Research Statement

Generated August 27, 2026

Lars Vilhuber — Cornell University, ILR School and Department of Economics

My research program sits at the intersection of labor economics and the infrastructure that makes empirical economics possible and transparent. Since my Ph.D., it has developed along three connected lines: (i) the measurement of labor market dynamics using linked employer-employee data; (ii) the protection of confidentiality in the data that make such measurement possible, in labor and other empirical economics; and (iii) the reproducibility and transparency of empirical research, in economics and beyond. These are not discrete agendas. They are connected to each other, each arising out of a practical obstacle encountered while pursuing one research agenda, leading to the search for a solution. I firmly believe that scientific research needs to be transparent and reproducible, and have practiced what I believe since my Ph.D. I also strongly believe that teamwork is crucial for robust empirical science, and have found collaborators in many adjacent (and not so adjacent) disciplines: computer and information scientists, statisticians, and social scientists in other disciplines. I have contributed to the peer-reviewed research literature in each of the areas I have worked in, but have also endeavored to create persistent public infrastructure, institutional change, and educational improvements.

Labor market dynamics and linked employer-employee data

The core theme in my labor economics research is worker mobility. My dissertation work was on worker mobility and employer-provided training in Germany and the United States (Vilhuber 1999, 2001).[1] My post-graduation work at York University and the U.S. Census Bureau looked into administrative data on workers, linked to firms, social programs, and each other. As one of the original research employees at the U.S. Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program, I helped research and then build the infrastructure for continuous measurement of gross employment and job flows, what became the Quarterly Workforce Indicators (Abowd, Stephens, Vilhuber et al. 2009; Abowd and Vilhuber, *Journal of Econometrics* 2011) and OnTheMap (Machanavajjhala et al, 2008), relying fundamentally on the interconnection of workers in the workplace. My research focused on job displacement, worker flows, and mobility. Contributions include the role of employer characteristics in escaping low-wage work (Holzer, Lane, and Vilhuber, *ILR Review* 2004), the composition of worker flows before displacement (Lengermann and Vilhuber 2002), inside knowledge and re-employment wages in a simple search model (Bowlus and Vilhuber, 2002), the link between human capital, mass layoffs, and firm deaths (Abowd, McKinney, and Vilhuber 2009), and the labor-market consequences of the housing price collapse (Abowd and Vilhuber, *AEA P&P* 2012). A recurrent methodological concern in this work is that measurement choices are not innocuous: coding errors in personal identifiers materially affect published economic statistics (Abowd and Vilhuber, *Journal of Business & Economic Statistics* 2005), uncertainty about supposedly deterministic "local labor market" definitions can change empirical conclusions (Foote, Kutzbach, and Vilhuber, *Applied Economics* 2021). I have also occasionally looked at some more micro-level questions related to workers in the economy, such as procedural justice in hiring (Cloutier and Vilhuber, *Journal of Managerial Psychology* 2008), or the role of criminal records in employment-related outcomes (Wells et al, 2020).

[1] For all references, see https://www.larsvilhuber.org/publications/.

Confidentiality protection of detailed economic data

Working with employer-employee microdata, and attempting to make the information contained therein broadly available to economics, made the confidentiality constraint impossible to ignore. I have treated it as a multi-faceted problem to solve: publish more data by innovating in the statistical methods to protect privacy, and broaden access to secure facilities when that is not effective. In various collaborations with John Abowd and other coauthors, I developed and evaluated various methods that often combined both paths of access. We adopted and scaled a novel noise infusion approach for the Quarterly Workforce Indicators (Abowd et al, 2009). We worked with computer scientists and geographers on the first ever translation of the hitherto theoretical concept of differential privacy into a real world statistical data product (Machanavajjhala et al, 2008). We showed in McKinney et al. (*JSSAM* 2020) that the impact on analysis done with such protected data is generally small, and degrades precisely where we would expect protection to be most needed: for small groups.

When access to detailed public data is insufficient to answer research questions, researchers need secured access to microdata. Curating, documenting, and "publishing" snapshots of the QWI infrastructure in the Census Bureau's Federal Statistical Research Data Center network was a natural next step (McKinney and Vilhuber, 2011; Vilhuber and McKinney, 2014; Vilhuber, 2018). The updating and curation are maintained to this day by staff at the Census Bureau.

The use of synthetic data – data produced to mimic the statistical properties of confidential data – was a different path. I have explored variants of synthetic data in various contexts (Drechsler and Vilhuber, *Statistical Journal of the IAOS* 2014; Miranda and Vilhuber, *SJIAOS* 2014; Pistner, Slavković, and Vilhuber 2018), but also built, operated, and obtained funding for the Synthetic Data Server, which for a decade gave outside researchers a usable path to confidential Census Bureau data.

A second strand quantifies the trade-off itself. With Abowd, Schmutte, and Sexton, I argued that privacy and statistical accuracy are public goods that markets and agencies will under-supply (*AEA P&P* 2019), documented the utility cost of formal privacy for national employer-employee statistics (Haney et al., 2017), and considered the statistical tradeoffs due to noise infusion (McKinney et al., *JSSAM* 2020). The empirical case for these concerns was made concrete in the simulated reconstruction and reidentification attack on the 2010 U.S. Census (Abowd et al., *Harvard Data Science Review* 2025), which was awarded the 2026 Caspar Bowden PET Award for "Outstanding Research in Privacy Enhancing Technologies". Recent work extends differential privacy to the setting of randomized controlled trials, seeking to find practical means by which researcher teams can protect the data they collect through often expensive RCTs, while at the same time protecting the privacy of their respondents better than they currently do, using traditional techniques (Webb et al, 2026). Importantly, we also assess when such efforts fail.

I have also written and contributed to synthesis pieces, which are needed to bring new methods to the discipline. This includes working papers giving an overview of methods for firm-level data (2013), conference papers arguing for involvement of economists in debates about privacy protection (*AEA P&P* 2019), chapters on disclosure limitation in linked data (2021), and discussing non-statistical methods of protecting confidential data (2024). We produced introductory teaching materials on formal privacy for economists (2019). I also relaunched and then served as managing editor of the *Journal of Privacy and Confidentiality*, from 2018 to 2024.

Reproducibility and the credibility of empirical economics

Confidential data and reproducible research are often perceived to be in tension, and resolving that tension has become a large part of my program. I first documented that in economics, the glass is half-empty - or possibly half-full. In our study of the reproducibility of economics research (Herbert et al., *Canadian Journal of Economics* 2024), we found that a substantial share of published articles could not be reproduced from their deposited materials. But a significant fraction was able to be reproduced, and part of the problem was identified

as data not being accessible. This analytical work has continued in large collaborative efforts (Brodeur et al., *Nature* 2026, Brodeur et al., *PNAS* 2026), though some of the nuances are lost as it uses a broader brush (and has more co-authors).

The original findings of our study suggested there was room for improvement in the discipline. My experience working with confidential data taught me that it was possible to make progress even in those difficult environments. In 2018, I was appointed as the AEA's inaugural Data Editor, and was given the opportunity to act on those observations. Building on the experience from our study, I designed and continue to run the verification process behind the Association's Data and Code Availability Policy. Under my leadership, the LDI Replication Lab, on behalf of the AEA, has evaluated more than 3,000 manuscripts and their replication packages, shared more than 4,500 reports on potential improvements of those packages with authors, and through that mechanism, reached more than 5,000 authors. Informal evidence from various evaluations suggests that this has made a material impact on the average quality of the materials published, both in the AEA's journals and more broadly.

Since 2018, I have convened a now monthly meeting of the most active data and reproducibility editors in economics, management, and finance. Together, we have developed the standards that now define how an economist prepares transparent research: the template README for social science replication packages (Vilhuber et al. 2020, 2022), and the common model policy used by those and other journals (Koren et al., 2022). Complementary technical work explored ways to provide more structured transparency to the journal case, through better ways to reference external linked objects (Lagoze and Vilhuber, *IJDC* 2021), through better ways to make research with confidential data reproducible (Vilhuber, *HDSR* 2025). Some of these strands have combined in the NSF-funded and now Sloan-funded Transparency Certified (TRACE) project, which developed and implemented methods by which computational workflows that cannot be easily re-executed can nevertheless be transparently published and certified by trusted institutions (in progress, but see Li et al., *IJDC* 2025).

My reflective and agenda-setting work appears in *HDSR* (2020), the *Journal of Econometrics* (2023), *Revue économique* in France (2025), and in *Perspektiven der Wirtschaftspolitik* in Germany (2026). I edit the "Reinforcing Reproducibility and Replicability" column at *HDSR*, have convened webinar series on the broader topic of reproducibility in various social sciences (Vilhuber et al., *HDSR* 2023, and other publications in that issue).

Infrastructure, training, and institutions as research output

I consider durable infrastructure and trained people as first-class outputs. The data pipeline I prepared as a post-doc at York University persisted for several years after my departure. The statistical pipeline behind the Quarterly Workforce Indicators is still active, more than 20 years after its inception. Education and dissemination are an integral part of what I do. I have built documentation systems (CED²AR), data-access mechanisms, and convening structures such as the 2016-2017 Practical Privacy workshops and the 2022-2023 Conference on Reproducibility and Replicability in Economics and the Social Sciences webinar series. The *Handbook on Using Administrative Data for Research and Evidence-Based Policy* (Cole, Dhaliwal, Sautmann, and Vilhuber 2021) is a reference for researcher-government data partnerships. I relaunched the *Journal of Privacy and Confidentiality*, running it for six years, and ensuring a robust handover in 2024.

My educational activities do not occur in traditional semester-long course settings. The LDI Replication Lab I run has taught over 200 undergraduates and numerous graduate students and pre-docs not just how to conduct reproducibility checks, but also how academic science works. What they needed to know could be, but was not, taught in any of the other classes they may have had in economics, statistics, or computer science. The approach is documented in the *Journal of Statistics and Data Science Education* (Vilhuber et al. 2022). Through regular workshops at conferences and academic departments, I teach the cutting-edge reproducibility methods that inspection of over 3,000 replication packages has allowed me to distill into a curriculum that does not have

a regular course counterpart. My students range from graduate students to more senior economists, but also data librarians or information scientists. I give up to 25 workshops and talks every year. Every report I have written as data editor, which is a unique interaction with authors in a non-traditional way, has also served to educate the recipients on how they could do even better next time, with links to examples and tutorials throughout.

The knowledge I have gained is also conveyed through committees and convenings. I was on the board of the Canadian research data center network for two three-year terms, plus an extension, and I have been chair of the French research data system's scientific committee for twelve years. For many years, I was counselor to the Cornell branch of the US research data center network, and have contributed to others as well. My open science and transparency experience has been called upon by various working groups in Germany, France, Canada, and the US, through various National Academies panels. I was a member for four years, and chaired the American Statistical Association's Committee on Privacy and Confidentiality for another year.

Various funders have generously supported the work that I and my collaborators have done. As PI or co-PI, I have managed approximately $10 million in awards, primarily from the National Science Foundation and the Alfred P. Sloan Foundation. This includes focused research, infrastructure support, and convenings such as conferences or research networks. As one of the PIs of the NSF-Census Research Network (NCRN), we coordinated jointly with NSF the efforts of more than two dozen PIs, and disseminated the output from more than $25 million in NSF research funding. We received the American Statistical Association's 2017 Statistical Partnerships Among Academe, Industry, and Government (SPAIG) Award for this work.

Future directions

Going forward, the three strands will continue to be intertwined, with a fourth strand emerging. The work as Data Editor has given me unprecedented insight into how economists conduct empirical research, and several exciting avenues of inquiry are emerging. Not surprisingly, economists are not uniform in their approach, but we do not regularly self-inspect, and have an astonishingly small amount of data on our own production function. The wealth of methods that are distilled into replication packages provide some insight, and some data points.

Infrastructure building is ongoing. New projects aim to support the persistence of one of the key pieces of software infrastructure most economists use every day: the Statistical Software Components (SSC) archive, better known as the place where you install your Stata packages from. With support from the SSC's founder, Kit Baum, (and a Sloan grant) we will set it on a sustainable path for the next 20 years. Extending the NSF-funded TRACE project, we are working with research infrastructure institutions, in central banks, journals, health data archives around the world, to implement trusted computational artifacts, expanding the scope of transparent research when data are confidential far beyond the current status.

Finally, the original quest to better understand labor market dynamics will continue. Occupational transitions within and across firms are an important part of understanding the AI disruptions, but so are the changes in trade flows that are emerging in more recent years in the wake of pandemics, newer types of international conflict, and the changing legal landscape. I continue to be intrigued by the question of how to measure these changes, and by how the humans that are the source of the data are affected by them.