

RESEARCH ARTICLE

negative results in social science

*David Lehrer^a, Janine Leschke^b, Stefan Ihachimi^c,
Ana Vasiliu^d and Brigitte Weiffen^e*

^aDepartment of Political Science, University of Helsinki. P.O. Box 54 (Unioninkatu 37), FIN-00014, Finland

E-mail: david.lehrer@helsinki.fi

^bEuropean Trade Union Institute, Research Department, Blv. du Roi Albert II, 5 box 4, B-1210 Bruxelles, Belgium

E-mail: jleschke@etui-rehs.org

^cMax Planck Institute for Demographic Research, Konrad-Zuse-Strasse 1, D-18057 Rostock, Germany

E-mail: lhachimi@jspurc.org

^dNational Centre for Sustainable Development, Dr. Burghilea 36, RO-73102 Bucharest, Romania

E-mail: vasiliu@jspurc.org

^eDepartment of Political Science, University of Tübingen, Melanchthonstr 36, D-72074 Tübingen, Germany

Corresponding author: E-mail: britta.weiffen@uni-tuebingen.de

doi:10.1057/palgrave.eps.2210114

Abstract

Do academic publication standards reflect or determine research results? The article proposes minimal criteria for distinguishing useful 'unpublishable' results from low-quality research, and argues that the virtues of negative results have been overlooked. We consider the fate these results have suffered thus far, review arguments for and against their publication and introduce a new initiative – a journal to disseminate negative results and advance debate on their recognition and use.

Keywords methodology; negative results; philosophy of social science; publication bias

NOT ALL RESEARCH RESULTS ARE PUBLISHED. SO WHAT?

In the social sciences, as in other sciences, conventions of academic method lead to selective reporting of results. Potentially useful information is lost. To be acceptable to one's scientific peers, research findings must be original, replicable, significant and relevant to the existing body of theory. Little room is left for the result of a sound research process that does not fulfil all of the above criteria. Pressured to 'publish or perish', researchers produce acceptable results ever more efficiently and do not dwell too long over results that are puzzling. This does not mean that apparently outlandish results, which eventually prove to be stimulating and seminal, never appear in print. But these are not regular events, and publishing such results can be difficult.

The quest for publishable results contributes to the social sciences' oft-criticised segmentation and overspecialisation. The addressed community becomes ever-smaller and the employed methodology more sophisticated. One may require years of experience in a particular subfield to replicate a result. In the social sciences, this situation is even more intractable than in the 'hard sciences,' as no fixed set of theory or methods exists. We find sociology divided into qualitative and quantitative approaches, economics transformed from a qualitative and historical discipline into 'social physics', and political science fragmented into 'separate tables' each with its own etiquette and menu. In this context the odd, unexpected or 'negative' result cannot be justified only by the soundness of the method applied.

Hence, we know more and more about less and less. We favour detailed documentation of what is already known over a rough view of uncharted territory; the 'one-best' path forward trumps a broadened perspective on the whole labyrinth.

'To be acceptable to one's scientific peers, research findings must be original, replicable, significant and relevant to the existing body of theory.'

Once we have found a single plausible explanation, rather than redoubling our efforts at discovery, we tend to stop searching and prepare to publish ... leaving it to other researchers to propose models alternative to the ones we adopt (Scheines, 2002; Granger and Jeon, 2004; Bartels, 1997). Moreover, we rarely take publication standards into account in describing our findings or assessing those of others: joint consideration of publication standards, methodological standards and theory development is a rare event and not a standard element of literature review.

The unsolved problem of negative and unpublished results informs discussion of knowledge cumulation in social science.

Studies of publication bias in political science, economics, and the hard and clinical sciences deal directly with questions of whether or not academic publication standards reflect or determine research results. This literature has focused on measuring the effect of bias on the publication of quantitative studies (Scargle, 2000; Gerber *et al*, 2001). Less work has been carried out on bias in the publication of qualitative studies, and on individual scholars' experience with their own unpublishable results (Gans and Shepherd, 1994). The task of determining what social science researchers should do about publication bias has thus far remained secondary.

Debates on the possibility and utility of meta-analysis in the social sciences necessarily address questions of negative

and unpublished results. Comprehensive meta-analysis is not possible if relevant findings remain unpublished. Gerber and Malhotra (2006) extend this concern to the level of individual results, asking if publication bias affects not only the aggregate conclusions of a long-term retrospective on a field but also the credibility of individual results. Others have argued that the effects of what is called publication bias on the body of results that are published may not be an impediment to useful meta-analysis (Sigelman, 1999), and that resolving the issue of publication bias for the sake of meta-analysis may not be worthwhile when elegant alternatives can be construed (King and Zeng, 2006). Rather than using meta-analysis to open a window onto unpublished results, shifting research practices to report a wider range of results may instead provide direct evidence of publication bias and its effects to supplement the indirect evidence exposed by meta-analysis.

Compelling arguments, from various quarters, for widening the range of formats of academic publication have in common a concern with the interplay between standards of publication and standards of evidence. Analyses of the economics of scientific research and publication (Zamora, 2002) and of how social science results are used (Baert, 2005), and projects to reconsider standards for evidence in social science research (Cartwright, 2004), have all contributed arguments in favour of revamped publication criteria and formats.

We argue that the model of communicating research in which only immediately relevant 'conclusive' and 'interesting' results are published may not be the best one. Under such a model, crucial but marginal forms of engagement with previous research results (such as replication) or with future ones (such as advancing tantalising hypotheses as 'negative results' rather than only building

on the predictably defensible ones) are inherently unrewarded by publishing practices and remain underprovided. In consequence, information on past efforts is not put to effective use; mistakes are replicated unknowingly by multiple researchers; accepted truths are insufficiently scrutinised; standards of evidence, theory and method are applied inconsistently across subfields; inordinate attention is placed on originality and innovation to the detriment of collaborative advancement of knowledge; sophisticated analytic techniques become ends in themselves; and the research process is unnecessarily shrouded in a twilight of unavailable data sets and non-replicable results. In contrast, a model in which social scientists engaged jointly with negative results might be one in which the discipline's hidden web of 'questionable' findings and 'uninteresting' evidence would become manifest. Under that model of research reporting, both what we know and what we do not know might take on new dimensions of application and interest.

DEFINING NEGATIVE RESULTS

Because the social sciences, unlike other sciences, do not rely on a single method such as experimentation, negative results in social science may be arrived at by diverse paths. We therefore propose a processual, rather than a generic, definition: negative results are findings that are validated *outside* the research context in which they were generated, but *not* by the standards of the heuristic process that generated them. As the examples below illustrate, this tension between the research process and some of its output may be one recognisable feature of negative results.

From this it may follow that both mistakes and unexpected findings are 'negative results'. However, the same criteria

that invalidate any result apply. Negative results are results that carry their own, often overlooked, merit. Negative results may be unexpected, but one could conceivably look for good negative results on purpose. An editorial context that welcomes negative results might create incentives for researchers to find them, with both 'positive' and 'negative' consequences for research.

The term 'negative results' may be an oxymoron – 'results' are generally good and 'positive' things. A carefully constructed referential system validates certain statements as good, useful and worthy of dissemination. The same system of reference discards statements that fall outside this set. Yet these other findings are also an output of the research process; they are conclusions not within the range of the validation system used to select some of the output as 'results'.

Selection is part of the research process: some output is discarded in order to move the task or programme forward along pre-set lines. Research also changes course as contrary evidence is found along the way. 'Positive' results are validated by reference to the research project's goals, the horizon and history of the field for which the output of the research is intended, and the beliefs, experience and expectations of the researcher. This selection process may be motivated by the expectations of the researcher and her intended audience, but ought we to restrict access to negative findings for those working within other subfields and paradigms, who may not have assumed or explained them away in the first place?

It is tempting to seek a simple marker or procedure to recognise 'negative results.' We suspect, however, that enforcement of 'result recognition recipes' is precisely what increased reporting of negative results might diminish. Bringing the epistemic victims of significance, confirmatory and publication bias to light

'Ersatz results are empirical findings that bear no clear relationship to any theory.'

as 'negative results' may be the best way to highlight, define and unravel the constraints by which they are produced.

HOW TO SPOT A NEGATIVE RESULT

Negative results are generally unused findings that may open new perspectives on the stylised facts of various social science subfields or paths to new research programmes. It may be difficult to distinguish useful negative results from failed low-quality research. Applying traditional quality criteria is not straightforward: to be valuable, negative results must meet certain scientific standards, but at the same time, their defining feature is that they run counter to established criteria of publishable research.

The following typology differentiates negative results based on the relationship they bear to the research process that generated them. We identify a broad range of negative results; criteria for determining which of these may contain information of value to other researchers are discussed in the subsequent section.

Inconclusive results in part confirm and in part reject theoretic expectations. *Non-results* bear a relationship to existing theory and hypotheses, but neither confirm nor reject researchers' assumptions. *Confutative results* appear to contradict previous findings and established theories. *Ersatz results* are empirical findings that bear no clear relationship to any theory. Examples of negative results are not easily found in the literature, as they are almost by definition not reported. Most of the examples below are therefore generic or hypothetical.¹

Inconclusive Results

Inconclusive results may be unstable and highly sensitive to model choice. They arise when analysis leads to diverse outcomes depending on what data are used, which cases or periods are observed or which methods are employed. They appear to offer contradictory evidence in part for and in part against the initial assumptions or the hypothesis under scrutiny.

One problem in evaluating inconclusive results lies in ensuring that they are of good quality and derive from a research process that meets high scientific standards. Reviewers must determine that results' inconclusiveness does not stem from using inappropriate data or from applying methods incorrectly.

Such determinations are not without pitfalls. The more sophisticated the method, the more contested its correct implementation is. Some procedures routinely applied by many researchers are deemed highly questionable by others. One example is the Beck and Katz (1995, 1996) approach to pooled time series cross-section analysis and the critique raised by Wilson and Butler (2007). Selection of method may not be a matter of right or wrong, but of degree of justifiability and of appropriateness to the specific research topic. Whether or not a specific analytic approach is justifiable may be difficult to judge by researchers working outside the substantive field, who are not familiar with the technique in question.

The reporting of inconclusive results might therefore enable reconciliation of contradictory evidence. It might lead researchers to reformulate or refine questions, or open debate on the applicability and limitations of particular methods and approaches.

Case I₁: Inconclusive results due to data collection method

In the late 1980s and the early 1990s, the National Opinion Research Center (NORC)

conducted a study on the Social Organization of Sexuality via face-to-face, paper-and-pencil interviews (Laumann *et al*, 1994). A surprising finding was the reporting discrepancy between men and women (men report more, women fewer partners of the opposite sex), a puzzle for which the research team 'like others, have no good answer'. Tourangeau and Smith (1996) attempted to resolve this inconsistency by an experiment on data collection mode and question format and context. They assumed that data collection problems were likely when sensitive questions concerning sex or drug use were asked. They sought to resolve the inconsistency in reporting by using different interview techniques for men and women; by manipulating some questions (using open and closed questions and questions permitting and restricting full reporting); and by computer-assisted self-administration. In combination, these changes led to decreases in the number of partners reported by men and increases in the number reported by women, eliminating discrepancies. Presser and Stinson (1998) reached similar findings in their study of religious attendance, and concluded that some kind of social desirability pressure was associated with interviews. Conventional survey participants substantially over-reported their attendance as compared to church attendance counts. The divergence was corrected using self-administration and time-use surveys.

Case I₂: Inconclusive results due to sample size

Gerber *et al* (2001) demonstrate that the strength of reported effects depends on the employed sample size. For larger samples even small coefficients are reported due to their significance, whereas in small samples models do not hold. This means that larger-N studies may suggest stronger effects than exist in reality and that, as Sigelman (1999) argues, studies

that fail to find a significant effect may largely be those with inadequate sample size. On the one hand, samples that are too small may fail to produce significant results. On the other hand, when the sample size becomes large enough (almost) every variation becomes statistically significant.

Case I₃: Inconclusive results due to data analysis technique

It has been widely debated whether economic globalisation leads to increases or decreases in social expenditure. The efficiency hypothesis states that increased internationalisation induces downward pressure on government spending, while the compensation hypothesis claims that increasing factor mobility leads to higher demand for social security and therefore to increased spending levels. Empirical research on the relationship between globalisation and the welfare state has led to contradictory results. Kittel and Winner (2005) suggest that this contradictory evidence might result from different data analysis techniques: a mere technical change, for example, from a simple pooled OLS regression to a two-way fixed-effects specification, leads to a change in the influence of the dependency ratio (defined as the share of citizens aged above 60 years and below 19 years) and in the share of imports from low-wage countries on total government expenditure, from a highly significant negative to an equally significant positive effect. Formerly significant positive effects of trade openness and partisan composition of government then become insignificant.

Prognosis. In the above cases, publication of negative results and discussion of their likely causes led to greater awareness of the effects of data-gathering techniques, sample size and analytic methods on findings.

Non-Results

Non-results are results that at first glance say nothing. A finding may be termed a non-result when the independent variable in question, contrary to prior expectations, is not significant. Here, 'significant' refers not only to statistical significance strictly defined, but more generally to the ability to confirm or reject a hypothesis or assumption. This might happen in a case study in which a phenomenon previously considered to be influential does not play a role at all, or in a content analysis of interviews in which no general pattern fitting or contradicting the original research question emerges. Non-results may also be due to the omission of important variables. In these cases results may look 'right' at first, but not when replicated or compared to other studies.

The problem with non-results is that of 'proving' that an examined variable has no influence on the phenomenon to be explained or that there is no traceable pattern in the material analysed. In statistical terms, this would mean that our analyses rendered evidence supporting the null hypothesis. However, as we do not know the probability of being mistaken in accepting the null hypothesis, we can never take it for granted.

A non-result might lead us to reformulate the hypothesis, for example, to focus on another aspect of the examined phenomenon – as non-results may be due to the fact that the indicator for this phenomenon is too crude. It could also lead to a strategy of disaggregation: disaggregating the dependent or independent variables into more clearly defined ones, or disaggregating the sample by switching from large-N research to closer examination of a few crucial cases.

These strategies are often beyond the scope of actual research projects and therefore may not be pursued further. Yet it might be useful to the scientific community to know that a research path

has been followed up to a certain point and that, if one wishes to explore the matter again, the research design may have to be adjusted.

Case N₁: Non-results due to omission of important control variables

Assume a (hypothetical) study testing the effectiveness of financial incentives in fostering employment activation among social assistance recipients. The treatment group receives a back-to-work bonus; the control group does not. The outcome is that the treatment group's activation rate does not differ from the control group's. This failure to find a correlation between the hypothesised independent and dependent variables is a non-result if the similar outcomes are due to additional characteristics of the treatment and control group participants, such as higher job-search intensity in the control group or less need of a job among the treatment group due to working partners. If treatment and control groups vary in characteristics that are important for outcomes and these characteristics are not controlled for in a study, then negative results that are actually non-results may emerge.

Case N₂: Non-results due to lack of significant effects

A hypothesis has been formulated that fits prior research in the democratisation literature. The assumption is that a positive relationship exists between absolute spending on external assistance to non-governmental organisations in post-communist countries and patterns of democratic transition and consolidation in those countries. Although the analysis is based on the best available data and measurement and the methods used are sound, no significant effects emerge in the analysis. Since the results do not advance the hypothesis the researcher is testing, she might discard them.

'As long as there is a concise explanation of why the old studies may need updating... then we are not dealing with negative results.'

Prognosis. Were the above results to be published, they might save other researchers from travelling needlessly down the same paths.

Confutative Results

Confutative results clearly contradict established paradigms or stylised facts. These results may emerge as a by-product of a research programme, or from explicit questioning and replication of existing studies.

In the case of replication, previous results are reanalysed, possibly demonstrating that a third variable's influence has been neglected, that biased data were used or that results were determined by outliers or influential cases. The replication then shows that the results of the older study are not robust (or that a spurious correlation has been reported). This does not necessarily disqualify the earlier study as mistaken. Old studies are updated all the time. At any given time, competing explanations for an observation and competing theories for a phenomenon coexist. It is less usual that such competing theories and explanations come to inform each other.

As long as there is a concise explanation of why the old studies may need updating (e.g. bad data quality or new available methods) and the new analysis generates a convincing result, then we are not dealing with negative results. When, moreover, a new, alternative explanation is identified and demonstrated, this 'positive' result might fit into

an ordinary journal. In other cases, however, either in a replication study or by chance and as a by-product of research on another topic, we may find empirical evidence that contradicts current theoretic tenets but that fails to lead to new theoretic developments. Sound empirical evidence that questions established paradigms, even if no alternative theoretic framework is offered, is worth presenting to the scientific community, since it might reorient research.

Case C₁: Results that contradict current theories and that are not easily explicable

Assume that all prior evidence on monetary policy supports the hypothesis that independent central banks maintain price stability more effectively than central governments do. As a by-product of a research project using good data and sound methodology, a researcher finds that contrary to existing evidence in other countries, the exercise of central government authority in his own country – counter to the independence hypothesis – seems to have been even more effective in controlling inflation than was the central bank in its period of greatest independence. The researcher finds no compelling explanation for this phenomenon and discards the finding. Had he published it, he might have incited other researchers to investigate the phenomenon, which might be explained by specific institutional constellations in the country under observation.

Case C₂: Confutative result due to the 'wrong' independent variable being significant

A carefully formulated hypothesis states that conservative parties in power promote pension reform under certain conditions. A researcher using good quality data and sophisticated methods finds no significant influence of party dominance on proclivity to reform, but instead finds a

positive tendency of parliamentary systems, rather than presidential ones, to undertake reform (regardless of the ideology of the party in power). As the finding contradicts her hypothesis, the researcher might discard the finding or shift focus to another policy field.

Case C₃: Replication with more elaborate data analysis that contradicts authoritative findings

Suppose a (now-famous) researcher has published a groundbreaking study that generated a well supported but totally new result in her field. In subsequent years the result of her study has frequently been used as the basis of other studies. A doctoral student replicating the study using a more sophisticated method obtains a result different from the original one yet logically more convincing. The popularity of the earlier result may dispose the neophyte to view his own result as unpublishable. Yet perhaps, if widely discussed and further advanced, the new evidence might supplant the old result.

Prognosis. Were the above results to be published, they might lead researchers to reformulate hypotheses generally accepted in the literature.

Ersatz Results

Ersatz or 'theory-free' results fail to fit the theory their finders happened to be interested in, or any existing theory at all. Like caffeine-free cola or alcohol-free beer, they seem to be something we were looking for, but not quite. Clearly, some results have emerged from the research process, either a statistically significant association or a systematic pattern in the material analysed. However, these patterns appear to lack context and do not relate to any theoretic expectations or assumptions.

The question is how to judge whether these (unintended) results are just

'accidents' or really meaningful.² They may of course be spurious correlations – associations that emerge because an underlying explanatory variable has been neglected. However, they may also indicate interrelationships between phenomena that had previously been ignored. It might therefore be worthwhile to explore the connections indicated by such results further, and to consider alternative theoretic contexts.

Case E₁: Ersatz result due to omission of important control variables

Let us say a study examines the health of urban and rural populations in developing countries. Data analysis reveals a new positive correlation between intensity of pollution and frequency of allergies: pollution and allergies are both more intense in urban areas, a result that had not been expected or predicted. The researchers, lacking an adequate theoretic context to interpret the findings, might conclude that pollution influences the distribution of allergies in an important way, when in fact the difference between urban and rural populations had only been due to a detection phenomenon. The rural population, lacking access to doctors, was less likely to be diagnosed with allergies. Due to the omission of an important variable (medical access), a correlation was found where actually there was none.

Case E₂: Ersatz result that is out of context, yet suggestive

Imagine a study investigating the propensity toward conflict of neighbouring states. The aim is to test the hypothesis that pairs of states with lower average domestic political regime duration are more prone to dyadic conflict. Data analysis indeed supports this hypothesis. Additionally, a strong negative correlation emerges between propensity toward conflict and the extent to which state pairs' shared borders are mountainous. Since

this result does not appear to fit existing evidence or theory, the researcher might simply discard it and concentrate on the evidence that confirms her hypothesis.

Prognosis. Were the above results to be published and discussed within an enlarged theoretic context, they might open new paths of research.

WHERE DO NEGATIVE RESULTS COME FROM?

Results such as the above become 'negative' due to bias, intended or unintended, on the part of the researcher himself or in the research community. We identify three interrelated types of bias in social science research.

Significance bias confounds statistical and substantive significance by awarding priority to statistically significant findings (Boruch, 2006; Gelman and Stern, 2005), and may operate both at the level of the individual investigator and at the level of peer review. It may describe not only a tendency to favour results that pass conventional tests of significance such as the five-per cent rule and to discard those falling even marginally below this threshold, but also the pressure felt by many researchers to produce such results.

Confirmatory bias describes a researcher's tendency to give priority to findings consistent with established theory and/or confirming her own hypothesis. It is particularly relevant to the researcher's own classification of scientific results as positive or negative. Mahoney (1977), in adducing some of the only available empirical evidence of confirmatory bias, defines it as the tendency to emphasise experiences that support one's views and to ignore or discredit those that do not. Such bias may condition the judgement of one's own or others' research. Since views are often influenced by findings already published, the danger is that

new findings that do not fit mainstream theory or the reference framework a certain discipline has generated will be discarded. MacCoun (1998) argues that once a particular theory about the world becomes widely accepted, it filters attention to and interpretation of incoming data. A very practical reason underlying the tendency to reproduce and verify results that are already established and acknowledged is the fact that revolutionary findings may be more difficult to publish than incremental ones and are more prone to being criticised or dismissed. Confirmatory bias may remain undetected because researchers are not aware of it or because objective criteria for measuring its influence are lacking.

Publication bias, on the other hand, reflects a preference of editors, reviewers and readers (and therefore authors) for results that clearly 'say' something (either confirming or disconfirming prior findings), and that do so forthrightly. New data, new theory, innovative use of existing methods, strong arguments based on evidence that is clearly indicative and carefully measured – all these are elements of the formula for bringing scientific results to market. It is not a bad formula for ensuring the quality of published findings. The question is whether or not it works too well: are useful results, and the information they contain, inappropriately excluded from scholarly discussion?

What can and ought to be done to limit or eliminate bias in social science research? According to MacCoun, some corrective strategies are already built into scientific practice through methodological training and professional socialisation, peer review, replication of prior work, competition of theories and meta-analysis. Yet many of these strategies are prone to bias themselves. Mahoney (1977), Fölster (1995) and Travis and Collins (1991) question the reliability of the peer-review process and assert that

reviewers tend to evaluate studies more favourably when they support their own views or conform to the current theoretic mainstream. Mahoney draws the conclusion from his experimental study that there is an apparent and counterproductive prejudice against 'negative' or disconfirming results. In their quasi-experimental study of methodological issues in meta-analysis, Bryant and Wortman (1984) indicate concern that studies accepted for meta-analysis might conform more to the investigator's expectations and beliefs than rejected studies. Likely bias against results that are not clearly disconfirming but simply 'fuzzy' has been still less studied. Some sources of bias can be avoided if methodological standards are rigorously applied; if research strategies and paths taken are clearly disclosed; and if diverse hypotheses are taken into account. Furthermore, as Bryant and Wortman suggest, bias in research could also be uncovered and countered if data were made available more often.

SELECTING GOOD NEGATIVE RESULTS

The question we have been asked by colleagues when discussing the possibility of collecting and publishing negative social scientific research results has not been 'what are they good for?' (as research instruments), but rather 'how are you going to establish selection criteria?' Both questions will ultimately be answered by the progress of the project of publishing negative results itself.

The technical standards for selecting good negative results are familiar ones: logical consistency, methodological correctness, replicability and engagement with relevant phenomena of social and political life, all sufficient to pass double-blind review by scholars sensitive to the challenges of a new form of publication.

There is indeed ambiguity in defining specific quality standards for selecting negative results, because they are essentially results that do not fit the conventional constraints of their particular research field. Making such results widely available through specialised forms of publication depends crucially on separating methodological mistakes or empirical flukes from truly unexpected and interesting insights into new or previously explored research territory. Clearly, this task will be no less an ongoing challenge for negative results than it is for 'positive' ones.

Inspired by frequently cited procedures for assessing measurement validity (Adcock and Collier, 2001), we define three criteria that may not be fulfilled by all negative results on their first appearance as by-products of other research, but which, if they are met, ensure those results' relevance to a wider research community: content, consistency and connectivity.

Content

Negative results may be unused or discarded results. They may be inconclusive, or contradict current theoretical assumptions, or emerge as secondary conclusions not relevant to the main topic of research. Thus they may *not* have been substantively interesting in the context of their emergence. To be useful negative results in a wider scientific context, however, they must be substantively interesting for other potential research questions.

Consistency

Negative results are generally unexpected in that they may contain contradictory elements, or counter established paradigms, or fail to shed light on initial assumptions. For these reasons they often appear to be implausible or illogical. To qualify as good negative results, however, they must (in spite of their apparent

'The technical standards for selecting good negative results are familiar ones.'

or actual implausibility) derive from a rigorous, transparent and appropriate research process. Transparent documentation and replicability are important to demonstrate that such results are not false observations, but rather plausible and consistent (if as yet inexplicable) findings.

Connectivity

Negative results often contradict researchers' primary expectations. Established hypotheses are questioned or refuted, or the new result poses a puzzle that existing theory cannot solve. Yet such results must not appear from nowhere. The research must begin with clearly defined goals and some connection to or motivation by previous work in the field.

One might object to the requirement of connectivity to prior research that it obstructs radical innovation and approaches not building on earlier findings. If all research must be classified as replication, modification or amplification of previous results, then the reporting of new phenomena, observations and interpretations is indeed impeded. The same selection problem, within conventional scientific peer review, has been criticised by many. For these reasons, the criterion of connectivity may be interpreted less strictly than the other two.

When reporting a negative result, the researcher may not be required to discuss previous findings extensively or to explain in full detail how her own research aims to improve them. One might, however, outline the original research question, why it

was of interest, and the context in which the negative finding arose. The specific reasons for considering the result 'negative' (implying also a statement of its relationship to existing findings) and its potential further use might be detailed.

WHY PUBLISH NEGATIVE RESULTS?

Publishing and discussing negative results might aid social scientists in refining theory or methods. Their dissemination could enhance quality control, incite wider discussion of the relationship between theory, methods and evidence, and lead to reinterpretation or contextualisation of specific empirical phenomena or even to the rethinking and reframing of theories. While negative results journals have begun to appear in the 'hard' and clinical sciences, to our knowledge this essay marks the first time that a journal of negative results has been proposed in the social sciences, and proposes the first systematic effort within the social sciences to recover the useful information lost through selective reporting of results.³

Negative results have been published, but not nearly often enough! Literature reviews record the moment and manner of changes in mainstream theories, and old research is brought back all the time. Such changes of perspective do not cause preceding results to be discarded as mistakes, and this is one reason why the word 'mistake' should not be associated with negative results. The findings that cause such shifts might begin their respective careers as negative results and could prove to be some of the most valuable research. If the heightened burden of proof required of such findings has inhibited their exposure, a designated forum for their dissemination may be required.

Unresolved puzzles may be the most recognisable species of negative results

because they are sometimes published as free-standing findings that subsequently engender their own line of research. For example, the *Journal of Economic Perspectives* describes the contents of its 'Anomalies' series as follows: 'An empirical result qualifies as an Anomaly if it is difficult to 'rationalize,' or if implausible assumptions are necessary to explain it within the paradigm' (Thaler, 1987–2001).

Other places where negative results may be found include the long footnotes of published articles that explore alternative hypotheses and present additional evidence; the concluding lines of articles and the last chapters of books that present partial conclusions that might inform future research – ideas that have emerged from the same heuristic process as the main work, but that have not been ranked among its current main results; and the secondary conclusions that emerge from research programmes (not only in empirical research) that were not sought by the programme or pursued by the team that produced the observation – ancillary findings sometimes mentioned in books and articles, that would never be included in an abstract.

All of these may nonetheless unmistakably be results, recognisable by their sound methodology and potential to inform further research. They may have little precedent to build upon. They may become the focus of the research projects from which they issued, or they may appear 'weak' and 'out of context' in relation to other findings. While negative results ought to have the potential to inform exhaustive research projects, the possibility of publishing a broader range of results might enable researchers to embark on particularly tempting investigations knowing that strong enough evidence to reach a definitive conclusion may not be at hand or lies just beyond economic constraints. Given the chance,

negative results may define a new perspective on current 'positive' ones.

Colleagues' responses to the proposal to collect and publish negative results have ranged from 'Finally a format for publishing that study I did' to 'A good researcher always finds a way to publish his findings'. It is as yet uncertain how many negative results are out there. We can, however, identify three distinct poles in the social science debate on the publication of results, both positive and negative. These are 'too much', 'not enough' and 'just right'.

The Publication Proliferation Hypothesis posits that too much social science research is already published: There are too many journals, filled with research of at times suspect quality. Anything that can be published already has been; the file-drawer problem is a chimera; 'truth will out.'

The Biased Reporting Hypothesis brackets concerns about the volume of trees, pixels or bandwidth destroyed by publication, focusing instead on the content of what is published. It conjectures that quantity is not the issue: what gets published is systematically, and perhaps inappropriately, selected from the available field of findings that *ought* (in an ideal research community) to be published.

The Efficient Discipline Hypothesis asserts that we already live in an ideal research community, or someplace close to it. There is indeed selection, and it works. What ought to be published, *is* published, and what is unfit is filtered out by peer review, or by researchers' own incentives to discard useless findings. Enough is published for science to advance, and advance it does.

In our search for evidence to confirm or disconfirm the hypothesis that (despite the vast quantity of social science research already published) selective reporting leads to a great deal of potentially useful information being lost, existing

'The Biased Reporting Hypothesis brackets concerns about the volume of trees, pixels or bandwidth destroyed by publication, focusing instead on the content of what is published.'

literature could help only so much: what is systematically excluded from print hardly shows up in a literature review! We went slightly further afield by conducting an online survey of approximately 175 social scientists in spring 2006.⁴ Respondents' reactions to the proposal to select and disseminate negative results reflect attitudes present in each of the three arguments: the majority contends that publication standards are the main criterion by which research results are defined as 'positive', and that publication bias (i.e. the application of publication criteria to the selection of research interests and the reporting of results) is enforced primarily by researchers themselves in the course of the research process. The difficulty of reporting 'unpublishable' negative results in citable, peer-reviewed form is widely recognised, as is the potential impact of such an enterprise should it succeed: respondents expect that disseminating negative results alongside 'positive' ones would fuel advancement in theory and methodology, expose the limits of specific analytic methods, and disseminate empirical evidence that does not conform to theoretic expectations.

The survey's conclusions echo longstanding concerns about how social scientific research is validated and used. Refraining from exposing 'outlier' results and those results that cannot easily be policed within one subfield alone may (even if unwittingly) promote a caricature

of science as a binary procedure in which methods are expected to 'clinch' evidence and deliver certainty (Cartwright, 2006). In reality, multiple results are possible, and a given result may be arrived at via more than one path. Multiple interpretations can be 'fitted' to a set of qualitative observations, just as variables are included, excluded and resized to improve the fit of a quantitative model. Why not report this variety in procedure, including those steps that may not have led to a preferred result?

A NEW JOURNAL

We introduce the *Journal of Spurious Correlations: Qualitative and Quantitative Results in the Social Sciences (JSpurC)*⁵ as a forum for increasing transparency in research (Begley, 2006; Breslow, 2006). The journal was established by a group of social scientists in Europe and the US to provide a legitimate venue for exploring pure and applied methodological questions in the social sciences in the company of colleagues. While a number of the present organisers are political scientists, such an initiative may be relevant to other social science disciplines as well, and to a range of methodological approaches beyond the 'quantitative'.

The journal was launched at the General Conference of the European Consortium for Political Research (ECPR) in Budapest in September 2005. It is an affiliated project of the Research Committee on Logic and Methodology of the International Sociological Association. It draws participating editors, advisory board members and contributors from around the world. In August 2006, the journal editors organised the short course *Rethinking Publication* at the Annual Meeting of the American Political Science Association in Philadelphia. In December 2006, the journal was selected by the

New York Times as one of the noteworthy ideas of the year (Skloot, 2006).

Other fields are now developing initiatives like this, particularly in the 'hard' sciences. Some of these are quite recent. In social science disciplines, as in the hard sciences, airing and discussing researchers' 'mistakes' could enhance quality control and community building. An initiative to recoup and recycle lost information could also catalyse (much-needed) wider discussion of the relationship between theory, method and evidence in research, and of how scholarly disciplines should handle their 'mistakes'.

In the physical and natural sciences, efforts to rescue and air negative results are now coming into their own. New journals are answering calls for such fora sounded in *Nature* and the *New Scientist* (Kotze *et al*, 2004). In the clinical and applied sciences (medicine, pharmacology, clinical psychology, cognitive and computer science and software engineering), the value of publishing negative results has spurred the creation of new journals and online trial and data registries and the expansion of existing ones.

While social science methods differ from those of the more experimental and applied sciences, problems of the censoring of published results and impacts of this censoring are similar. In the social and policy sciences, *Political Analysis*, the journal of the Society for Political Methodology of the American Political Science Association, includes a section on 'Replications and Extensions' of previously published work. *Empirical Economics* and the *Journal of Applied Econometrics* also include sections replicating prior studies. Such activities are important to the advancement of social science theory and methods, but are distinct from what *JSpurC* will do.

Rather than replication or review of prior publications, *JSpurC* will focus on the publication of original negative

results. The journal's remit is not limited to the publication of spurious regression results! *JSpurC* embraces all manner of findings and submissions on the frontier of social scientific research that advance its mission and that might not find a place within existing social science journals. These results are peer reviewed and vetted for research quality and scientific and heuristic value. While the journal's mission resonates with recent trends in the natural sciences, it does not aim to remake political or social science in the image of the other sciences.

Negative results will be presented in articles of approximately 2,500 words, without the extensive theory-building or literature review found in articles reporting 'positive' results – as many researchers have some negative results to share, but few may be inclined to make these the basis of an overly long article. Articles in the negative results series will be

'Disseminating negative results might show what did not work; augment our knowledge of what is not there; save other researchers from wasting time trying operations that have already failed...'

accompanied by expert commentary on both substance and method, and by author response when appropriate. The journal is less interested in results that have been rejected by other journals, and more interested in those that have not even been submitted.

Dissemination of negative results may inform ongoing discussion of empirical, methodological and theoretic topics in



social science research. In addition to publishing negative results, the journal will include substantive articles and editorials considering, *inter alia*, questions of how and in what way the concept of negative results applies to work generated within non-quantitative or non-positivist paradigms; what roles researcher expectations and the histories of our disciplines play in determining how research results are interpreted; whether or not choice of methods is also choice of results; and whether or not so-called negative results, in yet another instance of the social sciences aping the hard sciences, only represent a symptom of the social sciences' lack of a unified epistemology of their own.

What could social scientists learn from reporting a wider range of results?

Perhaps that negative results are not so 'negative' after all; perhaps that 'positive' results are not so positive. Disseminating negative results might show what did not work; augment our knowledge of what is *not* there; save other researchers from wasting time trying operations that have already failed; and demonstrate the limits and upper bounds of previously reported results. Most important, it might help to increase the speed and transparency with which new ideas are introduced, tested and filtered in social science research.

We invite you to dig into your file drawer, hard drive or wastebasket for the negative, spurious, anomalous, outlier or otherwise 'unpublishable' results that are a necessary by-product of social science research. Send them to us!

Notes

1 We draw particular inspiration from the typologisation of negative results in computer science in Prechelt (1997).

2 Debates on grounded theory, data mining and exploratory data analysis (Begg and Berlin, 1988; Simon, 1988) must be reserved for a future article.

3 We are indebted to such pioneering efforts in medical, computer and other sciences such as the *Journal of Negative Results in Ecology & Evolutionary Biology* (www.jnr-eeb.org), the *Journal of Negative Results in Biomedicine* (www.jnr-bm.com) and the *Forum for Negative Results of the Journal of Universal Computer Science* (www.jucs.org).

4 A summary of the survey results is available on the website www.jspurc.org and from the authors upon request.

5 www.jspurc.org.

References

- Adcock, R. and Collier, D. (2001) 'Measurement validity: a shared standard for qualitative and quantitative research', *American Political Science Review* 95: 529–546.
- Baert, P. (2005) 'Towards a pragmatist-inspired philosophy of social science', *Acta Sociologica* 48(3): 191–203.
- Bartels, L.M. (1997) 'Specification uncertainty and model averaging', *American Journal of Political Science* 41: 641–674.
- Beck, N. and Katz, J.N. (1995) 'What to do (and not to do) with time-series cross-section data', *American Political Science Review* 89: 634–647.
- Beck, N. and Katz, J.N. (1996) 'Nuisance vs. substance: specifying and estimating time-series cross-section models', *Political Analysis* 6: 1–34.
- Begg, C.B. and Berlin, J.A. (1988) 'Publication bias: a problem in interpreting medical data', *Journal of the Royal Statistical Society* 151(Series A): 419–463.
- Begley, S. (2006) 'New journals bet 'negative results' save time, money'', *Wall Street Journal*, 15 September.
- Boruch, R. (2006) 'The null hypothesis is not called that for nothing: statistical tests in randomized experiments', *Journal of Experimental Criminology*, forthcoming.

About the Authors

David Lehrer is a doctoral student at the University of Helsinki and a lecturer at Humboldt University Berlin. His research on the comparative political economy of post-communism has been supported by the US State Department, the German Marshall Fund of the United States and the Harvard Center for International Development.

Janine Leschke is a researcher at the European Trade Union Institute in Brussels. She wrote her PhD thesis on flexible employment and social insurance coverage in the Labour Market Policy and Employment unit of the Social Science Research Centre, Berlin (WZB). Her research interests include comparative welfare state analysis, labour market and social policies, and quantitative analysis.

Stefan Lhachimi is a Ph.D. candidate at the Max Planck Institute for Demographic Research, Germany. He holds a Master in Public Policy from the Terry Sanford Institute of Public Policy, Duke University, a Master in Political Science from the Free University Berlin and a Bachelor in Statistics from Humboldt University Berlin.

Ana Vasiliu is a consultant to private enterprises in transition economies and on international development projects. She studied business administration in Romania and economics at Cambridge and at Brown University, and has been a visiting fellow of the Center for Business and Government of the Kennedy School of Government, US.

Brigitte Weiffen is a Ph.D. candidate in sociology at the University of Bonn, and is currently working as a research fellow at the Centre for Peace and Conflict Studies, University of Tübingen. Her research interests include comparative democratisation, international organisations, Latin American politics and research methods.