The BESTEPS Guide:

7 Lessons towards Effective Economics PhD Research

Edited by: Anna Chesa-Llorens* and Mengwei Lin 林孟蔚†

This version: September 22, 2025

This *Guide* is intended to offer early-career economists, especially PhD students, insights from leading researchers on how to take the "best steps" at each key stage of a research project. It is based on a series of lectures held at the University of Barcelona (UB) School of Economics under the BESTEPS project.[1] Each lecture is self-contained and can be read independently. Taken together, the seven lectures aim to offer a practical and comprehensive guide to conducting research effectively during the PhD. This version collects the first three lectures by **Antonio Cabrales**, **Libertad González**, and **Joan Llull**.[2]

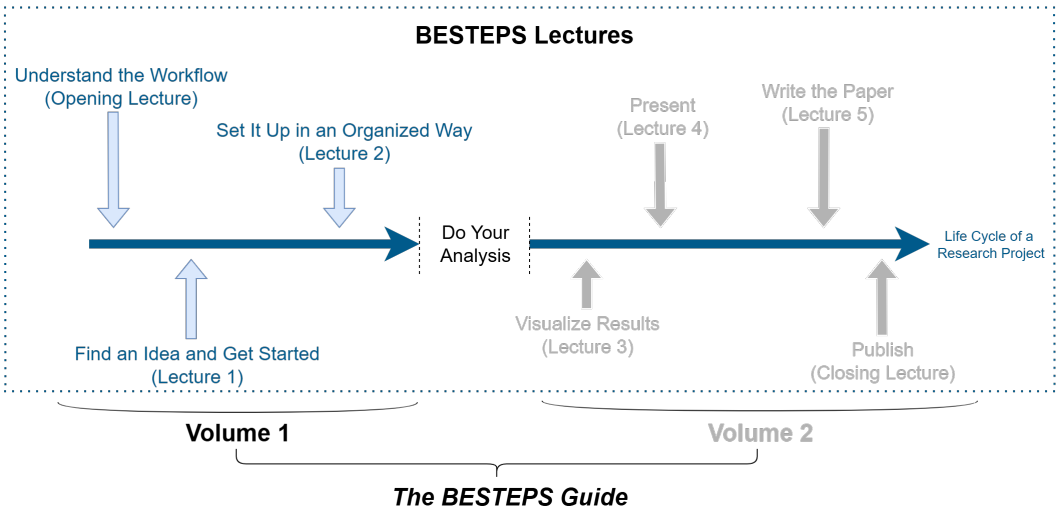

Figure 1: The BESTEPS Project

*University of Barcelona & Barcelona Institute of Economics (IEB). Email: annachesa@ub.edu

†University of Barcelona & Barcelona Institute of Economics (IEB). Email: mengwei.lin@ub.edu

[1]BESTEPS (Barcelona Essential Skills Training for Economics PhD Students) is a training program funded by the UB Doctoral School and held at the UB School of Economics. It is part of the training of the PhD in Economics program at the University of Barcelona. See Figure 1 for an overview of the BESTEPS project and the target of each lecture.

[2]The lecturers generously agreed to publish these notes as a public good, but did *not* review the draft. We are grateful to the following fellow young economists across fields and from various institutions for carefully reading the draft and providing thoughtful feedback: Bastian Bruestle (Zurich), Haochi Chen (Dublin), Zhenghao Chen (UCL), Alessandro Ciardo (Barcelona), Britta Hecker (IAB), Guillermo Martínez (CEMFI), Taoyi Sun (Purdue), Robert Włodarski (Cambridge), Jessie Ziyue Zhu (Maastricht). Names on this page are in alphabetical order. Any errors or omissions are our sole responsibility.

Contents

Preface

I am thrilled to share a few words to support the **Barcelona Essential Skills Training for Economics PhD Students (BESTEPS)**, a brilliant initiative led by an outstanding team of PhD students.

BESTEPS aims to provide a comprehensive set of skills that are essential at various points of doctoral research in Economics. It covers a wide range of topics, from the early stages of the research process to the final production of innovative and relevant research papers. These skills are not typically included in master's or PhD curricula, despite being crucial for a successful academic career and for enhancing the competitiveness of PhD students in the job market and beyond.

In terms of format, BESTEPS is unique in being tailored to the specific needs of Economics PhD students and being highly interactive. The program is structured as a course comprising several sessions, each focusing on a particular aspect of the process of creating, conducting, and publishing academic research. These sessions feature leading researchers sharing their experience. BESTEPS provides a solid foundation for PhD students to approach their research careers in a structured and effective way.

The stars aligned, and this proactive initiative coincided with a funding call launched by the UB Doctoral School to support training activities led by PhD students. The outcome was so successful that we feel compelled to share it—that is why we have put together this *BESTEPS Guide*. We are confident that PhD students from other programs, as well as other junior researchers, will benefit from these notes.

I want to take this opportunity to warmly thank the lecturers, Antonio Cabrales, Libertad González, and Joan Llull, as well as those yet to come. They have done (or surely will do) a wonderful job sharing their experience in a helpful and enlightening way. They truly exceeded our already high expectations.

I would also like to express my deep gratitude and congratulations to the exceptional team of PhD students who designed the BESTEPS proposal and have meticulously organized every detail of the program with rigor, dedication, and vision. The team, led by Mengwei Lin, includes Daniel Favre, Marcella De Giovanni, Sergio Blanco, Ruixi Zhang, Anna Chesa, Sofía Balladares, Thompson Ogajah, Natalia Carralero, Zhenning Han, Cecilia Ramírez, Jingwen Tan, Àlex Izquierdo, and many more. Their proactive spirit, commitment to academic excellence, and ability to turn an idea into a high-impact initiative are truly commendable. My sincere thanks also go to Jordi Roca and Judit Farré, whose administrative support has been essential to this program.

Last, to all the young researchers reading this *Guide*: we hope you find it as inspiring and useful as we intended. The research journey can be challenging, but you are not alone; there is a community of scholars ready to support you, and resources like BESTEPS are here to help you grow. Stay curious, stay resilient, and don't hesitate to reach out, ask questions, and seek connections. Your ideas matter, and your work has the potential to shape the future of economic research. Keep going—you've got this!

Concepció Patxot
Director, PhD in Economics Program
University of Barcelona

VOLUME ONE

Setting Up Your Research Project

The first part of *The BESTEPS Guide* is designed to support early-career researchers in taking the very first steps of their PhD journey. It focuses on the foundational phase of any research career: setting up a research project. This includes understanding the workflow, finding a promising idea, and developing the mindset and habits with data management and coding practices.

The three lectures in this volume provide complementary perspectives from three experienced economists: Antonio Cabrales, Libertad González, and Joan Llull. In the **Opening Lecture**, Antonio provides a broad overview of the research process, guiding early-career researchers through each stage and offering essential principles for dealing with uncertainty and maintaining focus. In **Lecture 1**, Libertad zooms in on the very beginning of the journey, giving practical, step-by-step guidance on how to generate questions and formulate research ideas. Finally, in **Lecture 2**, Joan focuses on a core competency of today's researchers: doing organized and replicable work, and emphasizing the mindset and practices of reproducibility, transparency, and research integrity.

Volume 1 of *The BESTEPS Guide* focuses on helping you start and structure a research project. Volume 2, to be released soon, will turn to the research product, exploring how to produce meaningful outputs and communicate research effectively. For a comprehensive overview of the BESTEPS program, please refer to Figure 1.

1 Opening Lecture: Navigating the Research Process in Economics

Lecturer: ***Antonio Cabrales****. Antonio is a Professor of Economics at UC3M, and formerly a Professor and Head of the Department of Economics at UCL. He is also the Executive Vice President of the European Economic Association, a Fellow of the Econometric Society, and a Founding Fellow of the Royal Economic Society. Antonio's research spans a broad range of fields within economics, including game theory and mechanism design, experimental economics, the economics of networks, and the economics of education, among others.*

In this lecture, Professor Cabrales provides a comprehensive framework for navigating the research process in economics, outlining the key stages of a study, highlighting critical points, common challenges, and practical strategies to overcome obstacles. He also shares what he would have liked to know when he started this journey.

Contents of this lecture:

1.1 Introduction: Research as a Power Law

Research follows a power law distribution: there are far more failures than successes. The likelihood of finding the perfect research idea early in your journey is low. Instead of aiming for immediate breakthroughs, you should embrace the iterative nature of the process. Anticipate challenges and surround yourself with people who can help you navigate them. This is why focusing on creating a safe space and a good work-life balance at the beginning of the PhD will make your experience way more fulfilling.

The following parts of this lecture will cover suggestions for idea generation and evolution, choosing the right methodology, writing and communicating, managing challenges along the way, and preparing for the Job Market.

1.2 Step 1: Idea Generation and Evolution

The first step in doing research is to come up with a research idea[3]. A key strategy for this is to maximize your exposure to diverse topics and perspectives in order to seek inspiration. One useful approach is to attend as many seminars as possible, even those outside your immediate field of interest. The value of knowledge is often unpredictable, so what may seem unrelated today might become crucial in the future.

Ideas emerge from multiple sources: reading both classic and recent academic literature, attending conferences, engaging in discussions with peers, seeking guidance from advisors, conversing with scholars from different fields, and even casual talks with family or following news and podcasts. The key is to explore widely, without the need for in-depth analysis at this stage. There is a nice way of doing this in a more "mechanical" way, that follows two steps: *first, think divergently; then, think convergently.*

Divergent Thinking: Generating Ideas

Creativity thrives when there are no immediate constraints. To enhance your ability to generate research ideas, practice brainstorming without self-criticism. The critical mind should not interfere with the generative process too early.

One useful technique is to write down subtopics within a broad research area. This exercise can help you identify which themes are most promising or personally engaging. The more ideas you produce, the better your chances of refining one into a viable research project.

Exercise 1. Practicing divergent thinking
Since generating ideas is usually challenging at this stage of your career, here are some quick exercises you can perform, either individually or in pairs, to develop your ability to generate research questions:

1. **Brainstorming**. As a starting point, brainstorm individually by writing down all research questions you can think of within 2 minutes, without judging their quality. Then pair up, read each other's questions, and try to generate new ones from them within another 2 minutes.

2. **Mind mapping**. First, pick a broad topic in economics (e.g., climate, gender, immigration, education, taxation, inflation, matching, networks)–preferably one you are *not* currently working on–and then map out all possible subtopics and angles within this topic. Doing so will help you discover new and potentially interesting research ideas.

Convergent Thinking: Refining Research Questions

Once you have generated a list of questions, it's time to evaluate them critically. When assessing a research question, imagine a colleague asking: *Why should I care?* Then you can consider some key

[3]The next lecture covers this step extensively.

questions, including:

- Is there a gap in the literature?
- Does the topic address an important social problem or policy issue?
- Can the research contribute to resolving a theoretical debate?
- Is the insight from this study applicable beyond its immediate context?...

Admittedly, not all PhD students have complete freedom in choosing their research topics. However, even if you are assigned a topic from the beginning, you can still approach it from an innovative angle. Flexibility in framing your research question is essential for making a meaningful contribution. Moreover, you need to find an angle that motivates you, as you will likely spend a considerable amount of time working on it.

Exercise 2. Practicing the Elevator Pitch

Once you have a good idea, you need to both understand it and be able to explain it to others. This step is crucial, since at this stage you need to get as much feedback as possible. A strong research idea should be explainable in 30 seconds. You should prepare a concise "elevator pitch". To craft your pitch, develop a brief summary answering the following questions:

- Who am I?
- What do I do?
- How do I do it?
- Why does it matter?

This pitch should be accessible to non-experts and should highlight the novelty and importance of your work. A good practice is to try to make someone among your family and friends to understand it.

Research Evolution and Sunk Cost Fallacy

You may think you have a great idea and you were also able to turn it into an "elevator pitch". Yet, research ideas are fluid and will likely change over time. Be open to iteration and refinement, it is part of the process! At the same time, be mindful of the sunk cost fallacy. If a project has been stuck for over a month with little progress, or if you do not anticipate any progress soon, reconsider whether it is worth continuing.

Keeping a steady flow of new ideas prevents stagnation. Therefore, you should continuously alternate between divergent and convergent thinking processes, placing special emphasis on divergent thinking. Then, once one idea seems likely to fail and a better outside option arises, do not hesitate to pivot. *It is crucial to maintain a pipeline of fresh ideas rather than clinging to a single project at all costs.*

1.3 Step 2: Choosing the Right Methodology

Once you have identified a concrete research question, the next step is to select an appropriate method to answer it. The gold standard in empirical research is the randomized controlled trial (RCT), but this is not always feasible (particularly at early stages of a research career). The optimal method will depend heavily on your specific question (and data availability, if empirical analysis is involved). To answer a question empirically, you might be using:

- Field experiment/RCT (maybe at a lower scale).
- Natural experiment/Instrumental Variables (IV).
- Lab experiments (including online and survey-based experiments, or lab in the field).
- Web scraping
- Administrative data

If you decide to take theoretical approaches, you will have

- Game theory
- Mechanism design
- Matching/coalitional analysis/networks
- General equilibrium
- Computational methods

These lists are non-exhaustive. The guiding principle should be to select a method that is as complex as necessary, *but no more*. Over-complicating a study to showcase technical skills can be counterproductive if it does not align with the research objectives. Nevertheless, exploring new techniques at this stage can be valuable and signal that you are able to work with more than one.

Exercise 3. Designing your methodology
You are interested in understanding "how variations in climate in recent years are affecting and will affect violence in schools in the future". Now take 5 minutes to try to design a study to answer that question. It can be theoretical, empirical, or a mix. It can be a general or partial equilibrium.

1.4 Step 3: Writing and Communicating

In the end, as PhD students, you are expected to produce research papers. Even at the early stage, once your idea and execution plan are ready, you need to start writing - otherwise, ideas stay in your mind or in an "Idea # 34543" folder. Effective writing is essential for academic success, though it can sometimes be quite tough. Some useful resources include:

- *The Little Book of Research Writing* (Varanya Chaubey).
- *A Guide for the Young Economist* (William Thomson).
- Papers by well-regarded writers in your field (e.g., Drew Fudenberg)

Clarity is key. Your audience may not be intimately familiar with your topic, so clearly explain why your research matters in an accessible and engaging way. While AI tools can assist with drafting, always refine the output to avoid overly abstract or convoluted phrasing. If your writing seems confusing to you, it will be even more to people who have not spent so many hours on your topic. Remember, you want to reach as many people as possible, so give writing an important slot in your agenda!

Exercise 4. Writing a clear abstract

Here are two versions of an abstract, a clear one and a not-so-clear AI-created one. Compare and discuss.

> **Version 1:** We examine experimentally the impact of communication on trust and cooperation. Our design admits observation of promises, lies, and beliefs. The evidence is consistent with people striving to live up to others' expectations so as to avoid guilt, as can be modeled using psychological game theory. When players exhibit such guilt aversion, communication may influence motivation and behavior by influencing beliefs about beliefs. Promises may enhance trustworthy behavior, which is what we observe. We argue that guilt aversion may be relevant for understanding strategic interaction in a variety of settings, and that it may shed light on the role of language, discussions, agreements, and social norms in these contexts.

> **Version 2:** This study engages with the experimental investigation of the implications that various forms of communication may have on cooperative behaviors and trust-related dynamics. Through a methodological framework that allows for the discernment of commitments, misrepresentations, and cognitive expectations, we analyze the extent to which individual actions align with hypothesized psychological motivations. The findings are interpretable within a framework where actors exhibit tendencies consistent with the minimization of anticipated emotional disutility, specifically in relation to internalized social expectations, a concept that can be linked to psychological game-theoretic modeling. Under conditions where such behavioral patterns are present, communication appears to exert an influence on decision-making processes by shaping second-order beliefs regarding strategic interaction. Notably, the presence of verbal commitments appears to correspond with an increase in behaviors that could be categorized as trustworthy. The broader theoretical relevance of guilt aversion is discussed in relation to its potential explanatory power for a range of strategic contexts, encompassing considerations of discourse, informal agreements, normative structures, and linguistic conventions.

Hints: Pay attention to how the second version over-complicates sentences, adds vague or overly abstract phrasing, and weakens key points.

1.5 Challenges along the Way

Doing a PhD is tough - nobody should tell you otherwise. However, a positive aspect is that everyone around you is experiencing similar challenges, creating a great opportunity to share thoughts and support each other. Here are some common challenges and tips to help you overcome them.

1. Changing directions

Hitting roadblocks is part of the research process. Have you recently encountered a challenge that forced you to shift focus? If so, reflect on what led to the change and what you learned from the experience. Flexibility is a crucial skill for a researcher, and self-knowledge is something you will be learning along the way.

2. Time management

Time allocation is a common pitfall for PhD students. We often misjudge how much time should be spent on different activities. Rationalize your schedule by prioritizing research and carefully managing external demands (meetings, calls, emails, etc.). The administrative burden is something PhD students often underestimate when starting their studies. Keep in mind that substantial time will likely be devoted to tasks such as applying for grants, managing conference budgets, or invigilating exams. Having a good organizing strategy is crucial, and remember that you can always ask senior PhD students how they handle these challenges. Some examples of productivity strategies include:

- **Dedicated Research Blocks**: Aim for two or three uninterrupted 90-minute research sessions every day - no chats, no smartphone, no distractions, nothing else! You can adapt the duration and frequency according to your schedule and maximum attention span.
- **Identify Your Chronotype**: Align your schedule with your natural energy levels, and plan deep work during peak productivity hours.
- **Learn to Negotiate**: Many academics struggle with negotiation. *Getting to Yes* by Roger Fisher and William Ury is a helpful resource. Effective negotiation skills can help you set appropriate meeting times or manage shared office spaces efficiently.

Exercise 5. Negotiating and "Managing Others"
Scenario: Your office-mate spends large amounts of time on Zoom talking very loudly. It is very distracting. You do not have the luxury of a nice room at home to work, and you need to be at the office. Now get in pairs, and negotiate about the issue. You have 5 minutes.

3. Dealing with Bullying and Harassment

Unfortunately, bullying and harassment occur in academia. If you encounter such behavior, it is important to speak out. Here are some actions that you can take:

1. Confirm that what you are experiencing is bullying.
2. Seek support - most universities have mechanisms in place.
3. Consider both formal and informal complaint routes.
4. Know what to expect after filing a complaint, including potential outcomes, and prepare yourself accordingly.

The most important thing is to feel safe in your workspace during your PhD (and in all aspects of life), so take this as seriously as it deserves. Find someone you trust to talk to, and act promptly!

1.6 Bonus: Preparing for the Job Market - It is too early, and it is not too early

The job market is that scary thing that happens when you are about to finish your PhD. Being ready for it is important, and there are some things that must be considered since the beginning, even though it seems far away. This is because success on the job market requires continuous intellectual engagement, so here are some suggestions that may be interesting to bear in mind since the beginning of your PhD:

- Maintain an idea notebook to capture spontaneous research thoughts.
- Set a weekly goal to generate new research questions.
- Always be prepared to answer: Why is this important?
- Network widely and be open to adjusting your research focus based on feedback.

Overall, the journey you are starting may feel frustrating at times, but it is also one of the most rewarding experiences you can have. By embracing experimentation, critical thinking, and adaptability, you can navigate the PhD effectively and lay the foundation for a fulfilling research career. Two final suggestions would be: 1) Try to concentrate on the activity, not the results. 2) Be ready to explore new ideas and build coalitions.

This was a broad overview of the entire process, so at this point, you may feel a bit overwhelmed - but don't worry! The next lectures will zoom in on specific stages of the journey and help you clarify any doubts about this exciting experience.

2 Lecture 1: Starting a Feasible and Promising Project

Lecturer: ***Libertad González****. Libertad is a Professor of Economics at UPF and BSE, and a research fellow at IZA, CReAM, and CESifo. Previously, she was the President of the European Society for Population Economics and a holder of an ERC Consolidator Grant. Her research spans labor and public economics, focusing on migration, fertility, gender inequalities, and child health and development.*

In this lecture, Professor González focuses on the step of idea generation. As you already know, this is a key step of the process - without an idea, there is essentially no project to pursue. Idea generation is often one of the most anxiety-inducing stages, so getting concrete advice on how to handle it can be particularly helpful! Professor González presents a 6-step process that she thinks are especially effective for developing successful ideas that can become promising papers afterwards.

Contents of this lecture:

To begin with, we need to overcome the tendency to wait for the "*big idea*". A common mistake among young researchers is to delay starting a project while reading broadly across topics in search of a perfect idea. Instead, it is essential to start working on something - feasibility and significance often only become clear after some time and effort. Even if the initial idea does not ultimately work out, the process will deepen your understanding of the literature, methods, and data, making the effort worthwhile!

2.1 Step 1: Select a Field (or two)

The first step in the idea-generation process is to choose a research field, as well as an advisor. This decision is typically shaped by a combination of factors, including your background and training, personal interests, the expertise of faculty members, and the resources available in your department, among others.

Note that your research does not have to align exactly with your advisor's work - you have the freedom to explore your own interests within your chosen fields. In many cases, the fields may already be partially

defined when you enter the PhD program, or it may take shape during your coursework.

It is essential to be aware of the various possible fields and ensure the chosen one genuinely motivates you. Remember, *a field is broader than a topic*: for example, labor economics is a field, while gender or migration are specific topics that can be explored within various fields, including labor economics.

2.2 Step 2: Select a Topic

Within your field of interest, identifying specific topics that intrigue you is crucial. Since research involves a long-term commitment, it is essential to select a topic that excites and motivates you, just as it was when choosing your field. Once again, consider your personal interests, the connections between these interests and your field, and the topics covered in your field courses, for instance. Importantly, the topics you choose do not need to be traditionally associated with your field. You can be innovative here!

Exercise 1: List three topics that interest you
Hints: These topics do not have to be "classic" economics topics. They can be migration, fertility, gender inequalities in economic outcomes, child health and development, poverty, just to name a few examples.

Some more examples: Tax evasion, tax incidence, corporate income tax and automation; populism, social norms, identity and beliefs; climate change, inequalities, subjective well-being.

2.3 Step 3: Find and Evaluate a Research Question

Once you have a topic, the next step is to start formulating research questions within that topic. Remember: a topic is not a question! You can always start out broad from a topic and progressively narrow it down until you arrive at a concrete, *researchable* question.

Here is an example. Say, you define a topic of interest - *migration*. Next, you narrow it down a little bit to, for example, *the economic impacts of migration*, which is still a topic. If you then focus on a specific hypothesized relationship or outcome, you may arrive at a research question such as: *What is the effect of immigration on the wages of native workers?*

Here, for young economics researchers, a practical *tip* is to avoid overly broad "why" questions. Instead, try to formulate more focused questions around specific mechanisms or relationships (Do you have any hypotheses in mind?).

Second, once you have a research question, you need to evaluate it. That is, you need clear criteria to judge whether your question is worth pursuing. Typically, a good research question should be:

- **Relevant**: People care about the question, or you can persuade them to care. At the very least,

YOU should care about it.

- **Answerable**: We don't know the full answer, and so can contribute to closing a gap in the literature.

A common question among early-stage researchers is: Where do research questions come from? One important source is the courses you take. At the PhD level, taking courses is no longer only about acquiring knowledge. Rather, it should serve as an opportunity for you to come up with your own research questions. Taking courses in different fields may help you combine new insights gained in class and arrive at novel research questions.

If you are no longer taking courses, seminars and conferences become key sources of inspiration. Listening to others present their work can often spark your own ideas. Lastly, non-academic sources, such as news articles, podcasts, or even casual conversations with friends and family, can also be surprisingly fruitful in generating good research questions.

Exercise 2: Write down two research questions based on your chosen topics[4]

Okay! So now you've completed the first three steps: selecting a field, choosing a topic, and formulating a question. What next?

At this point, it is tempting to begin reading the literature extensively. Indeed, eventually you will need to have a deep understanding of the existing literature in order to identify your own contribution. However, reading too early can be limiting: your thinking may become overly shaped by what has already been done, and this might lead to a contribution that is too incremental.

Instead, you should first think independently about how you would answer your question. That is to say, now is the time to develop a plan. This brings us to the next step.

2.4 Step 4: Develop a Plan - Approach and Literature

1. The Approach: methods and identification strategies

"An idea = a question + the proposed method to answer it." A well-defined plan that details your method is essential *after* defining your question but *before* touching the data. That said, building such a plan is not always straightforward. A helpful way to get started is by answering the following questions:

- How would you conduct the study with unlimited resources? (What would be the ideal way to answer your question?)
- What skills and methods will be required?

[4] You can also revisit Exercise 1 from the Opening Lecture to deepen your understanding of this step.

A very useful exercise is to imagine the ideal randomized controlled trial (RCT) that would allow you to answer your question. Try to understand what the randomization units would be, what the treatment would look like, and what outcomes you would explore (This exercise tends to be especially helpful for applied microeconomists but can also be useful for other economists.).

This thought experiment will help you visualize the structure of your idea, and, if the ideal RCT is not feasible, find the second-best way to answer the question. For empirical micro research, such second-best approaches include experiments (lab, field, or survey-based experiments) or quasi-experimental methods (Instrumental Variables, Difference-in Differences, Regression Discontinuity Designs, etc.).

What NOT to Do

1. Don't just find some data and start playing with it and running regressions!
2. Remember: simply exploring the data and conducting exploratory regressions is *not* an identification strategy.

Be ambitious and proactive. If the data you need is not available, look for ways to obtain it or consider generating your own. If you lack certain technical skills, take the initiative to learn them. If the project is time-consuming, start early. If funding is needed, explore potential sources and seek advice from advisors.

But most importantly, leverage your *unique* perspective: Your background, training, experience, and network provide you with a comparative advantage. Think about research settings and perspectives that you are especially well-positioned to explore.

Key Questions for identifying an approach

1. What would an ideal Randomized Controlled Trial (RCT) look like for answering your question?
2. If an RCT is not feasible, can you identify a natural experiment or quasi-experimental setting?
3. Or, which methods are appropriate (e.g., theoretical models, structural models, causal inference techniques, experiments)?
4. What assumptions will be needed to answer the question credibly?

2. Conducting the Literature Review

Finally! Once you have drafted a research plan, conducting a literature review helps you understand what others have done and identify your potential contribution. You should expect to find papers addressing similar questions - this is entirely normal (after all, many people are doing research!). The key is to identify what you can do differently or better.

However, bear in mind that you should avoid doing the following (as they do not constitute a meaningful contribution):

- Simply replicating a study using data for a different country, without a clear motivation for what new insights could be obtained.
- Claiming to "add to the literature on something" without a clear innovation.
- Writing a model (per se) without empirical relevance or support.
- Claiming to be the first to study something.
- Reporting mixed results in the current literature without a compelling explanation of how your project improves on it.

A true contribution should provide new knowledge that was previously unknown, improve existing methodologies, or offer more credible identification.

2.5 Step 5: Seek Feedback and Refine the Idea

Okay, now you have a clear question and a structured plan to answer it. However, before moving on to data analysis, it is important to discuss your idea with supervisor(s), peers, and other faculty members. Their input can help you refine both your research question and your approach.

One of the key roles of your *supervisor(s)* is to leverage their experience and expertise to help you evaluate whether an idea is worth pursuing. So listen carefully to their suggestions! (Fortunately, at UB we also have our *Research Ideas Coffee Session Series*, which provides an excellent opportunity to receive feedback from your peers. Make sure you participate actively and benefit from it!)

It is worth repeating that this step is crucial. Keeping your ideas to yourself might cause you to miss valuable perspectives or overlook potential issues. Feedback is an iterative process and may lead you to revisit earlier steps. That's not a waste of time - on the contrary, it's part of becoming an expert in the research process (which is actually what a PhD is for). So don't be afraid to ask for help and comments!

The Feedback Loop

After receiving feedback, you will often find yourself at a crossroads. Your advisor, or other people you consult, may give you one of three types of responses:

- **Great idea!** If the project is well-motivated, feasible, and methodologically sound, go ahead and proceed to execution (proceed to Step 6).
- **The project is not worth pursuing.** Perhaps the contribution is too minor, the identification strategy is not credible, or the question is infeasible. (Go back to Steps 2 or 3, and refine your topic or question.)

- **Unclear.** If they are unsure, talk to a few more people, or go back to Step 4 and work further on your plan. Then repeat Step 5.

Remember: iteration is part of the process. Hearing "no" or "not sure" is not failure - it's a prompt to think deeper and improve your idea.

2.6 Step 6: GO! Move Forward with Your Project

Arriving at this point is great news. With the "OK" from your advisor, you now have something to work on! While the execution stage is beyond the scope of this lecture, it is worth noting that there will be many back-and-forths, so make sure to ask for feedback frequently. You can present at internal seminars, talk with additional faculty members, visitors, and more senior students in close fields. From here on, the focus shifts to other key aspects of the research process, which will be addressed in the upcoming lectures.

2.7 Bonus: Do's and Don'ts for Junior Researchers

To wrap up the lecture, here is some extra advice from Professor González. First, a few common practices that may be unhelpful during the PhD, or when pursuing an academic career in general:

- Shying away from discussing your project: Engaging with others helps refine your ideas and improve research quality (we've already emphasized how important feedback is!).
- Rushing into data analysis without a clear plan: Always develop a well-defined identification strategy first.
- Starting with a dataset or policy change rather than a research question: Research should be driven by important questions, not just available data (though data and context can certainly inspire good questions).
- Following trends instead of personal interests: Study what you are genuinely passionate about.

And also, some general advice that should benefit your PhD journey:

- Keep your advisor(s) regularly updated.
- Get to know your data REALLY well (read the codebooks, explore descriptive statistics, ...).
- Start from a *relevant* question.
- Work on a topic that you truly care about.
- Think hard about your (potential) contribution to the literature.

The research process is iterative and requires persistence, creativity, and flexibility. Your initial idea will likely evolve over time, and that's perfectly fine. What matters is keeping your research question and approach clear. This will help you develop a solid project and ensure a meaningful contribution to the field. And most importantly, HAVE FUN!

3 Lecture 2: Doing Organized and Replicable Research

Lecturer: ***Joan Llull****. Joan is a Professor of Economics at IAE-CSIC and BSE. He is the Data Editor of Econometrica, Quantitative Economics, and Theoretical Economics. He was awarded an ERC Starting Grant and an ERC Consolidator Grant. His research is mainly in labor economics and structural microeconometrics, focusing on migration, inequality, human capital accumulation, technological progress, and labor market frictions.*

In this lecture, Professor Llull will present the replication package as a conceptual and practical framework for doing organized research through better managing data and coding. He provides an overview of replication packages and goes through concrete steps in building a replication package, which is not just to make it easier for others to reproduce your work, but also to help yourself keep track and build on you own work.

Contents of this lecture:

- Introduction: Reproducibility and Replication Packages
- Part 1: Managing Data
- Part 2: Coding Organized Projects
- Bonus: Main Takeaways and Ten Simple Rules

3.1 Introduction: Reproducibility and Replication Packages

Organizing the project according to the standard and structure of a replication package is not just about preparing for journal requirements or increasing replicability for others, but also having a disciplined framework for organizing the entire research workflow. Rather than treating it as a final-stage obligation, we should use the logic of a replication package to guide our work from Day 1: organizing and documenting raw data, scripts, and outputs with the end product in mind[5].

Reproducibility and Replicability in Economics

Reproducibility, defined as obtaining the same results using the same data and same methods, is what most journals verify before publication. It is the cornerstone of transparent and credible empirical research, allowing the profession to build incrementally and confidently on prior work. However, reproducibility is just the first step. The broader profession may later explore replicability, a more demanding standard: whether the same results can be recovered using new data, different methods, or both.

[5] Although it's not covered in this Lecture, you can also find information on how to track version control of a co-authored project here.

The responsibility of ensuring reproducibility falls not only on journals but also on us, the researchers. Reproducibility is directly tied to the impact of your research. Papers with accessible and well-documented code are easier for other researchers to engage with, cite, and extend. If you want your research to really have an impact, you should allow others to reproduce it and build on it.

In economics, a number of initiatives are actively promoting reproducibility and replicability. Journals now often require data and code availability statements, and some have appointed dedicated Data Editors to conduct pre-acceptance reproducibility checks. The Social Science Data Editors have developed shared standards, such as the DCAS and a template README file, to harmonize expectations across journals. Public platforms like the Replication Wiki curate over 16,000 links to published replication packages, while organizations such as the Institute for Replication (I4R) carry out post-publication checks and host Replication Games, bringing students and researchers together to test findings.

Replication Package: a Framework for Organizing a Project

When we refer to "replication packages", we are actually taking about "reproduction packages", in the sense that these packages typically allow readers to use the same data and follow the same methods to obtain the same results of a paper.

A replication package is not just a technical necessity; it's a matter of integrity and communication. Think of it as a **teaching tool** for your work, offering a clear step-by-step guide to how you reached your findings. Also, remember, future-you will thank you when revisiting the project for revisions or extensions. In contrast, an opaque or confusing package will frustrate users, diminish the impact of your research, and even prevent yourself from understanding the previous exercises.

It is essential to develop the habit of **creating a replication package from Day 1 of a project**. This offers several advantages.

1. It helps your future self: We often need to revise and update, and having the project organized towards a replication package makes this easier.
2. It fosters scientific openness, and you will be requested to publish your package at some point.
3. Packages that pass pre-acceptance checks often speed up the publication process (surprisingly, about 75% of submissions fail at least one check).
4. Replication packages increase the visibility and impact of your research: Data from the Econometric Society shows that a typical package is downloaded around 100 times per year.
5. A well-organized package is less likely to contain unnoticed errors and boosts the credibility of your research.

Structure of the Replication Package

But what exactly should we do to organize the project folder as a replication package? Or, what does a replication package contain? An ideal replication package typically contains **16 standard elements**[6] according to the Data and Code Availability Standard (DCAS), which aims at harmonizing the requirements from all journals. Broadly, a replication package (and your project folder) consists of **data, code, and documentation**[7]:

- **Data:** Include a Data Availability Statement, raw datasets, and, where relevant[8], the processed datasets used for analysis.
- **Code:** Provide all scripts needed to transform raw data into analysis-ready data, and to produce the final tables, figures, and other results.
- **Documentation:** This ties everything together. It should explain how to run the scripts, how to access external data (when it cannot be shared), and how any experiments or surveys were conducted, etc. At a minimum, include a **README file**.

Next, we will review the **best practices** for creating a good replication package (and for organizing your project), divided into two parts. First, we'll look at how to manage your data. Second, we'll walk through how to organize your project files on your computer. Each part is presented as a step-by-step guide.

3.2 Part 1: Managing Data

At its core, data management is about enabling replication. Even with flawless code, replication is impossible without the underlying data. Therefore, the first part of this lecture is devoted to how to manage data, divided into three steps: knowing the principles, dealing with different types of data, and citing and enabling access.

Step 1: Know the Principles

The guiding rule is simple: **Make your data available.** In case this is not possible due to legal restrictions, cite the data properly and describe how others can access it.

1. **General Principle**
    - As a default, include all datasets necessary to replicate your results.
    - Exceptions apply only when you lack permission to redistribute the data, or when your exact data extract is already available in a trusted repository (see below).

[6] See DCAS for a full list.
[7] See Section 3.3 for a detailed recommended folder structure in practice
[8] Meaning when creating the analysis-ready data is a lengthy process.

2. **Best Practices for Sharing Data**: To ensure others can fully understand and reproduce your work, share:

   - Copies of your datasets in non-proprietary formats (e.g., `.csv`, `.dta`, `.txt`, `.parquet`).
   - Raw data, along with cleaning code and documentation.
   - Comprehensive metadata and data dictionaries describing variable names, units, coding schemes, and data sources. This is especially crucial for datasets with abbreviations, survey codes, or administrative data.

   If you cannot share certain datasets, explicitly state this in both the README and metadata files, and explain the access conditions (e.g., through a formal request process).

3. **Trusted Repositories and the FAIR Principles**: If you don't upload the data yourself but it already exists in a repository, the repository should adhere to the FAIR principles:

   - **Findable:** Searchable via metadata and a persistent identifier (e.g., DOI).
   - **Accessible:** Downloadable using standard open protocols.
   - **Interoperable:** Compatible across different software environments and integrable with other datasets.
   - **Reusable:** Distributed under a clear license, with documented provenance and in standard formats.

   Trusted repositories include Zenodo, Dataverse, and OpenICPSR. Platforms such as GitHub, Dropbox, or personal websites do NOT meet FAIR standards.

Step 2: Dealing with Different Types of Data

1. **"Raw Data"**: Understanding the meaning of raw data is essential in empirical work:

   - Raw data **includes:** Datasets saved in a different format or with unused variables removed.
   - Raw data does **not include:** Datasets where observations were dropped or variables were transformed.

   This matters because raw data should allow others to recreate your cleaned dataset exactly from the original extract. The best practice in managing raw data is to keep it in read-only mode to prevent accidental changes. Besides, for historical or scanned data sources (e.g., old PDFs or digitized books), include them in a dedicated `Documentation` folder.

2. **Non-Redistributable Data**: If you cannot share certain data, you must request an exemption — typically during journal submission. Common reasons include using confidential administrative data, commercial restrictions, or public datasets with non-redistribution clauses. Even in these cases:

- Others should be able to apply for access under the same terms.
- The data extract should be archived (e.g., institutionally) to preserve reproducibility.
- Ideally, there should be no public alternative data that could yield equivalent insights.

A good practice when sharing real data isn't feasible is to provide **synthetic or simulated data**:

- This enables users to run the code and understand the workflow, even if they cannot replicate the exact results.
- Particularly useful for complex data structures such as panels, spatial data, or matched samples.

3. **Self-Collected Data**: If the data is collected by yourself, you cannot restrict it for your exclusive use after publication. You must:
   - Share relevant materials (e.g., questionnaires, methodologies).
   - Archive the raw, anonymized survey data (rather than just the cleaned version).
   - Follow IRB guidelines for anonymization and data publication.

Step 3: Citing Data and Enabling Access

A well-crafted Data Availability Statement (DAS) in the README file (more details about the file will come later) is essential. It provides transparency about every dataset used, including those you do not include in the replication package. Your DAS should cover:

- **Provenance**: The origin of the data — who collected it, when, and for what purpose.
- **Access information**: Where the data is hosted, with working links when possible.
- **Instructions**: If the data is not directly downloadable, explain how others can obtain it (e.g., by submitting a form, emailing the provider, paying a fee, or applying for access).
- **Primary data**: If you collected the data, describe your data collection procedures in detail, including sampling design, timing, questions, and instruments.
- **No data sharing?** Then your DAS must be especially detailed, making it clear what data is unavailable and why.
- **Clarity**: Ensure that readers can easily match datasets listed in your DAS with the files provided in your replication package.
- **Citations**: Include a references section in your README file, just as you would in a paper, to properly cite data sources and providers.

Ultimately, data sharing is fundamentally about **trust and transparency**. It demonstrates to the profession that your findings are not the result of a black box. By sharing the pieces (data, code, and documentation) you allow others to reconstruct the machine and confirm that it works. It also means thinking beyond immediate publication, toward the long-term reproducibility and accessibility of your work.

3.3 Part 2: Coding Organized Projects

As discussed in the Introducation, the replication package can serve as the central framework for organizing a research project — not something assembled at the end, but a structure that takes shape from Day 1 and evolves as the project develops. Delaying the creation of replication materials until after a paper is conditioanlly accepted often leads to disorganization and increases the risk of errors.

Step 1: Set up an Organized Project Folder

Imagine a future replicator (a reviewer, a PhD student, or even your future self) who encounters your code without prior context. Will they understand your logic, structure, and reasoning? Would they be able to replicate your results confidently?

This mindset of **computational empathy** encourages you to design your replication package with clarity and user-friendliness in mind. Every decision you make, from file organization to code comments, should aim to reduce the effort required for someone else to reproduce your findings.

In this spirit, establishing a clear and consistent folder structure lays the foundation of a replication package and for your own efficient research. It helps reduce the risk of losing track of steps, simplifies the revision process, and fosters good habits from the beginning. Here's an example of a recommended folder structure for a Stata user:

```
Project_Folder (Future Replication_Packge)/

 1_Data/
    Raw/        # Downloaded, unaltered data (e.g. CPI indices, shapefiles, surveys)
    Cleaned/    # Final datasets used for analysis (often left empty for replication uploads)

 2_Code/
    2025_01/    # Versioned by month; create a new folder when starting a new round of work
       Do.files/         # Main scripts organized by purpose
          0_Main.do
          1_Clean.do
          1_CL_undoc.do  # A special cleaning step, if applicable
```

```
      2_Descriptives.do
      2_DS_cohort.do
   Log.files/        # Logs for each run
   Tables/           # Tables generated during analysis
   Figures/          # Figures generated during analysis
   Stata_packages/   # Custom Stata ado directory for package reproducibility
2025_02/
...

3_Documentation/
   README.txt   # Instructions and metadata for replicators

4_Output/        # Optional: If tables and figures are too large, store final outputs here
                     # Otherwise, keep them within the corresponding Code subfolders
```

Here, separating raw data from cleaned data prevents confusion and accidental overwriting. Naming code folders by year and month (e.g., `2025_01`) helps track the evolution of the project. When moving forward, simply copy the code folder from the previous month to make a new one for the following month.

Remember, all we do here to set up an organized project folder from the very beginning isn't just about replication; it's about working efficiently and transparently from the very beginning.

Step 2: Code in an Organized Way

Zooming in on the `Code` folder, computational empathy is again central here. You should include all the code for data transformation and analysis. This makes explicit all steps and decisions, and allows readers to inspect your methods and reuse them more easily.

1. **Programs to include in the folder**

   - Data cleaning scripts that process raw files into usable formats.
   - Scripts for merging or reshaping datasets, and for constructing derived variables or indices.
   - Estimation and model-solving codes.
   - Simulation and visualization codes.
   - Code that generates tables, figures, and supplementary appendices.

   If you use manual software such as ArcGIS, you should also provide detailed step-by-step instructions for your exercises. However, it is recommended to use scripts whenever feasible.

2. **Have a master file**
   Moreover, the `Code` folder is not just a collection of scripts, but a complete, transparent record

of your computational workflow. To achieve this, begin by having a "**master file**" that runs everything, typically named `0_Main.do` (or `.jl`, `.py`, `.m`, `.R`, etc.). Ideally, this script is the one that:

- Sets up the computational environment by defining paths, loading required packages, and setting global macros or variables.
- Sequentially calls all other scripts in the correct order.
- Automates the entire analysis pipeline, from processing raw data to generating final tables and figures, so that a single run can reproduce all your results.

3. **Pay attention to paths**
   Always use relative paths. Set global paths at the top of main scripts (for example, in Stata: `global path "C:/MyPath/ReplicationPackage"`), and avoid absolute paths that tie your package to a specific computer, and write relative paths like `"$path/Data/Raw/census.csv"` so others can easily rerun your code by changing the first main line in the master code that defines the path. When writing paths, use **forward slashes (`/`)** to enhance cross-platform compatibility.

Bonus Tips: Extra good practices that show computational empathy

- Use subfolders generously to organize code, data, and outputs neatly.
- Keep individual scripts short and split long ones into sub-files.
- Prefer open, non-proprietary formats like `.csv` or `.xlsx` for data files.
- Keep track of dataset versions, dates, and download links.
- Give meaningful names to programs, variables, files, outputs, etc: use `avg_age`, not `v22`.

Step 3: Keep Track as the Project Moves On

As your project evolves, it's crucial to maintain clear, organized, and reproducible records of every step. This means not only writing clean, well-commented code but also ensuring that others (and your future self) can understand, replicate, and trust your entire workflow. Remember, computational empathy!

1. **Document every step to run the code**
   Every script you write along the way should be thoroughly documented. Consistently comment (and indent) your code to explain the logic and structures.

   In particular, build the **README file** as you go to explain how to reproduce your work. Basically, it should include:

   - The Data Availability Statement.

- A list of all datasets (included and omitted).
- Computational requirements (e.g., software versions).
- Descriptions of each script or program.
- Instructions for replicators (including run order).
- Citations for any reused data or code.

Following the Social Science Data Editors' README template can save time and help prevent errors. With the README file, you are able to provide a high-level overview of the program structure, list manual actions, and link outputs to scripts and files systematically. You should also cite third-party libraries and data in the README.

2. **Describe software and hardware requirements**
   Even perfect code can fail replication if others don't know what environment is needed to run it. Anticipate these questions:

   - Does the code run on a standard laptop, or does it require a workstation?
   - How long will it take to run? Seconds, minutes, hours, days?
   - What software and packages are needed (and their versions)?

   To make these clear for the replicator, in the README file, you should:

   - Specify precise software versions (e.g., Stata 17 SE, R 4.3.1, Python 3.10 with pandas 2.1.0).
   - List all critical libraries and packages - if the code breaks without them, name them; document their versions, as well as installation instructions.
   - List hardware used (e.g., Mac M2 with 16GB RAM) and note if it can run on less powerful ones.
   - Estimate running time by code section (e.g., data cleaning: 2 hours; regressions: 10 minutes).
   - If your package is too time-consuming: prepare a package that allows for a partial replication (e.g., include already-clean data in an intermediate file).

3. **Automate and save everything**
   Automation not only ensures consistency but also saves time when revisiting or updating the analyses. Manual steps, like copying regression outputs into tables or figures, introduce errors and reduce reproducibility. Instead:

   - Automate the generation of all tables and figures directly from code.
   - Save all outputs automatically and compute all in-text results from code.
   - Record logs and diagnostics for every script, and include them in your replication package in an accessible format.

- Automate package installations if possible.

In summary, you need to keep track of your own practices and provide clear instructions, with the goal of guiding the replicator in your mind; don't assume that users will "figure it out".

Step 4: Wrap Up with the Replication Package

Upon completion of the project (especially for a journal submission), it's time to finalize and polish your replication package. This step goes beyond simply re-running your code: it's about ensuring that your entire workflow is clear, reproducible, and ready to be shared with the world.

1. **List all exhibits that can be reproduced in the README**
   Transparency requires spelling out exactly what outputs your package will reproduce. Include a table that maps exhibits (e.g., tables and figures) to the corresponding scripts, output files, and line numbers. Indicate whether each exhibit requires confidential data. This gives replicators a clear roadmap of what to expect.

Table 1: List of Reproducible Exhibits

Exhibit	Script	Output File	Line Number	Confidential Data Needed?
Table 1	`regressions.do`	`table1.tex`	line 45	No
Figure 3	`plots.jl`	`figure3.pdf`	line 101	Yes

2. **Include all supporting materials**
   Now, before finishing your package, double-check whether it includes all the supporting materials. You should (whenever applicable):

   - Include survey instruments or experimental instructions
   - Original questionnaires, translations, sampling schemes, and reward structures.
   - If you used software (Qualtrics, SurveyCTO, zTree), share configuration files.
   - Include ethics approvals (IRB protocol number, institution).
   - Share pre-registration links (permanent URL or DOI).

   Remember, you should not include the manuscript itself unless journal policies explicitly allow it.

   At this point, also double-check if you have the complete and up-to-date README file. Then make sure that you use standard formats (PDF, TXT, etc.) for documentation files where possible.

3. **Tag a permissible license**
   Remember, sharing the replication package doesn't mean readers can use it without permission. A license clarifies what others can do with your code and data:

   - **For data:** Use a Creative Commons license (e.g., CC-BY for reuse with attribution, or CC0 for public domain).

- **For code:** Use standard open-source licenses such as MIT or BSD.

You should document the license clearly in your README file. Clear licensing not only boosts reuse, citation, and impact but also reduces legal ambiguity. At a minimum, ensure that your license allows replication, but check repository or journal policies—some impose specific licenses.

4. **Finally, re-run everything once again**
   Once you believe everything is in place, before zipping and sending it, perform a full dry run of your replication package. This process often reveals hidden issues such as hard-coded paths, undocumented steps, or missing dependencies. The goal is to confirm that your package runs seamlessly from start to finish, without any manual intervention. This can help you detect simple mistakes such as typos, file misnaming, and version mismatches.

   - Run the main script from top to bottom.
   - Ask a non-involved colleague or RA to test the package without your guidance - this mimics the true experience of an external replicator.
   - Run it in a clean environment (different machine if possible).
   - Confirm that everything runs from the package only.

   This final check helps you have fewer and faster iterations with Data Editors and leads to a faster publication process.

If your research process has been organized from the start, the replication package basically builds itself - just zip the project folder, and you're done! Magic! Now, the last thing to do is archive your replication package in a reliable, long-term, and citable repository. Choose a repository that issues a DOI, making your replication package easy to cite. If the journal does not specify a platform, you can consider OSF (Open Science Framework), Dataverse, Zenodo, and AEJ Data and Code Repository (for AEA journals). Remember, you are the one responsible for the package uploaded there, not the journal!

3.4 Bonus: Main Takeaways and Ten Simple Rules

Over the past years, we have seen a solid trend for Economics journals to ensure reproducibility, which is a necessary step to encourage replication. This lecture has explained the importance of reproducibility and how to build a replication package for an economics research project. It is worth repeating by the end of the lecture that documentation (README file) is essential. Also, keep the raw data, carefully read the terms of use, and, if necessary, apply for exemptions at first submission.

Let's now conclude this lecture with the "10 simple rules for creating a replication package" (Vilhuber, Koren, Csóka, Connolly, and Llull):

Rule 1: Computational Empathy!

Rule 2: Make your data available.

Rule 3: Cite data and describe how others can access.

Rule 4: Describe software and hardware requirements.

Rule 5: Provide code for data transformation & analysis.

Rule 6: Explain how to reproduce your work.

Rule 7: List all exhibits that can be reproduced.

Rule 8: Include all supporting materials

Rule 9: Tag a permissible license.

Rule 10: Re-run everything once again.

In the end, a complete and well-documented replication package is more than just a part of the submission. As researchers, following the steps to build your replication package the day you start your project creates a perfect opportunity to ensure an organized research process. In this sense, a replication package serves as a framework for organizing a research project!