Much promise, little practice: counterfactual impact analysis in Austrian RTI evaluation

Jakob Kofler, Jürgen Streicher, Verena Régent, Isabella Wagner

Submitted version (preprint) – not yet peer reviewed

This is a preprint published in fteval Preprints, the open repository of the Austrian Platform for Research and Technology Policy Evaluation (fteval). It has been screened for formal quality and thematic fit, but has not been peer reviewed. The content is the sole responsibility of the author(s).

Open review: This preprint has been submitted to the fteval Journal for Research and Technology Policy Evaluation and is undergoing open peer review. Review reports are published with their own DOIs and linked to this preprint. If the article is accepted, the final version will appear in the journal under a separate DOI and will be linked here. This preprint remains available in any case, together with its review reports.

How to cite this version:

Kofler, Jakob und Streicher, Jürgen und Régent, Verena und Wagner, Isabella (2026) *Much promise, little practice: counterfactual impact analysis in Austrian RTI evaluation*. Working Paper. fteval Preprints. (Submitted)

Please check the DOI for the most recent version before citing: 10.22163/fteval.2026.769

License: CC BY 4.0

ABSTRACT

Counterfactual impact analyses are widely regarded as the reference standard for demonstrating whether policy measures actually work. How widespread they are in the evaluation of Austrian research, technology and innovation (RTI) policy, which conditions facilitate or impede their use, and what follows for practice is the subject of a cooperative study conducted by the Austrian Platform for Research and Technology Policy Evaluation (fteval) with member institutions between July 2025 and July 2026. This article presents its findings. A systematic review of Austrian evaluation reports identified seventeen counterfactual impact analyses in fifteen publications; six of these meet the methodological minimum standard for robust causal evidence. By international comparison the absolute count is high, though in domestic practice counterfactual designs remain the exception rather than a standard instrument. Three structural bottlenecks explain the limited uptake: small sample sizes, funding accumulation, and temporal lags in secondary data. Three developments are loosening them: regulatory and disciplinary pressure, an expanded methodological toolkit, and improved data access through the Austrian Micro Data Center (AMDC). Fifteen stakeholder interviews show that limited uptake reflects institutional routines and procurement conditions rather than methodological rejection. We conclude that method choice should remain a deliberate, context-specific decision, and that triangulation with qualitative and theory-based approaches is necessary to obtain a holistic view of the impacts of RTI programmes, to understand how these impacts come about, and to identify the levers for increasing a programme's effectiveness and efficiency.

Keywords: counterfactual impact analysis, evaluation practice, research and innovation policy, Austria, method choice, data infrastructure

1. Introduction

Evaluations are a central instrument for reviewing and improving research, technology and innovation (RTI) policy. A recurring challenge is attribution: can observed changes be traced to the measure under examination, or would they have occurred anyway? Counterfactual impact analyses offer one methodological route to that question by comparing supported units with a constructed comparison group. In other policy fields, notably labour market and education policy, such designs are long established. In RTI policy they remain comparatively rare.

In 2024 the German Commission of Experts for Research and Innovation (EFI) published a review documenting the extent to which counterfactual methods are used in German RTI evaluation practice (Büchele et al., 2024). The Austrian Federal Ministry of Women, Science and Research (BMFWF) saw the value of a comparable exercise for Austria and supported the idea of an independent situational analysis. fteval subsequently issued a call for a cooperative study, carried out by a consortium of KMU Forschung Austria, WPZ Research, Joanneum Research and fteval.

This article summarises the study, which is published in full in German (Kofler et al., 2026). The study replicates the EFI approach for the Austrian context and extends it in four directions: a more differentiated assessment of methodological quality, which complements the three-level validity rating of the EFI coding scheme with the standardised Maryland Scientific Methods Scale; a review of the academic literature on causal methods in RTI evaluation; three international reference cases (SkatteFUNN in Norway, ZIM in Germany, SBIR in the USA)[1]; and fifteen qualitative interviews with Austrian commissioners, evaluators and ecosystem actors. Our purpose is descriptive and diagnostic rather than advocatory: we ask where Austrian practice currently stands, what shapes method choice, and what would need to change for counterfactual designs to be a realistic option where they genuinely add value.

2. Where Austria stands

The review drew on the fteval repository, the central archive for Austrian RTI evaluation reports, which covers three decades and held 796 entries at the time of the search. The repository can only include evaluations that have been published, and in Austria evaluation reports are not yet published as a matter of course. To identify as many evaluations with counterfactual analyses as possible, we therefore supplemented the search with a community survey among fteval members, both commissioners and evaluators, which yielded five further evaluations not available in the repository.

The selection was guided by the PRISMA 2020 statement (Page et al., 2021). A search string comprising German and English method-specific terms (such as difference-in-differences, matching, regression discontinuity or instrumental variable) and concepts of causal inference (such as counterfactual, treatment effect or selection bias), linked by OR, returned 264 repository entries; together with the five survey entries, 269 records entered screening. In a first step, a tool based on a large language model (Google NotebookLM) pre-classified the full texts against predefined relevance criteria, and this pre-selection was then validated manually, which left 45 records. In a second step, we screened these full texts manually and excluded studies

[1] The literature review and the reference cases are not discussed further in this article; they are documented in the full report (Kofler et al., 2026), available at https://repository.fteval.at/id/eprint/841/1/KWA-Studie_Report_fin.pdf.

that did not actually apply a causal method, for instance because they only discussed such methods or compared supported units with a control group without statistical inference. The final corpus, with a cut-off date of 15 October 2025, comprises fifteen publications that apply a counterfactual impact analysis. Because one fteval journal article reports three separate randomised trials of the Austrian Research Promotion Agency (FFG), these fifteen publications contain seventeen counterfactual analyses.

The temporal distribution is uneven. Three pioneering studies appeared between 2000 and 2004. Between 2005 and 2015, only one further study was identified. Thirteen of the seventeen date from the past decade.

For a research system of Austria's size, seventeen is a considerable figure. Büchele et al. (2024) identify fourteen evaluations with at least partly causal designs for Germany over 2009 to 2023, or around 17 per cent of the studies they examined. Collin et al. (2022) find eleven among 110 Swedish innovation policy evaluations, around 10 per cent. In evaluations of EU cohesion policy in the research and innovation field, the share is roughly 3 per cent (Santos & Coad, 2023). Measured against the size of the Austrian RTI system, uptake is therefore comparatively high in absolute terms. Relative to the 796 repository entries, the seventeen analyses correspond to around 2 per cent; because the repository also contains reports other than evaluations, and survey and selection procedures differ, this share is not directly comparable with the figures above.

The number of studies alone says little about the robustness of the causal evidence they provide. We assessed each study against the Maryland Scientific Methods Scale (SMS), which classifies designs in five levels and treats Level 3 as the minimum threshold for speaking of causal evidence (What Works Centre for Local Economic Growth, 2016). Eleven of the seventeen analyses sit at Level 2: seven use matching approaches, chiefly propensity score matching (PSM), and four use panel or regression models such as fixed-effects estimation or the generalised method of moments. These designs can adjust only for those observable characteristics that are captured in the available data and included in the model, so even observable differences are not necessarily fully accounted for; panel models additionally absorb time-constant differences between firms. None of them, however, can address unobserved, time-varying factors such as intrinsic innovation orientation or management quality, which typically correlate with the probability of applying for and receiving support.

Six analyses reach Level 3 or above. One regional policy evaluation applies conditional difference-in-differences (DiD) with an upstream matching step. One study exploits the internal scoring of applications to the FFG Basisprogramme in an instrumental-variable estimation that follows fuzzy regression-discontinuity logic. Four use randomised designs: three are purpose-built policy experiments and one is an ex-post evaluation of the Öko-Scheck, a small grant for environmentally friendly innovation projects, which makes analytical use of a lottery for selecting applications. According to the FFG, the lottery was introduced both because it was perceived as a fairer system and because of its administrative simplicity (Firpo, 2024). All four were published from 2022 onwards. The three policy experiments concern low-threshold administrative micro-interventions, such as feedback to applicants or crowdfunding incentives, rather than research funding as such; in the Öko-Scheck, by contrast, access to the funding itself was allocated by lottery.

3. What constrains uptake

Three interlocking factors explain the limited number of counterfactual evaluations.

The first factor is small sample sizes at programme level. Austrian evaluations are, with few exceptions, conducted at programme level. Because of limited budgets or thematic focus, many programmes support a statistically small number of organisations, with correspondingly few comparable non-supported firms. Programmes are also adjusted regularly, not least in response to earlier evaluations, which further reduces the pool of comparable cases and makes the effect of a specific programme design hard to isolate.

The second factor is funding accumulation in a dense support system. The Austrian funding landscape is well developed and organisations frequently draw on several programmes in the same period. Isolating the effect of any one of them is correspondingly difficult, particularly where case-level data are not accessible across instruments. The clearest example is the *Forschungsprämie*, Austria's volume-based tax credit for company R&D, for which case-level data have not so far been available to evaluators, leaving it unclear where and to what extent an organisation has already received support.

The third factor is the long time lag before effects become visible. Counterfactual designs depend on secondary data, since funding programmes collect no information on non-supported firms. Those data appear with a delay of up to two years: the Community Innovation Survey (CIS) for 2024 was published only in July 2026. RTI policy typically targets long-term effects such as the market introduction of innovations, expected at the earliest one to two years after project completion. Evaluations rarely run longer than a year and ex-post evaluations are usually commissioned during or immediately after a programme's final call. A simple calculation illustrates the consequences: if a programme approves its first cases in 2020, projects run two years, and market introduction follows two years later, the first effects should appear in the CIS 2024, published in mid-2026. Around 6.5 years elapse between the start of funding and the visibility of effects in secondary statistics.

Closely linked to this is restricted access to microdata. The two surveys most relevant for RTI, the R&D survey and the CIS, are conducted by Statistics Austria, but under the Federal Statistics Act firm-level data were until recently very difficult to obtain and barely usable for evaluation. This partly explains the limited causal validity of many earlier counterfactual analyses.

4. What is changing

Against these constraints, three developments are reshaping the picture.

The first development is growing regulatory and disciplinary pressure. Under State Aid Modernisation and the revised General Block Exemption Regulation, the European Commission requires an impact analysis using causal methods once a support measure exceeds an average volume of EUR 150 million per year. Because few Austrian programmes reach the EU threshold, the direct effect on the number of counterfactual evaluations has so far been marginal. The requirement nonetheless has a signalling effect beyond its formal scope: courts of audit, for instance, are increasingly calling for counterfactual impact assessments. In parallel, the applied econometrics literature has moved steadily towards counterfactual designs in the wake of the credibility revolution (Angrist & Pischke, 2010).

The second development is an expanded methodological toolkit. Synthetic control approaches and dynamic matching address precisely the Austrian bottlenecks of small case numbers and continuously running measures, and machine-learning-supported procedures such as double machine learning improve the handling of high-dimensional covariates. The three randomised policy experiments since 2022 represent a further, and in the Austrian context

new, route: developing policy design experimentally, so that evidence on what works feeds into the design of a measure while it is being shaped, rather than being reconstructed after the fact.

The third development is improved access to data. The Austrian Micro Data Center (AMDC) provides regulated online access to microdata for scientific purposes and makes the linkage of statistical, register and administrative data technically and legally possible. Some interviewees described it as a genuine game changer. It does not, however, turn counterfactual designs into standard practice: access is still perceived as administratively demanding, costly, and legally and institutionally complex, and registers currently have to be linked afresh for each project, which further increases cost and effort.

Two evaluations published after our cut-off suggest the trend is continuing and that the AMDC is beginning to matter in practice. The evaluation of the FFG Basisprogramme for 2008 to 2021 links FFG funding data at firm level with AMDC data and estimates effects using PSM combined with a staggered-adoption DiD estimator (Dinges et al., 2026). The evaluation of the AplusB Scale-up programme of Austria Wirtschaftsservice (aws) combines multi-stage matching with a doubly robust panel DiD and documents pre-trend and robustness checks (Frantz et al., 2026). On our reading of their method chapters, both would code at Level 3. Notably, the FFG study also estimates interaction effects between the Basisprogramme, the *Forschungsprämie* and other instruments, addressing the funding-accumulation problem head-on. For the Forschungsprämie, however, the study relies on firms' reported use of the premium in the R&D survey rather than on the premiums actually granted, for which case-level data remain unavailable to evaluators. The FFG study is, to our knowledge, the first large Austrian programme evaluation to link administrative funding data with official microdata at this scale.

5. What stakeholders say

The fifteen interviews covered three groups: commissioners and intermediaries, such as ministries and funding agencies (six); evaluators (four); and advisory and ecosystem actors (five), such as data providers and interest groups, which neither commission nor conduct evaluations but are active in the wider RTI system.

Relevance is acknowledged across all three. Commissioners and intermediaries expect counterfactual analyses above all to provide robust and communicable statements on the effects of RTI measures. They particularly value the possibility of expressing effects in figures, which helps to justify measures to policymakers, the wider public and other stakeholders, and they expect a better assessment of efficiency and, as a key benefit, estimates of additionality and deadweight. Some in this group stress, however, that the quality of the analysis is decisive and that results in the RTI field need to be interpreted with caution. Evaluators broadly share the expectation of robust quantitative evidence but attach a stronger methodological caveat: the full impact spectrum of a measure can rarely be captured quantitatively, so qualitative and theory-based methods remain necessary for dimensions such as processes, networking, institutional change or unintended effects. They also warn that legitimation pressure can lead to causal results being overstated in political debate or communication beyond what the data and context support. Advisory and ecosystem actors emphasise the objectifying function of robust evidence, and are the only group to name learning and programme development explicitly as expected benefits.

Calls for tender in Austria generally do not prescribe methods, so method choice effectively rests with evaluators. Evaluators report that proposals containing counterfactual elements are often received well when plausibly justified, even where not required, and describe the

necessary competence as available in their organisations. Commissioners, for their part, sometimes describe their own capacity to assess causal designs in detail as limited.

The binding constraints lie elsewhere. Counterfactual analyses require substantial preparatory work, and the appropriate estimator can often only be chosen after a closer look at the data, which the timelines and budgets of Austrian evaluation contracts do not always accommodate. The Austrian evaluation market is small, specialised and competitive, and procurement follows public tendering logic with limited budgets and short durations. The result is not resistance to the method but a set of incentives favouring designs that are calculable and deliverable within short periods.

One further barrier concerns the use of results. Condensing complex effects into a few figures makes them susceptible to undifferentiated use in political debate, and unfavourable or lower-than-expected values are attributed directly to responsible actors. That prospect can reduce the willingness to plan such designs early or to communicate their results openly.

6. What follows from this

The picture that emerges is neither one of neglect nor one of a system that simply needs to catch up. Methodological competence and interest are both growing, and the absolute number of counterfactual evaluations is high by international comparison. What limits uptake is a combination of structural features of the Austrian RTI system, data conditions, and procurement routines.

Easing these constraints would widen the scope for counterfactual designs, but it would not settle where they should be used. The more useful question is therefore not how to maximise the number of counterfactual analyses, but where they genuinely add explanatory value. On the evidence of our interviews and the international literature, they are best suited to clearly delimited measures with a sufficient number of supported and comparable units, traceable intervention logics and reliable data access. Where measures are intended to produce complex, long-term or systemic effects, where programmes involve few individual projects, or where the object is coordination or networking, counterfactual designs are not the appropriate method: complex and long-term effects arise from many factors acting together, small programmes rarely offer enough cases for robust estimates, and coordination or networking outcomes can hardly be isolated as the effect of a single measure, let alone fully quantified.

Counterfactual analyses isolate the existence and magnitude of an effect. They can rarely reconstruct the mechanism behind it empirically. They are therefore always one component of an evaluation, and one that benefits directly from the other methods applied alongside it, since effects become easier to interpret and to validate. Triangulating them with qualitative, descriptive or theory-based approaches is necessary to obtain a holistic view of a programme's impacts, to understand how these impacts come about, and to identify the levers for increasing its effectiveness and efficiency. Impacts that are difficult to quantify are not, for that reason, any less important.

The practical implication for the Austrian system is therefore twofold. Conditions should be improved so that counterfactual designs become feasible where they promise added value, because at present they are frequently ruled out by circumstance rather than by judgement. And the decision to use them, or not, should remain an explicit and justified one, taken case by case against the evaluation question, the intervention logic, the available data and the resources at hand.

7. What each actor can do

For evaluators. We recommend that evaluators broaden their methodological competence through training while staying connected to methodological developments. Recent advances should be applied specifically to Austrian conditions: synthetic control and dynamic matching address small case numbers and continuously running measures. Maryland SMS Level 3 should serve as a reference minimum standard where the context permits, with identifying assumptions such as parallel trends documented transparently and supported by robustness checks. Estimates should be reported together with their uncertainty and interpreted alongside the other evaluation findings, not as conclusive proof of a programme effect.

For commissioners. Internal competence is needed to judge where counterfactual approaches are applicable and what their results warrant. Evaluation should be planned from the programme design stage onwards, with relevant impact indicators identified early and data collected from the start. Budgets should be realistic: counterfactual analyses are methodologically demanding, data-intensive and often more costly than alternatives, so where a specific programme promises analytical added value, that cost belongs in the evaluation budget. Complete access to all funding data should be guaranteed so that the effects of individual programmes can be isolated.

For ecosystem actors and data providers. Access to the AMDC should become easier and less costly, with streamlined approval processes and shorter processing times. RTI-relevant datasets should be linked on a continuously updated basis rather than afresh for every project; Germany's EVALDAT project is testing exactly this, and Austria's microdata access to public registers gives such an approach even greater potential here. Awareness-raising is needed for a differentiated use of results, so that qualitative findings are not displaced and causal estimates are located within a broader, multi-method picture of impacts. This applies with particular force to interest groups and advisory bodies, which translate evaluation findings into economic policy debate, and to policymakers who use them as a basis for decisions.

REFERENCES

Angrist, J. D., & Pischke, J.-S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives, 24*(2), 3–30. https://doi.org/10.1257/jep.24.2.3

Büchele, S., Buenstorf, G., Cantner, U., Dreier, L., Meurer, P., & Neumann, L. P. (2024). *Commissioned project evaluations of research and innovation policy in Germany: A review* (Studien zum deutschen Innovationssystem, No. 11-2024). Expertenkommission Forschung und Innovation (EFI). https://hdl.handle.net/10419/285357

Collin, E., Sandström, C., & Wennberg, K. (2022). Evaluating evaluations of innovation policy: Exploring reliability, methods, and conflicts of interest. In K. Wennberg & C. Sandström (Eds.), *Questioning the entrepreneurial state: Status-quo, pitfalls, and the need for credible innovation policy* (Vol. 53, pp. 157–173). Springer. https://doi.org/10.1007/978-3-030-94273-1_9

Dinges, M., Dachs, B., Kerlen, C., Pintar, N., Reiter, C., Toepel, K., & Weinlinger, P. (2026). *Evaluierung der FFG-Basisprogramme für den Zeitraum 2008–2021*. AIT, Kerlen Evaluation. https://doi.org/10.22163/fteval.2026.758

Firpo, T. (2024). *Öko-Scheck evaluation report: Effects of the introduction of randomization to select projects*. Humboldt Innovation.

Frantz, F., Dudenbostel, T., Wiscovitch, M., & Alt, R. (2026). *Evaluierung des AWS AplusB Scale-up Programms*. Technopolis. https://doi.org/10.22163/fteval.2026.760

Kofler, J., Streicher, J., Régent, V., & Wagner, I. (2026). *Chancen & Grenzen kontrafaktischer Wirkungsanalysen für die österreichische FTI-Politik: Eine Situationsanalyse*. fteval Kooperationsstudie 2025–2026. https://repository.fteval.at/id/eprint/841/

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article n71. https://doi.org/10.1136/bmj.n71

Santos, A. M., & Coad, A. (2023). Monitoring and evaluation of transformative innovation policy: Suggestions for improvement. *Socio-Economic Planning Sciences*, *90*, Article 101714. https://doi.org/10.1016/j.seps.2023.101714

What Works Centre for Local Economic Growth. (2016). *Guide to scoring evidence: Using the Maryland Scientific Methods Scale*. https://whatworksgrowth.org/wp-content/uploads/16-06-28_Scoring_Guide.pdf

AUTHOR INFORMATION

Jakob Kofler
KMU Forschung Austria, Gusshausstrasse 8, 1040 Vienna, Austria
Email: j.kofler@kmuforschung.ac.at
ORCID: https://orcid.org/0000-0002-6059-9989

Jürgen Streicher
Joanneum Research, Vienna, Austria
Email: Juergen.Streicher@joanneum.at
ORCID: https://orcid.org/0000-0002-1836-0149

Verena Régent
WPZ Research GmbH, Vienna, Austria
Email: verena.regent@wpz-research.com
ORCID: https://orcid.org/0000-0002-3870-0112

Isabella Wagner
Österreichische Plattform für Forschungs- und Technologiepolitikevaluierung (fteval), Vienna, Austria
Email: wagner@fteval.at
ORCID: https://orcid.org/0000-0001-5009-2407

ACKNOWLEDGEMENTS

The underlying research was supported by the Austrian Federal Ministry of Women, Science and Research (BMFWF) and the Austrian Platform for Research and Technology Policy Evaluation (fteval). We thank all contributors to this study: Sascha Sardadvar (WPZ Research GmbH) and Peter Kaufmann (KMU Forschung Austria) for their contributions and quality assurance, Egor Burda for his preparatory searches, and all interview partners for the valuable information.