig, L., Gilbert, M., Reiss, F., Löffler, C., Simon, A., Hurrelmann, K., Walper, S., Schlack, R., Hölling, H., Wieler, L. H., & Erhart, M. (2023). Three years into the pandemic: Results of the longitudinal German COPSY study on youth mental health and health-related quality of life. *Frontiers in Public Health*, *11*, 1129073. https://doi.org/10.3389/fpubh.2023.1129073

Ravens-Sieberer, U., Erhart, M., Devine, J., Gilbert, M., Reiss, F., Barkmann, C., Siegel, N., Simon, A., Hurrelmann, K., Schlack, R., Hölling, H., Wieler, L. H., & Kaman, A. (2022). Child and adolescent mental health during the COVID-19 pandemic: Results of the three-wave longitudinal COPSY study. *Journal of Adolescent Health*, *71*(5), 570–578. https://doi.org/10.1016/j.jadohealth.2022.06.022

Ravens-Sieberer, U., Kaman, A., Erhart, M., Devine, J., Hölling, H., Schlack, R., Löffler, C., Hurrelmann, K., & Otto, C. (2023). Quality of life and mental health in children and adolescents during the first year of the COVID-19 pandemic: Results of a two-wave nationwide population-based study. *European Child & Adolescent Psychiatry*, *32*(4), 575–588. https://doi.org/10.1007/s00787-021-01889-1

Ravens-Sieberer, U., Kaman, A., Erhart, M., Otto, C., Devine, J., Löffler, C., Hurrelmann, K., Bullinger, M., Hölling, H., & Schlack, R. (2021). Seelische Gesundheit und psychische Belastungen von Kindern und Jugendlichen in der ersten Welle der COVID-19-Pandemie – Ergebnisse der COPSY-Studie. *Bundesgesundheitsblatt – Gesundheitsforschung – Gesundheitsschutz*, *64*(12), 1512–1521. https://doi.org/10.1007/s00103-021-03291-3

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press.

Rutter, M. (1979). Protective factors in children's responses to stress and disadvantage. In M. W. Kent & J. E. Rolf (Eds.), *Primary prevention of psychopathology: Vol. 3. Social competence in children* (pp. 49–74). University Press of New England.

Rutter, M. (1987). Psychosocial resilience and protective mechanisms. *American Journal of Orthopsychiatry*, *57*(3), 316–331. https://doi.org/10.1111/j.1939-0025.1987.tb03541.x

Saldern, M. von, & Littig, K.-E. (1987). *LASSO: Landauer Skalen zum Sozialklima für 4. bis 13. Klassen*. Beltz.

Schwartz, C. E., & Sprangers, M. A. G. (1999). Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. *Social Science & Medicine*, *48*(11), 1531–1548. https://doi.org/10.1016/S0277-9536(99)00047-7

Schwarzer, R., & Jerusalem, M. (1995). Generalized Self-Efficacy scale. In J. Weinman, S. Wright, & M. Johnston (Eds.), *Measures in health psychology: A user's portfolio. Causal and control beliefs* (pp. 35–37). NFER-NELSON.

Seligman, M. E. P. (2011). *Flourish: A visionary new understanding of happiness and well-being*. Free Press.

Seligman, M. E. P., Ernst, R. M., Gillham, J., Reivich, K., & Linkins, M. (2009). Positive education: Positive psychology and classroom interventions. *Oxford Review of Education*, *35*(3), 293–311. https://doi.org/10.1080/03054980902934563

Seligman, M. E. P., Steen, T. A., Park, N., & Peterson, C. (2005). Positive psychology progress: Empirical validation of interventions. *American Psychologist*, *60*(5), 410–421. https://doi.org/10.1037/0003-066X.60.5.410

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

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

Spinath, B., Stiensmeier-Pelster, J., Schöne, C., & Dickhaeuser, O. (2002). *SELLMO: Skalen zur Erfassung der Lern- und Leistungsmotivation [SELLMO: Scales for the assessment of learning and achievement motivation]*. Hogrefe.

Spinath, B., Stiensmeier-Pelster, J., Schöne, C., & Dickhäuser, O. (2012). *SELLMO: Skalen zur Erfassung der Lern- und Leistungsmotivation* (2nd ed.). Hogrefe.

Stice, E., Shaw, H., Bohon, C., Marti, C. N., & Rohde, P. (2009). A meta-analytic review of depression prevention programs for children and adolescents: Factors that predict

magnitude of intervention effects. *Journal of Consulting and Clinical Psychology*, *77*(3), 486–503. https://doi.org/10.1037/a0015168

Stiglbauer, B., Gnambs, T., Gamsjäger, M., & Batinic, B. (2013). The upward spiral of adolescents' positive school experiences and happiness: Investigating reciprocal effects over time. *Journal of School Psychology*, *51*(2), 231–242. https://doi.org/10.1016/j.jsp.2012.12.002

Suldo, S. M., & Shaffer, E. J. (2008). Looking beyond psychopathology: The dual-factor model of mental health in youth. *School Psychology Review*, *37*(1), 52–68. https://doi.org/10.1080/02796015.2008.12087908

Taylor, R. D., Oberle, E., Durlak, J. A., & Weissberg, R. P. (2017). Promoting positive youth development through school-based social and emotional learning interventions: A meta-analysis of follow-up effects. *Child Development*, *88*(4), 1156–1171. https://doi.org/10.1111/cdev.12864

Van Breukelen, G. J. P. (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. *Journal of Clinical Epidemiology*, *59*(9), 920–925. https://doi.org/10.1016/j.jclinepi.2006.02.007

Vossen, H. G. M., Piotrowski, J. T., & Valkenburg, P. M. (2015). Development of the Adolescent Measure of Empathy and Sympathy (AMES). *Personality and Individual Differences*, *74*, 66–71. https://doi.org/10.1016/j.paid.2014.09.040

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

Wagner, G., Zeiler, M., Waldherr, K., Philipp, J., Truttmann, S., Dür, W., Treasure, J. L., & Karwautz, A. F. K. (2017). Mental health problems in Austrian adolescents: A nationwide, two-stage epidemiological study applying DSM-5 criteria. *European Child & Adolescent Psychiatry*, *26*(12), 1483–1499. https://doi.org/10.1007/s00787-017-0999-6

Waters, L. (2011). A review of school-based positive psychology interventions. *The Australian Educational and Developmental Psychologist*, *28*(2), 75–90.

https://doi.org/10.1375/aedp.28.2.75

Waters, L., & Loton, D. (2019). SEARCH: A meta-framework and review of the field of positive education. *International Journal of Applied Positive Psychology*, *4*(1-2), 1–46. https://doi.org/10.1007/s41042-019-00017-4

Weare, K., & Nind, M. (2011). Mental health promotion and problem prevention in schools: What does the evidence say? *Health Promotion International*, *26*(suppl 1), i29–i69. https://doi.org/10.1093/heapro/dar075

Werner, E. E., & Smith, R. S. (1992). *Overcoming the odds: High risk children from birth to adulthood.* Cornell University Press.

White, M. A., & Murray, A. S. (2015). *Evidence-based approaches in Positive Education: Implementing a strategic framework for well-being in schools*. Springer. https://doi.org/10.1007/978-94-017-9667-5

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis* (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-24277-4

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

World Health Organization. (2022). *World mental health report: Transforming mental health for all*.

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Chapman; Hall/CRC. https://doi.org/10.1201/b15166

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C. S., Hinojosa, C. P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan, R., Buontempo, J., Yang, S. M., Carvalho, C. M., … Dweck, C. S. (2019). A national experiment reveals where a growth mindset improves

achievement. *Nature*, *573*(7774), 364–369. https://doi.org/10.1038/s41586-019-1466-y

Co-funded by
the European Union

„Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or OeAD-GmbH. Neither the European Union nor the granting authority can be held responsible for them.“

Appendix

Appendices

Appendix A: Scale Overview

Table A1 defines every scale abbreviation used in this manuscript and gives its source, item count, response format, and internal consistency in the present sample. It is the reference for the abbreviations appearing in all other tables.

Table A1
Scale Overview and Psychometric Properties

Scale	Full Name	Items	Response Scale	α pre	α post
SWB	School-Related Self-Efficacy	5	1–4	0.79	0.79
SWS	Flourishing Scale	8	1–7	0.86	0.88
AMES	Empathy and Sympathy	12	1–5	0.82	0.82
ASKU	General Self-Efficacy (Short)	3	1–5	0.69	0.73
CC	Classroom Climate (self-developed)	6	1–4	0.79	0.88
IE	Locus of Control (IE-4)	6	1–5	0.51	0.54
ITIS	Implicit Theories of Intelligence	8	1–6	0.73	0.76
SES	Self-Esteem (Rosenberg)	10	1–4	0.85	0.88
learning	Learning Goal Orientation (SELLMO)	5	1–4	0.62	0.62
subperform	School Performance Satisfaction	1	1–4	–	–

Note. This table defines every scale abbreviation used throughout the manuscript. Items = number of items in the scale; Range = response scale range. Alpha values are Cronbach's alpha in the present sample. Self-esteem (SES) and locus of control (IE) show low internal consistency and their total scores should be interpreted with caution.

Appendix B: Sensitivity Analyses

To assess whether the findings depend on decisions about which participants to retain, the primary group contrast (Model M1) was recomputed under six progressively stricter inclusion criteria. The criteria were: (1) the standard attention-check criterion, at least 2 of 3 checks correct; (2) all 3 pre-test attention checks correct; (3) all 3 attention checks correct at both occasions; (4) pre-post response-pattern correlation at or above .50; (5) correlation at or above .70; and (6) maximum strictness, combining criteria 3 and 4. Table A2 reports the main effects and Table A3 the interactions.

Table A2

Full Sensitivity Analysis Results: Model M1 (Binary Group Effect) Across All Exclusion Criteria and Scales

Criterion	Scale	N	Beta	SE	p	d	Sig
AC 2 of 3 (main analysis)	SWB	524	-0.006	0.047	0.891	-0.014	
AC 2 of 3 (main analysis)	SWS	525	0.045	0.071	0.524	0.063	
AC 2 of 3 (main analysis)	AMES	525	-0.055	0.057	0.336	-0.111	
AC 2 of 3 (main analysis)	ASKU	524	-0.019	0.064	0.769	-0.028	
AC 2 of 3 (main analysis)	CC	526	0.038	0.089	0.672	0.068	
AC 2 of 3 (main analysis)	IE	525	-0.003	0.050	0.946	-0.007	
AC 2 of 3 (main analysis)	ITIS	527	-0.070	0.071	0.320	-0.104	
AC 2 of 3 (main analysis)	SES	528	0.038	0.045	0.401	0.086	
AC 2 of 3 (main analysis)	learning	527	-0.048	0.040	0.222	-0.122	
AC 2 of 3 (main analysis)	subperform	524	0.039	0.072	0.589	0.062	
AC 3 of 3 at pre-test	SWB	408	-0.021	0.053	0.694	-0.046	
AC 3 of 3 at pre-test	SWS	409	0.027	0.068	0.692	0.040	
AC 3 of 3 at pre-test	AMES	409	-0.029	0.064	0.645	-0.059	
AC 3 of 3 at pre-test	ASKU	408	-0.041	0.071	0.565	-0.062	
AC 3 of 3 at pre-test	CC	409	0.078	0.099	0.429	0.136	

Table A2

Full Sensitivity Analysis Results: Model M1 (Binary Group Effect) Across All Exclusion Criteria and Scales

Criterion	Scale	N	Beta	SE	p	d	Sig
AC 3 of 3 at pre-test	IE	409	-0.011	0.058	0.852	-0.020	
AC 3 of 3 at pre-test	ITIS	410	-0.099	0.077	0.199	-0.148	
AC 3 of 3 at pre-test	SES	411	0.053	0.048	0.270	0.118	
AC 3 of 3 at pre-test	learning	410	-0.083	0.045	0.069	-0.205	
AC 3 of 3 at pre-test	subperform	408	-0.005	0.077	0.950	-0.008	
AC 3 of 3 at both occasions	SWB	355	-0.026	0.055	0.628	-0.057	
AC 3 of 3 at both occasions	SWS	356	0.039	0.070	0.576	0.060	
AC 3 of 3 at both occasions	AMES	356	-0.039	0.063	0.534	-0.079	
AC 3 of 3 at both occasions	ASKU	355	-0.038	0.076	0.618	-0.056	
AC 3 of 3 at both occasions	CC	355	0.033	0.104	0.750	0.058	
AC 3 of 3 at both occasions	IE	356	-0.015	0.062	0.814	-0.028	
AC 3 of 3 at both occasions	ITIS	356	-0.104	0.084	0.215	-0.157	
AC 3 of 3 at both occasions	SES	357	0.045	0.052	0.379	0.100	
AC 3 of 3 at both occasions	learning	356	-0.080	0.048	0.096	-0.197	
AC 3 of 3 at both occasions	subperform	355	0.000	0.069	0.999	0.000	
Response agreement r >= .50	SWB	495	0.005	0.048	0.909	0.012	
Response agreement r >= .50	SWS	496	0.029	0.067	0.665	0.040	
Response agreement r >= .50	AMES	496	-0.062	0.057	0.277	-0.125	
Response agreement r >= .50	ASKU	496	0.000	0.064	0.995	0.001	
Response agreement r >= .50	CC	497	0.022	0.089	0.806	0.039	
Response agreement r >= .50	IE	496	-0.008	0.051	0.880	-0.015	
Response agreement r >= .50	ITIS	498	-0.082	0.070	0.248	-0.121	
Response agreement r >= .50	SES	499	0.037	0.046	0.416	0.085	

Table A2

Full Sensitivity Analysis Results: Model M1 (Binary Group Effect) Across All Exclusion Criteria and Scales

Criterion	Scale	N	Beta	SE	p	d	Sig
Response agreement r >= .50	learning	498	-0.047	0.041	0.255	-0.120	
Response agreement r >= .50	subperform	495	0.056	0.069	0.422	0.088	
Response agreement r >= .70	SWB	367	-0.007	0.057	0.905	-0.015	
Response agreement r >= .70	SWS	368	0.016	0.080	0.840	0.022	
Response agreement r >= .70	AMES	369	-0.146	0.057	0.011	-0.299	*
Response agreement r >= .70	ASKU	368	-0.037	0.073	0.612	-0.056	
Response agreement r >= .70	CC	369	-0.018	0.099	0.852	-0.033	
Response agreement r >= .70	IE	368	0.025	0.060	0.682	0.046	
Response agreement r >= .70	ITIS	370	-0.129	0.085	0.131	-0.196	
Response agreement r >= .70	SES	371	0.019	0.051	0.708	0.043	
Response agreement r >= .70	learning	370	-0.056	0.044	0.205	-0.145	
Response agreement r >= .70	subperform	368	0.047	0.076	0.533	0.075	
Maximum strictness	SWB	343	-0.019	0.054	0.729	-0.040	
Maximum strictness	SWS	344	0.048	0.070	0.492	0.075	
Maximum strictness	AMES	344	-0.040	0.065	0.541	-0.080	
Maximum strictness	ASKU	343	-0.034	0.076	0.651	-0.050	
Maximum strictness	CC	343	0.024	0.104	0.819	0.042	
Maximum strictness	IE	344	-0.007	0.062	0.907	-0.014	
Maximum strictness	ITIS	344	-0.104	0.081	0.201	-0.156	
Maximum strictness	SES	345	0.039	0.052	0.452	0.086	
Maximum strictness	learning	344	-0.072	0.047	0.129	-0.181	
Maximum strictness	subperform	343	-0.002	0.070	0.983	-0.002	

Note. Each row represents one model re-estimation with the indicated exclusion criterion. Positive

Beta and d values indicate that the control group outperformed the intervention group. Scale abbreviations are defined in Appendix A. * $p < .05$. ** $p < .01$. *** $p < .001$.

Of the 60 main-effect tests, 1 reached significance, all on empathy (d = -0.30 to -0.06 across criteria), and that estimate does not hold up across neighbouring criteria. Under the strictest criterion—all three attention checks correct at both occasions combined with a pre-post response-pattern correlation of at least .50—no construct approaches significance and the estimates range from d = -0.18 to d = 0.09. The null main effect is therefore not an artefact of the inclusion rules. These are uncorrected tests; the correction for multiple outcomes reported in the Results applies here as well.

The interaction coefficients behave differently: they are of similar magnitude under every criterion, with the mean ranging only from d = -0.21 to d = -0.29, and the strictest criterion gives -0.28. The moderation pattern is therefore not created by any particular inclusion decision. This does not make it substantive: differential regression to the mean would survive these filters just as a genuine moderation would, which is why the argument against interpreting it is made on other grounds in the Discussion.

Table A3

Sensitivity Analysis: Model M2 Baseline × Group Interaction p-Values Across Inclusion Criteria

Scale	C1	C2	C3	C4	C5	C6
AMES	0.302	0.249	0.085	0.231	0.651	0.083
ASKU	0.042*	0.151	0.156	0.036*	0.068	0.088
CC	0.294	0.252	0.157	0.669	0.728	0.339
IE	0.009*	0.061	0.011*	0.003*	0.016*	0.002*
ITIS	0.677	0.410	0.324	0.746	0.844	0.307
SES	0.178	0.195	0.315	0.201	0.406	0.235
SWB	0.004*	0.003*	0.006*	0.000*	0.001*	0.002*
SWS	0.539	0.850	0.833	NA	0.931	0.758

Table A3

Sensitivity Analysis: Model M2 Baseline × Group Interaction p-Values Across Inclusion Criteria

Scale	C1	C2	C3	C4	C5	C6
learning	0.052	0.110	0.091	0.034*	0.002*	0.053
subperform	0.027*	0.187	0.297	0.010*	0.074	0.254

Note. Each cell shows the p-value for the pre-test x group interaction on the scale in that row under the criterion in that column. C1 = attention check 2 of 3 (main analysis); C2 = all 3 pre-test checks correct; C3 = all 3 checks correct at both occasions; C4 = pre-post correlation >= .50; C5 = correlation >= .70; C6 = maximum strictness (C3 and C4 combined). Scale abbreviations are defined in Appendix A. * p < .05.

The interactions were the most robust result in the study. Self-esteem and locus of control reached significance under all eight criteria, flourishing, general self-efficacy, and learning orientation under seven or eight, and classroom climate under three. Whether this robustness reflects a substantive effect or the differential measurement instability documented in Appendix J is the open question discussed in the Discussion.

Appendix C: Complete Interaction Probing Results

The Results section illustrates interaction probing for self-esteem, the construct with the strongest interaction. Table A4 gives the same analysis for every construct: the pre-test score is re-centered at each of five percentiles of its own distribution, and the group effect is re-estimated at that point. The percentile at which the estimate changes sign marks the baseline level above which the intervention group no longer differs favourably from the control group.

Table A4

Interaction Probing: Estimated Group Effects at Baseline Percentiles Across Constructs (Model M2)

Construct	Percentile	Pre Value	IG Effect	SE	t	p	Sig
SES	10%	2.1	0.070	0.075	0.937	0.349	
SES	25%	2.5	0.037	0.056	0.658	0.511	
SES	50%	2.9	0.003	0.048	0.069	0.945	
SES	75%	3.3	-0.030	0.057	-0.526	0.599	
SES	90%	3.6	-0.055	0.072	-0.772	0.441	
SWS	10%	4.5	-0.018	0.129	-0.142	0.887	
SWS	25%	5.5	0.006	0.079	0.082	0.934	
SWS	50%	6.0	0.019	0.074	0.254	0.799	
SWS	75%	6.5	0.031	0.088	0.355	0.723	
SWS	90%	6.9	0.040	0.107	0.380	0.704	
AMES	10%	2.6	-0.015	0.085	-0.176	0.861	
AMES	25%	3.1	-0.057	0.063	-0.906	0.366	
AMES	50%	3.5	-0.092	0.058	-1.593	0.112	
AMES	75%	3.8	-0.120	0.065	-1.839	0.067	
AMES	90%	4.2	-0.147	0.080	-1.843	0.066	
ITIS	10%	3.6	-0.036	0.110	-0.329	0.742	
ITIS	25%	4.0	-0.049	0.090	-0.543	0.587	
ITIS	50%	4.5	-0.065	0.077	-0.847	0.397	
ITIS	75%	5.0	-0.081	0.088	-0.929	0.353	
ITIS	90%	5.5	-0.098	0.116	-0.846	0.398	
IE	10%	1.8	0.130	0.081	1.595	0.112	
IE	25%	2.2	0.059	0.062	0.937	0.349	
IE	50%	2.5	-0.013	0.054	-0.238	0.812	

Table A4
Interaction Probing: Estimated Group Effects at Baseline Percentiles Across Constructs (Model M2)

Construct	Percentile	Pre Value	IG Effect	SE	t	p	Sig
IE	75%	2.8	-0.084	0.059	-1.416	0.158	
IE	90%	3.2	-0.155	0.077	-2.027	0.043	*

Note. IG effect = estimated group coefficient when pre-test is re-centered at the given percentile. * $p < .05$. ** $p < .01$. *** $p < .001$.

Appendix D: Per-School Change Scores

Table A5 reports mean pre-post change by school for four representative outcomes, together with each school's size and the group composition of its participating classes. It is included for transparency about how unevenly the sample is distributed across sites: two schools contribute more than a third of all students, while several contribute fewer than fifteen. These are descriptive means without covariate adjustment and are not comparable across schools; they are reported so that readers can see the heterogeneity underlying the aggregate results.

Table A5
Pre-Post Change Scores for Selected Outcomes by School

School	n	Group(s)	SES Change	AMES Change	SWB Change	CC Change
School_07	134	CG/IG	0.089	0.150	0.090	0.038
School_03	95	IG	0.062	0.060	0.039	-0.210
School_09	53	IG	0.030	0.248	-0.069	-0.288
School_14	53	IG/CG	-0.002	-0.131	0.004	-0.365
School_02	39	CG/IG	0.039	0.096	-0.006	-0.137

Table A5
Pre-Post Change Scores for Selected Outcomes by School

School	n	Group(s)	SES Change	AMES Change	SWB Change	CC Change
School_04	31	CG/IG	0.023	-0.099	0.006	0.258
School_12	29	IG/CG	0.031	0.018	-0.010	-0.384
School_10	26	IG	0.154	0.106	0.262	-0.186
School_06	22	CG/IG	-0.023	-0.030	0.245	0.076
School_11	22	IG/CG	-0.141	0.265	0.136	-0.318
School_05	14	CG/IG	-0.107	-0.128	-0.108	0.167
School_13	8	IG/CG	-0.112	-0.115	0.100	0.417
School_08	2	IG	0.000	-0.708	-0.200	1.833

Note. School identifiers are anonymized. The 13 schools listed are those contributing to the analysis sample; one of the 14 participating schools contributed no linkable pairs and therefore does not appear. Values are mean unadjusted pre-post change scores; positive values indicate an increase from pre-test to post-test. Schools differ in size, group assignment, and implementation level, so these descriptive means are not directly comparable and are reported for transparency rather than inference. Scale abbreviations are defined in Appendix A.

Appendix E: Registration Deviations

The study was registered on the Open Science Framework (OSF pm6uv, linked to project hb79r; Open-Ended Registration, registered 07.09.2025). The registration describes the design, target constructs, and procedures but does not contain preregistered hypotheses, primary outcome designations, or a statistical analysis plan. All analytical decisions were therefore made post hoc, and the inferential framework should be considered exploratory in its analytical specification. Table A6 summarizes the key deviations between the registered plan and the actual

implementation.

Table A6
Summary of registration deviations

Aspect	Registered	Actual	Severity
Outcome measures	"Newly developed, comprehensive scale"	Battery of established, validated instruments (Rosenberg SES, ASKU, IE-based, ITIS, AMES, Flourishing Scale, self-developed CC, self-developed learning)	Substantive (psychometrically improved, but entire measurement level replaced)
Classroom climate	LASSO [Landauer Skalen zum Sozialklima; Saldern and Littig (1987)]	Self-developed items conceptually based on the FEESS (Rauer & Schuck, 2003)	Substantive
Academic performance	Objective grade data from school records	Self-reported satisfaction with grades (1 item)	Substantive
Geographic scope	Austria, Germany, and Switzerland	Austria and Germany only	Moderate
Qualitative component	Mixed-methods with teacher and student interviews	Quantitative only; interviews conducted but reported separately	Expected

Aspect	Registered	Actual	Severity
Analytic approach	Not specified (Open-Ended Registration)	Multilevel modeling, equivalence testing, Bayesian analysis, FDR, Fisher z, RCI, factor analysis	N/A (no analysis plan was registered)
Sample	Grades 1–9, 40 classes from 14 schools, incl. teachers	Grades 5–9, 58 classes from 14 secondary schools; primary grades and teacher data analyzed separately	Moderate
Socioeconomic moderation	Differential effectiveness by socioeconomic background	No direct socioeconomic indicator was collected; parental country of birth is used as a proxy for migration background, which is related to but not interchangeable with socioeconomic position	Substantive (the registered moderator was not measured)

Note. Severity reflects how far the executed study departed from the registered plan. The registration was an Open-Ended Registration and contained no hypotheses, primary outcome designation, or analysis plan, so no deviation from a preregistered analysis was possible.

Appendix F: Complete Questionnaire Items

All items were administered in German. The questionnaire was organized into five parts (*Teil 1–5*). Part 1 collected demographic information (gender, age, home language, siblings, school performance satisfaction). Part 2 assessed subjective happiness and flourishing. Part 3 assessed self-efficacy and school-related self-efficacy. Part 4 assessed self-esteem, classroom climate, locus of control, and growth mindset. Part 5 contained the comprehension check and open-ended feedback. Items marked with (R) are reverse-coded. Four control items were embedded across the questionnaire and are marked in the tables below. Three of them, marked (AC), are directed-response attention checks with an unambiguously correct answer; these determined participant exclusion, with at least two of the three required. The fourth, marked (SD) and embedded in the school-related self-efficacy scale, is a social desirability item: it has no correct answer and was therefore *not* used for exclusion. Neither the (AC) nor the (SD) items enter any scale score.

Subjective School Performance Satisfaction (subperform)

Single item, 4-point scale: 1 = gar nicht zufrieden, 2 = wenig zufrieden, 3 = ziemlich zufrieden, 4 = sehr zufrieden.

Table A7
Subjective School Performance Satisfaction item

Code	Item Text
subperform[1]	Ich bin mit meinen Schulnoten …

Learning Goal Orientation (learning)

5 items from the SELLMO battery (Spinath et al., 2002); 4-point scale: 1 = Stimmt gar nicht, 2 = Stimmt nicht, 3 = Stimmt, 4 = Stimmt genau.

Table A8
Learning Goal Orientation (SELLMO) items

Code	Item Text
learning[1]	Ich bin neugierig und moechte oft neue Dinge lernen.
learning[2]	Wenn ich etwas nicht gleich kann, uebe ich weiter, bis es besser klappt.
learning[3]	Ich lerne gerne neue Dinge, auch ausserhalb der Schule.
learning[4]	Neue Themen interessieren mich meistens.
learning[5]	Ich lerne nur, wenn mich jemand daran erinnert. (R)

Subjective Happiness Scale (HAP / SWS context)

3 items adapted from Lyubomirsky and Lepper (1999); 7-point scales.

Table A9
Subjective Happiness Scale items

Code	Item Text	Response Anchors
HAP1[1]	Ich sehe mich generell als …	1 = nicht sehr glueckliche Person … 7 = sehr glueckliche Person
HAP2[1]	Im Vergleich zu den meisten, die gleich alt sind wie ich, betrachte ich mich als …	1 = weniger gluecklich … 7 = viel gluecklicher

Code	Item Text	Response Anchors
HAP3[1]	Wie sehr beschreibt dich diese Aussage? Manche Menschen sind generell sehr gluecklich. Sie sind fast immer froehlich und lachen sehr oft.	1 = nein, das passt gar nicht … 7 = ja, das bin ich
HAP3[2]	Wie sehr beschreibt dich diese Aussage? Manche Menschen sind generell nicht sehr gluecklich. Auch wenn nichts Schlimmes oder Bloedes passiert ist sind sie nicht sehr froehlich. (R)	1 = nein, das passt gar nicht … 7 = ja, das bin ich

General Self-Efficacy Short Scale (ASKU)

3 substantive items + 1 attention check (Beierlein et al., 2012); 5-point scale: 1 = trifft gar nicht zu, 2 = trifft wenig zu, 3 = trifft etwas zu, 4 = trifft ziemlich zu, 5 = trifft voll und ganz zu.

Table A10
General Self-Efficacy Short Scale (ASKU) items

Code	Item Text
ASKU[1]	In schwierigen Situationen kann ich mich auf meine Faehigkeiten verlassen.
ASKU[2]	Die meisten Probleme kann ich aus eigener Kraft gut meistern.

Code	Item Text
ASKU[3]	Auch anstrengende und komplizierte Aufgaben kann ich in der Regel gut loesen.
ASKU[CNTRL]	Ein Elefant passt in meine Schultasche. (AC)

School-Related Self-Efficacy (SWB)

5 substantive items + 1 social desirability check (Jerusalem & Satow, 1999); 4-point scale: 1 = Stimmt gar nicht, 2 = Stimmt eher nicht, 3 = Stimmt eher, 4 = Stimmt genau.

Table A11
School-Related Self-Efficacy (SWB) items. SD = social desirability check (not used for exclusion).

Code	Item Text
SWB[1]	Ich bin sicher, dass ich die Anforderungen in der Schule schaffen kann, wenn ich mich anstrenge.
SWB[2]	Ich habe die Faehigkeiten, die ich fuer die Schule brauche.
SWB[3]	Ich habe genug Interesse und Ausdauer, um die Sachen fuer die Schule zu schaffen.
SWB[4]	Wenn in der Schule etwas schwierig ist sehe ich das gelassen, da ich in meine Faehigkeiten vertrauen kann.
SWB[5]	Ich schaffe es meistens leicht meine Ziele in der Schule zu erreichen.
SWB[CNTRL]	Ich habe noch nie gelogen. (SD)

Flourishing Scale (SWS)

8 items adapted from Diener et al. (2010; German translation: Esch et al., 2013); 7-point scale: 1 = Stimme ueberhaupt nicht zu, 2 = Stimme nicht zu, 3 = Stimme eher nicht zu, 4 = Weder noch, 5 = Stimme eher zu, 6 = Stimme zu, 7 = Stimme voll und ganz zu.

Table A12
Flourishing Scale (SWS) items

Code	Item Text
SWS[1]	Mein Leben hat einen Sinn und ich habe wichtige Ziele.
SWS[2]	Meine Freunde und meine Familie sind fuer mich da und machen mich gluecklich.
SWS[3]	Ich finde die Sachen, die ich jeden Tag mache, spannend und interessant.
SWS[4]	Ich helfe anderen dabei, dass es ihnen besser geht.
SWS[5]	Sachen die mir wichtig sind kann ich gut.
SWS[6]	Ich bin ein guter Mensch und habe ein gutes Leben.
SWS[7]	Ich glaube, dass meine Zukunft gut wird.
SWS[8]	Meine Freunde und Familie schaetzen und respektieren mich.

Adolescent Measure of Empathy and Sympathy (AMES)

12 items (Vossen et al., 2015); 5-point scale: 1 = Nie, 2 = Fast Nie, 3 = Manchmal, 4 = Oft, 5 = Immer. Items are grouped into three subscales: Cognitive Empathy (CE), Affective Empathy (AE), and Sympathy (SY).

Table A13
Adolescent Measure of Empathy and Sympathy (AMES) items. CE = Cognitive Empathy; AE = Affective Empathy; SY = Sympathy.

Code	Subscale	Item Text
AMES[1]	CE	Ich kann leicht erkennen, was andere fuehlen.
AMES[2]	SY	Ich habe Mitleid mit einem Freund der traurig ist.
AMES[3]	CE	Ich kann oft verstehen, wie sich andere Menschen fuehlen, sogar bevor sie es mir sagen.
AMES[4]	SY	Ich habe Mitleid mit jemandem der unfair behandelt wird.
AMES[5]	AE	Wenn ein Freund wuetend ist, werde ich auch wuetend.
AMES[6]	SY	Ich mache mir Sorgen um Tiere, die verletzt sind.
AMES[7]	AE	Wenn ein Freund traurig ist, werde ich auch traurig.
AMES[8]	CE	Ich erkenne, wenn ein Freund wuetend ist, auch wenn er/sie versucht es zu verstecken.
AMES[9]	AE	Wenn ein Freund Angst hat, fuehle ich mich auch aengstlich.

Code	Subscale	Item Text
AMES[10]	CE	Ich erkenne, wenn jemand so tut, als waere er/sie gluecklich, die Person es aber eigentlich nicht ist.
AMES[11]	SY	Ich mache mir Sorgen, wenn andere Menschen krank sind.
AMES[12]	AE	Wenn Menschen um mich herum nervoes sind, werde ich auch nervoes.

Self-Esteem Scale (SES)

10 substantive items + 1 attention check (Rosenberg, 1965; German adaptation: Collani & Herzberg, 2003); 4-point scale: 1 = Stimme ueberhaupt nicht zu, 2 = Stimme nicht zu, 3 = Stimme zu, 4 = Stimme voll und ganz zu.

Table A14
Self-Esteem Scale (SES; Rosenberg) items

Code	Item Text
SES[1]	Insgesamt bin ich mit mir zufrieden.
SES[2]	Manchmal denke ich, dass ich gar nichts gut kann. (R)
SES[3]	Ich habe viele gute Eigenschaften.
SES[4]	Ich kann die meisten Sachen genauso gut wie die meisten anderen.

Code	Item Text
SES[5]	Ich finde, ich habe nicht viel, worauf ich stolz sein kann. (R)
SES[6]	Manchmal fuehle ich mich nutzlos. (R)
SES[7]	Ich bin wertvoll – wenigstens genauso viel wert wie andere.
SES[8]	Ich wuenschte, ich koennte mehr Wertschaetzung mir gegenueber haben. (R)
SES[9]	Alles in allem habe ich oft das Gefuehl, zu versagen. (R)
SES[10]	Ich habe eine positive Einstellung zu mir selbst.
SES[CNTRL]	Ich habe 25 Finger an meinen Haenden. (AC)

Classroom Climate (CC)

6 items from the FEESS battery (Rauer & Schuck, 2003); 4-point scale: 1 = Stimmt nicht, 2 = Stimmt eher nicht, 3 = Stimmt eher, 4 = Stimmt genau. Stem: "Bei uns in der Klasse …"

Table A15
Classroom Climate (FEESS) items

Code	Item Text
CC[1]	…lacht keiner den anderen aus.
CC[2]	…wird niemand von anderen Kindern geaergert.
CC[3]	…hoeren wir einander zu.
CC[4]	…halten wir alle gut zusammen.
CC[5]	…macht sich keiner ueber den anderen lustig.
CC[6]	…helfen wir uns gegenseitig.

Code	Item Text

Locus of Control (IE)

6 items from the IE-4 short scale (Kovaleva et al., 2012); 5-point scale: 1 = Stimme sehr zu, 2 = Stimme zu, 3 = Teils/teils, 4 = Stimme nicht zu, 5 = Stimme ueberhapt nicht zu. Items IE[1]–IE[3] assess internal locus of control; items IE[4]–IE[6] assess external locus of control.

Table A16
Locus of Control (IE-4) items. Items marked (R) are reverse-coded for computing the composite score (higher = more internal locus).

Code	Subscale	Item Text
IE[1]	Internal	Ich uebernehme gerne Verantwortung.
IE[2]	Internal	Wenn ich mich anstrenge, kann ich die meisten Sachen auch schaffen.
IE[3]	Internal	Wenn es Probleme gibt, finde ich meistens Wege sie zu loesen.
IE[4]	External	In der Schule und zu Hause bestimmen meistens andere, was in meinem Leben passiert. (R)
IE[5]	External	Erfolg haengt oft weniger von Leistung ab, sondern mehr von Glueck. (R)

Code	Subscale	Item Text
IE[6]	External	Bei wichtigen Entscheidungen richte ich mich oft danach, was andere tun. (R)

Implicit Theories of Intelligence Scale (ITIS)

8 substantive items + 1 attention check (Dweck, 1999; German factorial validation: Gnambs et al., 2020); 6-point scale: 1 = Stimme ueberhaupt nicht zu, 2 = Stimme nicht zu, 3 = Stimme eher nicht zu, 4 = Stimme eher zu, 5 = Stimme zu, 6 = Stimme voll und ganz zu. Items are categorized as Entity (fixed mindset) or Incremental (growth mindset) beliefs. Higher composite scores indicate a stronger growth mindset.

Table A17
Implicit Theories of Intelligence Scale (ITIS) items. Entity items are reverse-coded (R) for the composite score.

Code	Type	Item Text
ITIS[1]	Entity	Man hat eine bestimmte Menge an Intelligenz und man kann nicht wirklich viel daran aendern. (R)
ITIS[2]	Entity	Intelligenz ist etwas an einem, das man nicht sehr stark veraendern kann. (R)
ITIS[3]	Entity	Man kann Neues lernen, aber nicht wirklich schlauer werden. (R)

Code	Type	Item Text
ITIS[4]	Incremental	Jeder kann viel schlauer werden.
ITIS[5]	Incremental	Man kann immer viel schlauer werden.
ITIS[6]	Incremental	Egal wie schlau man schon ist, man kann immer noch schlauer werden.
ITIS[7]	Entity	Egal was man macht, man bleibt gleich schlau. (R)
ITIS[8]	Incremental	Sogar wie schlau man von Natur aus ist, kann man veraendern.
ITIS[CNTRL]	–	Schnee ist meistens weiss. (AC)

Comprehension Check

Single item, 4-point scale: 1 = Ich habe alle Fragen gut verstanden, 2 = Ich habe die meisten Fragen gut verstanden, 3 = Ich habe einige Fragen gut verstanden, 4 = Ich habe fast keine Fragen gut verstanden.

Table A18
Comprehension check item

Code	Item Text
check	Du hast es geschafft! Sag uns nun bitte noch wie gut du die Fragen verstanden hast.

Note. The demographic section additionally collected information on country of birth (student, mother, father; each with options *Ja*, *Nein*, *Ich weiss es nicht*), which served as a proxy for migration background. An open-ended feedback item at the end invited students to comment on their experience with the questionnaire.

Appendix G: Responder Analysis

The analyses in the main text compare group means. A mean can stay flat while individual students move in opposite directions, so this appendix asks a different question: what proportion of students changed by more than measurement error would produce, on at least one outcome?

Although no mean-level effects favored the IG, an individual-level responder analysis revealed that a higher proportion of IG students showed reliable improvement (Reliable Change Index) on at least one of the nine multi-item scales: 43.4% of IG students improved reliably on at least one outcome, compared to 37.9% of CG students. This 5.5 percentage-point difference suggests that the curriculum may facilitate individual-level change for a subset of students, even though these gains are offset at the group level by students who show no change or decline. This finding should be interpreted with caution, as it does not account for clustering and may partly reflect the greater overall instability observed in IG response patterns.

Appendix H: Factor-Level Analysis

An exploratory factor analysis was conducted on the pre-test scores across the outcome scales. The Kaiser-Meyer-Olkin measure of sampling adequacy was .87, and Bartlett's test of sphericity was significant ($\chi^2 = 1566.1$, $df = 45$, $p < .001$), confirming the data's suitability for factor analysis. The Kaiser criterion retained 3 components, but a three-factor extraction produced an ultra-Heywood case (an estimated communality above unity), which makes that solution inadmissible. The reported solution is therefore the 2-factor one, the most differentiated structure the data admit; it explains 51.2% of the total variance (eigenvalues: 3.97, 1.15; Table A19).

Factor 1 is a broad positive self-evaluation dimension, defined by flourishing, school-related self-efficacy, self-esteem, and general self-efficacy, with weaker contributions from satisfaction with school performance, classroom climate, and learning orientation. Factor 2 is a motivational dimension carried mainly by implicit theories of intelligence, with locus of control loading negatively on both factors. Empathy is the one construct the common structure fails to capture: its highest loading is only 0.21 and its communality is 0.04, meaning that almost none of its variance is shared with the remaining scales. Psychometrically this confirms that empathy stands apart from the other outcomes, but it does so by being poorly represented in this solution rather than by forming a factor of its own, as an earlier three-factor extraction had suggested.

Table A19

Exploratory Factor Analysis of Pre-Test Scores With Oblimin Rotation

Scale	F1: Self-evaluation	F2: Motivation	Communality
SWB	0.769	0.051	0.630
SWS	0.853	-0.049	0.691
AMES	0.002	0.205	0.042
ASKU	0.698	0.006	0.491
CC	0.460	-0.198	0.167
IE	-0.341	-0.323	0.322

Table A19
Exploratory Factor Analysis of Pre-Test Scores With Oblimin Rotation

Scale	F1: Self-evaluation	F2: Motivation	Communality
ITIS	0.001	0.651	0.425
SES	0.759	0.002	0.577
learning	0.390	0.284	0.335
subperform	0.465	0.122	0.283

Note. Factor loadings after oblimin rotation; communality = proportion of a scale's variance explained by the retained factors. A three-factor extraction was inadmissible (ultra-Heywood case), so the most differentiated admissible solution is reported. Scale abbreviations are defined in Appendix A. This analysis requires item-level data and is not reproducible from the anonymized publication dataset; it is regenerated by scripts/analysis/recompute_precomputed.R (see README).

Table A20 presents the multilevel model results for the factor scores. Consistent with the construct-level analyses, no significant main effects emerged for any factor under either the binary or the dose-response coding. The Baseline × Group interaction reached nominal significance for Factor 1 (β = -0.140, p = .044) but not for Factor 2 (β = -0.126, p = .102). The factor level therefore shows the same picture as the construct level: no aggregate group difference, and a baseline-dependent pattern that is confined to the broad self-evaluation dimension and does not survive the multiplicity correction applied to the construct-level interactions.

Table A20
Factor-Level Multilevel Models: Main Effects and Baseline Interactions

Factor	Model	Beta	SE	t	p	Sig
F1: Self-evaluation	Binary main effect	0.015	0.072	0.202	.840	
F1: Self-evaluation	Dose main effect	0.002	0.026	0.067	.947	

Table A20
Factor-Level Multilevel Models: Main Effects and Baseline Interactions

Factor	Model	Beta	SE	t	p	Sig
F1: Self-evaluation	Baseline × Group interaction	-0.140	0.069	-2.024	.044	*
F2: Motivation	Binary main effect	-0.052	0.067	-0.781	.435	
F2: Motivation	Dose main effect	0.002	0.024	0.096	.923	
F2: Motivation	Baseline × Group interaction	-0.126	0.077	-1.637	.102	

Note. Factor scores from the EFA with oblimin rotation, computed at both occasions from the pre-test loadings so the two time points share a metric. Binary main effect = intervention group (IG) versus control group (CG); dose main effect = reported teaching hours; Baseline × Group interaction = pre-test score by group. β = unstandardized coefficient. * $p < .05$. ** $p < .01$. *** $p < .001$. Requires item-level data; regenerated by scripts/analysis/recompute_precomputed.R (see README).

Appendix I: Implementation Heatmap

Mean pre-post change scores are shown separately for each implementation level and outcome construct in Figure A1. The figure is descriptive: the values are *unadjusted* change scores, whereas the inferential implementation analyses are the baseline-adjusted multilevel models reported in Table 8 and Table 13.

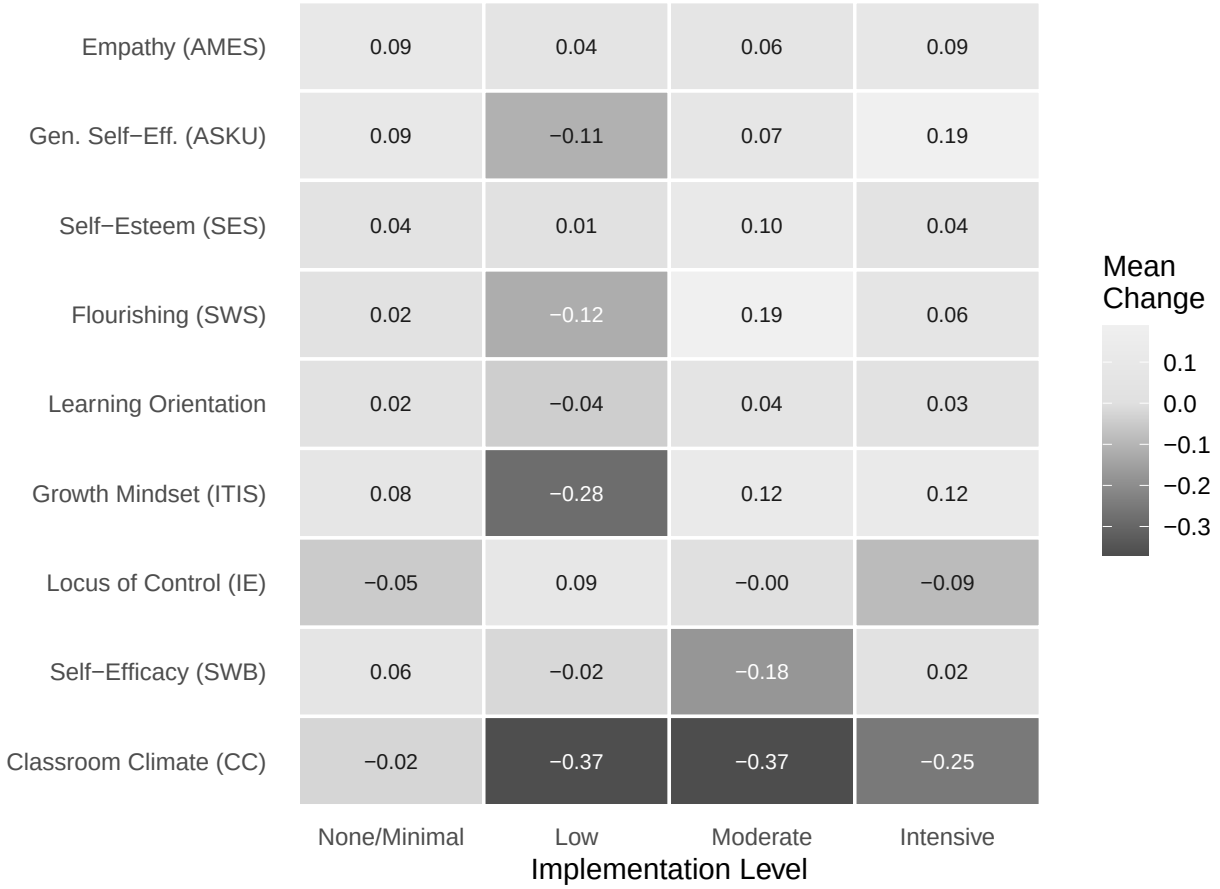
Appendix J: Model Assumption Checks

The multilevel models used throughout this manuscript rest on assumptions that, if violated, can distort both estimates and standard errors. This appendix documents whether they hold.

Each primary model was refitted for this appendix from the published dataset, so the

Figure A1

Heatmap of mean unadjusted *pre-post change scores by implementation level and outcome construct. Lighter values indicate improvement, darker values indicate decline. No consistent dose-response gradient is visible; moderate implementation shows the most negative changes for several constructs. Inferential tests are the baseline-adjusted models in Table 8 and Table 13.*



diagnostics below refer to exactly the models reported in the Results. Normalized residuals were used throughout, which rescale the raw residuals by the estimated variance structure and are therefore the appropriate quantity for assessing multilevel model fit.

Two assumptions held well. Residual distributions were close to symmetric and only mildly heavy-tailed, with the largest absolute skewness across outcomes at 1.60 and the largest absolute excess kurtosis at 7.42; departures of this size have negligible consequences for fixed-effect estimates at the present sample size. Linearity also held: the likelihood-ratio test for an added quadratic pre-test term was significant for 2 of the 9 outcomes, so modeling the pre-post

relation as linear is adequate.

The homoscedasticity assumption did not hold. The Breusch-Pagan test was significant for 5 of the 9 outcomes, and the direction is systematic rather than random: the ratio of residual standard deviations between the intervention and control groups ranged from 1.00 to 1.19, exceeding 1 for every outcome. Residual variability was thus consistently greater in the intervention group. This is the same phenomenon that the Fisher z-tests detect as lower pre-post stability in the intervention group, appearing here in the residual variance of the outcome models. Because a model assuming a common residual variance can misstate standard errors when this assumption fails, each primary model was refitted allowing the residual variance to differ by group.

Allowing a separate residual variance per group changed no conclusion: across the 9 outcomes for which both models converged, 0 changed significance status at $\alpha = .05$. The reported fixed-effect results are therefore not an artifact of the equal-variance assumption. Table A21 reports the full diagnostics, Table A22 the comparison of the two variance specifications, and Figure A2 the corresponding quantile-quantile plots.

Table A21

Assumption Diagnostics for the Primary Model of Each Outcome

Construct	n	Resid. skew	Resid. kurtosis	BP p	SD ratio	Linearity p	RE skew	RE kurtosis
SWB	469	-0.058	-0.150	0.000	1.194	0.052	0.135	-0.335
SWS	470	-1.596	7.417	0.374	1.003	0.001	-0.368	0.110
AMES	470	-0.455	1.366	0.000	1.014	0.677	-0.051	-0.114
ASKU	468	-0.103	0.153	0.001	1.042	0.837	-0.782	0.417
CC	471	-0.122	0.210	0.583	1.035	0.270	-0.595	0.018
IE	469	0.591	1.080	0.832	1.023	0.587	1.103	1.456
ITIS	471	-0.161	0.289	0.034	1.000	0.941	-0.309	-0.153
SES	472	-0.307	0.190	0.809	1.055	0.031	-0.321	0.443

Table A21
Assumption Diagnostics for the Primary Model of Each Outcome

Construct	n	Resid. skew	Resid. kurtosis	BP p	SD ratio	Linearity p	RE skew	RE kurtosis
learning	471	-0.051	0.106	0.019	1.027	0.594	-0.565	1.632

Note. All diagnostics refer to Model M1 refitted on the published dataset. Residual skewness and excess kurtosis are computed on normalized residuals; both are zero for a normal distribution. BP *p* is the *p*-value of a Breusch-Pagan test regressing squared residuals on fitted values; a small value indicates heteroscedasticity. SD ratio is the residual standard deviation of the intervention group divided by that of the control group; 1 indicates equal spread. Linearity *p* is the likelihood-ratio test comparing the model against one with an added quadratic pre-test term; a small value indicates a non-linear pre-post relation. RE skew and RE kurtosis describe the distribution of the estimated class-level intercepts. Locus of control (IE) is reported as missing because the model did not converge; the same non-convergence is noted for that construct throughout. Scale abbreviations are defined in Appendix A.

Table A22
Group Effect Under Equal Versus Group-Specific Residual Variance

Construct	p (equal variance)	p (group-specific)
SWB	0.786	0.830
SWS	0.829	0.828
AMES	0.141	0.144
ASKU	0.916	0.906
CC	0.939	0.923
IE	0.700	0.694
ITIS	0.396	0.394

Table A22
Group Effect Under Equal Versus Group-Specific Residual Variance

Construct	p (equal variance)	p (group-specific)
SES	0.920	0.902
learning	0.292	0.293

Note. Both columns report the *p*-value of the intervention-versus-control contrast in Model M1. The left column assumes one residual variance for all students, as in the models reported in the Results; the right column allows the residual variance to differ between the intervention and control groups. Locus of control (IE) is missing because the model did not converge under either specification. Scale abbreviations are defined in Appendix A.

Appendix K: FDR-Corrected Interaction Tests

The Baseline × Group interaction was tested for every construct under each of the three group codings. Table A23 lists all of these tests with their uncorrected and Benjamini-Hochberg corrected *p*-values, so that the effect of the correction can be inspected test by test rather than only in summary.

Table A23
FDR-Corrected p-Values for All 30 Baseline × Group Interaction Tests

Construct	Model	p (uncorrected)	p (FDR)	Sig (uncorr.)	Sig (FDR)
SWB	M4: broad binary	.002	.028	*	*
AMES	M6: dose	.002	.028	*	*
SWB	M2: strict binary	.004	.034	*	*
IE	M4: broad binary	.005	.034	*	*
SWB	M6: dose	.006	.034	*	*
learning	M6: dose	.010	.049	*	*

Table A23
FDR-Corrected p-Values for All 30 Baseline × Group Interaction Tests

Construct	Model	p (uncorrected)	p (FDR)	Sig (uncorr.)	Sig (FDR)
learning	M2: strict binary	.012	.051	*	
IE	M2: strict binary	.015	.055	*	
learning	M4: broad binary	.028	.090	*	
ASKU	M4: broad binary	.031	.090	*	
subperform	M4: broad binary	.033	.090	*	
subperform	M2: strict binary	.043	.107	*	
ASKU	M2: strict binary	.065	.150	NA	
IE	M6: dose	.102	.219	NA	
SES	M4: broad binary	.147	.294	NA	
SWS	M6: dose	.195	.365	NA	
CC	M2: strict binary	.207	.365	NA	
SES	M2: strict binary	.255	.420	NA	
AMES	M2: strict binary	.266	.420	NA	
subperform	M6: dose	.336	.505	NA	
CC	M4: broad binary	.374	.534	NA	
AMES	M4: broad binary	.418	.570	NA	
CC	M6: dose	.447	.583	NA	
ASKU	M6: dose	.481	.587	NA	
ITIS	M6: dose	.489	.587	NA	
SES	M6: dose	.602	.695	NA	
ITIS	M4: broad binary	.653	.726	NA	
ITIS	M2: strict binary	.711	.762	NA	
SWS	M2: strict binary	.748	.774	NA	

Table A23
FDR-Corrected p-Values for All 30 Baseline × Group Interaction Tests

Construct	Model	p (uncorrected)	p (FDR)	Sig (uncorr.)	Sig (FDR)
SWS	M4: broad binary	.839	.839	NA	

Note. Each row is one Baseline × Group interaction test. M2 = strict binary coding (intervention group [IG] vs. control group [CG]); M4 = broad binary coding; M6 = dose coding (reported teaching hours as a continuous predictor). p (uncorrected) = raw p-value; p (FDR) = Benjamini-Hochberg corrected value. Asterisks mark $p < .05$ before and after correction. Scale abbreviations are defined in Appendix A.

Figure A2

Quantile-quantile plots of the normalized residuals from the primary model of each outcome. Points following the diagonal indicate residuals consistent with normality; systematic departure at the extremes indicates heavier or lighter tails than normal.