THE ECONOMICS NETWORK

Improving economics teaching
and learning for over 20 years

The effective incorporation of research in undergraduate econometrics

CONTENTS

Steve Cook, Swansea University

`s.cook at swan.ac.uk`

& Duncan Watson, University of East Anglia

`Duncan.Watson at uea.ac.uk`[1]

Edited by Peter Dawson, University of East Anglia

PUBLISHED

November 2025

https://doi.org/10.53593/m4412a

Summary

This chapter examines the integration of research into undergraduate econometrics teaching through a focus on replication and reproduction (R&R) of published studies. Rather than limiting students to the passive reception of research findings, the R&R approach fosters active engagement by requiring them to work directly with real data and methods in order to reproduce results and interrogate techniques. The discussion situates R&R within broader debates on the role of research in teaching, highlighting its potential to make scholarship more accessible and relevant to students. Practical case studies and classroom-ready materials from the Economics Network Ideas Bank are presented, illustrating how this approach can enhance learning, develop quantitative skills, and build student confidence. Finally, the chapter shows how R&R aligns with wider educational goals, including employability, self-efficacy, and sustained engagement.

1. Introduction

The inclusion of research within teaching has received much attention in higher education. This is reflected in the frequent use of ‘research-led teaching’ as both a strategic goal and a marketing tool for universities. In this chapter, we explore how research can be incorporated into the undergraduate teaching of econometrics.

While this chapter touches on a range of issues, its central theme is the promotion of replication and reproduction (R&R) as a means of incorporating research into the teaching of econometrics. Although R&R is not presented here as a novel concept, given its presence in textbooks and the availability of data from published studies in software packages[2], the chapter offers a fresh contribution by:

i. synthesising debates on the embedding of research in teaching;
ii. collating discussions on the nature of R&R;
iii. providing linked practical examples or resources; and
iv. assessing the potential pedagogical value of R&R.

One key benefit we highlight is the ability of R&R to counter a potential ‘distancing’ effect of research-led teaching where students passively view the work of *others*. In contrast to this, R&R encourages active engagement as students work directly with data and retrace the steps of published studies thereby making them participants, rather than observers, in the research process. This supports a transition from being recipients of research findings to active developers of research understanding.

Before going further, it is useful to refer to Cook et al. (2025) who propose a highly simplified approach to the teaching introductory econometrics within an R&R framework. While that study shares the present chapter’s replication-based approach, its use of simplification is not required for the present study. As we shift focus from introductory econometrics to the higher-level econometrics considered in this chapter, students’ existing knowledge can now support deeper engagement with published research and the complexities of real-world data.[3]

The chapter begins in Section 2 with a review of key literature on embedding research in teaching, outlining both its benefits and the different ways in which it can be approached. Section 3 explores the concept of R&R more fully, discussing its role in the academic research arena and its growing use as a teaching method. Section 4 then turns to practical application, presenting three econometrics case studies from the Economics Network’s Ideas Bank (Cook and Watson, 2025a, b, c) based on research published in *Oxford Bulletin of Economics and Statistics*, *Journal of Econometrics* and *Energy Economics*. Section 5 connects R&R to wider themes in educational research, including cognitive load theory, self-efficacy, employability, and the development of quantitative skills. Section 6 concludes with final reflections.

2. Incorporating research in teaching

Interest in ‘incorporating research in teaching’ (IRT) has sparked a wide-ranging literature exploring the topic from multiple angles.[4] Prominent within this literature is the Boyer Commission (1998) which advocated embedding research in undergraduate education. Subsequent work has reinforced this view, highlighting how IRT can help students develop broad, transferable skills (Brew 2013, Ruth et al. 2023, Wood 2003). A growing body of studies also point to more specific benefits, particularly around employability (Boyd et al., 2010; Brew, 2013; Bowyer and Akpinar, 2022). Others link IRT to enhanced student engagement (Boyd et al., 2010), academic performance (Parker, 2008), well-being (Walkington and Ommering, 2022) and greater self-efficacy, alongside more positive attitudes towards research (Wessels et al., 2021). These advantages are echoed in the case studies compiled by Ansell and Marshall (2017) which showcase a range of practical interventions and outcomes.

The discussion so far makes a strong case for IRT. A key issue that follows is the relationship between the research activity of instructors and the effectiveness of IRT. Numerous questions arise, including: Is it sufficient for instructors to have only a general awareness of research? Is IRT more effective when instructors are research active? Does IRT become even more beneficial if instructors draw directly on their own research? This leads naturally to consideration of the concept of pedagogical content knowledge (PCK), introduced by Shulman (1986), which refers to the integration of subject expertise and teaching skill.[5] PCK highlights the value of research for instructors: staying current with developments in the field ensures teaching content is relevant and up to date, while pedagogical expertise supports effective delivery and assessment. However, it could be argued that being ‘research aware’ alone is not enough. Several studies suggest this, pointing to the benefits of direct involvement in research and instructors drawing on their own research. For example, Clark and Hordosy (2019) find that instructors who use their own research generate greater enthusiasm, which in turn boosts student engagement. Healey et al. (2010) report that students see staff research as beneficial to their own development, especially in building research skills. Lindsay et al. (2002) similarly highlight the enhanced credibility, enthusiasm and depth that research-active lecturers bring to the classroom. Taken together, these findings point to the advantages of instructors being research-active, and support the call of Vicens and Bourne (2009) for lecturers to be ‘*…shameless in bringing your research interests into your teaching*’.

Although the benefits of IRT are well documented, this may appear surprising given the often uneasy relationship between teaching and research in higher education (Hattie and Marsh, 1996; Schapper and Mayson, 2010; Macfarlane, 2011; Bamber et al., 2023). Still, with growing evidence in its favour, attention turns to how IRT can be implemented in practice. As Brew and Mantai (2017) note, a wide range of approaches is available. One influential framework is the 2×2 model proposed by Healey and Jenkins (2009) which classifies IRT along two dimensions: whether the focus is on research content or process, and whether students are treated as an audience or active participants. This produces four modes of engagement: research-led (content, audience), research-oriented (process, audience), research-based (process, participation), and research-tutored (content, participation). Drawing upon Healey and Jenkins (2009), a summarised version of this framework is presented in Table One below. Importantly, this model highlights a key limitation of research-led teaching: students may remain passive consumers of the findings of others.

Table One: The 2x2 classification of Healey and Jenkins (2009)

	Students as audience	Students as participants
Focus on content	Research-led	Research-tutored
Focus on process	Research-oriented	Research-based

Levy and Petrulis (2012) offer a complementary 2×2 model distinguishing between inquiry into existing versus new knowledge, and between tutor- versus student-led approaches. Other scholars extend these typologies. For example, Brew and Mantai (2017) critique such frameworks and propose a 'wheel' model focused on context and learning outcomes. Continuing with alternative terminology, Ansell and Marshall (2017) prefer the term 'research-informed teaching', while Cook and Watson (2023) employ 'research-driven teaching' to reflect a more comprehensive approach. Centred on R&R, the model of Cook and Watson (2023) integrates both the content of research and the methods used in its production. Crucially, it enables students to engage as recipients of research knowledge and as active participants in research activity, while exposing them simultaneously to research content and methods.

This chapter champions the 'research-driven' model. By revisiting published empirical studies, students work with original data, to reproduce and replicate results, and to explore alternative approaches. This shifts them from passive observers to active participants in research. In doing so, students engage with all four quadrants of Healey and Jenkins' framework. The approach is illustrated in case studies by Cook and Watson (2025a, b, c), examined in Section 4. Before that, however, Section 3 takes a closer look at the concept of R&R, discussing its associated terminology, prominence in research, and adaptation as a teaching tool.

3. Replication and reproduction

The reproduction and replication of research findings have a long history in academic research, with renewed attention sparked by the widely reported 'replication crisis'. Dubbed 'repligate' by Machery and Doris (2017), numerous papers are prominent in the literature associated with this crisis (see, inter alia, Ioannidis, 2005; Open Science Collaboration, 2015; Nosek et al. 2015, 2022). As noted by Machery (2020), an increased interest in the replicability of research findings sparked by the replication crisis has spread across disciplines since these debates (see, inter alia, the following studies and their references: Persaud et al., 2024; Rode et al., 2024).

Economics is clearly one discipline where interest in R&R is prominent. While this is illustrated, for example, by the relatively recent research of Camerer et al. (2016) and Chang and Li (2017), which highlight the challenges of reproducing published findings, there is also a collection of work pre-dating this. As Dewald et al. (1986) note, calls for greater transparency go back, at least, to Frisch's (1933) editorial in the first edition of *Econometrica* where the publication of data used in research was emphasised. Notable further examples of interest in R&R include the special edition of *Oxford Economic Papers* where R&R and more general re-evaluation of previous research were presented (Hendry and Morgan, 1989; Spanos, 1989; Thomas, 1989; Wulwick, 1989) and the 'Data Storage and Evaluation Project' launched in 1982 by the *Journal of Money, Credit and Banking* (Dewald et al., 1986). This interest has continued with, inter alia, the special issue of *Energy*

Economics dedicated to replication (Tol, 2019) and the contributions to R&R in the *American Economic Review Papers and Proceedings* of 2017 (e.g., Berry et al., 2017; Chang and Li, 2017; Duvendack et al., 2017). An ongoing commitment to R&R is also apparent with, for example, the *Journal of Applied Econometrics* introducing a dedicated replication section (Pesaran, 2003) which built upon the momentum resulting from its prior creation of a data archive in 1994. In addition, an alternative perspective on R&R within Economics is provided by the *Journal of Applied Econometrics* Experiment (see Magnus and Morgan, 1997), which examines the robustness of, and variability in, empirical findings through a field experiment.

Reflecting on the R&R literature in economics, two key issues stand out as warranting further discussion.[6] The first concerns the extent to which replication exercises are actually undertaken and the factors that encourage, or discourage, this activity. While much attention has been paid to whether findings can be reproduced, less is known about the extent to which R&R is actually undertaken, with Berry et al. (2017) noting the difficulty in estimating the volume of replication studies being published. With regard to factors encouraging or discouraging R&R studies, Chang and Li (2017) explore ways to make replication in economics more feasible.

The second issue relates to the terminology surrounding R&R, which is marked by variation and inconsistency not only in economics but across disciplines. As the National Academies of Sciences, Engineering and Medicine report (NASEM, 2019) observes, the lack of standardisation regarding the use of 'replication' and 'reproduction' complicates efforts to assess research credibility. This problem is echoed in the work of Nosek and Errington (2020) and Machery (2020)– the latter suggesting that the traditional distinction between direct and conceptual replication should be dropped in favour of a 'resampling' approach.

Economics contributes significantly to this terminological variation. Barba (2018), in a review of economics papers, highlights that many authors use 'reproduce' and 'replicate' interchangeably. Clemens (2017) also notes the variation in terminology employed in relation to R&R within economics, while providing further classifications in an attempt to introduce clarity and consistency. More specifically, Clemens (2017) defines replication as including both 'verification' and 'reproduction' tests, while robustness is taken to include reanalysis and extension. This work follows Pesaran (2003) where a distinction is made between 'narrow sense' replication (reproducing original results using the same data) and 'wide sense' replication (testing findings using new data). These distinctions mirror NASEM's (2019) proposal: reproducibility involves consistent results using the same data and methods, whereas replicability involves consistent results across independent studies using new data. In short, while extensive terminology exists with a variety of terms employed, a long-standing and continuing interest in R&R within economics is apparent.

So far, our discussion of R&R has focused on its place within the research community where revisiting findings is an activity for researchers. However, R&R has also gained attention in teaching contexts, as shown in recent studies by Ball et al. (2022), Janz (2016), Stojmenovska et al. (2019) and Smith et al. (2021). Predating these more recent studies, an earlier call to employ R&R within the teaching of Economics comes from Pesaran's (2003, p.11) statement when introducing the Replication Section of the *Journal of Applied Econometrics*: '*We also hope that this* [the replication section] *will encourage students and teachers of applied econometrics to replicate the published work in the classroom*'. This classroom use of R&R has been taken a step further by Wagge et al. (2019) and the Collaborative Replications and Education Project (CREP) where replication is encouraged for assessment design.

Building on this momentum, Cook and Watson (2023) extend the conversation by proposing a three-part framework for using replication in teaching comprising ‘direct replication’, ‘step replication’, and ‘flexible replication’. These forms of replication can be summarised as below:

- Direct replication: This refers to the familiar concept of the reproduction of published findings using the original data and methods employed in a study.
- Step replication: This presents an alternative challenge, requiring students to retrace the analytical steps that may be hidden within automated software outputs. For example, software might generate results for a two-step procedure at the click of a button; in step replication, students are expected to manually execute and understand each step of that process.
- Flexible replication: This adds a further layer to the notion of replication via its use resources that allow results to be repeatedly generated and subsequently reproduced. Examples of this are provided in Cook (2016, 2019).

Together, these approaches illustrate the growing pedagogical interest in R&R and its potential to deepen student engagement with research. The next section turns to concrete examples that put these ideas into practice through case studies developed to support this chapter.

4. Case studies and resources

To demonstrate how R&R can be integrated into the teaching of undergraduate econometrics, we draw on three associated case studies available from the Economics Network Ideas Bank (Cook and Watson, 2025a, b, c). Each focuses on unit root analysis but introduces distinct extensions and teaching opportunities beyond this core topic. While these case studies have been developed alongside this chapter, a further similar case study employing R&R (Cook, 2020) is available to readers as an additional resource.

A natural starting point for R&R in the context of unit root analysis is Nelson and Plosser (1982), a seminal study that examined the orders of integration of fourteen major U.S. macroeconomic time series. The influence of this study has been substantial, prompting a wide range of follow-up studies using both the original and extended versions of the dataset. Among the most notable contributions are Perron’s (1989, 1997) analyses of structural breaks and their impact on inferences. Other contributions include work by Abadir et al. (2013), Leybourne (1995), Lu and Podivinsky (2003), Lucas (1995), Lumsdaine and Papell (1997), Phillips (1991), and Rudebusch (1992).

Our first case study (Cook and Watson, 2025a) draws upon Leybourne (1995), in which the original Nelson-Plosser data are used to introduce the maximum augmented Dickey-Fuller (ADF) test. Designed to improve test power, the maximum ADF test provides a platform for students to explore key concepts such as statistical power, higher-powered testing and the use of simulation to assess the properties of tests. In doing so, the case study goes beyond basic reproduction to open up broader discussions around test design and interpretation.

The second case study (Cook and Watson, 2025b), based on Leybourne et al. (1998), offers a different perspective on unit root testing by focusing on structural change and the issue of empirical size. However, unlike Perron’s (1989, 1997) work, which examines breaks under the alternative hypothesis and the resulting loss of test power, Leybourne et al. (1998)

investigate how structural breaks can affect unit root tests when they occur under the null. This case study also moves beyond reproduction to consider replication (using the terminology of NASEM, 2019) via the use of a revised version, or later vintage, of the data employed in the original study.[7] Alternatively expressed in the terminology of Pesaran (2003), a ‘wide-sense’ replication rather than ‘narrow-sense’ replication is undertaken. This use of updated data allows students to explore how findings can shift over time, highlighting the evolving nature of empirical research and providing a foundation for broader discussions around data revision, real-time data, measurement systems, progressive modelling, and the reliability of empirical inference (see, for example: Cook, 2008; Croushore and Stark, 2003; Egginton et al., 2002; Garratt and Vahey, 2006; Mankiw and Shapiro, 1986; Mankiw et al., 1984; Patterson and Heravi, 1991).

The third case study (Cook and Watson, 2025c) is based on the reproduction of Holmes and Otero (2019). Like the previous case studies, it involves unit root testing but adds depth by examining relationships between time series. Specifically, this case study requires students to use R&R to undertake unit root analysis of not only individual price series, but also differentials – in this case, the difference between spot and futures prices – and extend their analysis through cointegration testing using the Johansen (1988) procedure. This case study therefore supports the development of further skills by moving from the consideration of individual series to the exploration of relationships between series via unit root analysis of differentials and the use of cointegration techniques.

As a final point, a common feature of these case studies is that they do not provide the data they consider, but instead present information on its source. As such, the development of data retrieval skills, which are highlighted as important in quantitative skills training in Economics (see, for example, QAA 2023), is supported by requiring students to access the data and undertake the steps required to transfer them into relevant econometric software.

5. Pedagogical concerns

This section explores the pedagogical benefits of using R&R, building on the themes introduced in Section 2. Rather than offering an exhaustive review, the aim here is to provide an overview of key ideas drawn from the broader literature.

As highlighted by Cook et al. (2019), Cook and Watson (2023) and Cook et al. (2025), the use of R&R in teaching can assist engagement with several key themes in pedagogical research. These include: *effective* active learning (Mayer, 2004, 2021); self-efficacy (Bandura, 1978; Zahaciva et al., 2005); anxiety towards quantitative methods (Dreger and Aiken, 1957; Dowker et al., 2016); cognitive load theory (Sweller et al., 1998, 2019; van Merriënboer and Sweller, 2005); and learning models such as productive failure (Kapur, 2008, 2012, 2015; Loibl et al., 2017), impasse-driven learning (VanLehn, 2003) and the expertise reversal effect (Cooper and Sweller, 1987; Kalyuga et al., 2001, 2003; Kirschner et al., 2006; Sweller and Cooper, 1985).

R&R supports these areas by offering students a clear, structured objective – namely, the replication and reproduction of published findings. This reduces cognitive load by providing focus and scaffolding, while also supplying a concrete context through which productive failure, impasse-driven learning, and other active learning strategies can be employed. Because R&R requires critical engagement rather than passive behaviour, it encourages cognitive, rather than simply behavioural, activity and thereby supports *effective* active learning (Mayer 2004, 2021). Also, when students succeed, they not only improve their skills but also build confidence, helping to alleviate anxiety and strengthen self-efficacy.

R&R also has clear links to employability, as the replication of empirical research requires students to work with data, software, and published studies. The supports the development of skills directly aligned with the growing emphasis on data literacy and quantitative competence in social science education (MacInnes et al., 2016; Mansell, 2015). With further reports such as Quantifying the UK Data Skills Gap (DSIT & DDCMS, 2021) and Data Science in the New Economy (World Economic Forum, 2019) underscoring the increasing importance of these capabilities from an employment perspective, the benefits of R&R are further emphasised.

In the context of economics specifically, a survey of professional economists by Anand et al. (2019) identified 'Evaluating Econometric Work Done by Others' as a leading area where further support could be offered during undergraduate education. Additional training needs flagged in the same survey included 'Doing Econometrics' and 'Using Econometrics Software'. Clearly, these findings provide further support for the use of R&R, with these activities figuring prominently in the active revisiting of published empirical research.

6. Conclusion

This chapter has advanced the case for replication and reproduction (R&R) as far more than a methodological add-on: it is a catalyst for transforming undergraduate econometrics teaching. By revisiting the historical roots of R&R, engaging with the pedagogical literature, and presenting practical strategies for implementation, we have shown how R&R can turn students from passive recipients of knowledge into active participants in the research process. Its benefits extend beyond the classroom, equipping students with the confidence, quantitative expertise, and data literacy demanded in today's academic and professional landscapes. Ultimately, R&R provides a distinctive bridge between research and teaching – one that not only enriches learning but also redefines what it means to study econometrics in a research-driven environment.

Related resources

1. Cook, S. 2006. Understanding the construction and evaluation of forecast evaluation statistics using computer-based tutorial exercises. Economics Network Ideas Bank. https://doi.org/10.53593/n143a
2. Cook, S. 2019. Forecast evaluation using Theil's Inequality Coefficients. Economics Network Ideas Bank. https://doi.org/10.53593/n3168a
3. Cook, S. 2020. Unit root analysis. Economics Network Ideas Bank. https://doi.org/10.53593/n3341a
4. Cook, S. and Watson, D. 2025a. Replication and Reproduction I: Leybourne (1995, *Oxford Bulletin of Economics and Statistics*) and the maximum Dickey-Fuller test. Economics Network Ideas Bank. https://doi.org/10.53593/n4409a
5. Cook, S. and Watson, D. 2025b. Replication and Reproduction II: Leybourne *et al.* (1998, *Journal of Econometrics*) and the Dickey-Fuller test in the presence of

breaks under the null. Economics Network Ideas Bank. https://doi.org/10.53593/n4410a

6. Cook, S. and Watson, D. 2025c. Replication and Reproduction III: Holmes & Otero (2019, *Energy Economics*), unit root testing of differentials and cointegration analysis. Economics Network Ideas Bank. https://doi.org/10.53593/n4411a

References

1. Abadir, K., Caggiano, G., and Talmain, G. 2013. Nelson–Plosser revisited: The ACF approach. *Journal of Econometrics* 175, 22-34. https://doi.org/10.1016/j.jeconom.2013.02.006
2. Anand, P., Roope, L., and Ross, A. 2019. How economists help central government think: Survey evidence from the UK Government Economic Service. *International Journal of Public Administration* 42, 1145-1157. https://doi.org/10.1080/01900692.2019.1575668
3. Ansell, M. and Marshall, S. 2017. What does research-informed teaching look like? https://www.advance-he.ac.uk/knowledge-hub/what-does-research-informed-teaching-look Advance HE
4. Ball, R., Medeiros, N., Bussberg, N. and Piekut, A. 2022. An invitation to teaching reproducible research: Lessons from a symposium. *Journal of Statistics and Data Science Education* 30, 209-218. https://doi.org/10.1080/26939169.2022.2099489
5. Bamber, M., McCormack, J. and Lyons, B. 2023. Conceptualising 'within-group stigmatisation' among high-status workers. *Work, Employment and Society* 37, 757-775. https://doi.org/10.1177/09500170211041287
6. Bandura, A. 1978. Self-efficacy: toward a unifying theory of behavioral change. *Advances in Behaviour Research and Therapy* 1, 139-161. https://doi.org/10.1037/0033-295X.84.2.191
7. Barba, L. 2018. Terminologies for reproducible research. *arXiv*, 1802.03311. Available: https://arxiv.org/pdf/1802.03311.
8. Berry, J., Coffman, L., Hanley, D., Gihleb, R. and Wilson, A. 2017. Assessing the rate of replication in Economics. *American Economic Review* 107, 27-31. https://doi.org/10.1257/aer.p20171119
9. Bolt, J. and van Zanden, J. 2024. Maddison style estimates of the evolution of the world economy: A new 2023 update. *Journal of Economic Surveys* 39, 639-659. https://doi.org/10.1111/joes.12618
10. Bowyer, D. and Akpinar, M. 2022. Experiences of academics and undergraduate students on research-based learning: A tale of two institutions. *Innovations in Education and Teaching International* 61, 45-56. https://doi.org/10.1080/14703297.2022.2149606
11. Boyd, W., O'Reilly, M., Bucher, D., Fisher, K., Morton, A., Harrison, P., Nuske, E., Coyle, R. and Rendall, K. 2010. Activating the teaching-research nexus in smaller universities: Case studies highlighting diversity of practice. *Journal of University Teaching and Learning Practice* 7(2). https://doi.org/10.53761/1.7.2.9

12. Boyer Commission. 1998. Re-inventing undergraduate education: A blueprint for America's research universities. New York: Carnegie Foundation for University Teaching. https://files.eric.ed.gov/fulltext/ED424840.pdf

13. Brew, A. 2003. Teaching and Research: New relationships and their implications for inquiry-based teaching and learning in higher education. *Higher Education Research & Development* 22, 3-18. https://doi.org/10.1080/0729436032000056571

14. Brew, A. 2013. Understanding the scope of undergraduate research: a framework for curricular and pedagogical decision-making. *Higher Education* 66, 603-618. https://doi.org/10.1007/s10734-013-9624-x

15. Brew, A. and Mantai, L. 2017. Academics' perceptions of the challenges and barriers to implementing research-based experiences for undergraduates. *Teaching in Higher Education* 22, 551-568. https://doi.org/10.1080/13562517.2016.1273216

16. Camerer, C., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M. and Wu, H. 2016. Evaluating replicability of laboratory experiments in economics. *Science* 351, 1433-1436. https://doi.org/10.1126/science.aaf0918

17. Carlson, J., Daehler, K., Alonzo, A., Barendsen, E., Berry, A., Borowski, A., Carpendale, J., Chan, K., Cooper, R., Friedrichsen, P., Gess-Newsome, J., Henze-Rietveld, I., Hume, A., Kirschner, S., Liepertz, S., Loughran, J., Mavhunga, E., Neumann, K., Nilsson, P., Park, S., Rollnick, M., Sickel, A., Schneider, R., Suh, J., van Driel, J. and Wilson, C. (2019). The Refined Consensus Model of Pedagogical Content Knowledge in Science Education. In: Hume, A., Cooper, R. and Borowski, A. (eds) Repositioning Pedagogical Content Knowledge in Teachers' Knowledge for Teaching Science. Singapore: Springer. https://doi.org/10.1007/978-981-13-5898-2_2

18. Chamberlain, J. 2016. Ensuring the criminological skills of the next generation: a case study on the importance of enhanced quantitative method teaching provision. *Journal of Further and Higher Education* 41, 448-459. https://doi.org/10.1080/0309877X.2015.1117602

19. Chang, A. and Li, P. 2017. A pre-analysis plan to replicate sixty Economics research papers that worked half of the time. *American Economic Review* 107, 60-64. https://doi.org/10.1257/aer.p20171034

20. Clark, T. and Hordosy, R. 2019. Undergraduate experiences of the research/teaching nexus across the whole student lifecycle, *Teaching in Higher Education*, 24, 412-427 https://doi.org/10.1080/13562517.2018.1544123

21. Clemens, M. 2017. The meaning of failed replications: A review and proposal. *Journal of Economic Surveys* 31, 326-342. https://doi.org/10.1111/joes.12139

22. Cook, S. 2008. Cross-data-vintage encompassing. *Oxford Bulletin of Economics and Statistics* 70, 849-865. https://doi.org/10.1111/j.1468-0084.2008.00533.x

23. Cook, S. 2016. Modern econometrics: Structuring delivery and assessment. *Cogent Economics and Finance* 4. https://doi.org/10.1080/23322039.2016.1152705

24. Cook, S. and Watson, D. 2023. The use of online materials to support the development of quantitative skills. In: Nind, M. (ed.), The Handbook of Teaching and Learning Social Research Methods, Cheltenham: Edward Elgar. pp. 274-286.

25. Cook, S., Dawson, P. and Watson, D. 2025. Bridging the Quantitative Skills Gap: Teaching simple linear regression via simplicity and structured replication, Economics Network Handbook for Economics Lecturers. https://doi.org/10.53593/n4229a

26. Cooper, G. and Sweller, J. 1987. The effects of schema acquisition and rule automation on mathematical problem-solving transfer. *Journal of Educational Psychology*, 79, 347–362. https://doi.org/10.1037/0022-0663.79.4.347

27. Counsell, A., Cribbie, R. and Harlow, L. 2016. Increasing literacy in quantitative methods: The key to the future of Canadian psychology. *Canadian Psychology* 57, 193-201. https://doi.org/10.1037/cap0000056

28. Croushore, D. and Stark, T. 2003. A real-time data set for macroeconomists: Does the data vintage matter? *Review of Economics and Statistics*, 85, 605-617. https://www.jstor.org/stable/3211700

29. Davidson, J., Hendry, D., Srba, F. and Yeo, S. 1978. Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the United Kingdom. *Economic Journal* 88, 661-692. https://www.jstor.org/stable/2231972

30. Department for Science, Innovation and Technology and Department for Digital, Culture, Media and Sport [DSIT & DDCMS]. 2021. Quantifying the UK Data Skills Gap. London: HMSO. https://www.gov.uk/government/publications/quantifying-the-uk-data-skills-gap/quantifying-the-uk-data-skills-gap-full-report

31. Dewald, W., Thursby, J. and Anderson, R. 1986. Replication in Empirical Economics: The Journal of Money, Credit and Banking Project. *American Economic Review* 76(4), 587-603. https://www.jstor.org/stable/1806061

32. Dickey, D. and Fuller, W. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74, 427-431. https://doi.org/10.1080/01621459.1979.10482531

33. Dowker, A., Sarkar A. and Looi, C. 2016. Mathematics anxiety: what have we learned in 60 years? *Frontiers in Psychology* 7, 508. https://doi.org/10.3389/fpsyg.2016.00508

34. Dreger R. and Aiken L. 1957. The identification of number anxiety in a college population. *Journal of Educational Psychology* 48, 344-351. https://doi.org/10.1037/h0045894

35. Duvendack, M., Palmer-Jones, R. and Reed, W. 2017. What is meant by replication and why does it encounter resistance in economics? *American Economic Review* 107, 46-51. https://doi.org/10.1257/aer.p20171031

36. Egginton, D., Pick, A. and Vahey, S. 2002. Keep it real!: A real-time UK macro data set. *Economics Letters*, 77, 15-20. https://doi.org/10.1016/S0165-1765(02)00094-0

37. Ericsson, N. 2004. The ET interview: Professor David F. Hendry. *Econometric Theory* 20, 745-806. https://www.jstor.org/stable/3533545

38. Fraser, S. 2016. Pedagogical Content Knowledge (PCK): Exploring its usefulness for science lecturers in Higher Education. *Research in Science Education* 46, 141–161. https://doi.org/10.1007/s11165-014-9459-1

39. Frisch, R. 1933. Editor's note. *Econometrica* 1, 1-4. https://www.jstor.org/stable/1912224

40. Garratt, A. and Vahey, S. 2006. UK real-time macro data characteristics. *Economic Journal*, 116, F119-F135. https://www.jstor.org/stable/3590486

41. Hattie, J. and Marsh, H. 1996. The relationship between research and teaching: A meta-analysis. *Review of Educational Research* 66, 507-542. https://doi.org/10.3102/00346543066004507

42. Healey, M. and Jenkins, A. 2009. Developing undergraduate research and inquiry. York: Higher Education Academy. https://www.advance-he.ac.uk/knowledge-hub/developing-undergraduate-research-and-inquiry

43. Healey, M., Jordan, F., Pell, B. and Short, C. 2010. The research-teaching nexus: a case study of students' awareness, experiences and perceptions of research. *Innovations in Education and Teaching International* 47, 235-246. https://doi.org/10.1080/14703291003718968

44. Hendry, D. 1995. Dynamic Econometrics. Oxford: Oxford University Press.

45. Hendry, D. 1986. Using PC-GIVE in econometrics teaching. *Oxford Bulletin of Economics and Statistics* 48, 87-98. https://doi.org/10.1111/j.1468-0084.1986.mp48001007.x

46. Hendry, D. 2024. A brief history of general-to-specific modelling. *Oxford Bulletin of Economics and Statistics*, 86, 1-20. https://doi.org/10.1111/obes.12578

47. Hendry, D. and Ericsson, N. 1991. Modeling the demand for narrow money in the United Kingdom and the United States. *European Economic Review* 35, 833-881. https://doi.org/10.1016/0014-2921(91)90039-L

48. Hendry, D. and Morgan, M. 1989. A re-analysis of confluence analysis. *Oxford Economic Papers*, 41, 35-52. https://www.jstor.org/stable/2663181

49. Holmes, M. and Otero, J. 2019. Re-examining the movements of crude oil spot and futures prices over time. *Energy Economics* 82, 224-236. https://doi.org/10.1016/j.eneco.2017.08.034

50. Ioannidis, J. 2005. Why most published research findings are false. *PLoS Med* 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

51. Janz, N. 2016. Bringing the Gold Standard into the classroom: Replication in University teaching. *International Studies Perspectives* 17, 392-407. https://doi.org/10.1111/insp.12104

52. Johansen, S. 1988. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* 12, 231-254. https://doi.org/10.1016/0165-1889(88)90041-3

53. Kalyuga, S., Chandler, P., Tuovinen, J., and Sweller, J. 2001. When problem solving is superior to studying worked examples. *Journal of Educational Psychology* 93, 579-588. https://doi.org/10.1037/0022-0663.93.3.579

54. Kalyuga, S., Ayres, P., Chandler, P., and Sweller, J. 2003. Expertise reversal effect. *Educational Psychologist* 38, 23-31. https://doi.org/10.1207/S15326985EP3801_4

55. Kapur, M. 2008. Productive failure. *Cognition and Instruction* 26, 379-424. https://doi.org/10.1080/07370000802212669

56. Kapur, M. 2012. Productive failure in learning the concept of variance. *Instructional Science* 40, 651-672. https://doi.org/10.1007/s11251-012-9209-6

57. Kapur, M. 2015. Learning from productive failure. *Learning: Research and Practice* 1, 51-65. https://doi.org/10.1080/23735082.2015.1002195

58. Kirschner, P., Sweller, J. and Clark, R. 2006. Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential and inquiry-based teaching. *Educational Psychologist* 41, 75-86. https://doi.org/10.1207/s15326985ep4102_1

59. Levy, P. and Petrulis, R. 2011. How do first-year university students experience inquiry and research, and what are the implications for the practice of inquiry-based learning? *Studies in Higher Education*, 37, 85-101. https://doi.org/10.1080/03075079.2010.499166

60. Leybourne, S. 1995. Testing for unit roots using forward and reverse Dickey-Fuller regressions. *Oxford Bulletin of Economics and Statistics* 57, 559-571. https://doi.org/10.1111/j.1468-0084.1995.tb00040.x

61. Leybourne, S., Mills, T. and Newbold, P. 1998. Spurious rejections by Dickey-Fuller tests in the presence of a break under the null. *Journal of Econometrics* 87, 191-203. https://doi.org/10.1016/S0304-4076(98)00014-1

62. Lindsay, R. Breen, R. and Jenkins, A. 2002. Academic Research and Teaching Quality: The views of undergraduate and postgraduate students. *Studies in Higher Education* 27, 309-327. https://doi.org/10.1080/03075070220000699

63. Loibl, K., Roll, I. and Rummel, N. 2017. Towards a theory of when and how problem solving followed by instruction supports learning. *Educational Psychology Review* 29, 693-715. https://doi.org/10.1007/s10648-016-9379-x

64. Lu, M. and Podivinsky, J. 2003. The robustness of trend stationarity: An illustration with the extended Nelson–Plosser dataset. *Econometric Reviews* 22, 261-267. https://doi.org/10.1081/ETC-120024075

65. Lucas, A. 1995. An outlier robust unit root test with an application to the extended Nelson-Plosser data. *Journal of Econometrics* 66, 153-173. https://doi.org/10.1016/0304-4076(94)01613-5

66. Lumsdaine, R. and Papell, D. 1997. Multiple trend breaks and the unit-root hypothesis. *Review of Economics and Statistics* 79, 212-218. https://www.jstor.org/stable/2951453

67. Machery, E. 2020. What is a replication? *Philosophy of Science* 87, 545-567. https://doi.org/10.1086/709701

68. Machery, E. and Doris, J. 2017. An open letter to our students: Doing interdisciplinary moral psychology. In Voyer, B. and Tarantola, T. (eds.) Moral psychology: A multidisciplinary guide, pp.119-143. Berlin: Springer. https://doi.org/10.1007/978-3-319-61849-4_7

69. Macfarlane, B. 2011. Prizes, pedagogic research and teaching professors: lowering the status of teaching and learning through bifurcation. *Teaching in Higher Education* 16, 127-130. https://doi.org/10.1080/13562517.2011.530756

70. MacInnes, J., Breeze, M., de Haro, M., Kandlik, M. and Karels, M. 2016. Measuring Up: International Case Studies on the Teaching of Quantitative Methods in the Social Sciences. London: The British Academy.

71. Maddison, A. 1995. Monitoring the world economy 1820-1992. Paris: OECD.

72. Magnus, J. and Morgan, M. 1997. The design of the experiment. *Journal of Applied Econometrics* 12, 459-465. https://www.jstor.org/stable/2285119

73. Mankiw, G. and Shapiro, M. 1986. News or noise?: An analysis of GDP revisions. *Survey of Current Business* 66, 20-25.

74. Mankiw, G., Runkle, D. and Shapiro, M. 1984. Are preliminary announcements of the money stock rational forecasts? *Journal of Monetary Economics* 14, 15-27. https://doi.org/10.1016/0304-3932(84)90024-2

75. Mansell, W. 2015. Count Us In: Quantitative Skills for a New Generation. London: British Academy.

76. Mathieson, S. 2019. Integrating research, teaching and practice in the context of new institutional policies: a social practice approach. *Higher Education* 78, 799-815. https://doi.org/10.1007/s10734-019-00371-x

77. Mayer, R. 2004. Should there be a three-strikes rule against pure discovery learning? *American Psychologist* 59, 14-19. https://doi.org/10.1037/0003-066x.59.1.14

78. Mayer, R. 2021. Multimedia Learning (3rd edition). Cambridge: Cambridge University Press.

79. National Academies of Sciences, Engineering and Medicine (NASEM). 2019. Reproducibility and replicability in science. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303

80. Nelson, C. and Plosser, C. 1982. Trends and random walks in macroeconomic time series. *Journal of Monetary Economics* 10, 139-162. https://doi.org/10.1016/0304-3932(82)90012-5

81. Ng, S. and Perron, P. 1995. Unit root tests in ARMA models with data-dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association* 90, 268-281. https://doi.org/10.1080/01621459.1995.10476510

82. Ng, S. and Perron, P. 2001. Lag length selection and the construction of unit root tests with good size and power. *Econometrica* 69, 1519-1554. https://doi.org/10.1111/1468-0262.00256

83. Nind, M. 2020. A new application for the concept of pedagogical content knowledge: teaching advanced social science research methods. *Oxford Review of Education* 46, 185-201. https://doi.org/10.1080/03054985.2019.1644996

84. Nosek, B. and Errington, T. 2020. What is replication? *Philosophy of Science* 87, 545-567. https://doi.org/10.1086/709701

85. Nosek, B., Hardwicke, T., Moshontz, H., Allard, A., Corker, K., Dreber, A., Fidler, F., Hilgard, J. Struhl, M., Nuijten, M., Rohrer, J., Romero, F., Scheel, A., Scherer, L., Schönbrodt, F. and Vazire, S. 2022. Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology* 73, 719-748. https://doi.org/10.1146/annurev-psych-020821-114157

86. Nosek, B., Alter, G., Banks, G., Borsboom, D., Bowman, S., Breckler, S., Buck, S., Chambers, C., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R,. Goroff, D., Green, D., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Paluck, E., Simonsohn, U., Soderberg, C.,

Spellman, B., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E., Wilson, R. and Yarkoni, T. 2015. Promoting an open research culture. *Science* 348 (6242) 1422-1425. https://doi.org/10.1126/science.aab2374

87. Open Science Collaboration 2015. Estimating the reproducibility of psychological science. *Science* 349. https://doi.org/10.1126/science.aac4716

88. Parker, J. 2018. Undergraduate research, learning gain and equity: the impact of final year research projects. *Higher Education Pedagogies* 3, 145-157. https://doi.org/10.1080/23752696.2018.1425097

89. Patterson, K. and Heravi, S. 1991. Data revisions and the expenditure component of GDP. *Economic Journal* 101, 887-901. https://doi.org/10.2307/2233861

90. Perron, P. 1989. The Great Crash, the oil price shock, and the unit root hypothesis. *Econometrica* 57, 1361-1401. https://doi.org/10.2307/1913712

91. Perron, P. 1997. Further evidence on breaking trend functions in macroeconomic variables. *Journal of Econometrics* 80, 355-385. https://doi.org/10.1016/S0304-4076(97)00049-3

92. Pesaran, M. 2003. Introducing a replication section. *Journal of Applied Econometrics* 18, 111. https://doi.org/10.1002/jae.709

93. Persaud, D., Ward, L. and Hattrick-Simpers, J. 2024. Reproducibility in materials informatics: Lessons from 'A general-purpose machine learning framework for predicting properties of inorganic materials'. *Digital Discovery* 3, 281-286. https://doi.org/10.1039/d3dd00199g

94. Phillips, P. 1991. To criticize the critics: An objective Bayesian analysis of stochastic trends. *Journal of Applied Econometrics* 6, 333-364. https://doi.org/10.1002/jae.3950060402

95. Quality Assurance Agency for Higher Education. (QAA). 2023. *Subject Benchmark Statement: Economics*. QAA. https://www.qaa.ac.uk/docs/qaa/sbs/sbs-economics-23.pdf

96. Rudebusch, G. 1992. Trends and random walks in macroeconomic time series: A re-examination. *International Economic Review* 33, 661-680. https://doi.org/10.1016/0165-1889(88)90043-7

97. Rode, J., Jonuschies, I., Matthiesen, S. and Gericke, K. (2024). Replication studies in engineering design - a feasibility study. *Proceedings of the Design Society* 4, 115-124. https://doi.org/10.1017/pds.2024.14

98. Ruth, A., Brewis, A., Beresford, M. and Stojanowski, C. 2023. Research supervisors and undergraduate students' perceived gains from undergraduate research experiences in the social sciences. *International Journal of Inclusive Education*, 1-18. https://doi.org/10.1080/13603116.2023.2288642

99. Schapper, J. and Mayson, S. 2010. Research-led teaching: moving from a fractured engagement to a marriage of convenience. *Higher Education Research and Development* 29, 641-651. https://doi.org/10.1080/07294360.2010.489236

100. Scott Jones, J. and Goldring, J. 2015. 'I'm not a quants person'; key strategies in building competence and confidence in staff who teach quantitative research methods'. *International Journal of Social Research Methodology* 18, 479-494. https://doi.org/10.1080/13645579.2015.1062623

101. Shulman, L. 1986. Those who understand: Knowledge growth in teaching. *Educational Researcher* 15(2), 4-14. https://doi.org/10.3102/0013189X015002004

102. Smith, L., Yu, F. and Schmid, K. 2021. Role of replication research in biostatistics graduate education. *Journal of Statistics and Data Science Education* 29, 95-104. https://doi.org/10.1080/10691898.2020.1844105

103. Spanos, A. 1989. Early empirical findings on the consumption function, stylized facts or fiction: A retrospective view. *Oxford Economic Papers* 41, 131-149. https://doi.org/10.1093/oxfordjournals.oep.a041890

104. Sweller, J. and Cooper, G. 1985. The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction* 2, 59-89. https://doi.org/10.1207/s1532690xci0201_3

105. Sweller, J., Van Merrienboer, J.J.G. and Paas, F. 1998. Cognitive architecture and instructional design. *Educational Psychology Review* 103, 251-296. https://doi.org/10.1023/A:1022193728205

106. Sweller, J., Van Merrienboer, J. and Paas, F. 2019. Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review* 31, 261-292. https://doi.org/10.1007/s10648-019-09465-5

107. Thomas, J. 1989. The early econometric history of the consumption function. *Oxford Economic Papers* 41, 131-149. https://doi.org/10.1093/oxfordjournals.oep.a041888

108. Tol, R. 2019. Special issue on replication. *Energy Economics* 82, 1-3. https://doi.org/10.1016/j.eneco.2018.12.017

109. Vicens, Q. and Bourne, P. 2009. Ten simple rules to combine teaching and research. *PLoS Computational Biology* 5(4): e1000358. https://doi.org/10.1371/journal.pcbi.1000358

110. Wagge, J., Brandt, M., Lazarevic, L., Legate, N., Christopherson, C., Wiggins, B., Grahe, J. 2019. Publishing research with undergraduate students via replication work: The collaborative replications and education project. *Frontiers in Psychology* 10: 247. https://doi.org/10.3389%2Ffpsyg.2019.00247

111. Stojmenovska D., Bol T. and Leopold T. 2019. Teaching replication to graduate students. *Teaching Sociology* 47, 303-313. https://doi.org/10.1177/0092055X19867996

112. VanLehn, K., Siler, S., Murray, C., Yamauchi, T., and Baggett, W. 2003. Why do only some events cause learning during human tutoring? *Cognition and Instruction* 21, 209-249. https://doi.org/10.1207/S1532690XCI2103_01

113. van Merrienboer, J. and Sweller, J. 2005. Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review* 17, 147-177. https://doi.org/10.1007/s10648-005-3951-0

114. Walkington, H. and Ommering, B. 2022. How does engaging in authentic research at undergraduate level contribute to student well-being? *Studies in Higher Education* 47, 2497-2507. https://doi.org/10.1080/03075079.2022.2082400

115. Wessels, I., Rueß, J., Gess, C., Deicke, W. and Ziegler, M. 2021. Is research-based learning effective? Evidence from a pre–post analysis in the social sciences, *Studies in Higher Education*, 46, 2595-2609. https://doi.org/10.1080/03075079.2020.1739014

116. Wood, W. 2003. Inquiry-based undergraduate teaching in the life sciences at large research Universities: A perspective on the Boyer Commission Report. *Cell Biology Education* 2, 112-116. https://doi.org/10.1187/cbe.03-02-0004

117. World Economic Forum. 2019. Data Science in the New Economy. https://www3.weforum.org/docs/WEF_Data_Science_In_the_New_Economy.pdf

118. Wulwick, N. 1989. Phillips' approximate regression. *Oxford Economic Papers* 41, 170-188. https://doi.org/10.1093/oxfordjournals.oep.a041891

119. Zahaciva, A., Lynch, S. and Espenshade, T. 2005. Self-efficacy, stress, and academic success in college. *Research in Higher Education* 46, 677-706. https://doi.org/10.1007/s11162-004-4139-z

[1] We are very grateful to the editor, Peter Dawson, for comments that have improved both the content and presentation of this chapter and the three associated case studies (Cook and Watson, 2025a, b, c).

[2] An example of the use of R&R in textbooks is provided by the consideration of Hendry and Ericsson (1991) in Hendry (1995). An example of the appearance of R&R in software is the provision of the data of Davidson et al. (1978) in PC-GIVE (see Ericsson, 2004; Hendry, 1986, 2024).

[3] Here 'higher-level' simply refers a level above introductory. As will be seen later, the specific case studies reviewed focus upon published research considering unit root and cointegration analysis which can both be viewed as 'higher-level'.

[4] As will be seen, a number of alternative terms are used in relation to the inclusion of research in teaching. IRT is employed here for convenience as a means of referring to 'incorporating research in teaching' or 'the incorporation of research in teaching'.

[5] Following its introduction and subsequent exploration for secondary-level education, PCK has gained traction in the Higher Education literature with recent research considering its alternative forms (Carlson et al., 2019), application to different disciplinary areas (Nind, 2020) and the factors impacting upon lecturer engagement with PCK (Fraser, 2016).

[6] Given our interest in econometrics, our reference to Economics arises via consideration of this as the broader discipline within which econometrics lies. We are not arguing that what follows is specific to Economics.

[7] The data employed in Leybourne et al. (1998) are available from Maddison (1995). These revised data are available from the 2023 Maddison Project Database (Bolt and van Zanden, 2024).

THE ECONOMICS NETWORK

Improving economics teaching
and learning for over 20 years

Replication and Reproduction I: Leybourne (1995) and the maximum Dickey-Fuller test

Steve Cook
Swansea University
s.cook at swan.ac.uk
and Duncan Watson
University of East Anglia
Duncan.Watson at uea.ac.uk

PUBLISHED
November 2025

https://doi.org/10.53593/n4409a

CONTENTS

This case study is the first in a set of materials on the effective incorporation of research in undergraduate econometrics edited by Peter Dawson of the University of East Anglia.

1. Introduction

Cook and Watson (2025) have promoted the use of replication and reproduction (R&R) as a means of incorporating research into the teaching of econometrics. Emphasising the direct engagement with research necessitated by R&R, Cook and Watson (2025) argue it will lead to numerous benefits and allow a range of pedagogical objectives to be addressed.

In this case study, the work of Leybourne (1995) (*Oxford Bulletin of Economics and Statistics*) is employed as vehicle for utilising R&R in the teaching of econometrics. The focus of Leybourne (1995) is unit root analysis and, more specifically, the introduction of a higher-powered version of the Augmented Dickey-Fuller (ADF) test (Dickey and Fuller, 1979). To illustrate the empirical application of the proposed maximum ADF test, Leybourne (1995) employs the data used in the seminal research of Nelson and Plosser (1982). While this involves the analysis of fourteen different U.S. macroeconomic time series, the present case study demonstrates the application of R&R using one of these series. The discussion of the employment series presented here can, of course, be extended to the remaining 13 series in the Nelson-Plosser dataset.

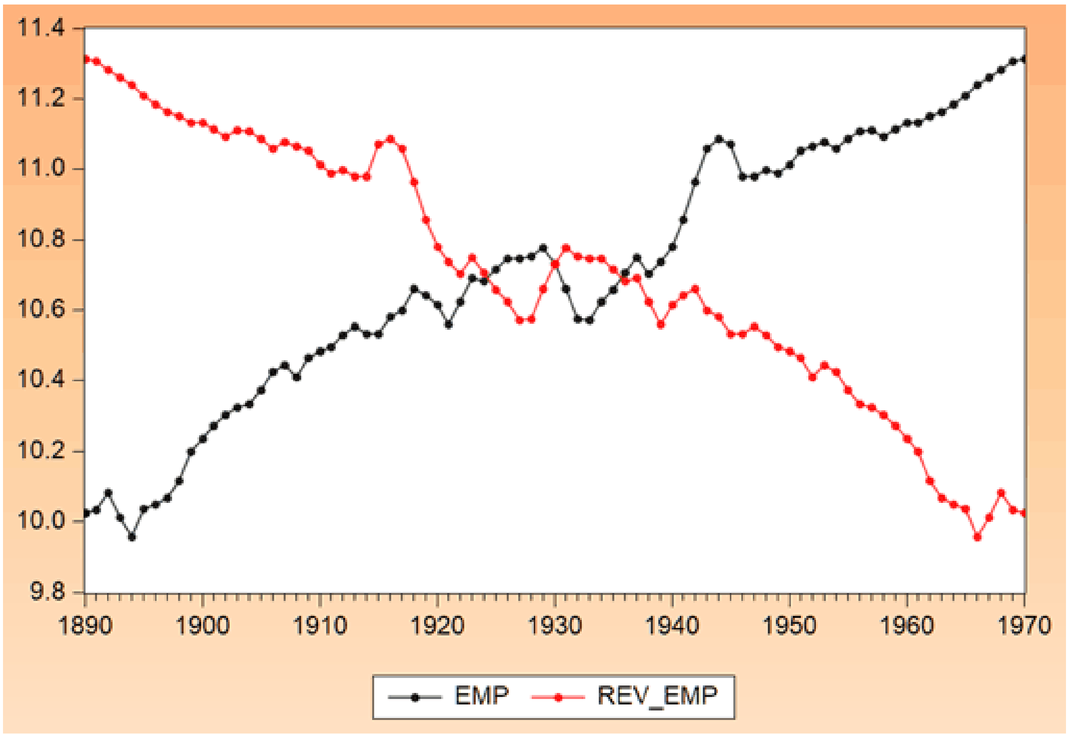
The intention of the present case study is to use data employed in Leybourne (1995) to reproduce the results presented for both the ADF test and the (then new) maximum ADF test. In the process of undertaking this exercise, learners are prompted to engage closely with the work presented by Leybourne (1995) and to consider, inter alia: the decisions made with regard to the inclusion of deterministic terms in, and lag optimisation of, the relevant testing equations; the creation of a 'reversed' realisation of a series; the derivation of a maximum ADF test statistic; and the use of alternative, non-standard finite-sample distributions. Beyond these empirical issues, the revisiting of Leybourne (1995) also exposes learners to the issue of test power and its analysis via Monte Carlo simulation. Results demonstrating the relative powers of the two (ADF and maximum ADF) tests are provided in Tables 3 and 4 of Leybourne (1995), offering concrete examples that support classroom discussions of test power with specific simulated results.[1]

2. Reproducing the results of Leybourne (1995)

To illustrate the inclusion of the research in teaching, we consider one of the series examined in Leybourne (1995). The series considered is the employment variable contained in the original Nelson-Plosser dataset.[2] Here, the natural logarithm of the employment series is denoted as EMP and is observed at an annual frequency for the period 1890-1970.

The maximum ADF test of Leybourne (1995) is so named because it considers two ADF test statistics: the ADF test statistic resulting from examination of a series of interest (the usually considered 'forward realisation') and the ADF test statistic arising from examination of the 'reversed realisation' of the series. In short, for analysis of EMP we first examine the series running from 1890 to 1970 and then construct its reversed realisation (denoted as REV_EMP), in which observations run from 1970 to 1890. The resulting maximum ADF test statistic is then given as the greater value of the two ADF test statistics obtained. To illustrate the series, both EMP and REV_EMP are shown in Figure One below.

Figure One: Employment and its reversed realisation



The results reported in Leybourne (1995) from application of the ADF and maximum ADF tests to the two employment series are presented in Table One below.[3]

Table One: Leybourne (1995) results

Leybourne (1995)			
p	τ_τ	τ_τ^r	τ_τ^{max}
1	-3.13	-2.97	-2.97

The challenge for learners is to reproduce the results shown in Table One. The analysis for the two (forward and reverse) employment series is based upon the following underlying testing equation, labelled (1) below:

$$(1) \quad \Delta y_t \; = \alpha + \beta t + \phi y_{t-1} + \sum_{j=1}^{p} \gamma_j \, \Delta y_{t-j} + e_t$$

The unit root null hypothesis can then be stated as $\mathrm{H}_0\!: \; \phi = 0$ and tested against the alternative hypothesis of (asymptotic) stationarity given as $\mathrm{H}_1\!: \; \phi < 0$ via the test statistic $\tau_\tau = \frac{\hat{\phi}}{s.e.(\hat{\phi})}$. Table One presents the calculated τ_τ statistic, the ADF test statistic based upon the reversed realisation (τ_τ^r) and maximum test statistic (τ_τ^{max} = max[τ_τ, τ_τ^r]).

Two key issues of importance in applying the above unit root tests are the choice of appropriate deterministic terms in the underlying testing equation and determining its degree of augmentation (p). Clearly, the series is trending so the deterministics to include are as given in (1) – that is, $\alpha + \beta t$. The value of p to employ is provided in Table One. However, flexibility is available to the lecturer to vary the challenge augmentation poses for reproduction. For example, to simplify the demands of the reproduction exercise, the value of p could be provided to learners. Going a step further, the value of p might not be disclosed and learners could instead be told of the decision rule employed by Leybourne. Alternatively, and more challenging still, students could be asked to consult the paper to find the rule employed. This would result in discovery of the use of the sequential t-statistic rule and require utilisation of the appropriate level of significance for its application. Clearly this latter option reinforces the earlier discussion of R&R requiring learners to engage with the studies they read.

Reproduction of the results of Leybourne (1995) is provided in Tables Two and Three below. Clearly when producing the results in Tables Two and Three a range of issues arise in addition to challenging knowledge of how to correctly specify appropriate testing equations. For example, the drawing of inferences requires appreciation of the fact that while the ADF test results for the original (forward) series will involve consideration of the Dickey-Fuller distribution, inferences for the maximum ADF test will require appropriate consideration of the tabulated values provided in the tables of Leybourne (1995). Applying linear interpolation to critical values of −2.87 and −2.84 for the maximum ADF test at the 10% level of significance for sample sizes of 50 and 100 respectively, the resulting 10% critical value of −2.85 for the sample of 79 observations employed in this analysis allows rejection against a maximum ADF test statistic of −2.97. This rejection at the 10% level can be compared to a marginal failure to reject at this level using the ADF test (for the forward series), where the relevant p-value is 10.7%.

Table Two: Reproducing Leybourne (1995)- the original EMP series

Null Hypothesis: EMP has a unit root Exogenous: Constant, Linear Trend Lag Length: 1 (Automatic - based on t-statistic, lagpval=0.05, maxlag=11)			
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		−3.128528	0.1070
Test critical values:	1% level	−4.078420	
	5% level	−3.467703	
	10% level	−3.160627	

Table Three: Reproducing Leybourne (1995)- the reversed realisation of EMP (REV_EMP)

Null Hypothesis: REV_EMP has a unit root Exogenous: Constant, Linear Trend Lag Length: 1 (Automatic - based on t-statistic, lagpval=0.05, maxlag=11)

	t-Statistic
Augmented Dickey-Fuller test statistic	−2.967223

As additional points, note that when undertaking the exercises in relation to Tables One to Three, the issues raised in Perron (1989, 1997) concerning misclassification of orders of integration as a result of structural change should be discussed. Therefore, discussion of potential power gains should be qualified by recognising breaks in the Nelson-Plosser series. Also, while the above activities provide an exercise in direct replication, using the terminology of Cook and Watson (2023), since results are automatically produced using econometric software, step replication could be performed via a more manual approach – creating lagged differenced terms, specifying and estimating the testing equation to be employed, and then identifying the relevant output to draw inferences.

3. Conclusion

Cook and Watson (2025) have championed the use of R&R as a means of introducing research into the teaching of econometrics, discussing a collection of potential benefits, including those relating to engagement, activity and pedagogical research. In this case study, the R&R has been illustrated using the study of Leybourne (1995). The above discussion has shown that the process of reproducing published empirical results develops both core econometric skills and transferable skills.

References

Cook, S. and Watson, D. 2023. The use of online materials to support the development of quantitative skills. In Nind, M. (ed.), The Handbook of Teaching and Learning Social Research Methods, Cheltenham: Edward Elgar. pp. 274-286.

Cook, S. and Watson, D. 2025. From provision to understanding: The effective incorporation of research in undergraduate econometrics. In The Handbook for Economics Lecturers. Economics Network. https://doi.org/10.53593/m4412a

Dickey, D. and Fuller, W. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74, 427-431. https://doi.org/10.1080/01621459.1979.10482531

Leybourne, S. 1995. Testing for unit roots using forward and reverse Dickey-Fuller regressions. *Oxford Bulletin of Economics and Statistics* 57, 559-571. https://doi.org/10.1111/j.1468-0084.1995.tb00040.x

Nelson, C. and Plosser, C. 1982. Trends and random walks in macroeconomic time series. *Journal of Monetary Economics* 10, 139-162. https://doi.org/10.1016/0304-3932(82)90012-5

Ng, S. and Perron, P. 1995. Unit root tests in ARMA models with data-dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association* 90, 268-281. https://doi.org/10.1080/01621459.1995.10476510

Perron, P. 1989. The Great Crash, the oil price shock, and the unit root hypothesis. *Econometrica* 57, 1361-1401. https://doi.org/10.2307/1913712

Perron, P. 1997. Further evidence on breaking trend functions in macroeconomic variables. *Journal of Econometrics* 80, 355-385. https://doi.org/10.1016/S0304-4076(97)00049-3

Footnotes

[1] The Monte Carlo results provided in Leybourne (1995) can be also employed to support discussions concerning the derivation and nature of critical values and the analysis of empirical size. The results concerning these issues are provided in Tables 1-2 and Tables 5-8 respectively.

[2] This data set is available from a number of sources including http://korora.econ.yale.edu/phillips/.

[3] These results are provided in Leybourne (1995, Table 9, p.570). It can be noted that different approaches to determining the degree of augmentation of ADF testing equations are employed in Nelson and Plosser (1982) and Leybourne (1995). While the former adopts an approach based upon the use of autocorrelation and partial autocorrelation functions, the latter employs the sequential t-statistic rule. It can also be noted that the different lag lengths obtained under these methods result in different calculated ADF test statistics, with the lag length of 2 in Nelson and Plosser (1982) resulting in an ADF test statistic of −2.66 in contrast to the value of −3.13 in Table One above.

THE ECONOMICS NETWORK

Improving economics teaching
and learning for over 20 years

Replication and Reproduction II: Leybourne et al. (1998) and the Dickey-Fuller test in the presence of breaks under the null

Steve Cook
Swansea University
s.cook at swan.ac.uk
and Duncan Watson
University of East Anglia
Duncan.Watson at uea.ac.uk

PUBLISHED
November 2025

https://doi.org/10.53593/n4410a

CONTENTS

This case study is the second in a set of materials on the effective incorporation of research in undergraduate econometrics edited by Peter Dawson of the University of East Anglia.

1. Introduction

Cook and Watson (2025) argue that replication and reproduction (R&R) provide an active learning environment by immersing learners in the research they read and have recently championed this as a means of incorporating research into the teaching of econometrics. This accompanying case study demonstrates the use of R&R through the work of Leybourne et al. (1998, hereafter referred to as LMN). However, rather than focusing solely on the exact reproduction of the original findings using the authors' data, it extends the analysis by considering results obtained using a more recent data vintage and a longer sample period. As such, the case study offers learners exposure to the progressive and evolving nature of econometric research while simultaneously developing higher-order skills in critical evaluation, methodological reflection and independent inquiry within an active learning framework.

Following Perron (1989), it has been recognised that structural change can result in the Dickey-Fuller (DF) test (Dickey and Fuller, 1979) failing to reject the unit root null hypothesis for otherwise stationary series. More recently, LMN have shown that structural change can also lead to misclassification in the 'other direction': the DF test may spuriously reject the null hypothesis when applied to unit root processes subject to a structural break in either the level or drift. This phenomenon is a key finding in LMN, whose simulation results demonstrate that the extent of size distortion depends on the location, size, and nature (i.e. a break in level or drift) of the structural break.[1]

The impact of a break under the null is illustrated in Tables 1 and 2 of LMN. For level breaks, oversizing is shown to be greatest when the break occurs at the very start of the sample. For drift breaks, however, the maximum distortion occurs when the break occurs early in the sample, but not at the very beginning. For example, LMN report the highest empirical size they detect – 93.6% at the 5% nominal level of significance – when the largest break in drift they consider occurs at observation 16 in a sample of 100 observations.[2] This striking result highlights the severity of the oversizing problem encountered by the DF test under structural change.

To supplement their simulation results, LMN provide an empirical analysis of the natural logarithm of the ratio of real per capita GDP for Denmark and Germany. Using data from Maddison (1995), they consider annual observations for this ratio over the period 1950 to 1994. The series is particularly relevant for their simulation analysis given its apparent break in trend early in the sample. To demonstrate the issue of spurious rejection, LMN show that applying the (augmented) DF (ADF) test to the full 1950-1994 sample leads to overwhelming rejection of the unit root hypothesis, whereas application to a post-break sample (1957-1994) results in non-rejection at conventionally considered levels of significance.

This case study is motivated by LMN's empirical analysis. While LMN's results are reproduced, the focus here is on exploring the impact of a more recent vintage of data and extended sample period on the subsequent inferences drawn. Using the updated data, the original analysis is repeated for both the 1950-1994 and 1957-1994 samples, as well as for extended samples covering the period up to 2022. The intention is therefore to move beyond direct reproduction of published research and to expose learners to the progressive and evolving nature of econometric modelling, where findings can be revisited using more recent data and longer sample spans.

To achieve its objectives, this case study proceeds as follows. In Section 2, the results of LMN are reproduced for the 1950-1994 and 1957-1994 samples employed by the authors. We refer to this as the ‘original’ data. In addition to this ‘reproduction’, the findings of LMN are revisited to consider alternative methods for determining the degree of augmentation of the ADF testing equation and their impact upon the inferences drawn. It should be noted that the results presented in this section using the original data are not intended to support a data-based reproduction exercise in class. Learners are not expected to replicate the results themselves given the restricted availability of the data employed. Instead, the material is provided for discussion and comparison with the findings of LMN, with some suggestions offered for its use. In Section 3, we turn to the main focus of the case study: an exploration of the impact of revised data and an extended sample period on the results obtained by LMN. Here, alternative open-source data are employed to generate results that can be compared with those of the original study. Section 4 offers concluding remarks.

2. Using the original data

We begin by considering the empirical analysis of the ratio of per capita GDP for Denmark and Germany presented by LMN. This analysis is based on annual data for the period 1950 to 1994 drawn from Maddison (1995). Using this source, we construct the relevant ratio, referred to here as ‘RATIO’. This series is presented in Figure One below.

Figure One: Plotting the LMN series

It can be seen that Figure One reproduces the plot of the ratio series presented in Leybourne et al. (1998, p.199, Fig. 1). The ADF test is applied to this series following the approach adopted by LMN – specifically, using the sequential t-statistic rule at the 5% level of significance to determine the degree of augmentation, with an intercept and trend term included as deterministic components. This approach produces the results shown in Table One below.[3] These again reproduce the findings reported by LMN, with the unit root hypothesis overwhelmingly rejected at conventionally employed significance levels (the p-value is 0.09%).

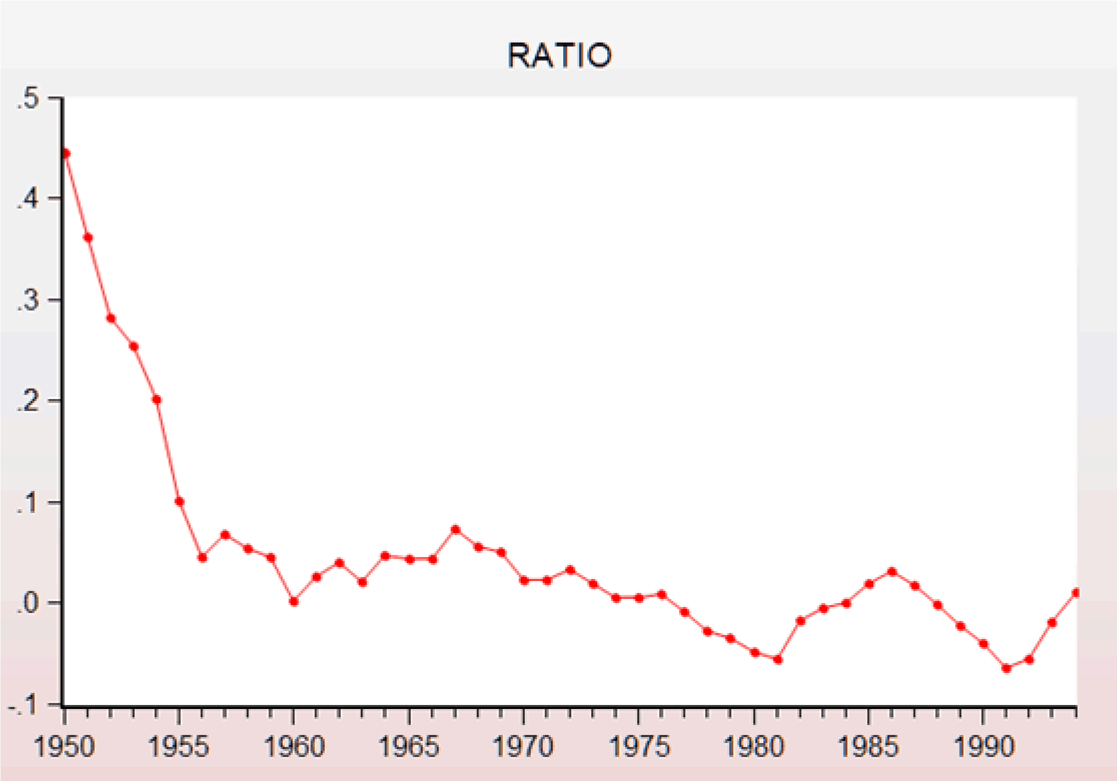


Table One: Reproducing the full sample results in LMN

Null Hypothesis: RATIO has a unit root Exogenous: Constant, Linear Trend Lag Length: 0 (Automatic - based on t-statistic, lagpval=0.05, maxlag=9)			
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		−5.071702	0.0009
Test critical values:	1% level	−4.180911	
	5% level	−3.515523	
	10% level	−3.188259	

With a focus on the impact of structural change, LMN continue their analysis by considering alternative break-incorporated unit root tests. In addition, they apply the ADF test to a post-break sample which yields an ADF test statistic of −2.20. Therefore, analysis of this post-break sample leads to non-rejection of the null at conventional significance levels, suggesting that the rejection observed in the full sample analysis may be spurious. Using the Maddison data, the post-break results of LMN can be reproduced, as shown in Table Two.

Table Two: Reproducing the post-break sample results in LMN

Null Hypothesis: RATIO has a unit root Exogenous: Constant, Linear Trend Lag Length: 0 (Automatic - based on t-statistic, lagpval=0.05, maxlag=9)			
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		−2.201191	0.4754
Test critical values:	1% level	−4.219126	
	5% level	−3.533083	
	10% level	−3.198312	

The above reproduction of the results of LMN has adopted their exact approach to the application of the ADF test. One subsequent issue to consider is how alternative methods of lag optimisation might affect the results obtained. That is, given the testing equation (1) below, we can examine different approaches to determining the value of p, along with their potential impact on the resulting test statistic and the inferences drawn.

$$(1) \quad \Delta y_t = \alpha + \beta t + \phi y_{t-1} + \sum_{j=1}^{p} \gamma_j \Delta y_{t-j} + e_t$$

In addition to the sequential t-statistic rule at the 5% level of significance employed by LMN, augmentation of the ADF testing equation can be carried out using several other approaches. Commonly employed alternatives include the sequential t-statistic rule at the 10% level of significance[4], the Akaike Information Criterion (AIC), the Schwarz Information Criterion (SIC), and the modified Akaike Information Criterion (MAIC). Results obtained using these alternative approaches for the full sample available are reported in Table Three below, with significance at the 5% and 1% levels indicated by '*' and '**', respectively.

Table Three: Replication of LMN

Optimisation Rule	(A)DF test statistic	p
t-statistic (0.05)	−5.07 **	0
t-statistic (0.10)	−2.00	6
AIC	−3.80 *	1
SIC	−3.80 *	1
MAIC	−5.07 **	0

It can be seen that while the modified AIC results in selection of the same lag length as the sequential t-test statistic rule at the 5% level of significance – namely, no lagged difference terms – other methods select different lag lengths. The SIC and AIC both choose one lag, while the t-statistic rule at the 10% level selects six lags. The key finding here is that the use of a longer lag length generates a test statistic that rejects the unit root null hypothesis at conventional significance levels. This highlights how lag augmentation affects the effective sample size used in empirical analysis, raising an issue that supplements standard discussions of lag selection in relation to serial correlation.

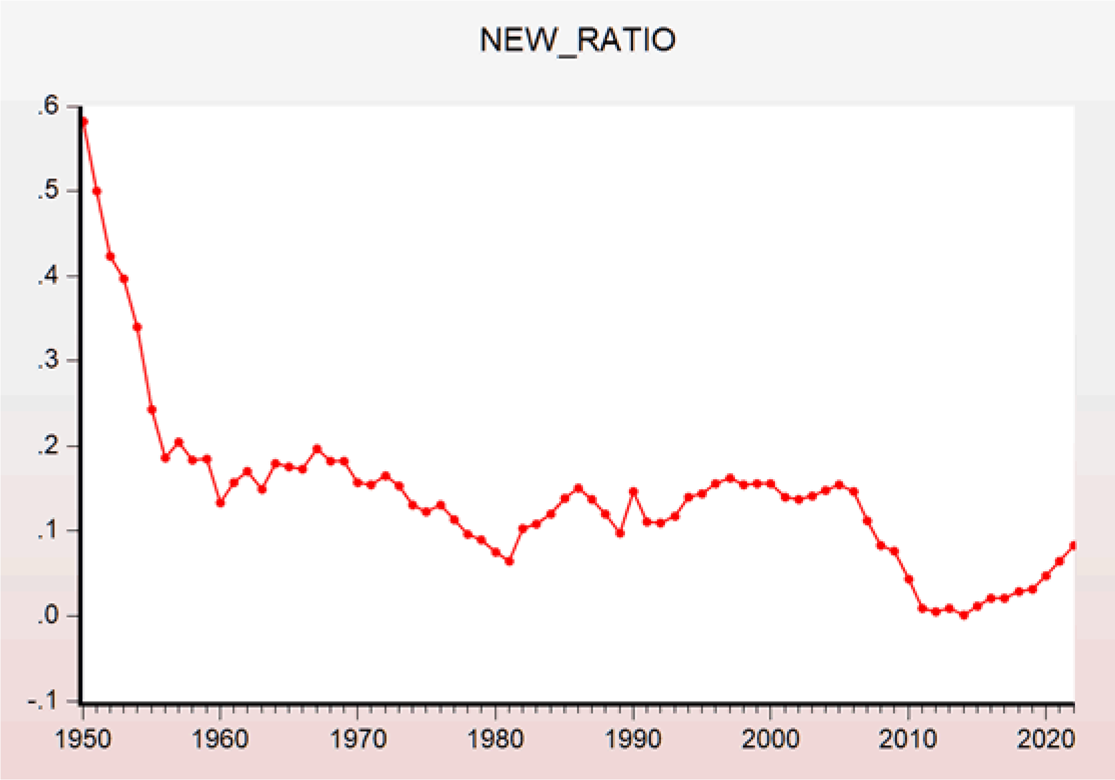
Given the restricted availability of the original data, the above results are not intended to support a data-based approach to the analysis of LMN– that is, learners are not expected to reproduce these results themselves, as the data are accessible via the Maddison text. Instead, the material is provided to support classroom discussion and to develop understanding in various ways. Here, a range of options is available. The output in Tables One and Two can be related to the test statistics reported in LMN to reinforce understanding of the analysis presented in the paper. Equation (1) can be considered alongside these results, as well as extended output containing complete estimation results for the ADF testing equations, to develop a clearer grasp of the underlying mechanics. Awareness of sample dependence and structural change can be deepened by comparing the plot of the series with the contrasting inferences in Tables One and Two. Table Three can be used to support discussion of alternative lag optimisation techniques and their effects on both inferences and the effective sample size used in estimation.

While these materials can be used in several ways to support discussion of LMN and its insights into unit root testing, the focus of this case study lies in a data-based re-evaluation of LMN's findings using revised data and an extended sample period. This is the subject of the following section.

3. Using revised data

The empirical findings of LMN can be replicated using a more recent vintage of data. The data employed here are taken from the 2023 Maddison Project Database (Bolt and van Zanden, 2024).[5] These series are not only revised relative to those used by LMN but also extend to 2022, thus allowing consideration of an extended sample period. The required per capita GDP ratio, constructed using these revised data, is labelled 'NEW_RATIO' and is presented in Figure Two below.

Figure Two: Revised data



Using the above data, four sets of results can be obtained. To assess the influence of data revision, this later vintage can be used to replicate results over the 1950-1994 and 1957-1994 sample periods originally considered by LMN. In addition, the availability of a longer data series allows for the construction of extended 'break-including' and 'post-break' samples covering 1950-2022 and 1957-2022, respectively. The results obtained from applying the ADF test to these four samples using the approach followed by LMN towards the specification of the ADF testing equation (i.e. including an intercept and trend as deterministic components and augmentation using the sequential t-statistic rule at the 5% level of significance) are reported in Tables Four to Seven below.

Table Four: Replicating LMN with revised data (1950-1994)

Null Hypothesis: RATIO has a unit root Exogenous: Constant, Linear Trend Lag Length: 0 (Automatic - based on t-statistic, lagpval=0.05, maxlag=9)			
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		−4.954714	0.0012
Test critical values:	1% level	−4.180911	
	5% level	−3.515523	

	10% level	−3.188259	

Table Five: Replicating LMN with revised data (1957-1994)

Null Hypothesis: RATIO has a unit root Exogenous: Constant, Linear Trend Lag Length: 0 (Automatic - based on t-statistic, lagpval=0.05, maxlag=9)			
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		−2.408953	0.3694
Test critical values:	1% level	−4.219126	
	5% level	−3.533083	
	10% level	−3.198312	

Table Six: Replicating LMN with revised data (1950-2022)

Null Hypothesis: RATIO has a unit root Exogenous: Constant, Linear Trend Lag Length: 0 (Automatic - based on t-statistic, lagpval=0.05, maxlag=11)			
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		−5.938186	0.0000
Test critical values:	1% level	−4.090602	
	5% level	−3.473447	
	10% level	−3.163967	

Table Seven: Replicating LMN with revised data (1957-2022)

Null Hypothesis: RATIO has a unit root Exogenous: Constant, Linear Trend

Lag Length: 0 (Automatic - based on t-statistic, lagpval=0.05, maxlag=10)			
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		−1.944131	0.6202
Test critical values:	1% level	−4.103198	
	5% level	−3.479367	
	10% level	−3.167404	

With the updated data, classroom activities can centre on reproducing the results in the four tables and interpreting the inferences. Consequently, discussions can move beyond an initial focus on unit root testing and structural change, as considered by LMN, to explore the implications of data revision (see, inter alia, Cook, 2008; Croushore and Stark, 2003; Egginton et al., 2002; Garratt and Vahey, 2006; Mankiw et al., 1984; Mankiw and Shapiro, 1986; Patterson and Heravi, 1991). This, in turn, enables appreciation of the evolving nature of research, where findings can be revisited and extended through the use of new data or alternative methods.

4. Conclusion

This case study has explored the use of R&R as a means of incorporating research into the teaching of econometrics. Rather than simply duplicating the results of a published article using the original data, the focus here has been on re-evaluating the findings of LMN using new data that are revised and extend over a longer sample period. By adopting this approach, the study creates an active learning framework that exposes learners to the progressive and evolving nature of econometric modelling where findings can be revisited as new data become available. At the same time, the results based on the original data used by LMN are also presented to support classroom discussion in multiple ways. The suggested use of this material reflects a three-part structure that integrates core methodological content (e.g. the structure and mechanics of the approaches considered), lecturer-provided resources, and published research.

References

Bolt, J. and van Zanden, J. 2024. Maddison style estimates of the evolution of the world economy: A new 2023 update. *Journal of Economic Surveys* 39, 639-659. https://doi.org/10.1111/joes.12618

Cook, S. 2008. Cross-data-vintage encompassing. *Oxford Bulletin of Economics and Statistics* 70, 849-865. https://doi.org/10.1111/j.1468-0084.2008.00533.x

Cook, S. and Watson, D. 2025. Incorporating research in the teaching of undergraduate econometrics. In The Handbook for Economics Lecturers. Economics Network. https://doi.org/10.53593/m4412a

Croushore, D. and Stark, T. 2003. A real-time data set for macroeconomists: Does the data vintage matter?, *Review of Economics and Statistics*, 85, 605-617. https://doi.org/10.1162/003465303322369759

Dickey, D. and Fuller, W. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74, 427-431. https://doi.org/10.1080/01621459.1979.10482531

Egginton, D., Pick, A. and Vahey, S. 2002. Keep it real!: A real-time UK macro data set. *Economics Letters* 77, 15-20. https://doi.org/10.1016/S0165-1765(02)00094-0

Garratt, A. and Vahey, S. 2006. UK real-time macro data characteristics. *Economic Journal* 116, F119-F135. https://doi.org/10.1111/j.1468-0297.2006.01067.x

Leybourne, S., Mills, T. and Newbold, P. 1998. Spurious rejections by Dickey-Fuller tests in the presence of a break under the null. *Journal of Econometrics* 87, 191-203. https://doi.org/10.1016/S0304-4076(98)00014-1

Maddison, A. 1995. Monitoring the world economy 1820–1992. Paris: OECD.

Mankiw, G., Runkle, D. and Shapiro, M. 1984. Are preliminary announcements of the money stock rational forecasts? *Journal of Monetary Economics* 14, 15-27. https://doi.org/10.1016/0304-3932(84)90024-2

Mankiw, G. and Shapiro, M. 1986. News or noise?: An analysis of GDP revisions. *Survey of Current Business* 66, 20-25.

Ng, S. and Perron, P. 1995. Unit root tests in ARMA models with data-dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association* 90, 268-281. https://doi.org/10.1080/01621459.1995.10476510

Patterson, K. and Heravi, S. 1991. Data revisions and the expenditure component of GDP. *Economic Journal* 101, 887-901. https://doi.org/10.2307/2233861

Perron, P. 1989. The Great Crash, the oil price shock, and the unit root hypothesis. *Econometrica* 57, 1361-1401. https://doi.org/10.2307/1913712

Footnotes

[1] Note that Leybourne et al. (1998) consider a single break which takes the form of either a break in level or break in drift.

[2] The prominent feature of the findings of Leybourne et al. (1998) is the oversizing, or spurious rejection, associated with the DF test when applied to unit root processes subject to structural change 'early' in the sample period considered. However, the simulation results provided also demonstrate under-sizing when breaks occur later in the sample.

[3] Note that application of the ADF test results in the use of a testing equation without any lagged differenced regressors- i.e. a DF test.

[4] Ng and Perron (1995) consider the sequential t-statistic rule at both the 5% and 10% levels of significance.

[5] These data are available from https://www.rug.nl/ggdc/historicaldevelopment/maddison/releases/maddison-project-database-2023

Views on request

THE ECONOMICS NETWORK

Improving economics teaching
and learning for over 20 years

Replication and Reproduction III: Holmes & Otero (2019), unit root testing of differentials and cointegration analysis

Steve Cook
Swansea University
s.cook at swan.ac.uk
and Duncan Watson
University of East Anglia
Duncan.Watson at uea.ac.uk

PUBLISHED
November 2025

https://doi.org/10.53593/n4411a

CONTENTS

This case study is the third in a set of materials on the effective incorporation of research in undergraduate econometrics edited by Peter Dawson of the University of East Anglia.

Summary

Cook and Watson (2025a) have recently championed the benefits of using reproduction and replication (R&R) as a means of incorporating empirical research in the teaching of econometrics. By requiring close engagement with published findings, they argue that

R&R offers a number of pedagogical benefits. This study provides an illustration of R&R in practice by drawing upon Holmes and Otero's (2019, *Energy Economics*) analysis of crude oil spot and futures prices. Using the data from this study, unit root and cointegration analysis are applied and discussed in the course of reproducing published findings.

1. Introduction

Building on the approach set out in Cook and Watson (2025a), the recent case studies of Cook and Watson (2025b, 2025c) have demonstrated how replication and reproduction (R&R) can be used to embed research directly within the teaching of undergraduate econometrics. This study extends that work by shifting the focus from unit root analysis of individual series to the examination of relationships between series – specifically, through the unit root testing of ratios, or differentials, and subsequent cointegration analysis. As in the earlier case studies of Cook and Watson (2025b, 2025c), the analytical framework is anchored in the R&R of a published study, with Holmes and Otero (2019, hereafter HO) providing the empirical context here. Rather than simply presenting econometric techniques, the R&R approach employed challenges students to actively reproduce published findings, thereby fostering deeper understanding of both methodology and interpretation.

2. Reproducing Holmes & Otero (2019)

HO provide an analysis of crude oil spot and futures prices. Their study examines a range of series, considering different maturities for futures contracts as well as alternative data frequencies (daily, weekly, and monthly). For the non-daily observations, both average and end-of-period figures are used, thus further increasing the number of series considered. These series are available from the supplementary data appendix provided with the original paper.[1] To illustrate how this paper can serve as a resource for integrating research into teaching, the present study focuses on daily data for two of the series considered by HO, namely, spot prices and 4-month futures prices. Expressed in natural logarithmic form, these series are denoted here as SPOT and CL4 respectively.

The augmented Dickey-Fuller (1979, hereafter referred to as ADF) test results reported for these series in Table 3 of HO (p. 231) are reproduced in Tables One and Two below. As Cook and Watson (2025a) note, reproducing empirical findings encourages closer engagement with the original research. Here, for example, the structure of the testing equation – as presented in equation (1) below – requires careful consideration, including discussion of its components and their roles. Similarly, the construction of the test statistic – as shown in equation (2) – can be examined in relation to the testing equation itself. The approach to augmentation adopted by HO, which uses the Schwarz Information Criterion (SIC) with a maximum lag of 30, is employed to reproduce their findings. However, this offers a basis for discussion of alternative approaches employing, for example, sequential t-testing or the Akaike Information Criterion (AIC), as well as use of the Schwert (1989) rule, to determine the maximum lag length.[2]

(1) $$\Delta y_t = \alpha + \beta t + \phi y_{t-1} + \sum_{j=1}^{p} \gamma_j \Delta y_{t-j} + e_t$$

(2) $$\tau_\tau = \frac{\hat{\phi}}{s.e.(\hat{\phi})}$$

Table One: Replicating SPOT unit root test results

Null Hypothesis: SPOT has a unit root Exogenous: Constant, Linear Trend Lag Length: 0 (Automatic - based on SIC, maxlag=30)			
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		−2.654313	0.2560
Test critical values:	1% level	−3.959311	
	5% level	−3.410428	
	10% level	−3.126974	

Table Two: Replicating CL4 unit root test results

Null Hypothesis: CL4 has a unit root Exogenous: Constant, Linear Trend Lag Length: 1 (Automatic - based on SIC, maxlag=30)			
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		−1.923580	0.6419
Test critical values:	1% level	−3.959311	
	5% level	−3.410428	
	10% level	−3.126974	

The results in Tables One and Two lead to non-rejection of the unit root hypothesis at conventional significance levels for both SPOT and CL4. HO then proceed to examine the long-run relationship between spot and futures prices. This is explored through unit root testing of spot-to-futures differentials (or ratios) and the application of the Johansen (1988) procedure to assess potential cointegration. The differential between the SPOT and CL4 series is constructed and referred to here as ‘DIFF’. The ADF test results for this series are reported in Table Three and reproduce the findings of HO (p. 232, Table 4) for this SPOT:CL4 differential.

Table Three: Replicating results for the differential

Null Hypothesis: DIFF has a unit root Exogenous: Constant Lag Length: 19 (Automatic - based on SIC, maxlag=30)			
		t-Statistic	Prob.*
Augmented Dickey-Fuller test statistic		−5.159380	0.0000
Test critical values:	1% level	−3.431083	
	5% level	−2.861749	
	10% level	−2.566923	

Again, this analysis can be used to challenge and deepen understanding of the application and interpretation of the ADF test. For example, reproducing the test statistic requires careful attention to the approach to augmentation and the inclusion of deterministic terms. Reviewing the full set of results from the ADF test also enables reproduction of the value of the estimate of the coefficient on the lagged differential regressor reported in the original study. These findings can be drawn upon for classroom discussion alongside other issues such as the use of alternative augmentation methods and their effects upon inferences, and the implications of detecting stationarity in a ratio constructed from non-stationary series.

Turning to the consideration of cointegration, the findings of HO for our chosen series are again reproduced – see Tables Four and Five below and Tables 5 and 6 in HO (pp. 233-234). This replication requires learners to engage closely with the original research, supporting the development of a deeper understanding of the Johansen (1988) procedure. A range of issues can be explored through this application, including: the nature and role of the underlying vector autoregressive (VAR) model; the structure and purpose of the vector error correction model (VECM); the impact of alternative information criteria; the connection between unit root testing of the differential and the Johansen analysis of the underlying series; the use of the Johansen approach in bivariate settings; comparisons with the Engle-Granger (1987) procedure; and the interpretation of non-zero eigenvalues in identifying cointegrating relationships.

Table Four: Reproducing Trace statistic results

Unrestricted Cointegration Rank Test (Trace)				
Hypothesized No. of CE(s)	Eigenvalue	Trace Statistic	0.05 Critical Value	Prob.** Critical Value
None *	0.009528	70.54988	15.49471	0.0000
At most 1	0.000232	1.669685	3.841465	0.1963

Trace test indicates 1 cointegrating equation(s) at the 0.05 level

* denotes rejection of the hypothesis at the 0.05 level

**MacKinnon-Haug-Michelis (1999) p-values

Table Five: Reproducing Maximum eigenvalue test results

Unrestricted Cointegration Rank Test (Max-eigenvalue)				
Hypothesized No. of CE(s)	Eigenvalue	Max-Eigen Statistic	0.05 Critical Value	Prob.** Critical Value
None *	0.009528	68.88020	14.26460	0.0000
At most 1	0.000232	1.669685	3.841465	0.1963

Max-eigenvalue test indicates 1 cointegrating equation(s) at the 0.05 level

* denotes rejection of the hypothesis at the 0.05 level

**MacKinnon-Haug-Michelis (1999) p-values

3. Conclusion

This study demonstrates how reproduction and replication can do more than reinforce econometric technique – they can transform the way students engage with research. By working through the empirical structure of HO, learners are exposed to the practical application of unit root testing, the construction and interpretation of differentials, and the logic of cointegration through the use of the Johansen procedure. Each stage prompts methodological reflection: from lag selection and deterministic components to the meaning of stationarity and long-run equilibrium.

Crucially, the study equips students to move beyond formulaic application and to develop a clearer sense of how econometric decisions shape economic conclusions. For lecturers, it offers a ready-made resource that connects abstract concepts with real-world data and published research, encouraging replication not just as a technical exercise but as a tool for critical learning.

References

Cook, S. and Watson, D. 2025a. From provision to understanding: The effective incorporation of research in undergraduate econometrics. Economics Network Handbook for Economics Lecturers. https://doi.org/10.53593/m4412a

Cook, S. and Watson, D. 2025b. Replication and Reproduction I: Leybourne (1995, *Oxford Bulletin of Economics and Statistics*) and the maximum Dickey-Fuller test. Economics Network Ideas Bank. https://doi.org/10.53593/n4409a

Cook, S. and Watson, D. 2025c. Replication and Reproduction II: Leybourne *et al.* (1998, *Journal of Econometrics*) and the Dickey-Fuller test in the presence of breaks under the null. Economics Network Ideas Bank. https://doi.org/10.53593/n4410a

Dickey, D. and Fuller, W. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association* 74, 427-431. https://doi.org/10.1080/01621459.1979.10482531

Engle, R. and Granger, C. 1987. Co-integration and error correction: Representation, estimation and testing. *Econometrica* 55, 251-276. https://doi.org/10.2307/1913236

Holmes, M. and Otero, J. 2019. Re-examining the movements of crude oil spot and futures prices over time. *Energy Economics* 82, 224-236 https://doi.org/10.1016/j.eneco.2017.08.034

Johansen, S. 1988. Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* 12, 231-254. https://doi.org/10.1016/0165-1889(88)90041-3

MacKinnon, J. G., Haug, A. and Michelis, L. 1999. Numerical distribution functions of likelihood ratio tests for cointegration. *Journal of Applied Econometrics* 14, 563-577. https://doi.org/10.1002/(SICI)1099-1255(199909/10)14:5<563::AID-JAE530>3.0.CO;2-R

Schwert, G. 1989. Tests for unit roots: A Monte Carlo investigation. *Journal of Business and Economic Statistics* 7, 147-159. https://doi.org/10.1080/07350015.1989.10509723

Footnotes

[1] https://www.sciencedirect.com/science/article/pii/S0140988317303018?via%3Dihub#s0030

[2] Note that the sample-based Schwert rule leads to a maximum lag length of 34 for the sample considered, rather than the value of 30 employed by HO.