Quantitative Macro: Lessons Learnt from Fourteen Replications

Robert Kirkby*

October 20, 2021

Abstract

I replicate all tables and figures from fourteen papers in Quantitative Macroeconomics, with an emphasis on incomplete market heterogeneous agent models. I report three main findings: (i) all (non-welfare related) major findings of the papers replicate, (ii) welfare findings based on linear approximation methods —1st-order perturbation, linear and log-linearization around steady-state, and linear-quadratic methods— should be treated as quantitatively suspect, (iii) decisions around methods for discretizing exogenous shocks have a large and unappreciated influence on results and should be prominently discussed in papers. While some smaller aspects of the papers do not replicate exactly, rather than nitpick in the body of this paper I instead describe some lessons learnt that may be useful for practitioners working with Quantitative Macroeconomic models. The replications use global methods allowing for non-linearities and I argue that these are important and need to be more widely used I provide a checklist that researchers can use when trying to check that their work will be more easily reproducible. Matlab codes implementing the replications using the VFI Toolkit are provided, and full results of all replications are given in the online appendix. I conclude with three core points for best practice: (i) codes be made directly available (e.g., on github, not only 'on request', and not just inside a zip file), (ii) report not just baseline parameters but also hyperparameters, equilibrium values, non-baseline parameters and initial conditions, and (iii) replication means rewriting codes from scratch, not just re-running available codes.

Keywords: Quantitative Macroeconomics, Numerical methods, Replication.

JEL Classification: E00; C68; C63; C15

*Thanks to the following who answered emails of questions about replicating their papers, often providing code or comments: Javier Díaz-Giménez, Mark Huggett, Ayse Imrohoroglu, Selahattin Imrohoroglu, Guido Lorenzoni, Carlos Urrutia. This is essentially a full list of those contacted; if people dislike replication of their work such mentality is not present among Macroeconomists. Further thanks to the following for helpful conversations and feedback, about specific replications or about replication in general: Chris Carroll, Alessandro Di Nola, Chris Edmond, Denise Manfredini, and Nathan Palmer. Special thanks to Shutao Cao for patiently listening to my innumerable issues around the minutiae of papers which I was struggling to replicate; most often due to errors on my part. Thanks to Rāpoi at Victoria University of Wellington for use of their computing facilities. Kirkby: Victoria University of Wellington. Please address all correspondence about this article to Robert Kirkby at <robertdkirkby@gmail.com>, robertdkirkby.com.

Contents

Imitation is the sincerest [form] of flattery.
— Colton, Charles Caleb (1824)

I quasi-replicate a number of classic papers in Quantitative Macroeconomics. The replications are quasi-replications in two senses: I do not attempt to use the same numerical methods to solve the model as the original authors, and I (only) replicate all figures and tables relating to the model.[1] My interest is not in nitpicking about where the original papers report a 'wrong number' (whether due to typo, coding error, etc.), and for this reason I relegate all the actual replicated tables and figures to the appendix. The focus of this paper is instead on the lessons to be learned from these replications and on providing some suggestions for best practice based based on the experience of performing the replications.

My main finding is that there is no replication crisis in the Macroeconomics of Quantitative Macroeconomics, but there is a minor crisis in the Quantitative. By this I mean that the major conclusions from the all the papers replicated are unchanged, but most of the papers contain some numbers that are incorrect by a magnitude that is quantitatively important.[2]

Replication is typically thought of as relating to data and statistics. So why replicate computational results from Quantitative Macroeconomics? The main reason is the exact same reason underlying the importance of replication to data and statistics: establishing the reliability of existing results. The need to do so follows directly from thinking of computational models as a form of laboratory in which we run experiments (Bona and Santos, 1997). A secondary use for replications follows as Economists often learn to write code by solving existing models and replication provides the needed reliable solutions for this.[3] If anything, simple mistakes may be more common when computing Quantitative Macroeconomic models than in other parts of Economics as they depend not only on using data and statistics but also require substantial coding. An additional reason is to understand the influence on Macroeconomics of the choice of which numerical methods are used to solve the model: I document some interesting examples of the importance of this.

Table 1 provides a list of the papers I replicate with a focus on general equilibrium heterogeneous agent models with incomplete markets. In defense of the selection I simply note that the replicated papers are well cited with a mean number of citations of 394 in Ideas Repec and of 1167 in Google

[1] Two main aspects of the papers therefore remain unreplicated: any tables or figures relating purely to the empirical data, and any results reported in the text but without appearing in any table or figure.

[2] This does not mean all papers in Quantitative Macroeconomics replicate. Two examples: Hatchondo, Martinez, and Sapriza (2010) show that some important, but not the main, findings of the sovereign debt papers of Aguiar and Gopinath (2006) and Arellano (2008) fail to replicate; they were numerical error. Takahashi (2014) shows that main finding of Chang and Kim (2007) fails to replicate as it was numerical error (reply of Chang and Kim (2014)).

[3] I personally become interested in the issues of numerical error and replication after a 'lost week' spent trying to understand why, when first learning to solve heterogeneous agent models, my codes would not replicate the results of the bottom right corner of Table 2 of Aiyagari (1994), something I now know is because the originals contained numerical error.

Table 1: Papers Replicated

Paper	Horizon	Eqm	Other
Hansen (1985)	∞	R.A.	
Imrohoroglu (1989)	∞	Partial	
Díaz-Giménez, et al (1992)	∞	Partial	
Hopenhayn & Rogerson (1993)	∞	GE	Entry-Exit
Huggett (1993)	∞	GE	
Aiyagari (1994)	∞	GE	
Hubbard, Skinner & Zeldes (1994)	80	Partial	Panel Data
Imrohoroglu, Imrohoroglu & Joines (1995)	65	GE	
Huggett (1996)	79	GE	
Conesa & Krueger (1999)	66	GE	Transition
Castaneda, Díaz-Giménez & Ríos-Rull (2003)	∞	GE	Inequality
Restuccia & Urrutia (2004)	2-∞	GE	Intergen.
Restuccia & Rogerson (2008)	∞	GE	Entry-Exit
Guerrieri & Lorenzoni (2017)	∞	GE	Transition

RA=Representative Agent General Eqm. GE=General Equilibrium.
Intergen.=Intergenerational linkages (in 2-period OLG).

Scholar as of early 2021 (and a minimum number of citations of 80 and 224, respectively). The codes implementing these replications are all available at github.com/vfitoolkit/vfitoolkit-matlab-replication. Note that this covers a range of 'model-types' including partial and general equilbrium; finite and infinite horizon, including overlapping generations; stationary equilibrium and transition paths; and agent entry and exit. And involves analysing a variety of model 'outputs' including time-series properties, cross-sectional distributions, aggregates, and panel-data.

Replication of these papers was performed using discretized value function iteration with simulations and agents distributions computed on discretized state space; these methods have known reliable convergence properties to the true solution under conditions that are applicable to a broad class of Macroeconomic models (Kirkby, 2017a, 2019) as well as performing well on accuracy in comparisons with other methods (Aruoba et al., 2006; Santos, 2000; Peralta-Alva and Santos, 2014) as long as sufficiently large grids are used.[4] These discretized grid methods would be inappropriate for the solution of state-of-the-art models where a trade-off between speed-and-accuracy has to be made. For replication however the appropriate focus is on accuracy and robustness at the expense of speed. Discretized grid methods combine a high accuracy, as long as large grids are used, with known convergence properties and a robustness to a wide range of model properties. While it is impossible to know for certain that the solutions of the replications given here are the true solutions I am confident that the replicated solutions are accurate as the answers given are insensitive to the grid sizes used; this 'insensitivity' is given a precise meaning below.[5]

Implementation of the replications makes use of the VFI Toolkit for Matlab (Kirkby, 2017b),

[4]To be precise what matters is not having a large size of the grid, but having small spacings between grid points.

[5]Reassurance may be taken from the replication of Imrohoroglu (1989): the original author Ayse Imrohoroglu has since run a second replication joint with Kanika Aggarwal and got the same results as the replication reported here (communication by email from Ayse Imrohoroglu).

which has the advantage that the outputs of most functions that make up the codes involved in the replications have been widely tested and hopefully therefore less likely to contain errors.[6]

Table 2 shows for each replication the quantiles of the percentage difference between the replication and original results.[7] It is based on all the entries of all the Tables from each paper. The absolute percentage difference between the replication value and the value in the original paper was calculated for every table entry, and the quartiles of these are reported. The main weakness of this is that it obviously misses any Figures. To ensure that the replication results are not driven by numerical error the replications were required to pass the test that a 'substantial'[8] increase in the grid size results in a change in the upper quartile of the absolute percentage difference was less that 5% between the results of the two replications (grid and substantially increased grid); note that this is much stricter than it first sounds as, e.g., many papers contain numbers like 0.1, so if this changed to 0.11 with the substantially increased grid this would be a change of greater than 5%. As a result is believed that the replication numbers are the accurate numbers however this cannot be known for certain as, e.g., a parameter that should be set to 2.4 could instead be set to 2.6 due to a typo. Comparison of the measure across papers should be taken as illustrative rather than definitive as papers that provide, e.g., greater breakdown of statistics across different subpopulations, will somewhat naturally be likely to display greater numerical error.

The only 'substantial' failure to replicate is the welfare results of early papers. This appears to be explained by the use of linear-quadratic methods, while we use non-linear methods to solve the models. For papers such as Imrohoroglu (1989) and Díaz-Giménez, Prescott, Alvarez, and Fitzgerald (1992) the methods used solved the policy function with enough accuracy that their findings on model statistics related to policies and stationary distributions replicate fine. However those same methods led to highly inaccurate welfare evalutations as the value functions were not accurately computed. This finding is not entirely novel, but it's importance is widely underappreciated. Kim and Kim (2007) show that 1st-order approximation methods deliver incorrect welfare results if even when using the correct (to 1st-order) optimal policies (although these can be largely avoided by putting the 1st-order solution into the unapproximated welfare function), while Judd, Maliar, and Maliar (2017) show further that 1st-order solution methods are simply incorrect for many Macroeconomic models, deriving *minimum* error bounds that are large enough to be troubling. I conjecture that this problem, inaccurate welfare results, is likely widespread in early Quantitative Macroeconomics papers and recommend that any welfare result from pre-2000 should be treated as quantitatively suspect until replicated. The continued widespread use of linear-quadratic methods

[6]Coding errors are a genuine concern. A recent replication by Bédécarrats, Guérin, Morvant-Roux, and Roubaud (2019) of Crepon, Devoto, Duflo, and Parienté (2015), an empirical analysis of a field experiment, found numerous coding errors, and that analysis would likely have contained way less lines of code than most quantitative Macroeconomics papers.

[7]Thanks to a referree for advising the addition of such a measure.

[8]Appendix 4 reports the grid and substantially larger grid for all the papers. Because many models have, e.g., a two-element grid representing employment and unemployment not all grids can be increased and so a general definition of 'substantial' is not attempted. All results reported in this paper are based on the substantially larger grids.

Table 2: Percentage Difference between Numbers in Replication and Original Paper

Paper	Absolute value of percentage difference	
Hansen (1985)	Lower Quartile	0%
	Median	2.9 %
	Upper Quartile	9.7%
Imrohoroglu (1989)	Lower Quartile	0.2%
	Median	15.1 %
	Upper Quartile	72.5%
Díaz-Giménez, Prescott, Alvarez & Fitzgerald (1992)	Lower Quartile	4.8%
	Median	32.9%
	Upper Quartile	105.9%
Hopenhayn & Rogerson (1993)	Lower Quartile	0.8%
	Median	13.1%
	Upper Quartile	32.0%
Huggett (1993)	Lower Quartile	0.1%
	Median	1.5%
	Upper Quartile	13.4%
Aiyagari (1994)	Lower Quartile	0.9%
	Median	2.4%
	Upper Quartile	7.0%
Hubbard, Skinner & Zeldes (1994)	Lower Quartile	9.7%
	Median	65.6%
	Upper Quartile	118.8%
Imrohoroglu, Imrohoroglu & Joines (1995)	Lower Quartile	0.7%
	Median	2.3%
	Upper Quartile	10.5%
Huggett (1996)	Lower Quartile	11.6%
	Median	26.4%
	Upper Quartile	48.5%
Conesa & Krueger (1999)	Lower Quartile	2.8%
	Median	10.8%
	Upper Quartile	43.5%
Castaneda, Díaz-Giménez & Ríos-Rull (2003)	Lower Quartile	2.3 %
	Median	5.2%
	Upper Quartile	15.3%
Restuccia & Urrutia (2004)	Lower Quartile	9.6%
	Median	25%
	Upper Quartile	54.9%
Restuccia & Rogerson (2008)	Lower Quartile	0%
	Median	0%
	Upper Quartile	2.9%
Guerrieri & Lorenzoni (2017)		Paper contains no Tables. Just Figures. (expect one table of parameter values)

Note: For all the entries of all the Tables from each paper: the absolute percentage difference between the replication value and the value in the original paper is calculated, then the quartiles of these are calculated. When both the relication and original values are zero this is considered to be a zero absolute percentage difference. Values of parameters and other things that are 'impossible' not to perfectly replicate are omitted from the calculations. All figures are omitted from this measure. The top decile was calculated but it is heavily influenced by, e.g., errors where the replication value is 0.003 and original is 0.001, which seems misleading and so is not reported. There were some 'Nonzeros' for which the replication is non-zero and the original is zero, the number of non-zeros were: 2 in Hansen (1985), 1 in Hopenhayn & Rogerson (1993), 1 in Aiyagari (1994), 12 in Hubbard, Skinner & Zeldes (1994), 1 in Imrohoroglu, Imrohoroglu & Joines (1995).

in Ramsey optimal policy where maximizing the welfare function is part of the computational exercise leaves some major open questions about the results of that literature until replication studies are undertaken in that area. Loosely related, first-order (and second-order) perturbation methods have also been shown to give incorrect solution to the Diamond-Moretensen-Pissarides model of search-labor markets (Piccione and Rubinstein, 2007).[9]

One topic that requires much greater discussion in Quantitative Macroeconomics papers is the discretiziation of shocks.[10] Many papers contain a substantial discussion of calibration and some robustness exercises to parameter values. The choice of discretiziation method by contrast rarely warrants more than a passing mention, often in a footnote, despite being vastly more important in most models than many parameters. In practice the discretization choices play a key role in determining income risk and the distributions of earnings and wealth. More subtle is the relationship between the exogenous shocks and market incompleteness. Note that in most incomplete market models the incompleteness arrives precisely because there are no assets with returns that span the space of idiosyncratic shocks. Hence when the idiosyncratic shocks are small the markets are largely complete, while when idiosyncratic shocks are large markets are very incomplete. The discretization of exogenous shocks, because it determines both the riskiness and range of the idiosyncratic shocks is therefore also determining the degree of market incompleteness that distinguishes heterogeneous agent models from representative agent models. Quantitative Macroeconomic papers would be much improved by treating these choices of shock discretization to the same level of discussion, analysis and sensititivity as any other modelling decision. As an example of their importance Guerrieri and Lorenzoni (2017) use the Tauchen method to discretize an AR(1) shock in a study of the credit crisis that followed the Great Financial Crisis of 2007.[11] Just changing the hyperparameter of the Tauchen method to other reasonable values can cause the zero-lower bound on interest rates to bind for decades, rather than the few years in the baseline model (and seen in reality).

Replication in Economics: Controversy about replication has raged in Psychology where a project by the Open Science Collaboration to repeat one hundred influential studies was able to successfully replicate the original results in only around 40% of cases (Collaboration, 2015).[12] Closer to home for Economists have been controversies about the results of Reinhart and Rogoff on the relationship between government debt and economic growth, and Miguel and Kremer (2004) on the

[9]The 'appropriate' level of approximation will always be context dependent. Our point here is that the level of approximation resulting from 1st-order pertubations, log-linear approximations, and linear-quadratic return functions is inappropriate for most dynamic stochastic Economic models. These methods create 'economically significant' numerical approximation error in model outcomes that are of interest.

[10]Numerical quadrature methods are standard for evaluating integrals/expectations, and discretizing shocks is required as part of this. Alternatives exist, like Monte-Carlo integration, but are rarely used as they are too slow.

[11]Specifically, Guerrieri and Lorenzoni (2017) use the Tauchen-Hussey method which is a combination of the Tauchen method with a specific formula for selecting the hyperparmeter to match the second moments.

[12]An EconTalk podcast on the study may interest readers. 'Around 40%' refers to the passage from the abstract stating that: '39% of effects were subjectively rated to have replicated the original result'. A similar effort now underway in Economics can be found at replicationnetwork.com

effects of de-worming on education in Kenya.[13] Within the field of lab experiments in Economics Camerer et al. (2016) try to replicate 18 studies published in American Economic Review and Quarterly Journal of Economics during 2011-2014, and conclude that replication is successful in 60-80% of the papers (depending on exact metric of 'success'). In a related study Dreber et al. (2015) find that prediction markets in which people can bet on which replications will succeed and fail did well in sense that when they predicted a replication would fail it did (when prediction markets predicted that the replication would succeed this was largely unrelated to outcome of replication); this suggests that informally the profession is aware of certain existing results that are unlikely to replicate. Ferraro and Shukla (2020) provide evidence that suggests empirical environmental economics suffers many of these issues and suggest a variety of was the profession might adapt and improve.

While replication is important it is not a panacea for all problems.[14] Even papers that were retracted due to known error continue to be cited; 20,000 articles listed as retracted by Replication Watch were still cited 85,000 times after retraction. Other loosely related issues include the bias of publication to only publish statistically signicant results (Brodeur et al., 2016). The problems don't just lie with the studies themselves, newspapers rarely report on null-findings and rarely do follow-ups to reporting on results that fail to hold in reproduction studies (Dumas-Mallet et al., 2017).[15] Replications are also often potentitaly difficult, expensive, and time-consuming: a recent effort to replicate 50 papers studying Cancer, with a budget in excess of $1.3 million, ended up replicating just 18. Certainly, the replications in the present paper consumed a lot of time.

We are not aware of any existing replication study in Quantitative Macroeconomics (beyond the two or three individual replications mentioned in the introduction). The closest is Chang and Li (2015) who look at research transparency or 'the basic goal of computational reproducibility' (in the words of Miguel (2021)). They take a very different approach and rather than try to replicate

[13] Reinhart and Rogoff originally argued that there was a cut-off for Government debt of around 90%-of-GDP, below which there was little relationship with economic growth and above which there was a strong negative correlation; but the statistical significance of the specific 90%-of-GDP cut-off was shown to be due to Excel error (Herndon, Ash, and Pollin, 2013); the broader negative correlation holds, only the cut-off failed to replicate. Miguel and Kremer (2004) argued, based on a randomized controlled trial, that de-worming of children in Kenya had large positive effects on school attendance and educational outcomes. Two studies, one a replication and another a re-analysis questioned some aspects of the results. At the end of the day the results of the original study appear to stand-up well (more: links, short video).

[14] For example it will not detect data fraud, which while very rare does occur; e.g Evidence of Fraud in an Influential Field Experiment About Dishonesty describes the use of fraudelent data by Shu, Mazar, Gino, Ariely, and Bazerman (2012); note that the paper reports multiple experiment results and the only one which uses fraudelent data which was performed by Ariely & Mazar. The original findings has already been rejected by a followup paper which lead to all authors of the original revising their views. Of course, it is possible the researchers simply received 'faulty' data from the company and did not notice; Facebook accidently gave researchers erroneous data via their Social Science One project which was spotted by one of the many research groups using it.

[15] "This year, a study looked at how newspapers reported on research that associated a risk factor with a disease, both lifestyle risks and biological risks. For initial studies, newspapers didn't report on any null findings, meaning those that had results without expected outcomes. They rarely reported null findings even when they were confirmed in subsequent work. Fewer than half of the "significant" findings reported on by newspapers were later backed by other studies and meta-analyses. Most concerning, while 234 articles reported on initial studies that were later shown to be questionable, only four articles followed up and covered the refutations. "(source)

the results of the papers, their interest is instead whether the original authors of the papers supply codes and when they do whether these codes can simply be run to computationally reproduce the results of the papers. A similar approach is taken by Gertler et al. (2018) who find that in 203 papers from top Economics journals while many provide code only in 37% of cases did it actually run, and in only 14% of cases was there both raw data and the code that generates the papers results (tables and figures) from this data. These approaches are in line with the AEA (American Economic Association) Code and Data Policy,[16] although the interest of the AEA policy is about ensuring that a study is reproducable, rather than whether a study been replicated. While subtle, the distiction is important as reproducable can be thought of as true even though the code or data-treatment contains errors and would fail to replicate; that original code runs and reproduces tables and figures in no way tests for the existence of errors in the code itself although it does make it much easier to detect and resolve them.

This current approach to replication in Quantitative Economics with it's focus on reproducibility obviously misses any issues of whether the original results were themselves correct, which is the main purpose of replication. While availability of code is important reproducibility is *not* replication. Replication necessarily involves writing new code as simply running existing codes includes replicating all the errors made in the original when treating the data and writing the code. Availablility of code is important because code often contains information unintentionally missing from a published paper. For example, papers simply forget to state some initial condition, or the weights used during calibration, or the formula for a certain moment, or parameter values of a counterfactual exercise, etc.

Zimmermann (2015) suggests the need for a Journal of Replication in Economics as a way to overcome the current status quo in which academics typically recieve little to no recognition or reward for performing replications. The area of Applied Econometrics is ahead in this area with the Journal of Applied Econometrics having a Replication Section since 2003. An online effort by ReplicationWiki, hosted by the University of Göttingen, aims to provide a clearinghouse for replications, on the assumption that people already perform replications and simply need some outlet for them. Nor can citations nor a large following literature be relied on as a substitute for replication: oestrogen receptor cycling in the field of breast cancer research was built on two papers each of which had more than 1000 citations over nearly 20 years, but has now been found to be completely incorrect with neither of the original papers being replicable (Holding, 2019). Christensen, Freese, and Miguel (2019) is a recent book that describes many of these issues, problems, and possible solutions, but with a focus on purely empirical work based on regressions and randomized controlled-trails.[17] It provides a good guide for those interested in improving the reproducibility of

[16]This AEA Code and Data Policy applies to all journals published by the AEA, including but not limited to: American Economic Review and American Economic Journal: Macroeconomics.

[17]Randomized controlled-trials (RCTs) provide a gold-standard, but not a silver bullet. One issue is whether randomization ends up truely random. An EconTalk podcast with James Heckman describes an RCT for a drug to treat AIDS. Participants randomly received the AIDS drug (treatment) or a placebo (control). Because at that time

their own work.

For Quantitative Macroeconomics researchers interested in trying to ensure that their own computational work is reproducible Section 3 presents a checklist, based on my experience with difficulties commonly encountered. This checklist is strictly intended as an aid for researchers, not as a requirement to be imposed. Naturally it will be incomplete but should help researchers who wish to make their work more transparent and reproducible avoid the oversights most common in the literature.

By making replication easier to perform it is hoped that issues such as robustness of model prediction and sensitivity to parameters and model specification will become easier to perform. The importance of developing computational modelling packages such as Dynare, EconARK, GDSGE, niqlow and VFI Toolkit should viewed as part of contributing to this.[18] The literature on empirical regressions has begun developing tools to address these issues of specification searching with a good overview provided by Chapter 7 of Christensen, Freese, and Miguel (2019).[19] Quantitative Macroeconomics would also benefit from such an approach, and simple replication of existing results is a first step on the road to being able to solve models easily enough to make this possible.

The rest of this paper simply describes some general lessons learnt from the process of replicating these papers. Much of what follows might be misread as picking on certain authors/papers by calling out their minor errors. This is far from my intention, which is to understand where common errors are being made and how the profession might do better. The best defense of my intentions is that any author/paper which appears in this work was one I have chosen to spend a few days of my life in replicating as I thought it was sufficiently important in the development of Quantitative Macroeconomics.[20] After all, [replication] is the sincerest form of flattery!

AIDS was a death-sentence the participants were so terrified that they met up outside the lab, put all their pills into a bowl, and then each took a handful containing a mixture of drug and placebo. The Doctors performing the trail were unaware that their randomization had failed. A second issue for many RCTs is lack of power to find effects due to small sample sizes. An example is documented for Microfinance initiatives by Dahal and Fiala (2019), who find that of all eight peer-reviewed RCT publications not a single one has sufficient sample size to have the enough power to find a statistically significant result of the likely (as indicated by point estimates) size of such a result. Note, the issue is not just the 'raw' sample size but also the compliance or take-up rate (what fraction of those offered microfinance loans actually use them); the problems with the AIDS study could be viewed as their having zero net compliance rate (no actual treatment of the treated, relative to control) and hence no statistical power.

[18] Important related efforts aim to develop the underlying libraries and tools, rather than direct modelling, such as QuantEcon.

[19] Blinder and Watson (2016) provide the odd case of a paper the second-half of which sets out to show that out of a few tens of possible specifications only a few lead to a statistically significant result (in explaining the 'D-G gap'). Rather than concluding that the most of the few statistically significant variables are likely the result of specification searching across various regressions leading to spurious significance, they instead present it as a robustness exercise. This has now been gamified. Riffing on the article entitled *Let's Take the Con Out of Econometrics* (Leamer, 1983) one might conclude that they claim to have turned the con into a pro! This is my personal opinion and the reader should obviously treat it as such; both authors have plenty of other good papers and I am a big fan of the other work of Mark Watson in particular, especially on understanding long-run relationships between variables.

[20] For many of the replications reported here I chose to spend a few days replicating, but actually ended up spending a few weeks and in some cases months.

1 Lessons Learned from Common Issues

Some of the main issues encountered during the replications provide lessons for best practice that Macroeconomists can learn from. However the one common pitfall from which there is nothing to be learned is that coding bugs do occour, this appears to have affected a small fraction of the numbers reported in the papers; as a friend expressed it, if you start with n bugs and squash one you are left with n bugs. The main issues and recommandations based on these are discussed. The recommendations are then summarised as a checklist in Section 3.

Issue: Graphing Probability Distributions. I recommend that researchers plot cumulative density functions, rather than probability density functions. Probabilitity density functions can mislead for two reasons: first, they obviously depend on the number of grid points used; second, they appear more sensitive to numerical error. Since many solution methods in quantitative economics involve discretizing shock processes this leads to very different looking probability density functions when the number of grid points used to discretize the shock changes; loosely, doubling the number of grid points would halve the probability mass at each point.[21] This issue is minimized but not entirely eliminated when using cumulative density functions.

One alternative approach is to parametrize the probability density — say as Chebyshev polynomials, or as a mixture of parametric probability distributions, etc. — but this approach is likely limited if the interest is in, eg., inequality and the shares of Total Income held by the Top 1% as the parametrization will implicitly impose some assumptions on these shares.[22] Comparing a number of alternatives I concluded that when probability density functions are plotted the best performance comes from graphing kernel-smoothed density functions estimated from the discretized probability mass function.[23]

Issue: Only baseline case parameters are provided. Papers essentially always provide all the parameter values for their baseline calibrations (a few do not report the final value of things such as general equilibrium prices that would be of much use for replications when trying to understand where differences may be arising). However a number of papers do not report all the parameter values for alternative calibrations, such as those used for 'policy experiments' or difference 'cases' (e.g., Castaneda, Díaz-Giménez, and Ríos-Rull (2003) and Hubbard, Skinner, and Zeldes (1994)). Such parameter values would be appropriate for inclusion in a technical computational appendix.

Issue: Naming variables. Many papers use different names for variables in their papers and code, complicating reading the code for anyone else. Ideally this would not occour, but a more

[21] Probability density functions can also be misleading in the sense that they are very sensitive to numerical error. For example Imrohoroglu (1989) graphs the probability density function, finding two spikes, and provides an interpretation of the intuition said to underlie the existence of these spikes. These spikes appear to have been numerical approximation error and disappear when the grid is made much finer.

[22] In theory they needn't impose any strong assumptions on the shares as the order of the polynomials approaches infinity. But in practice the polynomials are typically low-order, as otherwise most of the computational advantages to using them are lost.

[23] The codes replicating Imrohoroglu (1989) contain a commented out section comparing a few alternatives.

reasonable solution might be the provision of dictionaries anywhere this does occour.

Issue: Reporting parameter values Three main problems occour: First, the reported parameter is for a different time-period to the model (e.g., report the annual value, when model period is two months). Second, reported standard deviation is for the stochastic process, but equations describe it as being for the innovations to that process. Third, parameters that vary over life-cycle are only reported as a Figure (so exact values are unavailable).[24] To be more precise about the second of these, many papers will, e.g., have an AR(1) process and describe σ to be the standard deviation of the innovations, but then when reporting the calibrated variables instead report σ as the standard deviation of the AR(1) process itself. My own suggestion is to use a notation that always specifically emphasises when, e.g, a standard deviation is that for innovations ϵ to the AR(1) process z call it σ_ϵ, and when for the AR(1) itself call it σ_z. This simply helps to differentiate between the two standard deviations which are otherwise often and easily mixed up by accident during writing.

Issue: Calibration Details. Many papers will describe which moments were targeted by the calibration. But they will not provide details on how the calibration itself was implemented. While in earlier papers this was fine as most moments are targeted independently more recent papers often jointly target a number of moments. This typically will mean they have implemented a single-objective optimization that assigns each target moment a weight (multi-objective optimization is also a possibility but based on informal conversations seems rarely used by Economists). These weights are not typically reported (eg., Castaneda, Díaz-Giménez, and Ríos-Rull (2003) do not provide such detail). I suggest that papers should more often include a technical computational appendix which provides this kind of detail. Along the same lines the initial values from which such optimization takes place are almost never given. The availability of codes turned out to be important factor in mitigating this. For example Auerbach and Kotlikoff (1987) describe the calibrated values of their age-dependent parameter e, but do not explain that these are in fact the log values, and that one must take their exponential and then normalize them so that the age one value of e_1 is set to 1; Figure 5.2 made it clear that something was missing in the original description of the calibrated values of e and as their codes are available it was easy enough to find out what.

Issue: Availability of Codes. In a few cases the original codes are available from the authors website. In most cases however one had to contact the author directly, and even then some authors no longer had codes (to be fair some of these papers are from early 1990s). As an extreme example the codes for Aiyagari (1994) are unavailable online and the author is deceased. While there is an increasing requirement from journals to provide codes[25] the most obvious improvement would be an increased use of github to make codes publicly available; journals that already provide their own

[24] As concrete examples, these issues occour in Díaz-Giménez, Prescott, Alvarez, and Fitzgerald (1992), Restuccia and Urrutia (2004), and Huggett (1996) respectively.

[25] For example, providing codes is now required by all the 'Top 5' Economics journals.

online code repositories are a perfectly satisfactory substitute/complement. This issue appears to already be well recognized in Economics and is therefore likely 'already solved' as it were. Current approaches typically have journals provide codes in downloadable zip files making the process much more onerous that if each Journal simply uploaded all codes to its own github repository or similar; this would make them all instantly searchable and easily accessed and read. The importance of making codes available is the clearest lesson from the replications reported in this paper. Where authors provided codes (often on email request) these were able to resolve many other problems that arose during replication for many of the reasons described elsewhere in this paper.

Issue: Parameter Robustness and Numerical Approximation Errors. Many papers have a 'default' parametrization and have performed some kinds of tests to check that their numerical methods are performing well at minimizing numerical error. They then look at how changing parameters would change certain model outputs. Often these tests will, eg., induce further curvature into certain parts of the solution and this interacts with the numerical methods to worsen their performance. For example Aiyagari (1994) reports the degree of precautionary savings (eg., as the resultant interest rate) for various parametrizations. While the results relating to low-risk and low-risk-aversion are numerically accurate, those relating to high-risk and high-risk-aversion contain substantial numerical error.

The results of tests for the magnitude of numerical errors, such as Euler Equation residuals (Santos, 2000), are sensitive to the parameter values. This fact is known to be the case from the theory underlying such tests but the issue is often ignored in practice. One possibility would be that when measures of numerical accuracy are presented they should be reported across the range of parameter values that are made use of in the model. An alternative might be for the profession to move more towards the use of *adaptive* numerical methods, such as those in Brumm and Scheidegger (2017), which assess approximation errors and then update based on them as part of the solution method itself. Both of these suggestions are rather onerous so for the present simply having researchers more aware of this issue might be the best approach.

Issue: Welfare Evaluations. Some of the replicated papers used linear-quadratic methods (Díaz-Gímenez, 2001) to solve the value function problem. Replication of these papers often showed high accuracy in variables that depend on the stationary distribution and policy function. However the welfare calculations appear to contain substantial numerical error. It is suspected, but not known, that this relects that linear-quadratic methods perform fine for computing policy functions but provide a poor approximation of the actual value function itself. Since welfare calculations are based on the value function itself they were therefore erroneous. This illustrates how numerical errors in different aspects of the model can be very different. It is common practice to report the results of tests for the magnitude of numerical errors, such as Euler Equation residuals which look at the policy function. It is important to understand the conditions under which these also imply limited numerical errors elsewhere in the model (Santos, 2000; Fernandez-Villaverde, Rubio-Ramirez, and Santos, 2006; Kirkby, 2019). In the current instance of the errors in the value function

and linear-quadratic methods the theory relating the value function and Euler equation residuals (Santos, 2000) does not apply.[26]

Issue: Formulae for model statistics. Typically, when reporting model statistics papers provide a verbal description of how they are calculated, but rarely include an explicit equation. This lead to some difficulties in replication. For example, in Díaz-Giménez, Prescott, Alvarez, and Fitzgerald (1992) most statistics could be replicated exactly, but a few table entries could not, it seems likely this is simply because I was unable to turn the verbal descriptions into the precise formula. Another example: Restuccia and Urrutia (2004) calculate 'cross-sectional disparity' as the standard deviation of log earnings, but what is unclear from the written description is that in this two-period OLG model the 'cross-section' is computed conditional on age being 2, not across the whole model economy. One solution would be to put more formulas in Technical appendices, however this seems overly onerous given that the same issue can largely be solved by improved availability of codes.

2 Influence of Numerical Methods on Economics

The need for greater discussion in Quantitative Macroeconomics papers of the discretiziation of shocks —on par with the usual discussion of parameter choices and the sensititivity of results— stems from the large influence these have in many models on driving both modelling choices and quantitative results.

The main discretization methods used all relate to AR(1) shock processes with normally distributed innovations, namely the Tauchen and Rouwenhorst methods (Tauchen, 1986; Rouwenhorst, 1995). Both perform acceptably in most situations as long as sufficient grid-points are used although the later is to be preferred when shocks are highly persistent. When these are used the most important thing is that both grid-size and hyperparameters need to be reported, and some sensititivity/robustness analysis to these choices should be performed. The most common 'error' in the literature is simply to choose 'too few' grid-points and ignore the large quantitative impact of this in driving results. Variations of these exist (Tauchen and Hussey, 1991; Adda and Cooper, 2003; Floden, 2008) but I recommend against their use[27] as they typically perform worse than the

[26] It does not apply for two reasons: first the linear-quadratic methods themselves, secondly as there are periodically-binding constraints. This second reason is worth emphasising as it applies to almost all heterogeneous agent incomplete markets models: since we do not know where the periodically binding constraint actually binds the Euler equation residuals are not a valid measure of numerical error. Li (2015) explains this problem in detail and derives numerical bounds for Euler equations with periodically-binding constraints but they turn out to be insufficiently tight to be practically useful.

[27] The Tauchen-Hussey method (Tauchen and Hussey, 1991) in particular should no longer be used. It's poor performance is well documented and the existing alternatives are just as easy to implement (Toda, 2020). An indication of how widespread this method is comes from it's inclusion as a central part of the textbook and toolkit of Miranda and Fackler (2002), and it's inclusion in QuantEcon as a standard numerical quadrature method (the algorithm is often coded as the use of a function called *qnwnorm*).

Rouwenhorst method and lack the transparency of the Tauchen method.[28,29] The same is true for finite-horizon models with AR(1) shock processes with normally distributed innovations where the parameters are age-dependent: the natural extension of the Rouwenhorst method performs best, and the natural extension of the Tauchen method is transparent (Fella, Gallipoli, and Pan, 2019). The main point here though is not so much which method is used, but that these choices need to be discussed in the papers at least as much as any other calibration choice; they only become irrelevant with grid-sizes of a magnitude almost never seen in practice.

The focus of all of these common discretization processes on normally-distributed shocks also seems misguided. Given that discrete Markov processes will be used to compute the models, why run the data through the straight-jacket of an AR(1) process before it reaches the model? Why not go more directly from data to discrete Markov process? This approach allows much more general and realistic shock processes to be used, and is likely to be especially important in any attempts to model income risks, rare disasters (and more broadely the impacts of climate change), and asset prices. Several methods to do this already exist and the literature would be improved by their more widespread adoption; again, alongside more discussion in papers of these discretization choices and their impact on results. Some existing approaches include the quadrature method of Farmer and Toda (2017) which allows more non-parametric approximations, the approach of Castaneda, Díaz-Giménez, and Ríos-Rull (2003) who simply calibrate a four-state Markov directly, and the use of histograms to create 'bins' and then simply 'count-and-normalize' transitions to implement the maximum-likelihood estimator of a finite-state markov (Kirkby (2017b) explains this in detail for model of Hansen (1985)).

Beyond just the choice of discretizing shock processes, the reporting of various choices of numerical methods and hyperparameters would ideally also be more widely discussed in papers. But given the onerous nature of trying to test for sensititivity/robustness of these choices this is probably best left to replication studies using different methods.

One article I would have liked to replicate but did not is Kydland and Prescott (1982). The reason is itself an interesting example of the important role played by the choice of numerical methods, especially those that involve large amounts of approximation. The model of Kydland and Prescott (1982) contains a six-dimensional state variable, making it prohibitively complicated for the discretized value function iteration methods I use in our replications. The model can however be easily solved using the linear-quadratic value function iteration methods used by Kydland and Prescott (1982), which involves solving for six co-efficients, rather than a full six-dimensional object. This is because using linear-quadratic value function iteration methods means that the full distribution

[28]The Tauchen (1986) is transparent in the sense that it forces the researcher to specifically choose the hyperparameter value, henceforth Tauchen's q, which determines the the maximum and minimum grid points as being plus/minus $q/2$ standard deviations. This being forced to explicitly choose the hyperparameter means the researcher is aware of the choice, and likely aware of the role it plays in determining model results.

[29]The superior performance of the Rouwenhorst method is documented for stochastic Real Business Cycle by Kopecky and Suen (2010) and the Diamond-Mortensen-Pissarides search model by Piccione and Rubinstein (2007).

of the shocks does not matter for evaluating expectations of next periods value function, only their conditional mean.

The issue of the use of linear and log-linear, and first- and second-order perturbation in welfare evalutations has already been described in the Introduction. The results of Judd, Maliar, and Maliar (2017) showing that the minimum error bounds on linear, log-linear, and first-order approximations are large enough to be problematic for most Economic models should dissuade Economists from using them in any application. This is especially true thanks to the implementation of second-order and higher methods in many available codebases (including Dynare). Users should also be aware that first-order methods imply only the conditional mean matters for expectations, and that with second-order only the conditional mean and conditional second moment matter; this means they are, e.g., simply unusable for any study of the impact of rare events/disasters or conditional changes in volatility. Wherever possible Economists should be making greater use of global non-linear solution methods.[30]

3 A Checklist for Reproducibility

Table 3 is provided to act as a simple checklist that researchers interested in ensuring reproducibility of their work can use to avoid common omissions. The table is not intended to be comprehensive, but is intended to make it easier to avoid omissions that are common in the existing literature.

[30]Linear methods are sometimes the only way of solving large models, and I would not advocate abandoning them for doing so. But wherever a choice is feasible much greater use of global non-linear methods should occour. For example, there is no excuse for the use of linear-methods to solve mid-size representative agent DSGE models in Dynare given how easily second-order perturbation methods can be used instead.

Table 3: Checklist for Reproducibility

Item	Tick
Copy of codes uploaded, preferably to a third-party repository (github, OSF, dataverse)	
If not obvious from filenames, uploaded codes includes a Readme file explaining what to run.	
Readme file may also describe what software (and versions) were used. What hardware was used.	
Readme file may also give rough guidance on runtimes (a few minutes/hours/weeks).	
Parameters: In codes, parameters are stored in a data structure that can be exported as JSON.	
Parameters: Where parameter names differ between paper and codes a 'dictionary' is provided.	
Parameters: Include general equilibrium values, initial conditions, alternative calibrations.	
Parameters Bonus: Include hyperparamers for numerical methods used.	
Explicit formulae provided for all model statistics reported in paper.	
When codes/functions are taken from previous projects mention their source.	
Bonus: Codes contain easy to understand comments and variable names.	
Bonus: Codes make it clear which parts of code are generating which results in paper.	
Bonus: In paper describe the numerical methods used, even just 'same as paper X'.	

4 Conclusions

We end simply with an inculcation to the importance of reproducibility of results in Economic Science, and in Science more generally: "Non-reproducible single occurrences are of no significance to science." — Popper, Karl (1934, The Logic of Scientific Discovery)

References

J. Adda and R.W. Cooper. Dynamic Economics: Quantitative Methods and Applications. MIT Press, 2003.

Mark Aguiar and Gita Gopinath. Defaultable debt, interest rates and the current account. Journal of International Economics, 69:64–83, 2006.

S. Rao Aiyagari. Uninsured idiosyncratic risk and aggregate saving. Quarterly Journal of Economics, 109(3):659–684, 1994.

Christina Arellano. Default risk and income fluctuations in emerging economies. American Economic Review, 98(3):690–712, 2008.

S. Aruoba, Jesus Fernandez-Villaverde, and Juan Rubio-Ramirez. Comparing solution methods for dynamic equilibrium economies. Journal of Economic Dynamics and Control, 30(12):2477–2508, 2006.

Alan Auerbach and Laurence Kotlikoff. Dynamic Fiscal Policy. Cambridge University Press, 1987.

Florent Bédécarrats, Isabelle Guérin, Solene Morvant-Roux, and Francois Roubaud. Estimating microcredit impact with low take-up, contamination and inconsistent data. a replication study of crepon, devoto, duflo, and pariente (american economic journal: Applied economics, 2015). International Journal for Re-Views in Empirical Economics, 3, 2019. doi: https://doi.org/10.18718/81781.12.

Alan Blinder and Mark Watson. Presidents and the us economy: An econometric exploration. American Economic Review, 106(4):1015–45, 2016.

Jerry Bona and Manuel Santos. On the role of computation in economic theory. Journal of Economic Theory, 72:241–281, 1997.

Abel Brodeur, Mathias Lé, Marc Sangnier, and Yanos Zylberberg. Star wars: The empirics strike back. American Economic Journal: Applied Economics, 8(1):1–32, 2016.

Johannes Brumm and Simon Scheidegger. Using adaptive sparse grids to solve high-dimensional dynamic models. Econometrica, 85(5):1575–1612, 2017. doi: 10.3982/ECTA12216.

Colin F. Camerer, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu. Evaluating replicability of laboratory experiments in economics. Science, 2016. doi: 10.1126/science.aaf0918.

Ana Castaneda, Javier Díaz-Giménez, and Jose Victor Ríos-Rull. Accounting for the U.S. earnings and wealth inequality. Journal of Political Economy, 111(4):818–857, 2003.

Andrew Chang and Phillip Li. Is economics research replicable? sixty published papers from thirteen journals say 'usually not'. Finance and Economics Discussion Series of the Board of Governors of the Federal Reserve System (U.S.), 2015-83:1–25, 2015.

Yongsung Chang and Sun-Bin Kim. Heterogeneity and aggregation: Implications for labor-market fluctuations. American Economic Review, 97(5):1939–56, 2007.

Yongsung Chang and Sun-Bin Kim. Heterogeneity and aggregation: Implications for labor-market fluctuations: Reply. American Economic Review, 104(4):1461–1466, 2014.

Garret Christensen, Jeremy Freese, and Edward Miguel. Transparent and Reproducible Social Science Research: How to Do Open Science. University of California Press, 2019.

Open Science Collaboration. Estimating the reproducibility of psychological science. Science, 349: 6251, 2015.

Juan Carlos Conesa and Dirk Krueger. Social security reform with heterogeneous agents. Review of Economic Dynamics, 2(4):757–795, 1999.

Bruno Crepon, Florencia Devoto, Esther Duflo, and William Parienté. Estimating the impact of microcredit on those who take it up: Evidence from a randomized experiment in morocco. American Economic Journal: Applied Economics, 7(1):123–150, 2015.

Mahesh Dahal and Nathan Fiala. What do we know about the impact of microfinance? the problems of statistical power and precision. World Development, 118:123–150, 2019. doi: https://doi.org/10.1016/j.worlddev.2019.104773.

J. Díaz-Gímenez. Linear quadratic approximations: An introduction. In R. Márimon and A. Scott, editors, Computational Methods for the Study of Dynamic Economies, chapter 2. Oxford University Press, 2001.

Javier Díaz-Giménez, Edward C. Prescott, Fernando Alvarez, and Terry Fitzgerald. Banking in computable general equilibrium economies. Journal of Economic Dynamics and Control, 16: 533–559, 1992.

Anna Dreber, Thomas Pfeiffer, Johan Almenberg, Siri Isaksson, Brad Wilson, Yiling Chen, Brian A. Nosek, and Magnus Johannesson. Using prediction markets to estimate the reproducibility of scientific research. PNAS, 112(50):15343–15347, 2015. doi: 10.1073/pnas.1516179112.

Estelle Dumas-Mallet, Andy Smith, Thomas Boraud, and Francois Gonon. Poor replication validity of biomedical association studies reported by newspapers. PLOS One, 10.1371/journal.pone.0172650, 2017.

Leland E. Farmer and Alexis Akira Toda. Discretizing nonlinear, non-gaussian markov processes with exact conditional moments. Quantitative Economics, 8(2):651–683, 2017.

Giulio Fella, Giovanni Gallipoli, and Jutong Pan. Markov-chain approximations for life-cycle models. Review of Economic Dynamics, 34:183–201, 2019.

Jesus Fernandez-Villaverde, Juan Rubio-Ramirez, and Manuel Santos. Convergence properties of the likelihood of computed dynamic models. Econometrica, 74(1):93–119, 2006.

Paul Ferraro and Pallavi Shukla. Is a replicability crisis on the horizon for environmental and resource economics? Review of Environmental Economics and Policy, 14(2):339–351, 2020. doi: https://doi.org/10.1093/reep/reaa011.

Martin Floden. A note on the accuracy of markov-chain approximations to highly persistent ar(1) processes. Economics Letters, 99(3):516–520, 2008.

Paul Gertler, Sebastian Galiani, and Mauricio Romero. How to make replication the norm. Nature, 554:417–419, 2018. doi: 10.1038/d41586-018-02108-9.

Veronica Guerrieri and Guido Lorenzoni. Credit crises, precautionary savings, and the liquidity trap. Quarterly Journal of Economics, 132(2):1427–1467, 2017.

Gary Hansen. Indivisible labor and the business cycle. Journal of Monetary Economics, 16(3): 309–327, 1985.

Juan Carlos Hatchondo, Leonardo Martinez, and Horacio Sapriza. Quantitative properties of sovereign default models: Solution method. Review of Economic Dynamics, 13(4):919–933, 2010.

Thomas Herndon, Michael Ash, and Robert Pollin. Does high public debt consistently stifle economic growth? a critique of reinhart and rogoff. Political Economy Research Institute, University of Massachusetts Amherst, Working Paper Series, WP322:301–350, 2013.

Andrew Holding. Novelty in science should not come at the expense of reproducibility. The FEBS Journal, 2019. doi: 10.1111/febs.14965.

Hugo Hopenhayn and Richard Rogerson. Job turnover and policy evaluation: A general equilibrium analysis. Journal of Political Economy, 101(5):915–938, 1993.

Glenn Hubbard, Jonathan Skinner, and Stephen Zeldes. The importance of precautionary motives in explaining individual and aggregate saving. Carnegie-Rochester Conference Series on Public Policy, 40(1):59–125, 1994.

Glenn Hubbard, Jonathan Skinner, and Stephen Zeldes. Precautionary saving and social insurance. Journal of Political Economy, 103(2):360–399, 1995.

Mark Huggett. The risk-free rate in heterogeneous agent incomplete insurance economies. Journal of Economic Dynamics and Control, 17:953–969, 1993.

Mark Huggett. Wealth distribution in life-cycle economies. Journal of Monetary Economics, 38: 469–494, 1996.

Ayes Imrohoroglu, Selahattin Imrohoroglu, and Douglas Joines. A life cycle analysis of social security. Economic Theory, 6(1):83–114, 1995.

Ayse Imrohoroglu. Cost of business cycles with indivisibilities and liquidity constraints. Journal of Political Economy, 97(6):1368–1383, 1989.

Ayse Imrohoroglu and Edward C. Prescott. Evaluating the welfare effects of alternative monetary arrangements. Quarterly Review of the Federal Reserve Bank of Minneapolis, Summer:3–10, 1991.

Kenneth Judd, Lilia Maliar, and Serguei Maliar. Lower bounds on approximation errors to numerical solutions of dynamic economic models. Econometrica, 85(3):991–1012, 2017.

Jinill Kim and Sunghyun Henry Kim. Two pitfalls of linearization methods. Journal of Money, Credit and Banking, 39(4):995–1001, 2007.

Robert Kirkby. Convergence of discretized value function iteration. Computational Economics, 49(1):117–153, 2017a.

Robert Kirkby. A toolkit for value function iteration. Computational Economics, 49(1):1–15, 2017b.

Robert Kirkby. Bewley-Huggett-Aiyagari models: Computation, simulation, and uniqueness of general equilibrium. Macroeconomic Dynamics, 23(6):2469–2508, 2019.

Karen Kopecky and Richard Suen. Finite state markov-chain approximations to highly persistent processes. Review of Economic Dynamics, 13(3):701–714, 2010.

Finn Kydland and Edward C. Prescott. Time to build and aggregate fluctuations. Econometrica, 50(6):1345–1370, 1982.

Edward E Leamer. Let's take the con out of econometrics. American Economic Review, 73(1): 31–43, 1983.

Huiyu Li. Numerical policy error bounds for eta-concave stochastic dynamic programming with non-interior solutions. Computational Economics, 46(2):171–187, 2015.

Edward Miguel. Evidence on research transparency in economics. Journal of Economic Perspectives, 35(3):193–214, 2021.

Edward Miguel and Michael Kremer. Worms: Identifying impacts on education and health in the presence of treatment externalities. Econometrica, 72(1):159–217, 2004.

Mario J. Miranda and Paul L. Fackler. Applied Computational Economics and Finance. MIT Press, 2002.

Adrian Peralta-Alva and Manuel Santos. Analysis of numerical errors. In Karl Schmedders and Kenneth L. Judd, editors, Handbook of Computational Economics, volume 3, chapter 9. Elesevier, 2014.

Michele Piccione and Ariel Rubinstein. Equilibrium in the jungle! The Economic Journal, 117 (552):883–896, 2007.

Carmen Reinhart and Kenneth Rogoff. Growth in a time of debt. American Economic Review Papers and Proceedings, 100(2):573–578. doi: 10.1257/aer.100.2.573.

Diego Restuccia and Richard Rogerson. Policy distortions and aggregate productivity with heterogeneous establishments. Review of Economic Dynamics, 11:707–720, 2008.

Diego Restuccia and Carlos Urrutia. Intergenerational persistence of earnings: The role of early and college education. American Economic Review, 94(5):1354–1378, 2004.

G. Rouwenhorst. Asset pricing implications of equilibrium business cycle models. In Cooley T., editor, Frontiers of Business Cycle Research, chapter 10. Princeton University Press, 1995.

Manuel Santos. Accuracy of numerical solutions using the euler equation residuals. Econometrica, 68(6):1377–1402, 2000.

Lisa L. Shu, Nina Mazar, Francesca Gino, Dan Ariely, and Max H. Bazerman. Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. PNAS, 15:127–148, 2012.

Shuhei Takahashi. Heterogeneity and aggregation: Implications for labor-market fluctuations: Comment. American Economic Review, 104(4):1446–60, 2014.

George Tauchen. Finite state markov-chain approximations to univariate and vector autoregressions. Economics Letters, 20:177–181, 1986.

George Tauchen and Robert Hussey. Quadrature-based methods for obtaining approximate solutions to nonlinear asset pricing models. Econometrica, 59(2):371–396, 1991.

Alexis Akira Toda. Data-based automatic discretization of nonparametric distributions. Computational Economics, 57(4):1217–1235, 2020.

Christian Zimmermann. On the need for a replication journal. Federal Reserve Bank of St. Louis Working Paper Series, 2015-16:1–16, 2015.

A The Replicated Figures and Tables

This Appendix contains all of the replicated Figures and Tables from each of the papers replicated. It is organised as a subsection for each paper. We comment on the output only when it differs notably from the results of the original paper. A brief mathematical description of the baseline model being solved is given for each paper. For full descriptions of the models being solved, including their economic use and interpretation, please consult the papers themselves.

Codes which perform these replications, creating all the Tables and Figures from scratch, as well as a pdf with a brief model description and the full results, can be found at: https://github.com/vfitoolkit/vfitoolkit-matlab-replication

These codes were all implemented in Matlab, and for purposes of this paper were run in Matlab (versions between 2018a and 2020b) using the VFI Toolkit (vfitoolkit.com). They were run on a variety of computers all running Linux (Kubuntu is the best distro ;), with NVIDIA gpus (with 2gb to 40gb GDDR ram) and from two to twenty CPU cores and with memory of 16gb to 120gb.

The replication codes were written with robustness, transparency and ease to follow what is being done in mind, and with little to no concern for run-time (many unnecessary objects are computed). Most therefore take from a few days to a week to run.

To ensure accuracy the grid sizes were increased until a 'substantial' change in the grid size resulted in a 'very tiny' change in results. To be precise, the grid sizes between which the upper quartile of the absolute percentage differences for all numbers presented in tables was less than 0.05 (5%) were:

- Hansen (1985): [751,3001,91], [601,2501,91]
- Imrohoroglu (1989): [1501,2], [701,2]
- Díaz-Giménez, Prescott, Alvarez & Fitzgerald (1992): [2,1000,2,4,1], [2,800,2,4,1]
- Hopenhayn & Rogerson (1993): [601,33], [401,33]
- Huggett (1993): [1024,2], [512,2]
- Aiyagari (1994): [1024,27], [512,21]
- Hubbard, Skinner & Zeldes (1994): [1501,21,21],[751,15,15]
- Imrohoroglu, Imrohoroglu & Joines (1995): [1251,2],[1001,2]
- Huggett (1996): [2001,19],[1501,19]
- Conesa & Krueger (1999): INCOMPLETE

- Castaneda, Díaz-Giménez & Ríos-Rúll (2003): [151,2501,8],[101,2001,8]
- Restuccia & Urrutia (2004): [200,200,200,101,2,15],[150,150,150,51,2,15] (bhatprime,h,bhat,thetahat,s,b)
- Restuccia & Rogerson (2008): [200,3], [100,3]
- Guerrieri & Lorenzoni (2017): INCOMPLETE

The reported replication results reflect the first/larger of these grid sizes. These grids are chosen solely for accuracy without a concern for speed and so do not represent a sensible speed-accuracy trade-off for normal use; they are in some sense 'too large'. The ordering of the grid sizes is always the same as in the codes and reflects the ordering of the concepts of decision variables, endogenous states, exogenous states that underlies the algorithms used by the VFI Tookit. Note that in many models for some variables, in particular the exogenous variables, the grid size is not something that can be increased without changing the interpretation; e.g., two exogenous states that represent employment and unemployment. Some papers use different grids for baseline models and alternative models, the above reports the baseline model grid sizes. This criterion of an upper quartile difference of less than 5% is more demanding than it first sounds as, e.g., many papers have many numbers like 0.1 for which a change to 0.11 is a change of more thant 5%.