

# **Comparing LLM and human reviews of social science research using data from Unjournal.org**

Valentin Klotzbücher      David Reinstein      Tianmai Michael Zhang

2025-12-13

We study how a frontier language model evaluates social-science research compared to expert human reviewers in The Unjournal. Using the same structured rubric as human evaluators, we ask GPT-5 Pro to rate papers on overall quality, methods, evidence, communication, openness, and global relevance, and to produce a narrative assessment anchored in the PDF of each paper. We first compare its quantitative ratings to aggregated human scores across The Unjournal’s existing evaluations, then take a closer qualitative look at a small set of focal papers, including a high-profile mapping study of natural regeneration. For these focal cases, we examine where the model’s written review overlaps with and diverges from the human reports, and how both sides describe the main strengths, weaknesses, and policy relevance of the work. So far, the model reliably identifies many of the same methodological and interpretive issues that human experts emphasize, but it tends to translate these into more generous numerical ratings and narrower uncertainty intervals. We view this as an initial, work-in-progress probe of LLM-based peer review in a high-stakes, policy-relevant domain, and as the first step toward a broader benchmark and set of tools for comparing and combining human and AI research evaluations.

# Table of contents

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Data and methods</b>	<b>6</b>
2.1	Unjournal.org evaluations . . . . .	6
2.2	LLM-based evaluation . . . . .	6
	<b>References</b>	<b>8</b>
	<b>Appendices</b>	<b>9</b>
<b>A</b>	<b>LLM evaluation summaries and reasoning traces</b>	<b>9</b>
A.1	1_PDFsam_Nature published version 17 Nov 2021 . . . . .	9
A.2	Class Warfare Political Exclusion of the Poor and the Roots of Social Revolutionary Terrorism 1860 1950 . . . . .	14
A.3	sciadv.adh4030 . . . . .	20

# 1 Introduction

Is AI good at peer-reviewing? Does it offer useful and valid feedback? Can it predict how human experts will rate research across a range of categories? How can it help academics do this “thankless” task better? Is it particularly good at spotting errors? Are there specific categories, e.g. spotting math errors or judging real-world relevance, where it does surprisingly well or poorly? How does its “research taste” compare to humans?

If AI research-evaluation works it could free up a lot of scientific resources – perhaps \$1.5 billion/year in the US alone ([Aczel, Szaszi, and Holcombe 2021](#)) – and offer more continual and detailed review, helping improve research. It could also help characterize methodological strengths/weaknesses across papers, aiding training and research direction-setting. Furthermore, a key promise of AI is to directly improve science and research. Understanding how AI engages with research evaluations may provide a window into its values, abilities, and limitations.

In this project, we test whether current large language models (LLMs) can generate research evaluations that are comparable, in structure and content, to expert human reviews. The Unjournal systematically prioritizes “impactful” research and pays for high-quality human evaluations, including structured numeric ratings with credible intervals, claim identification and assessment, predictions, and detailed narrative reports. We use a frontier LLM (OpenAI’s GPT-5 Pro) to review the same social-science and policy-relevant working papers under essentially the same rubric.

For a first pass we focus on papers that already have completed Unjournal evaluation packages. For each of 47 such papers, the model reads the PDF that human evaluators saw and returns: (i) percentile ratings and 90% credible intervals on The Unjournal’s seven criteria, and (ii) two 0–5 journal-tier scores (“should” and “will” be published). In an additional, richer run on a small set of focal papers, we keep these quantitative outputs but also require a long diagnostic summary and high-effort reasoning trace. We then compare the model’s ratings, journal-tier predictions, and qualitative assessments to the existing human evaluations.

Future iterations will extend this design to papers that are still in The Unjournal’s pipeline, where no human evaluations are yet public. This will let us study out-of-sample prediction, reduce the risk of model contamination from published evaluations, and test LLMs as tools for triaging and prioritising new work.

## Our work in context

Luo et al. ([2025](#)) survey LLM roles from idea generation to peer review, including experiment planning and automated scientific writing. They highlight opportunities (productivity, coverage of long documents) alongside governance needs (provenance, detection of LLM-generated content, standardizing tooling) and call for reliable evaluation frameworks.

Eger et al. ([2025](#)) provide a broad review of LLMs in science and a focused discussion of AI-assisted peer review. They argue: (i) peer-review data is scarce and concentrated in CS/OpenReview venues; (ii) targeted assistance that preserves human autonomy is preferable to end-to-end reviewing; and

(iii) ethics and governance (bias, provenance, detection of AI-generated text) are first-class constraints.

Zhang and Abernethy (2025) propose deploying LLMs as quality checkers to surface critical problems instead of generating full narrative reviews. Using papers from WITHDRARXIV and an automatic evaluation framework that leverages “LLM-as-judge,” they find the best performance from top reasoning models but still recommend human oversight.

Pataramutaporn et al. (2025) asked four nearly state-of-the-art LLM models (GPT-4o mini, Claude 3.5 Haiku, Gemma 3 27B, and LLaMA 3.3 70B) to consider 1220 unique papers “drawn from 110 economics journals excluded from the training data of current LLMs”. They prompted the models to act “in your capacity as a reviewer for [a top-5 economics journal]” and make a publication recommendation using a 6-point scale ranging from “1 = Definite Reject...” to “6. Accept As Is...”. They asked it to evaluate each paper on a 10-point scale for originality, rigor, scope, impact, and whether it was ‘written by AI’. They also (separately) had LLMs rate 330 papers with the authors’ identities removed, or replacing the names with fake male/female names and real elite or non-elite institutions (check this) or with prominent male or female economists attached.

They compare the LLMs’ ratings with the RePEC rankings for the journals the papers were published in, finding general alignment. They find mixed results on detecting AI-generated papers. In the names/institutions comparisons, they also find the LLMs show biases towards named high-prestige male authors relative to high-prestige female authors, as well as biases towards elite institutions and US/UK universities.

There have been several other empirical benchmarking projects, including work covered in LLM4SR: A Survey on Large Language Models for Scientific Research and [Transforming Science with Large Language Models: A Survey on AI-assisted Scientific Discovery, Experimentation, Content Generation, and Evaluation](#).

Our project distinguishes itself in its use of *actual* human evaluations of research in economics and adjacent fields, past and *prospective*, including both reports, ratings, and predictions.<sup>1</sup> The Unjournal’s 50+ evaluation packages enable us to train and benchmark the models. Their pipeline of future evaluations allow for clean out-of-training-data predictions and evaluation. Their detailed written reports and multi-dimensional ratings also allows us to compare the ‘taste’, priorities, and comparative ratings of humans relative to AI models across the different criteria and domains. The ‘journal tier prediction’ outcomes also provides an external ground-truth<sup>2</sup> enabling a human-vs-LLM horse race. We are also planning multi-armed trials on these human evaluations (Brodeur et al. 2025) to understand the potential for *hybrid* human-AI evaluation in this context.

---

<sup>1</sup>Other work has relied on collections of research and grant reviews, including NLPEER, SubstanReview, and the Swiss National Science Foundation. That data has a heavy focus on computer-science adjacent fields, and is less representative of mainstream research peer review practices in older, established academic fields. Note that The Unjournal commissions the evaluation of impactful research, often from high-prestige working paper archives like NBER, and makes all evaluations public, even if they are highly critical of the paper.

<sup>2</sup>About verifiable publication outcomes, not about the ‘true quality’ of the paper of course.

## 2 Data and methods

We draw on two main sources:

- 1) Human evaluations from [The Unjournal's public evaluation data](#) (PubPub reports and the Coda evaluation form export).
- 2) LLM-generated evaluations using a structured JSON-schema prompt with `gpt-5-pro-2025-10-06` (knowledge cut-off: 30 September 2024).

### 2.1 Unjournal.org evaluations

We use The Unjournal's public data for a baseline comparison. At The Unjournal each paper is typically evaluated (aka 'reviewed') by two expert evaluators<sup>1</sup> who provide quantitative ratings on a 0–100 percentile scale for each of seven criteria (with 90% credible intervals),<sup>2</sup> two "journal tier" ratings on a 0.0 - 5.0 scale,<sup>3</sup> a written evaluation (resembling a referee report for a journal), and identification and assessment of the paper's "main claim". For our initial analysis, we extracted these human ratings and aggregated them, taking the average score per criterion across evaluators (and noting the range of individual scores).

All papers have completed The Unjournal's evaluation process (meaning the authors received a full evaluation on the Unjournal platform, which has been publicly posted at [unjournal.pubpub.org](#)). The sample includes papers spanning 2017–2025 working papers in development economics, growth, health policy, environmental economics, and related fields that The Unjournal identified as high-impact. Each of these papers has quantitative scores from at least one human evaluator, and many have multiple (2-3) human ratings.

### 2.2 LLM-based evaluation

Following The Unjournal's [standard guidelines for evaluators](#) and their [academic evaluation form](#), evaluators are asked to consider each paper along the following dimensions: **claims & evidence**, **methods, logic & communication**, **open science**, **global relevance**, and an **overall assessment**. Ratings are interpreted as percentiles relative to serious recent work in the same area. For

---

<sup>1</sup>Occasionally they use 1 or 3 evaluators.

<sup>2</sup>See their guidelines [here](#); these criteria include "Overall assessment", "Claims, strength and characterization of evidence", "Methods: Justification, reasonableness, validity, robustness", "Advancing knowledge and practice", "Logic and communication", "Open, collaborative, replicable science", and "Relevance to global priorities, usefulness for practitioners"

<sup>3</sup>a normative judgment about 'how well the research should publish' and "a prediction about where the research will be published"

each metric, evaluators are asked for the midpoint of their beliefs and their 90% credible interval, to communicate their uncertainty. For the journal rankings measure, we ask both “what journal ranking tier should this work be published in? (0.0-5.0)” and “what journal ranking tier will this work be published in? (0.0-5.0)”, with some further explanation. The full prompt can be seen in the code below – essentially copied from the Unjournal’s guidelines page.

We captured the versions of each paper that was evaluated by The Unjournal’s human evaluators, downloading from the links provided in The Unjournal’s Coda database.

We evaluate each paper by passing the PDF directly to the model and requiring a strict, machine-readable JSON output. This keeps the assessment tied to the document the authors wrote. Direct ingestion preserves tables, figures, equations, and sectioning, which ad-hoc text scraping can mangle. It also avoids silent trimming or segmentation choices that would bias what the model sees.

We enforce a JSON Schema for the results. The model must return one object for each of the named criteria including a midpoint rating and a 90% interval for each rating. This guarantees that every paper is scored on the same fields with the same types and bounds. It makes the analysis reproducible and comparisons clean.

We request credible intervals (as we do for human evaluators) to allow the model to communicate its uncertainty rather than suggest false precision; these can also be incorporated into our metrics, penalizing a model’s inaccuracy more when it’s stated with high confidence.

Relying on GPT-5 Pro, we use a single-step call with a reasoning model that supports file input. One step avoids hand-offs and summary loss from a separate “ingestion” stage. The model reads the whole PDF and produces the JSON defined above. We do not retrieve external sources or cross-paper material for these scores; the evaluation is anchored in the manuscript itself.

The Python pipeline uploads each PDF once and caches the returned file id keyed by path, size, and modification time. We submit one background job per PDF to the OpenAI Responses API with “high” reasoning effort and server-side JSON-Schema enforcement. Submissions record the response id, model id, file id, status, and timestamps.

We then poll job status and, for each completed job, retrieve the raw JSON object, and write the responses to disk.

# References

- Aczel, Balazs, Barnabas Szaszi, and Alex O Holcombe, “A billion-dollar donation: Estimating the cost of researchers’ time spent on peer review,” *Research integrity and peer review*, 6 (2021), 1–8 (Springer).
- Brodeur, Abel, David Valenta, Alexandru Marcoci, Juan P Aparicio, Derek Mikola, Bruno Barbaroli, Rohan Alexander, Lachlan Deer, Tom Stafford, Lars Vilhuber, and others, “Comparing human-only, AI-assisted, and AI-led teams on assessing research reproducibility in quantitative social science,” *Institute for Replication (I4R) Discussion Paper*, 195 (2025).
- Eger, Steffen, Yong Cao, Jennifer D’Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, Chenghua Lin, Nafise Sadat Moosavi, Wei Zhao, and Tristan Miller, “Transforming science with large language models: A survey on AI-assisted scientific discovery, experimentation, content generation, and evaluation,” *arXiv preprint arXiv:2505.05151*, (2025).
- Luo, Ziming, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du, “LLM4SR: A survey on large language models for scientific research,” *arXiv preprint arXiv:2501.04306*, (2025).
- Pataranutaporn, Pat, Nattavudh Powdthavee, Chayapatr Achiwaranguprok, and Pattie Maes, “Can AI solve the peer review crisis? A large scale cross model experiment of LLMs’ performance and biases in evaluating over 1000 economics papers,” 2025.
- Zhang, Tianmai M, and Neil F Abernethy, “Reviewing scientific papers for critical problems with reasoning LLMs: Baseline approaches and automatic evaluation,” *arXiv preprint arXiv:2505.23824*, (2025).

# A LLM evaluation summaries and reasoning traces

## ! Important

Current testing with GPT-5 Pro to generate detailed reasoning traces and assessment summaries for various UNRELATED papers. The full outputs are included below for reference.

## A.1 1\_PDFsam\_Nature published version 17 Nov 2021

### i Model assessment summary

This paper exploits a novel, high-frequency data source—8 million calls to 23 helplines in 19 countries—to describe mental health concerns during COVID-19 and relate call volumes/topics to infection rates and policies. The design has clear strengths (large scale, daily observability, cross-context variation, topic tagging, pre-period coverage for many helplines, and multiple checks) but several identification and measurement limitations temper causal claims. Data are heterogeneous: some helplines provide call-level records with topics and demographics, others provide only daily aggregates, and the US analysis uses “routed” calls rather than answered calls. The harmonization of diverse topic taxonomies into eight broad categories is reasonable but introduces non-classical measurement error; category definitions differ by helpline and may have changed over time, yet no validation audit or coder consistency checks are shown. Caller representativeness is unknown and could shift during the pandemic; although the paper focuses on within-helpline changes (which removes time-invariant selection), the composition of callers (e.g., first-time versus repeat, gender/age mix, access to phones/privacy) plausibly changed with lockdowns and school/work closures. Repeated calls by the same individual are counted each time, and while the share of first-time/sporadic callers is reported, dynamics in this mix are not systematically modeled. A major concern is supply constraints: during surges, unanswered calls rose and capacity expanded gradually. The paper argues answered/unanswered status is random with respect to topics, so composition should be unaffected, but this assumption is untested and may be unrealistic if call-handling protocols triage acute/suicidal risk or if patience to wait differs by problem type. For volume analyses, capacity constraints create a lower bound on demand; for topic shares, selective answering (even unintentionally) could bias estimates. The main descriptive result—35% peak increase in calls about six weeks post-outbreak—is based on an event-study that includes helpline, year, week-of-year, and day-of-week fixed effects. However, 4 of 21 helplines lack 2019 baselines; Figure 1 shows only four weeks of pre-trend, limiting a visual test of parallel trends; and other contemporaneous shocks could confound the timing. Topic analyses use a binary

post-outbreak indicator (and later a two-period split), which compresses rich dynamics; modeling flexible event-time effects by topic would better show pre-trends and persistence. The U.S. panel leverages state-week variation in infections, the Oxford stringency index, and an income-support index with two-way fixed effects. This is informative but still observational; policy indices are potentially endogenous to unobserved distress, and infection rates proxy for multiple time-varying shocks (risk salience, media attention, hospital strain). Week fixed effects absorb nationwide shocks, but local confounders (unemployment, closures of mental-health services, vaccination rollouts, media campaigns, center-specific staffing or technical changes) remain unaccounted for. The dependent variable for the U.S. is “routed” calls, not answered calls; routing and overflow rules, center hours, wait-time thresholds, or technical outages can change over time and across states, risking non-stationary measurement error. State-of-origin based on phone numbers introduces misclassification via mobile numbers from other states; while likely mean-zero, it adds noise that could attenuate effects. The choice to cluster standard errors at the “state–month” level is unconventional; two-way clustering by state and week (or Driscoll-Kraay) would better address spatial and temporal dependence. In the France/Germany topic-specific regressions (daily, with helpline fixed effects and seasonal controls but not week fixed effects), identification relies on within-country temporal variation; with only two helplines, inference about policy elasticities is vulnerable to coincident trends or specification choices. Reported elasticities are small and significant (e.g., stricter NPIs associated with more suicide-related calls; income support with fewer), but causal interpretation remains tentative. The claim that pandemic-related concerns “replaced rather than exacerbated” underlying anxieties leans on shares: several non-COVID topics fall in relative frequency. This does not rule out increases in absolute counts for some categories, especially when total calls rise; indeed, early-wave suicidality shares do not significantly change overall (Fig. 2b), though later waves see decreases in the two-country sample. The paper notes possible underreporting of violence because callers may be less able to safely phone during stay-at-home orders, but this is not directly tested; this could bias the inference that violence became less prevalent. Robustness is partly addressed through weighted/unweighted models, alternative outbreak definitions (infection threshold vs. SIP), period splits, and supplementary specifications; nonetheless, more explicit falsification tests (e.g., placebo “pseudo-outbreak” dates in 2019; alternative clustering schemes; inclusion of helpline-specific linear trends; controlling for local unemployment or mobility; instrumenting policies with neighboring states’ policies or political variables) would strengthen credibility. Transparency is good for methods and code (Stata do-files shared), but replication is constrained: most data are proprietary, require separate agreements, and cross-helpline compatibility rests on non-public coding manuals. Figures are generally clear and consistent, and the narrative is careful to couch policy results as associations, though some language (e.g., “implies that financial relief can allay distress”) could be toned down given identification caveats. Overall, the work convincingly documents time patterns and topic shifts in helpline activity during COVID-19 and provides suggestive evidence that policy stringency is associated with increased distress and that income support is associated with reductions. The main limitations are potential selection/measurement changes on both the demand and supply sides of helplines, compressed treatment timing in topic models, partial baseline coverage, endogeneity of policies, and replication barriers due to data access. Despite these, the paper usefully advances the monitoring toolkit for population mental health and supplies actionable correlational evidence relevant to crisis policymaking.

## Model reasoning trace

### **Analyzing helpline data**

I'm looking at the data for 17 helplines from 2019, specifically focusing on the same days. However, it seems they don't show pre-trends before week 0, just the 4 weeks prior. They have various methods like weighted and unweighted data. I noticed they didn't utilize difference-in-differences with controls but instead used event time with helpline fixed effects while considering year, week of the year, and day of the week. There's a lot to unpack here!

### **Evaluating NPI effects and data availability**

They're discussing trends related to the treatment effects of non-pharmaceutical interventions (NPIs) and income support, but they rely on Oxford indices, which might have measurement limitations. They consider policy stringency across states, suggesting that national policies could skew the scores, reducing within-state variation. The income support index for the US heavily weighs national policies, potentially limiting variation. They log data weekly and find small coefficients, showing positive but not statistically significant effects in the US. Data is accessible upon permission, and the code is openly available, which is great!

### **Exploring replication barriers and limitations**

It seems there's a replication barrier due to data sharing restrictions. They show various specifications in supplementary materials but aren't robust to heterogeneity among helplines. They account for capacity constraints, but I wonder if they adjusted for multiple hypotheses—it's not mentioned. The data on mental health across 19 countries is useful, demonstrating the positive impact of income support. However, generalizing findings to lower-income countries may be limited. There are also concerns about the variability in call supply capacity and whether it affects the results—like, are some callers more patient based on the topic?

### **Questioning assumptions in call patterns**

It seems the data isn't entirely random; they didn't test for call reasons that remain unaffected by supply. Call topics are based on answered calls, meaning changes in capacity could alter topic composition, but they assume call answering is random—which is questionable. Additionally, the data aggregates calls rather than unique individuals, so repeated calls could skew mental health observations. They mention that a large percentage of calls were from first-time or sporadic callers, but the range is inconsistent across helplines without adjustments for repeat callers. There are variations in how suicidal calls are classified, and mapping categories may be inconsistent due to changed classification codes during the pandemic. This inconsistency makes me wonder about their conclusions regarding the relationship between issues and call prevalence—there seems to be a lack of clarity!

### **Examining method choices and results**

I noticed that the abstract claims calls related to suicide are less prevalent than before. In the first wave, the figures show decreased relationships regarding livelihood and violence, but suicidality remained unchanged. Overall, they state a decrease across both first and later waves. Regarding methods, they applied OLS with a transformation that may be odd for zero counts. It's interesting that they cluster standard errors at the helpline-week level, which seems unusual, potentially to account for within-week correlation. For the US weekly state-level panel, clustering at the state-month level raises questions about the number of clusters being adequate.

### **Evaluating clustering methods**

I'm considering the clustering methods used, which involve 51 states over several months from 2019 to 2021. Clustering by state gives at least 51 clusters, but clustering at state-

month would create more. I'm wondering if two-way clustering is better, especially since they included both state and week fixed effects. Using cluster at state-month might not consider cross-state correlation across weeks, even though week fixed effects account for national shocks. So, I think state-level clustering would be the safer approach to avoid correlated errors.

### **Assessing methodological choices**

They clustered at the state-month level to address serial correlation, but it feels a bit unnatural. The variation in non-pharmaceutical interventions (NPIs) across states appears limited since national restrictions often overshadow local policies. It's noted that they show large standard errors without significance, while Germany and France had significant NPI effects. They share their Stata do-files and source data, but there's a potential confusion treating "fear of infection" as only pandemic-related issues. Also, they might not adequately account for seasonality and rely on some assumptions that could lead to misleading conclusions.

### **Analyzing comparability and findings**

It looks like they included week-of-year fixed effects to control for seasonality, which is good, but it reduces comparability since these factors are helpline-specific, interacted with "year," "week," and "day." There's also a concern about p-hacking, as they tested eight categories without adjusting for multiple tests, even though they note their results are robust with small effect sizes. The claim that "income support alleviates distress from lockdowns" might simply reflect general time patterns. For the US, they didn't find significant effects from stringency or income support, but they concluded that suicide-related calls increased with stricter policies and decreased when income support was provided.

### **Clarifying findings and relevance**

The statement about findings particularly refers to helplines in Germany and France, while the US data isn't included, which might overshadow the overall conclusions due to the limited sample size. There's also uncertainty regarding the interplay with changes in mental healthcare supply. They mention re-running the analysis as the situation matures. The global relevance is rated high given the focus on mental health, but the open science aspect is moderate since the data is restricted, although they do provide the code and a contact list.

### **Evaluating research quality**

They provide aggregated source data, which is great! The clarity in writing and figures earns a high score in "Logic\_communication," though there are no sample selection details for all helplines. The "Claims\_evidence" is rated moderate-high due to the potential illusions of substitution, making it a bit debatable. The methods show moderate-high quality since they used fixed effects, but there's an observational identification of policy effects. The contribution to knowledge is high with new cross-country data. Overall, this article is of high quality, likely earning an 85 percentile.

### **Assessing quality scores**

It's not the absolute best, but it's strong overall. I'd rate "Claims\_Evidence" around 78, "Methods" at 75, and "Advancing Knowledge" at 80. For "Logic & Communication," I think an 88 is fitting. However, "Open Science" might drop to about 60 due to data restrictions, while "Global Relevance" sits at 82.

I'll provide credible intervals, maybe  $\pm 10$  for the overall. For journal tier predictions, the normative "should" might be around 4.2 since it's in Nature, while "will" could range from 3.8 to 4.4. I'll propose "should" at 4.1 and "will" at 4.3, with intervals.

### **Preparing for journal ranking assessment**

The request is for "Journal ranking tier 'should' and 'will'" on a scale of 0.0 to 5.0, including

credible intervals. I think I'll aim for a midpoint score of 4.0, although that might feel a bit unnatural due to limitations in open science.

Next, I need to craft a summary of about 1,000 words in one compact paragraph. They want it to be specific, neutral, and concrete, so I'll aim for a final word count between 800 to 1100 words to stay on target.

### **Structuring content for analysis**

Let's outline the content for the analysis. I'll start with data and measurement details, mentioning helplines from 19 countries, with 8 million calls and various topics. I'll highlight strengths, like a high-frequency sample using JHU and Oxford indices, while noting limitations such as unknown caller representativeness and measurement errors.

I plan to address potential biases from repeated callers, the effect of concurrent events, and the lack of clear causation in conclusions about income support. I'll provide suggestions for improvements, like additional event studies, modeling supply constraints, and using placebo tests. Overall, I want a balanced approach.

### **Producing metrics for assessment**

I need to finalize some metrics for the assessment. Let's go with an overall score of 84; I noticed I initially thought 85, but it feels right. For credible intervals, I'll set a lower bound of 74 and an upper bound of 92.

For specific categories, I'll assess claims evidence at 76, methods at 74, advancing knowledge at 81, and logic communication at 88. Open science will be 55, and global relevance will be 84. I'll list the journal tier suggestions as well.

Finally, I must write an assessment summary of about 850-1000 words, ensuring I'm specific with references like "Fig. 1 shows a 35% increase in calls."

### **Finalizing assessment metrics**

I need to include some important figures in the assessment. For instance, "Fig. 3 shows two helplines over the long run, consistent and dynamic." Then, "Fig. 4 details the US Lifeline, showing weekly call volumes with negative elasticity related to infections; however, the stringency and income support effects aren't significant and not robust."

I'll also mention "Fig. 5, which includes Germany and France's topic-specific regressions, indicating a positive elasticity for stringency and a negative effect for income support."

Now, I'll produce the metrics object and ensure the JSON matches the specified schema, confirming that the bounds are correctly set, and the conditions hold true.

### **Crafting the final analysis**

I want to focus on avoiding heavy formatting in this final analysis. It's important that I don't mention the authors to keep the writing neutral. I'll aim for approximately 1,000 words, targeting a range of 850 to 1,000 to maintain specificity without overwhelming detail. Now, let's take the analysis and turn it into the final document. I'll work on this to ensure it's clear and accessible for the reader.

## A.2 Class Warfare Political Exclusion of the Poor and the Roots of Social Revolutionary Terrorism 1860–1950

### **i** Model assessment summary

This paper studies whether political exclusion of the poor predicts the emergence and activity of social-revolutionary terrorist groups (anarchist/leftist) across 99 countries from 1860–1950. The main outcome is the annual count of active social-revolutionary groups from a pre-1950 terrorism dataset; the key regressor is a V-Dem expert-coded index of political power distribution by socioeconomic position (sign-reversed so higher values mean greater exclusion). The empirical core is a two-way fixed-effects OLS design with Driscoll–Kraay standard errors, supplemented by an IV strategy that instruments local exclusion with the average exclusion in the same world region (excluding the focal country). The authors run numerous robustness checks: alternative dependent variables (per-capita counts, inverse hyperbolic sine, binary activity, group formation), alternative class-cleavage measures (civil-liberties discrimination, unequal access to public services/jobs/business, and a composite index), additional controls (natural disasters, education, state involvement), nonlinearity tests, and adjustments to address potential regional trending in the IV setup. The main finding—a positive association between exclusion of the poor and social-revolutionary terrorism—is persistent, and no comparable relationship appears for nationalist, right-wing, or religious terrorism.

Key strengths: (i) Timely and clearly framed question with plausible mechanism grounded in rational-choice and grievance/relative deprivation perspectives. (ii) Rare use of pre-1970 terrorism data, filling a gap and enabling historical comparisons. (iii) Careful panel specification with country and year fixed effects, attention to cross-sectional dependence, and transparent reporting of estimates. (iv) Extensive robustness work, including multiple operationalizations of both the outcome and the focal regressor, and comparison across ideologies that functions as a placebo test. (v) Replication materials and use of public data sources make reproduction feasible.

Main limitations and identification risks: (1) Measurement of the dependent variable. The outcome is the number of active groups rather than attacks or casualties; this may capture organizational proliferation more than intensity, and differences in how groups are counted or named. The data are likely sparse with many zeros and small counts, which raises distributional concerns for linear OLS. The paper partially addresses this via binary and asinh transformations and a group-formation onset measure, but does not estimate count models with fixed effects (e.g., conditional Poisson), nor assess zero-inflation explicitly. This choice could affect efficiency and possibly inference; showing that FE Poisson results align would further bolster credibility.

- (2) Measurement of the key regressor. The V-Dem class-power index relies on expert coding and item response models; measurement error—especially for the 19th century—may be nontrivial. The authors acknowledge potential attenuation; the IV approach mitigates this but hinges on strong assumptions. Relatedly, life expectancy is used as the main development proxy due to data constraints; while understandable, it may capture mortality crises and wars as well as development, and might mediate the effect of exclusion. Additional controls for industrialization/urbanization (only partially approximated later) would help reduce omitted variable bias.

- (3) Endogeneity and the IV strategy. Using regional average exclusion as an instrument raises exclusion-restriction concerns: cross-border diffusion of anarchist/leftist networks, region-wide shocks (empire formation/dissolution, war mobilization, decolonization, legal changes), and region-level democratization may directly affect local terrorism independent of local exclusion. The authors try to neutralize this by: removing potentially collinear controls, adding regional trends for democracy/autonomy and regional means of several covariates, and even including mean regional terrorism. First-stage F statistics are strong, and Anderson–Rubin confidence sets support non-weak inference. Nonetheless, with a single aggregated instrument, residual correlation between regional exclusion and local terrorism through unobserved regional shocks remains a live possibility. A leave-one-region-out analysis or alternative instruments (e.g., historical franchise reforms, suffrage expansions, or settler institutions) would further strengthen causal claims.
- (4) Dynamics and timing. Most specifications use contemporaneous exclusion. If political inclusion reforms produce lagged changes in group activity, or if terrorism prompts reforms with a delay, estimates could be biased. While a lagged dependent variable is tested in robustness, lag structures for exclusion are not explored systematically (e.g., distributed lags or leads to test pre-trends). Event-study style checks around major enfranchisement or labor-law reforms would clarify dynamics and mitigate reverse causality concerns.
- (5) Controls and potential over/under-control. Democracy and domestic autonomy may be conceptually close to political exclusion and could lie on the causal pathway; including them risks absorbing part of the effect of interest or inducing post-treatment bias, whereas omitting them risks omitted variable bias. The paper shows robustness across specifications, but a directed acyclic graph (DAG) or explicit mediation analysis could clarify which specifications should be interpreted as total vs. direct effects. Additional plausible confounders (union density, strike frequency, industrial employment share, urbanization rates) are largely unavailable for the full period but, where feasible, would be valuable sensitivity checks.
- (6) External validity and scope. Africa is excluded for data reasons, colonies and dominions are handled via a domestic autonomy index, and classification of groups may vary by archival coverage. These choices are well-motivated but limit generalizability and could introduce selection if data coverage correlates with both exclusion and terrorism. The authors partly address this by including autonomy and discrimination indices, but a discussion of coverage bias in the Old Terrorism Dataset and V-Dem for early periods would help readers calibrate inference.
- (7) Interpretation and magnitudes. The paper focuses on statistical significance. Substantive interpretation of effect sizes is limited; translating coefficients into probabilities of any group being active or into changes in expected group counts for realistic shifts in exclusion would help readers assess practical importance. Heterogeneity by baseline level of development or regime type is not explored beyond nonlinear checks; interactions might reveal where exclusion is most consequential.

Communication and transparency are generally strong: the argument is clear, data sources and construction steps are described, tables and figures are interpretable, and numerous appendices

document robustness. Replication materials are available via a data repository, which aids reproducibility. Overall, the paper provides credible, carefully triangulated evidence that political exclusion of the poor is linked to social-revolutionary terrorism in 1860–1950. The central identification step—regional-trends IV—remains the most debatable element, and the distributional choice for counts is a secondary methodological limitation. Neither seems fatal; however, future work could solidify causality with alternative instruments, dynamic designs, and count-model estimators, and reinforce mechanism tests by examining reform episodes and labor-movement covariates.

## Model reasoning trace

### **Creating assessment JSON**

I need to create a JSON that aligns with the schema for paper assessment with tiers version 2. First, I'll draft an assessment summary of about 1000 words, focusing strictly on the content of the provided PDF.

Next, I'll include metrics like percentile scores and credible intervals for several categories, which will involve boundaries from 0 to 100 and maintaining inequalities. I'll also determine tier recommendations based on the document's content, considering the empirical study's methodology and its thoroughness, including various robustness checks across a sample of 99 countries.

### **Reviewing replication data and issues**

I need to provide details about code and data replication. They mentioned that replication data and code can be found on Mendeley Data, which supports open science. The dataset comes from Tschantret's 2019 work and V-Dem, and they've included replicable code.

I have to identify potential issues: the dependent variable is the number of active groups, not attacks, due to data constraints. There are also questions of omitted variables, reverse causality, and the validity of the IV instrument relating to regional political exclusion. While they controlled for regional trends, concerns about exogeneity persist. The sample only includes 99 countries, omitting Africa and facing uncertainties with V-Dem expert coding and development proxies.

### **Analyzing dependent variable and methods**

They treat the dependent variable as a count using OLS, though counts with fixed effects might benefit from Poisson FE due to zero-inflation. They also explore transformations like inverse hyperbolic sine but don't use Poisson directly. The range of active groups goes up to 14, and OLS may be acceptable with robust errors.

However, they didn't address pre-trends or dynamic panel issues, and there are endogeneity concerns. The OLS coefficient suggests about 0.07 groups per unit of exclusion, but this interpretation isn't properly scaled. There are measurement issues related to ideology, and risks of sample selection bias persist due to potential misclassification and non-random missing data.

### **Examining lag structure and confounding factors**

There might be a lag structure present, but they didn't test for lags of exclusion, using only some in extensions. They mentioned including a lagged dependent variable in robustness tests, which is helpful. However, they didn't consider lags for explanatory variables or "placebo ideologies."

They didn't address potential confounding factors like unionization or labor conflict, and there

are no variables for strikes, likely due to limited data. Their theoretical foundation includes a rational-economic model and relative deprivation, which is good. While historical, they briefly touch on current inequalities, suggesting moderate relevance for policies today, but immediate guidance is limited. I'm glad they provide open science access via Mendeley for good replicability.

### **Assessing data reliability and methods**

Some data before 1900 relies on V-Dem version 10, which is replicable. When it comes to logic and communication, the clarity is good, and the figures, tables, and footnotes are well-structured. They're advancing knowledge by using a unique dataset focused on that early period and addressing ideology-specific effects, which is a valuable contribution—even if it's not surprising, it holds up across various checks.

Their methods involve OLS with fixed effects and Driscoll-Kraay errors, which is typical. However, the IV instrument might violate exclusion due to regional shocks driving terrorism directly. They incorporate year fixed effects to account for global shocks, but region-specific confounding potential still exists. While they add controls for regional democracy trends, strong assumptions underpin their identification. Another concern is the sparse country-year distribution of the dependent variable, which may lead to many zeros. Although the fixed effects plus Driscoll-Kraay might manage this, using a count panel with Poisson fixed effects could be more efficient; still, they explored a binary dependent variable for robustness.

### **Evaluating dependent variable and control biases**

The dependent variable is the number of active groups, and a single country-year can have multiple active groups merged. The theoretical focus seems to be on activity intensity rather than just group counts, which is good since they also test group formation as an onset. However, there's potential for overcontrol bias because they include life expectancy and democracy, which might act as mediators for political exclusion. While their goal is to disentangle effects, controlling for democracy could inadvertently overlap with the political exclusion variable, even though they aren't the same.

The instrument they use is region-level exclusion, but if transnational groups are influenced by that, it may compromise the exclusion restriction. They do try to account for regional trends and local terrorism, but it might not address everything entirely. They apply Driscoll-Kraay standard errors for clustering, which is solid, though missing data from Africa limits generalizability.

### **Assessing functional form and coding issues**

The time fixed effects are applied well. They assume a linear effect, but don't test for non-linearity regarding exclusion, which might overlook some important dynamics. They did explore alternative measures, which is a good step.

The dependent variable coding comes from Tschantret, but they don't adjust for group size or actual attacks, creating a potential mismatch. Their sample covers 1860-1950, including many colonies and addressing domestic autonomy and discrimination, which is commendable. Ethical considerations aren't applicable here.

I need to craft an assessment summary to highlight these strengths and weaknesses while providing numeric values. For overall scoring, I'd suggest around the 72nd percentile with a confidence interval of [60, 83].

### **Evaluating assessment scores**

The overall quality seems strong, but not top-tier.

For claims and evidence, they clearly assert that political exclusion drives social-revolutionary

terrorism, providing solid OLS and IV evidence along with robust checks. I'd rate this around 75 with a confidence interval of [62, 86].

Their methods include good fixed effects OLS with Driscoll-Kraay errors, although there are some concerns about IV exclusion and DV measurement. I'd give this about 68, with a CI of [55, 80].

The advancing knowledge aspect benefits from a unique historical dataset bridging the pre-1970 era, so I'd score it at moderate-high 70 with [58, 84].

In terms of logic and communication, it's very well-written, earning an 82 with [72, 90].

On the open science front, the data and code are available on Mendeley, enhancing replicability, so I rate this at 85 with [75, 92]. I have to mention that while the dataset for old terrorism is accessible, careful attention is needed for replication code, which is good.

Lastly, concerning global relevance, it's historically oriented but ties to current issues of inequality. I'd give it a 55 with [40, 70]. Overall, this could potentially reach a top field journal tier.

### **Assessing publication evaluation**

The paper appears in Defence and Peace Economics, which is a field journal. However, I think we should evaluate it normatively instead of focusing on publication standards. I'd rate its strengths around 3.4, classifying it as a "Top B-journal/Strong field," possibly reaching a 4, making my score approximately 3.6 with a confidence interval of [3.0, 4.2].

For predicted publication, considering its typical field relevance, I'd say 3.2 with a CI of [2.6, 3.8]. Sure, I need to provide ratings from 0-5, including decimals, and I'll keep it straightforward as the instructions suggest.

I aim for a robust summary of about 800-1100 words, ensuring credible intervals are at 90%. It's essential to provide the structure in JSON format, and I won't worry about identifying people in images since it doesn't apply here.

### **Crafting assessment summary**

Let's summarize the research question and dataset first. The focus is on how political exclusion may foster social-revolutionary terrorism, utilizing a unique dataset.

Next, there are several threats to consider, particularly around the dependent variable (DV) measurement—comparing groups versus attacks, low counts, and significant zeros. It's noted that they didn't use Poisson FE or negative binomial models, although results remained robust across transformations like binary and inverse hyperbolic sine (IHS).

For the independent variable measurement, they reverse the V-Dem "political power by socio-economic position," relying on expert ratings that might introduce measurement errors, especially for pre-1900 data.

Concerns about endogeneity loom as reverse causality and omitted variables may play a role. Their instrumental variable (IV) approach, based on regional political exclusion, raises worries about exclusion restrictions, particularly with transnational ideological diffusion.

Additionally, they don't test lags on exclusion, which calls into question the short-run vs long-run effects.

Lastly, while they account for domestic autonomy and regional discrimination in colonies, measuring the DV becomes complex, especially regarding terrorist groups and classification.

### **Assessing dataset and methodology**

They used a dataset covering 99 countries, but excluded Africa, which limits generalizability. When it comes to alternative measures, they tested additional class cleavage indicators, reducing the sample size, yet the results generally hold, except for state business opportunities.

For outliers and influential values, they addressed these using inverse hyperbolic sine (IHS) and binary measures, but didn't mention jackknife techniques, which could improve robustness. Their standard errors, Driscoll-Kraay, seem solid. However, potential high multicollinearity between exclusion and democracy wasn't discussed, even though it's likely present.

They did not convert their findings into meaningful effects, like probabilities of at least one group, which could enhance interpretation. On a positive note, they briefly show that other ideologies are unresponsive, which is good.

Their control set includes life expectancy as a development indicator, but this might act as a mediator or confounder since GDP data is missing. Urbanization and industrialization considerations were notably absent, aside from later references to state economic involvement and access to education.

Reproducibility looks good with data and code links provided, and the graphical presentation, with Figure 1 showing counts and Figures 2 and 3 included, is well done!

### Filling the JSON format for metrics

I need to structure the metrics in JSON format correctly.

Overall score is 72, with a confidence interval ranging from 60 to 82. For claims and evidence, the score is 75, with a range of 62 to 86. The methods score 68, with lower and upper bounds of 55 and 80 respectively. Advancing knowledge gets a score of 70, ranging from 58 to 82. Logic and communication rates an impressive 84, with a confidence interval of 74 to 92. Open science scores 86, possibly within the range of 76 to 94. Finally, the global relevance score is 56, with a range of 40 to 68.

As for the tiers, "should" proposes a score of 3.6 between 3.0 and 4.2, while "will" suggests a score of 3.2, ranging from 2.6 to 3.8. I need to ensure I'm matching the schema names like "tier\_should" and "tier\_will" to keep consistency.

Alright, it's time to write up the assessment summary without including any author names beyond general references. I'm aiming for about 900 to 1100 words, focusing on specific issues while avoiding heavy formatting — let's construct that!

## A.3 sciadv.adh4030

### i Model assessment summary

This paper combines quasi-experimental variation in the timing and duration of COVID-19 school closures across German federal states and school tracks with (i) nationally representative adolescent mental health survey data collected near the end of the first lockdown (COPSY, May–June 2020) and a pre-pandemic benchmark (BELLA, 2015–2017), and (ii) high-frequency crisis helpline data (2019–2020). The main identification uses a two-way fixed-effects design leveraging within-state, across-grade/track and within-grade/track, across-state differences in mandated weeks without in-person schooling, with age and sex controls and extensive fixed effects. The authors conclude that each additional week of school closure worsened health-related quality of life (HRQoL) and increased psychosomatic, behavioral-emotional, and depressive symptoms (no detectable effect on anxiety), that effects were stronger for boys, younger adolescents, and those with limited living space, and that family problems were the predominant issue revealed in helpline calls; they further argue school closures largely explain the overall deterioration of youth mental health observed between the pre-pandemic and first-wave samples, albeit with wide confidence intervals.

Key identification threats and data limitations are as follows. First, the main estimates come from a single cross-section during the first wave (COPSY) with variation in exposure generated by reopening schedules. While two-way fixed effects absorb state-level and grade/track-level differences, the design relies on the assumption that no other within-state, grade/track-specific shocks correlated with reopening sequences affected mental health. The logic is plausible since other policies were not grade-targeted; nevertheless, unobserved, grade-specific changes in school support, remote learning intensity, or local implementation that correlate with reopening priority could bias estimates. The paper mitigates this by adding state-level pandemic severity and policy stringency controls in a correlated random effects (Mundlak) framework; results are stable, but those controls cannot rule out grade-specific factors by state. Second, the pre-pandemic comparison relies on BELLA measured 2015–2017 in a different mode (computer-assisted, in-person) versus COPSY online in 2020. Secular trends and mode effects could confound the pre/post difference. The authors include time trend and seasonality checks and find little sensitivity, but the several-year gap and mode shift remain an important caveat when concluding that closures “largely explain” the deterioration. Relatedly, only two of the mental health scales are available pre-pandemic; the other three are standardized within COPSY, which complicates harmonized interpretation across measures.

Third, exposure measurement is imperfect. “Weeks of closure” is defined up to the (unknown) individual COPSY interview date and is top-coded at the COPSY start date in baseline; an alternative uses the end date; and another adjusts for holidays; results are robust. However, the end of closure is defined as “partial reopening,” which may not capture the intensity of in-person schooling (e.g., hybrid schedules varying across schools/classes), implying classical measurement error likely attenuating estimates and leaving possible heterogeneity unobserved. Fourth, inference and clustering: standard errors are clustered at the treatment level (state $\times$ grade $\times$ track), which is appropriate conceptually, but the effective number of clusters may be moderate; no wild-cluster bootstrap is reported. Fifth, the COPSY analytic sample is n~907 adolescents; sampling weights support representativeness on observables, but non-response/selection into online participation could bias results if correlated with both mental

health and closure exposure within fixed-effect cells.

Sixth, the helpline analyses are suggestive rather than causal. The linking of callers to exposure uses age as a proxy for grade and the receiving center's location as the state identifier, which may not precisely reflect the caller's residence or school context. Topic coding is counselor-reported and calls considered are those that progressed to counseling, creating selection that may vary over time. The analysis shows compelling descriptive patterns (family-problem calls spiking when schools are closed and diverging between early vs. late returners), reinforced by an ARMA model with fortnight dummies; however, other contemporaneous changes (e.g., local news cycles, staggered reopenings of out-of-school activities) could contribute, and the median split on closure weeks is a coarse exposure metric.

Seventh, heterogeneity analyses rely on interactions with discrete splits (e.g., living space per school-aged child above/below the median), which are interpretable but may mask continuous dose-response patterns; measurement of living space and the allocation of space among children are approximate. Eighth, the absence of an anxiety effect may reflect the timing (short-run, first wave) or measurement focus (generalized anxiety), and should be interpreted cautiously. Ninth, the large per-week effect magnitudes (e.g.,  $\sim 0.11$  SD per week on HRQoL) imply sizable cumulative impacts over the observed exposure range; while consistent with the pre/post differences, they depend on linearity over the 4.7–10.1 week window and may not extrapolate to later, longer closures—as the authors note. Tenth, external validity is limited: Germany's first-wave context combined low adolescent case rates, specific reopening priorities by grade/track, and particular broader restrictions; the authors appropriately caution about cross-country generalization.

Internal consistency and robustness are otherwise strong: estimates are stable across multiple specifications (e.g., alternative imputation of exposure, quadratic terms, parental-reported outcomes, absorbing state $\times$ track fixed effects), and adding state-level pandemic severity and detailed policy stringency measures does not change the closure effects; only one of 65 policy coefficients reaches 5% significance (on anxiety), limiting concerns about omitted, state-level policy confounds. Pre-pandemic mental health measures are uncorrelated with mandated closure duration, supporting quasi-exogeneity of exposure. Communication is clear and transparent; figures align with text; the logic linking design to assumptions is explicit; and the heterogeneity and helpline sections triangulate mechanisms (family stress). Replication barriers exist: the core microdata (COPSY/BELLA) and helpline data are restricted-access; the ordinance-derived school-closure dataset and aggregate policy data are available, but code availability is not explicitly documented. Overall, this is a careful, policy-relevant study with a credible identification strategy given available data, appropriately framed limitations, and strong but not definitive evidence that school closures directly worsened adolescent mental health in the short run and likely explain a substantial share of the observed deterioration during the first pandemic wave.

#### 💡 Model reasoning trace

##### Considering mental health trends

I need to mention selection into tracks while noting the missing within-track grade-specific pre-trends. It's vital to include pre-pandemic mental health data from 2015-2017, earlier than 2020. I should also consider potential secular trends and check time trends with quarter dummies. Measurement concerns arise from self-reported mental health, and there might

be reporting bias due to different standardizations. Plus, using sampling weights and being aware of finite cluster concerns is important for accuracy. I'll look into the survey start date assumptions and how they handle top-coding and variations.

### **Addressing potential confounding factors**

I see that they control for age and sex, incorporating school track-specific grade level fixed effects and state fixed effects. They identify variables across states and grade-tracks, which seems promising. However, potential confounding arises with additional state-level policy measures that aren't age- or grade-specific, especially if families with older kids differ. Despite this, they include fixed effects for state and grade-by-track while also considering state-level stringency measures. They interpret effect sizes per week as -0.107 SD for health-related quality of life, but that seems significant. Am I interpreting that correctly?

### **Examining effects of school closures**

It seems that over 4 additional weeks would lead to a decline of 0.428 SD, and the total decline they found compared to pre is 0.824 SD. They suggest that closures largely explain this deterioration but also caution that there's a wide confidence interval. If mandated closures average about 4.7 weeks, their per-week effect sizes might be overstated. The measured overall effect net of closures becomes insignificant, implying the closures alone may not fully account for the differences. I'm curious about how they interact closure measures with COPSY!

### **Analyzing closure effects in the study**

In Equation 3, they include variables like dscx, ct, and the interaction ct\*dscx. Beta3 measures the effect of closures during the pandemic, while Beta2 captures the COVID-19 effect net of closures. Interestingly, Beta2 is reported as insignificantly different from zero, suggesting closure explains only 31% of the decline in health-related quality of life. There's a moderate sample size of n=907, and variations in closure weeks are limited, which could lead to imprecision. Measurement modes also differ between pre-pandemic and pandemic surveys, contributing to potential biases.

### **Evaluating anxiety measurement and helpline data**

Anxiety, measured as SCARED, shows no notable effects, which might be due to measurement limitations tied to generalized anxiety. The crisis helpline data relies on call volumes, topics, and locations of receiving centers, but it's not fully random as it captures only calls that proceeded to counseling. Although confidentiality is maintained, the observational nature limits causal conclusions. They present moving averages and ARMA regression with dummies while identifying variations over short windows. There's a suggestion that family problem calls decreased after reopening, but linking these effects to closure exposures is challenging due to potential confounding factors, including top-coding closure weeks.

### **Exploring grade-specific confounders**

I'm considering the underlying assumption that there are no grade-specific confounders at the state level. It's possible that distance learning quality differed across grades or tracks, which ties back to the closure effects. There's also a concern that the intensity of remote learning might vary independently from closure weeks. Although closure weeks align with grade levels, grade-track fixed effects are included to account for this. However, if certain grades, particularly higher ones, reopened earlier, it could create a different slope in mental health changes. This might introduce bias and calls into question whether we can assume stable grade-specific effects across states. What if, for instance, ninth graders in some areas had greater access to remote learning resources?

### **Assessing grade-specific measures and data access**

There's a challenge in considering "other measures" that aren't grade-specific, as resource allocation may vary based on policy and the reopening of grades. Since reopening timelines might correlate with additional grade-specific supports, it's tough to exclude those factors. They work with a small number of states (16), but clustering may yield a significant number of clusters, potentially over 100. Interpreting the results as "school closures largely explain" the findings seems a bit strong, and they do provide cautions. Additionally, the dataset isn't publicly shareable, and access requires contact. Code availability is unclear, making open science somewhat limited. The study's global relevance is high for policy discussions, especially regarding closure tradeoffs, but its generalizability is uncertain, focused only on Germany's first wave.

### **Analyzing methods and measurement boundaries**

The use of two-way fixed effects seems reasonable, and I like that they incorporate Chamberlain-Mundlak correlated random effects to include factors like state-level policy stringency and pandemic severity. They have a variety of robustness checks, including imputation methods and using parental reports as outcomes, which is good! However, addressing pre-trends through within-sample event studies isn't possible with this cross-sectional dataset. Additionally, there are concerns about measurement boundaries. Standardizing scales like HBSC-SCL, SDQ, and SCARED within COPSY without a pre-pandemic baseline limits effect sizes. For KIDSCREEN-10 and CES-DC, standardization to BELLA makes sense for comparisons across time, but I wonder if they also standardized for COPSY-only regressions.

### **Evaluating standardization and measurement concerns**

KIDSCREEN-10 and CES-DC are standardized to a mean of 0 and SD of 1 in BELLA, and they use these scales in COPSY. If the standardization remains consistent, the effect sizes would reflect pre-pandemic levels. However, for other scales used only in COPSY, they standardize within that context, which feels a bit inconsistent, but they clarify this for comparison purposes. Sampling weights and clustering SEs at the treatment level are good practices. There's potential measurement error regarding closure exposure, as many grade-level reopenings were partial. This could lead to inaccuracies, but it might also introduce variability. They assume that other measures don't vary by grade level, though I'm curious if there were any grade-specific contact restrictions.

### **Assessing helpline linkage and study evaluation**

The helpline links calls based on the location of the receiving center, which might not match the caller's actual location. They recognize this is a regional center receiving calls from various areas, which is important. They use age as a proxy for grade since the exact grade isn't known. We should assign quantitative metrics related to fields like public health, social science, or economics. Overall, it's a strong paper with credible quasi-experimental design addressing a relevant policy question. Limitations include its cross-sectional nature, which makes it hard to fully separate the effects of the pandemic from closures. The helpline analysis provides suggestive evidence but isn't completely causal. I'd rate it around the 80th to 88th percentile compared to top field work.

### **Evaluating overall scores and claims**

I'm setting the overall midpoint at 84, with a confidence interval of 75-90. The major claim that "school closures largely explain deterioration" might be slightly overstated, but it's presented cautiously, so I'm giving it an 82 (74-88). For methods, the two-way fixed effects and correlated random effects checks are solid but not gold standard—I'd rate that an 80 (70-87). Advancing knowledge has strong policy relevance and adds to the literature, so maybe an 85

(78-92). However, global generalization remains uncertain, so I'm setting that at 83 (74-90). The writing and figures are clear and transparent, so that gets an 88 (80-94).

Open science is limited since the data isn't open, only a compiled school policy dataset is available. I'm unsure about the code, so I'll give that a low score of 55 (40-65). For global relevance, it's high because they provide quant estimates—I'd say 80 (70-90). Considering the quality, I'd rate the journal tier at 4.3, with a confidence interval of 3.8-4.7, since it's published in a good-tier journal like Sci Adv.

### **Clarifying journal tier metrics**

I'm considering the journal tier, which will likely fall between a top field or A-journal, so I'm thinking of setting that at 4.2 with a confidence interval of 3.5-4.8. Now, the schema expects numeric entries for 'tier\_should' and 'tier\_will' as objects with corresponding scores and confidence intervals, which are nested under the metrics as 'tier\_should' and 'tier\_will'. I see that these metrics should include score, ci\_lower, and ci\_upper. Next, I need to ensure that I provide 90% credible intervals within 0-100 for percentiles, making sure that the lower value is less than the mid, which is less than the upper.

Additionally, the assessment summary should aim for about 1000 words. I plan to produce around 800-1100 words and I'll be specific by mentioning internal coherence checks, sample sizes, and cluster counts. While the exact cluster count isn't known, I can discuss the possibility of a limited number of treated cells, as they cluster by state, grade, and track, suggesting an adequate count but with some caution.

### **Highlighting potential issues and structuring metrics**

I need to mention that the claim of no anxiety effect might stem from measurement timings early in the pandemic. While we can't rely solely on external factors, it's a plausible limitation. Another aspect to address is potential selection bias due to non-response in COPSY. Although weights are used to ensure representativeness, unobservable selection might still be an issue. Additionally, family climate is measured by parents, but there's no direct mental health effect found on them, which raises the possibility of measurement errors due to adaptation or under-reporting. The paper claims all necessary data for evaluations are available, but replicability isn't straightforward since code isn't provided.

To comply with the metrics, I need to produce sub-objects for each metric: overall, claims\_evidence, methods, advancing\_knowledge, logic\_communication, open\_science, and global\_relevance. Each one will need midpoint, lower\_bound, and upper\_bound values, and I should ensure no extra text beyond the JSON. Valid numeric fields will be required, using decimals where needed.

### **Creating the assessment summary**

I need to summarize the design as a two-way fixed effects approach, utilizing exogenous variations from federal states' reopening sequences by grade and track. The primary data comes from COPSY's wave from May to June 2020 and pre-pandemic BELLA measurements from 2015-2017, using various mental health measures while applying weights for robustness.

The findings show that an additional week of closure reduces health-related quality of life (HRQoL) while increasing psychosomatic issues, though no anxiety effects were found. There are some threats to validity, mostly from the cross-sectional design and potential measurement errors.

I must also note concerns around external validity, the representativeness of helpline calls, and the lack of longer-term follow-ups or robust coding for spatial data. Overall, there's internal consistency across measures, though transparency may be hindered by access constraints to

their data and code.

### Assigning scores for evaluation

I'm setting the scores now, starting with "Open Science." Given they don't provide code and the data is restricted, I think a score of 50 with a range from 35 to 60 makes sense. I previously considered 55, but 50 seems more fitting right now.

Overall, I'm aiming for a midpoint of 84 with lower and upper bounds of 75 and 90. Here's what I'll assign:

- Claims Evidence: 82 (72-88)
- Methods: 80 (70-87)
- Advancing Knowledge: 83 (73-90)
- Logic Communication: 88 (80-94)
- Open Science: 50 (35-60)
- Global Relevance: 80 (68-90)

Now I'll produce the 'tier\_should' score of 4.3 (CI 3.8-4.7) and 'tier\_will' score of 4.2 (3.5-4.8). I need to ensure this is all valid JSON, respecting the constraints on percentiles. Also, I should produce a detailed summary of around 1000 words, keeping in mind the desired verbosity.

### Preparing the summary

I want to craft a summary that falls around 900 to 1050 words. My goal is to keep it neutral, specific, and concrete, focusing solely on the content of the PDF. While I can mention the paper's title, I need to avoid bringing up "Science Advances" or the status of the paper as evidence. Also, to maintain neutrality, I should refrain from mentioning authors' names. Now, I'll proceed to produce the final JSON.