

Das Schulfach Selbstentwicklung — was die Evaluation zeigt

Interner Bericht für Global Educational Transformation

David Matischek, Universität Graz

11. August 2026

Inhaltsverzeichnis

1	Kurzfassung	1
2	Was untersucht wurde	2
3	Der Hauptbefund	2
4	Der Umsetzungsbefund	3
5	Was für GET spricht	4
5.1	Die Zielgröße ist richtig gewählt	4
5.2	Innerhalb derselben Schule dreht sich die Richtung	4
5.3	Kein Schaden	4
5.4	Ein sauberer Nullbefund ist ein Beitrag	5
6	Empfehlungen	5
7	Durchführung: was funktioniert hat und was nicht	6
7.1	Schulen gewinnen	6
7.2	Kontrollklassen	6
7.3	Erhebung vor Ort	7
7.4	Protokolle und Datenfluss	7
7.5	Team und Rollen	7
7.6	Zeitplanung	8
8	Empfehlungen für die nächste Erhebung	8
9	Was diese Studie nicht beantworten kann	9

1 Kurzfassung

Die Evaluation des Schulfachs *Selbstentwicklung* umfasst 528 Schüler:innen aus 59 Klassen an 13 Schulen in Österreich und Deutschland, jeweils zu Schuljahresbeginn und -ende befragt. Verglichen wurden 304 Schüler:innen aus Interventionsklassen mit 224 aus Kontrollklassen.

Das zentrale Ergebnis ist ein Nullbefund. Auf keiner der zehn erhobenen Größen — Wohlbefinden, Selbstwert, Selbstwirksamkeit, Empathie, Growth Mindset, Klassenklima, Kontrollüberzeugung, Lernorientierung, Leistungszufriedenheit — unterscheidet sich die Entwicklung der Interventionsklassen bedeutsam von der der Kontrollklassen. Der größte gemessene Unterschied liegt bei einer Effektstärke von 0.18; als praktisch relevant gilt in diesem Feld üblicherweise erst 0,20.

Ebenso wichtig: Es gibt keinen Hinweis auf Schaden. Auf keiner Größe entwickelten sich die Interventionsklassen signifikant ungünstiger. In der Präventionsforschung ist das keine Selbstverständlichkeit.

Der praktisch bedeutsamste Befund betrifft nicht die Wirkung, sondern die Umsetzung. Nur 51 % der Schüler:innen in Interventionsklassen waren in einer Klasse, in der das Fach tatsächlich wöchentlich stattfand. 27 % saßen in Klassen, in denen es gar nicht oder nur vereinzelt unterrichtet wurde. Das Fach, das evaluiert wurde, ist damit in weiten Teilen nicht das Fach, das konzipiert wurde.

2 Was untersucht wurde

Erhoben wurde mit etablierten, validierten Fragebogenverfahren zu zwei Zeitpunkten. Die Zuordnung zur Interventions- oder Kontrollbedingung erfolgte auf Klassenebene durch die Schulen selbst, nicht zufällig — das ist die wesentliche methodische Einschränkung und bedeutet, dass Unterschiede zwischen den Gruppen auch andere Ursachen haben können als das Fach.

Die Auswertung berücksichtigt, dass Schüler:innen in Klassen und Klassen in Schulen geschachtelt sind. Zusätzlich zur Frage „wirkt es?“ wurde geprüft, ob die Daten positiv für die *Abwesenheit* eines Effekts sprechen (Äquivalenztests und Bayes-Faktoren) und ob die Wirkung davon abhängt, wie viel tatsächlich unterrichtet wurde.

3 Der Hauptbefund

Tabelle 1: Gruppenunterschied in der Entwicklung über das Schuljahr

Konstrukt	Effektstärke d	p-Wert	Bewertung
Empathie	-0.18	0.14	kein Unterschied
Lernorientierung	-0.11	0.29	kein Unterschied
Growth Mindset	-0.10	0.40	kein Unterschied
Leistungszufriedenheit	+0.08	0.50	kein Unterschied
Kontrollüberzeugung	-0.04	0.70	kein Unterschied
Schulische Selbstwirksamkeit	+0.03	0.79	kein Unterschied
Wohlbefinden (Flourishing)	+0.02	0.83	kein Unterschied
Allgemeine Selbstwirksamkeit	+0.01	0.92	kein Unterschied
Klassenklima	+0.01	0.94	kein Unterschied

Tabelle 1: Gruppenunterschied in der Entwicklung über das Schuljahr

Konstrukt	Effektstärke d	p-Wert	Bewertung
Selbstwert	+0.01	0.92	kein Unterschied

Ein positives Vorzeichen bedeutet einen Vorteil der Interventionsgruppe, ein negatives einen Vorteil der Kontrollgruppe. Kein Wert erreicht statistische Signifikanz, und alle liegen deutlich unter der Schwelle praktischer Relevanz.

Dieser Befund ist belastbar: Er bleibt bestehen, wenn man strenger auswählt, wen man einbezieht, wenn man die Gruppen anders zuschneidet und wenn man nur jene Schüler:innen betrachtet, deren Zuordnung über beide Zeitpunkte zweifelsfrei dokumentiert ist. Bei 3 der zehn Größen zeigen Äquivalenztests, dass ein bedeutsamer Effekt statistisch ausgeschlossen werden kann; für alle zehn spricht der Bayes-Faktor mit einem Verhältnis zwischen 7:1 und 22:1 für die Annahme, dass kein Effekt vorliegt.

Was der Nullbefund nicht heißt: Er heißt nicht, dass das Fach wirkungslos *ist*. Er heißt, dass unter den Bedingungen dieses Schuljahres, mit diesen Messinstrumenten und dieser Stichprobengröße keine Wirkung nachweisbar war. Für kleine Effekte war die Studie zu klein.

4 Der Umsetzungsbefund

Dies ist aus unserer Sicht das Ergebnis mit dem größten praktischen Wert.

Tabelle 2: Tatsächliche Umsetzung in den Interventionsklassen

Umsetzung	Schüler:innen	Anteil
wöchentlich (wie konzipiert)	155	51 %
unregelmäßig	9	3 %
wenige Male im Schuljahr	29	10 %
gar nicht oder minimal	53	17 %
nicht dokumentiert	58	19 %

Von den Interventionsklassen haben 16 das Fach wöchentlich unterrichtet und 13 gar nicht oder nur wenige Male im Schuljahr. Bei einem weiteren Teil ließ sich die Umsetzung aus den Protokollen nicht rekonstruieren.

Entscheidend ist, was daraus folgt: **Wir haben geprüft, ob mehr Unterricht zu besseren Ergebnissen führt — und finden keinen solchen Zusammenhang.** Kein einziger Dosis-Wirkungs-Zusammenhang wird signifikant, alle liegen unter einer Effektstärke von 0.05. Klassen, die das Fach wöchentlich hatten, unterscheiden sich nicht von Klassen, die es nie hatten.

Das ist unbequem, aber wichtig: Der Nullbefund lässt sich **nicht** allein damit erklären, dass zu wenig unterrichtet wurde. Hätte das Fach in der umgesetzten Form gewirkt, müsste sich das bei den wöchentlich unterrichtenden Klassen zeigen. Es zeigt sich nicht.

Zugleich bleibt richtig: Was evaluiert wurde, war zu großen Teilen nicht das konzipierte Fach. Beide Aussagen gelten nebeneinander.

5 Was für GET spricht

5.1 Die Zielgröße ist richtig gewählt

Eine begleitende Bachelorarbeit (Hummel, 2026) hat an den Ausgangsdaten dieser Studie gezeigt, dass Growth Mindset das Wohlbefinden von Jugendlichen bedeutsam vorhersagt ($r = .27$). Wir haben das nachgerechnet und bestätigt. Das heißt: Die Größe, die GET adressiert, hängt in genau dieser Population tatsächlich mit dem zusammen, was GET erreichen will. **Was nicht gelang, war die Veränderung dieser Größe — nicht ihre Auswahl.**

5.2 Innerhalb derselben Schule dreht sich die Richtung

Neun Schulen führten sowohl Interventions- als auch Kontrollklassen. Vergleicht man nur innerhalb dieser Schulen — also Klassen desselben Hauses, mit denselben Rahmenbedingungen — kehrt sich das Bild:

Vergleich	Konstrukte zugunsten der Intervention
über alle Schulen	2 von 9
innerhalb der Schulen	6 von 9

Signifikant wird nichts, und die Teilstichprobe ist kleiner. Diese Analyse beweist keine Wirkung. Aber sie zeigt, dass der Vergleich über Schulen hinweg mit Unterschieden zwischen den Schulen vermischt ist — Schulen, die GET einführen, unterscheiden sich von Schulen, die es nicht tun. Für künftige Evaluationen heißt das: **Kontrollklassen sollten wenn möglich aus derselben Schule kommen.**

5.3 Kein Schaden

Auf keiner Größe entwickelten sich die Interventionsklassen signifikant schlechter. Eine frühere Auswertung hatte einen Nachteil bei Empathie nahegelegt; dieser Befund hat sich als Artefakt einer fehlerhaften Datenverknüpfung erwiesen und ist nach deren Korrektur verschwunden.

5.4 Ein sauberer Nullbefund ist ein Beitrag

Im Feld der schulischen Persönlichkeitsförderung dominieren Einzelstudien mit positiven Ergebnissen aus methodisch schwächeren Designs. Die große britische SEAL-Evaluation (Humphrey et al., 2010) fand über rund 40 Schulen ebenfalls keine Effekte, mit derselben Erklärung: sehr unterschiedliche Umsetzungsqualität. Auch die interne Pilotstudie (Hemmelmayr, 2025) fand außer einem Empathieeffekt nichts, der sich hier nicht bestätigt hat.

Dass zwei unabhängige Auswertungen mit verschiedenen Stichproben, Designs und Auswertenden zu demselben Bild kommen, macht den Befund belastbarer als jede Einzelstudie es wäre. GET kann das offensiv vertreten: Eine Organisation, die ihr eigenes Programm ergebnisoffen evaluieren lässt und einen Nullbefund veröffentlicht, ist glaubwürdiger als eine, die es nicht tut.

6 Empfehlungen

- 1. Umsetzung sichern, bevor erneut evaluiert wird.** Eine Wirksamkeitsstudie an einem Fach, das in der Hälfte der Klassen nicht stattfindet, kann die Frage nicht beantworten. Vor der nächsten Evaluation braucht es eine verbindliche Mindestdosis und deren laufende Erfassung — nicht rückblickend aus Protokollen, sondern begleitend.
- 2. Kontrollklassen aus derselben Schule.** Das ist der wirksamste Hebel gegen den Selektionseffekt und kostet nichts außer Absprache.
- 3. Einzelne Bausteine testen statt des Gesamtpakets.** Wenn das Paket als Ganzes keinen messbaren Effekt zeigt, ist die nächste sinnvolle Frage, welcher Baustein etwas bewirkt — Perspektivenübernahme, kooperatives Lernen, Emotionsarbeit. Das ist mit kleineren, kürzeren Studien machbar.
- 4. Erhebung technisch absichern.** Ein erheblicher Teil des Auswertungsaufwands ging darauf, dass die Zugangscodes bei der ersten Erhebungswelle nicht gespeichert wurden und nachträglich rekonstruiert werden mussten. Ein automatisch vergebener, gespeicherter Code löst das vollständig.
- 5. Längere Beobachtungszeit.** Acht Monate sind für Größen wie Selbstwert kurz; die Forschung zeigt Entwicklungsverläufe über Jahre. Eine Nacherhebung ein Jahr später wäre aussagekräftig und mit geringem Aufwand möglich.
- 6. Die Erwartung anpassen.** Universelle Programme für alle Schüler:innen erzielen international kleine Effekte. Ein Fach, das für alle gleich gut wirkt, ist unrealistisch; realistischer ist eines, das für einen Teil der Schüler:innen deutlich etwas bringt. Diese Frage lässt sich nur mit gezielterer Messung beantworten.

7 Durchführung: was funktioniert hat und was nicht

Dieser Abschnitt ist der eigentliche Zweck des Dokuments. Die Ergebnisse stehen im wissenschaftlichen Bericht; was hier folgt, ist das Erfahrungswissen, das sonst mit dem Personalwechsel verloren geht.

7.1 Schulen gewinnen

GET hat die Liste der teilnehmenden Schulen und Klassen geliefert. Diese Liste war der Ausgangspunkt, aber **nicht der Stand**: Bei jeder Schule musste einzeln nachgefragt werden, ob sie tatsächlich an der Studie teilnimmt. Zwischen „ist im Programm“ und „macht bei der Evaluation mit“ liegt ein Schritt, der aktiv geklärt werden muss.

Der Genehmigungsweg über die Bildungsdirektionen ist der längste Posten im Zeitplan. Realistisch sind **drei bis vier Monate**, bis alles steht; zwei Monate sind machbar, wenn man konsequent nachfasst. Wer diese Zeit nicht einplant, verliert den geplanten Erhebungszeitpunkt. In Deutschland kam eine Absage der Bildungsdirektion, in einem weiteren Fall wollte die Direktion selbst nicht — beides Gründe, die sich nicht abwenden lassen, die man aber früh genug erfahren muss, um umzuplanen.

Ansprechpartner: Oft war es die Direktion, häufig aber die von GET weitergebildete Lehrkraft. Am besten lief es, wenn **die Direktion von Anfang an informiert und als zentrale Ansprechperson eingebunden** war. Läuft der Kontakt nur über eine einzelne Lehrkraft, hängt das ganze Projekt an dieser Person.

Kontaktaufnahme zu rund 80 % telefonisch. E-Mails blieben oft lange unbeantwortet. Das ist keine Unhöflichkeit der Schulen, sondern der Alltag dort — wer auf E-Mail-Antworten wartet, wartet. Terminfindung für die Erhebungen lief ebenfalls fast durchgehend telefonisch und war mühsam, weil sie für jede Klasse einzeln geführt werden musste.

Die Studie bedeutet für Schulen, Klassen und Lehrkräfte **Mehraufwand ohne unmittelbaren Gegenwert**. Es war regelmäßig Überzeugungsarbeit nötig.

7.2 Kontrollklassen

Die Akquise der Kontrollklassen hätte ohne beharrliches Nachfragen und ohne Erklären, **warum** eine Kontrollgruppe unverzichtbar ist, nicht funktioniert. Das war vor allem Gesprächsarbeit mit den Direktionen. Wer diesen Teil unterschätzt, bekommt am Ende eine Studie ohne Vergleichsgruppe — und damit keine Studie.

Die Auswertung zeigt, wie folgenreich das ist: Der Vergleich über Schulen hinweg ist mit allem konfundiert, was Schulen unterscheidet. Nur in Schulen mit beiden Armen lässt sich sauber vergleichen, und dort dreht sich die Richtung der Befunde. **Kontrollklassen aus derselben Schule sind kein Nice-to-have, sondern der Unterschied zwischen aussagekräftig und nicht aussagekräftig.**

7.3 Erhebung vor Ort

Das Team ist selbst in die Schulen gefahren und hat vor Ort erhoben. Das ist **teuer und zeitintensiv**, hat aber die Datenqualität gesichert — standardisierte Durchführung, Ansprechpartner bei Rückfragen, kontrollierte Bedingungen.

Einweisung der Untersuchungsleitenden: mindestens eine Klasse gemeinsam mit einer erfahrenen Erhebungsleitung, danach selbstständig, außer die Person wollte oder brauchte noch Begleitung. Dieses Vorgehen hat sich bewährt und sollte beibehalten werden.

LimeSurvey war die größte technische Schwachstelle. Zwei Probleme traten regelmäßig auf:

1. **Limits bei gleichzeitigen Zugriffen** — besonders wenn mehrere Klassen derselben Schule an einem Tag im selben Raum auf denselben Geräten erhoben wurden.
2. **Sperre nach mehrfacher Falscheingabe des Tokens** — dann mussten die Schüler:innen einige Minuten warten.

Beides führte zu Wartezeiten und Unruhe im Klassenraum und zwang teilweise zum Ausweichen auf Papierfragebögen. Was geholfen hat: privater Browsermodus, andere Geräte, mobiler Hotspot oder ein anderer Internetzugang. Das sind Notlösungen, keine Lösung.

Für die nächste Erhebung ist eine eigene LimeSurvey-Instanz in Abstimmung mit UNIT geplant. Zusätzlich: pro Klasse eine eigene Umfrage statt einer einzigen für alle. Das entschärft beide Probleme auf einmal.

7.4 Protokolle und Datenfluss

Die Vorbereitung der Research-Protokolle hat viel Zeit gekostet — und sie wurden **nicht von allen Untersuchungsleitenden sorgfältig ausgefüllt**. Das ist rückblickend der teuerste organisatorische Fehler des Projekts, denn die Protokolle sind die einzige Quelle für die Klassenzuordnung und für den Umsetzungsgrad. Die lückenhafte Dokumentation ist der Grund, warum 58 Schüler:innen keinen belegten Umsetzungsgrad haben.

Die Ursache liegt nicht allein bei den Untersuchungsleitenden: Es hätte **von Anfang an klarer kommuniziert werden müssen, dass die Protokolle Teil der Daten sind** und nicht Begleitpapier.

Ein Laptop mit Protokollen ist ausgefallen, die darauf gespeicherten Protokolle waren verloren. Zu diesem Zeitpunkt gab es noch keinen geregelten Datentransfer. Der spätere Wechsel zu **uniSHARE** hat sich **sehr bewährt** und sollte von Beginn an eingesetzt werden — er war nur deshalb spät, weil das System noch nicht bekannt war.

Die Tokenausgabe selbst funktionierte gut. Problematisch war das **Zurückführen** der Protokolle in die zentrale Ablage.

7.5 Team und Rollen

Rolle	Person
Gesamtorganisation, Studienleitung	David Matischek
Wissenschaftliche Leitung	Emilia Iannilli
Projektmitarbeit	Sophie Hainig
Masterarbeit Lehrkräfteerhebung	Jennifer Platt
Erhebungsunterstützung	je zwei Praktikantinnen für Pre und Post
Unterstützung vor Ort	GET

Was gut funktioniert hat: Die Unterstützung durch GET bei den Erhebungen war notwendig und hat getragen. Die Einarbeitung neuer Erhebungskräfte im Tandem hat funktioniert.

Was sich geändert hat: In der Pre-Erhebung lag praktisch die gesamte Organisation bei einer Person. Das war zu viel. In der Post-Erhebung wurde stärker delegiert, und **das hat gut funktioniert**. Für künftige Teamleads ist das die wichtigste Lehre aus dem Projektverlauf: Delegation ist nicht das Risiko, sie ist die Voraussetzung dafür, dass die zweite Erhebungswelle nicht an einer Person hängt.

Der formale Weg an die Universität — Projekt offiziell anbinden, Anträge einreichen — hat ebenfalls lange gedauert und gehört in die Zeitplanung.

7.6 Zeitplanung

Der Zeitplan war eng, aber tragfähig. Der Engpass liegt eindeutig **vor** der Pre-Erhebung: Genehmigungen, Schulkontakte, Terminfindung und Materialvorbereitung brauchen mehr Vorlauf, als intuitiv angenommen. Mehr Zeit vor T1 hätte an mehreren Stellen gleichzeitig geholfen.

8 Empfehlungen für die nächste Erhebung

Über die inhaltlichen Empfehlungen hinaus, aus der Durchführung:

1. Rückwärts planen, ab T1. Bildungsdirektionen drei bis vier Monate, Universität und Anträge parallel, Schulkontakte und Terminfindung danach. Wer im September erheben will, beginnt im Frühjahr.

2. Direktion als zentrale Ansprechperson etablieren, auch wenn der Erstkontakt über eine Lehrkraft läuft. Telefonisch, nicht per E-Mail.

3. Kontrollklassen aus derselben Schule verlangen und den Grund erklären. Das ist Verhandlungssache und muss früh geführt werden.

4. Protokolle als Datenerhebung behandeln. Verbindliche Vorlage, feste Abgabefrist, zentrale Ablage in uniSHARE ab dem ersten Tag, und eine Person, die die Vollständigkeit prüft, solange sich Lücken noch schließen lassen.

5. Eigene LimeSurvey-Instanz mit UNIT, eine Umfrage pro Klasse. Plus ein vorbereiteter Plan B auf Papier, damit ein technischer Ausfall keine Erhebung sprengt.

6. Umsetzung laufend dokumentieren lassen. Die Lehrkräfte sollten mitschreiben, was tatsächlich unterrichtet wurde — kurz, aber regelmäßig. Rückblickende Rekonstruktion aus Protokollen liefert genau die Unschärfe, die diese Studie an mehreren Stellen begrenzt.

7. Qualitative Interviews von Beginn an einplanen. Diese Empfehlung kam aus der Soziologie und deckt sich mit dem, was die quantitativen Daten nicht leisten können. Sie beantworten, was ein Fragebogen nicht erfasst: wie das Fach erlebt wurde und was im Unterricht tatsächlich passiert ist.

8. Erhebung durch die Lehrkräfte selbst ist der naheliegende Weg, Kosten und Terminaufwand zu senken. Er ist tragfähig, wenn zwei Bedingungen erfüllt sind: eine **gute Einschulung** — Online-Meetings, Schulungsvideos, bei Bedarf eine Schulung vor Ort — und eine **standardisierte Durchführungsanleitung**, weil sonst die Vergleichbarkeit zwischen Klassen verloren geht, die das Team durch die eigene Anwesenheit bisher sichergestellt hat. Die Aufmerksamkeitsprüfungen im Fragebogen bleiben dabei das Sicherheitsnetz; in dieser Studie haben sie 67 Fälle ausgeschlossen.

9 Was diese Studie nicht beantworten kann

Sie misst Selbsteinschätzungen zu zwei Zeitpunkten. Sie erfasst nicht, wie das Fach erlebt wurde, was im Unterricht tatsächlich passierte, was Lehrkräfte beobachtet haben und ob einzelne Schüler:innen davon profitiert haben, ohne dass es sich im Klassenmittelwert zeigt. Die qualitativen Interviews, die parallel geführt wurden, adressieren einen Teil davon und sollten in die Gesamtbewertung einfließen.

Ein rechnerischer Nullbefund auf Gruppenebene und eine wertvolle Erfahrung für einzelne Jugendliche schließen einander nicht aus.



**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.“