Policy extrapolation: a review of problems and current solutions

Deliverable D2.3

Milestone 2

Coordinator of this work: Università di Teramo

Authors: Simone Busetti, Lorenzo Pagni

info@evaluatingfoodpolicy.it

Finanziato dall'Unione europea NextGenerationEU

PRIN 2022-PNRR-SH2-Prof. SIMONE BUSETTI-PI Unite (codice P2022XWBT8_01) Finanziato dall'Unione europea – Next Generation EU – CUP C53D23009980001

Finanziato dall'Unione europea
NextGenerationEU

1. Introduction

External validity and extrapolation are crucial concepts in scientific research, especially in social sciences and policy evaluation. External validity refers to the extent to which the results of a study can be generalized beyond the specific sample and context in which they were obtained (Cook & Campbell, 1979). This is a fundamental aspect of empirical research, as it determines the applicability and relevance of the findings to broader populations and different settings. High external validity means that the study's conclusions can be extended to other groups, settings, and times with confidence. Factors that influence external validity include the representativeness of the sample, the ecological validity of the study environment, and the robustness of the experimental design (Shadish et al., 2002).

Extrapolation, although for some authors a synonym of external validity, involves applying the findings of a study to populations or settings not directly examined or targeted by the study. It is a process that extends the inferences made from the study to different contexts. Extrapolation is particularly important in fields like medicine, education, and social policy, where direct experimentation on all possible contexts is impractical or unethical. The process of extrapolation requires careful consideration of the similarities and differences between the original study conditions and the new contexts to which the findings are being applied (Berk, 1983).

External validity and extrapolation are essential for policymakers because they ensure that the results from research studies can be applied to broader populations. For instance, a policy intervention that works well in a small, controlled study setting, may not have the same impact when applied to a larger, more diverse population. Policymakers rely on research findings to design and implement effective interventions, but without effective extrapolation, there is a risk that these interventions may not achieve the desired outcomes in real-world settings (Rossi et al., 2004). External validity allows researchers and policymakers to make informed predictions about how a policy might perform in different contexts, which is critical for effective and efficient policy design and implementation. It helps in predicting the success or failure of policies before they are widely implemented, saving time, resources, and potential negative impacts on the population.

Methodology

Google Scholar was used as the primary search engine. The following keywords were utilized: "External validity" yielding 1330 results, "Extrapolation" with 12100 results (most of which pertained to statistical and mathematical modeling, with relevant

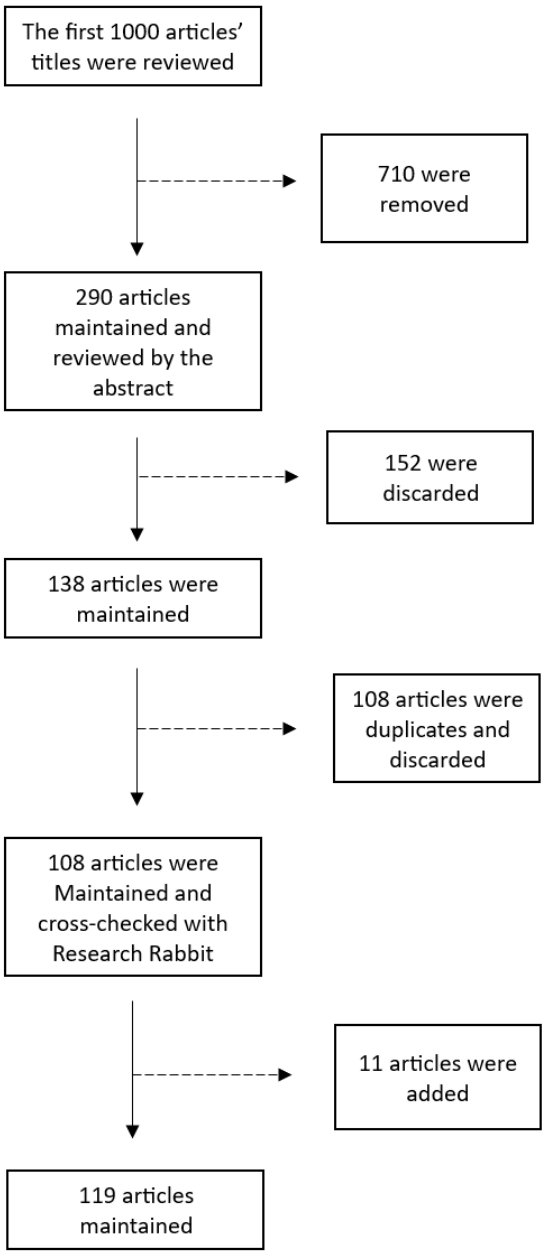

Figure 1 Diagram of the review process

contributions selected for species extrapolation), "External validity + Policy" with 47 results, "Extrapolation + Policy" with 13 results, "External validity + Policy program" and "Extrapolation + Policy program," both of which returned no results, "External validity + Intervention" with 28 results, and "Extrapolation + Intervention" with 5 results. Citations were preemptively excluded from the search.

Subsequently, the papers were analysed by asking ChatGPT-4 to answer the following six questions for each selected paper:

1. What are the main problems in extrapolation / external validity raised by the article/book?
2. What features of the new context and/or target population can affect extrapolation / external validity?
3. Does the article/book mention examples of problematic extrapolation/external validity? Which one?
4. Does the article/book mention examples of successful extrapolation/external validity? Which one?
5. What are the methods for fixing problems of extrapolation / external validity?
6. Consider the case of transferring an urban food security program from one city to another, what would this article suggest to do?

These questions aimed to capture the authors' understanding of external validity and extrapolation, existing challenges, and proposed methods. They also aimed to have

a practice-oriented view on the topic by searching for examples and contextualising the approach to the case of food security policy.

After having analysed the paper individually, we grouped papers to have a comparative analysis of their relevance to the case of food security programs; we submitted groups of ten papers and asked ChatGPT-4 to perform a comparative assessment:

7. Consider the case of transferring an urban food security program from one city to another, can you rate the relevance of the proposed approach of this article on a five-point scale (1 = not relevant, 5 = very relevant)?

The result was an evaluation of all papers that ranged from 3 to 5. At the end of the process, 28 papers received a relevance rating of 5 based on the specified criteria.

The next section includes all evaluations and insights derived from the reviewed papers. For each fo the 119 article and book, we produced a short sheet describing the topics covered by the paper and reporting its relevance for the case of food security programs (see deliverable 2.2).

All short sheets were read and analyzed, and the most relevant papers were selected for full reading (all those rated 5 plus all articles considered relevant after reading the short sheets). The following sections report the information gathered on definitions, challenges, methods, and examples of external validity and extrapolation.

3. Definitions of external validity and extrapolation

External validity and extrapolation are critical concepts in research to determine the generalizability of study findings and their applicability to broader contexts, populations, and times. In the next two sections, we report existing definitions of the two concepts, with the warning that some authors use them interchangeably, as two features of the same theoretical framework, or they use one or the other but with overlapping meaning. In our reading, the two terms can be considered synonymous. However, authors using external validity stress the representativeness and generalisability of findings to broader unspecified settings, while extrapolation is used for the process of applying findings to specifically identified contexts. For the objectives of the EFP project, we will use the term extrapolation.

3.1 External validity

External validity mainly refers to the extent to which the results of a study can be generalized beyond the specific conditions of the original research.

According to Cook (2014), external validity is about generalizing causal knowledge obtained about a treatment to other settings or units (i.e., persons). In his view, extrapolation is one function of the needed procedures to obtain external validity: the first is the *representation function* (specifying what the original sampling particulars–treatment, units, settings, times–represent as more general populations or categories), the second is the *extrapolation function* (drawing conclusions about persons, settings, times, and treatment and outcome variants that have different attributes from those observed at the sampling).

Rothwell (2005) describes external validity in the context of randomized controlled trials (RCTs) as the extent to which the results of an RCT can be generalized to the wider patient population, settings, and times. He highlights the importance of representativeness and applicability of findings beyond the study sample. External validity is considered a fundamental feature of the relevance of research findings, since only if externally valid they can meaningfully inform routine clinical practice and ensure the applicability of treatments to individuals outside the original study sample.

Tipton and Peck (2017) define external validity as the degree to which the results of an intervention can be generalized to other settings and populations. The authors focus on generalizability; one feature of external validity considering the ability to relate the sample of units and settings found in the original study to the set of units and settings in the population.

Burchett et al. (2011) consider *external validity* as a generic concept related to the likelihood that a study's findings could be generalized to other (unspecified or more general) samples or settings. In analysing external validity and the practical use of research findings, they also distinguish between applicability and transferability. *Applicability* refers to the likelihood that an intervention could be implemented in a new, specific setting; this entails the need to adapt, tailor or 'individualize' interventions and programmes to ensure their appropriateness for one's local setting. Applicability is a precondition to *transferability*, i.e., the likelihood that the study's findings could be replicated in a new, specific setting (i.e. that its effectiveness would remain the same).

Finally, Williams (2020) contends that external validity is often framed as a question of *generalizability*, i.e. whether the impacts of a policy evaluated in a specific context are the same in other, unspecified, contexts. Instead, the policy-relevant question on external validity is one of the *applicability* of evidence: how evidence gathered from another context can be applied to our specific context. In this respect, external validity cannot be judged per se; it can only be determined by an understanding of the specific features of the destination context that might interact with the mechanism

of the policy. Indeed, external validity failure (the policy has different effects in different contexts) happens when its mechanism (or theory of change) interacts with a difference in the new context.

3.2 Extrapolation

Similarly to external validity, extrapolation is generally used to mean the potential of applying the findings from a specific study or set of studies to different contexts, populations, or times. However, most uses of the term focus on applying study findings to a specific target (context, population, time) instead of looking at wider generalisability and representativeness.

Steel (2007) refers to extrapolation as the ability to transfer causal generalizations from one context to another. The importance of extrapolation is that evidence concerning the model or original population is more accessible than that for the target with which one is presently concerned. The extrapolation problem (and extrapolation failures) derive from the possible heterogeneity between the context of the original research and the one of the target context. In addition, the extrapolator's circle may plague strategies for extrapolation. The circle happens when, in order to know if findings from the source case can be extrapolated to the target case, one needs so much previous information about the target context (for instance, to assess its similarity with the source case and the effectiveness of a treatment) that knowledge about the source case becomes irrelevant.

Bardach (2004) defines the 'extrapolation problem' as a special case of the generalization or external validity problem. In his account, extrapolation refers to the practical activity of using evidence from a source site to solve the same policy problem by replicating and adjusting the intervention in the target site. The extrapolation problem arises because of the heterogeneity of the two sites, the faulty analysis of the source site and what should be transferred, and the assumption that in real settings, a strictly *faithful replication is almost always impossible* (hence, not even the treatment or intervention can or should be replicated as such).

Khosrowi (2022) argues that one should distinguish between problems of extrapolation and extrapolative inferences. Problems of extrapolation derive from having two populations where a causal effect learned in one shall be used to infer a causal effect in the other; the challenge is that the two populations might differ in causally relevant ways that can matter to the extrapolation of the effect. Extrapolation inferences regard assumptions about the similarities and differences between the two populations and how these latter can affect 'effect evidence' from one population to the other.

Bareinboim and Pearl (2013) define extrapolation as a synonym for *transportability,* i.e. the generalization of causal effects across populations. The authors provide a formal framework for deciding when and how causal effects can be extrapolated, emphasizing the need for understanding the mechanisms and conditions that underlie the original findings.

4. Challenges in external validity and extrapolation

The challenges of external validity and extrapolation are critical in the field of policy evaluation and research. Numerous issues can arise when attempting to generalize, transport or extrapolate research findings, starting with the selection of the study populations, the design of the study, contextual variation, the size of the intervention, implementation and scalability, and the mechanisms underlying the intervention. All these issues are particular examples of the fundamental problem of extrapolation: the difference between the original and target context.

4.1 Context

The issue of context is a central challenge in the discussion of external validity and extrapolation in policy evaluation. Context refers to the specific settings, conditions, and circumstances under which a study is conducted or a program implemented. Several authors have explored this issue, highlighting the complexities and challenges involved.

Pritchett and Sandefur (2015) address the challenge of context in evaluating the external validity and extrapolation of social programs in development economics. They use the Root Mean Squared Error (RMSE) to compare the performance in assessing the external validity of two kinds of evidence: observational data from the program's context and experimental evidence from different contexts. Using microcredit studies they show that non-experimental data within context often produce more reliable estimates while only if the number of experimental data increases significantly does its reliability improve. The greater performance of observational data 'within context' demonstrates the complex nature of social programs, the fundamental causal role of contexts in determining results and hence the difficulty in generalizing findings across different settings.

Pritchett and Sandefur highlight several challenges to external validity, two of which are worth mentioning here. First, 'we don't know what context means'. While hard sciences have parameters influencing relations that are known with engineering precision, social programs work in contexts characterised by a long list of unknown factors interacting in unknown ways. In order to increase external validity, there is a need for a solid theory of what context is for a given program. Second, programs

change across contexts in both their design and implementation, so that broad classes such as ‘microcredit’ or ‘pay-for-performance’ have doubtful construct validity (i.e. they may represent profoundly different programs). This means that programs change with contexts and that findings from multi-site studies may miss important program variations.

Similarly, Bold et al. (2013) investigate the challenges of external validity and extrapolation in the context of educational interventions in Kenya. They note that while randomized trials can provide internally valid measures of causal effects, their applicability to different contexts remains questionable. A significant challenge is the heterogeneous treatment response, suggesting that an intervention's effectiveness observed in one population may not translate to another. When implementing programs in heterogeneous populations, responses will necessarily vary. More so, even when populations are homogeneous, treatment effects can vary based on the implementing institution. For instance, the study compares the outcomes of a contract teacher program implemented by NGOs versus the government, revealing that the intervention's success is heavily influenced by the implementing subject (namely, NGOs show a better performance).

Vivalt (2020) addresses the issue of external validity with a new database of 15,024 estimates from 635 papers on 20 types of interventions in international development, gathered in the course of meta-analysis. On the one hand, the analysis shows that smaller studies tend to report larger effect sizes, as do programs implemented by non-governmental organizations (NGOs) or academics. On the other, a notable result is that studies of interventions that may have a more direct causal effect exhibit less heterogeneity in treatment effects. Taken together, these results suggest greater attention be paid to study characteristics, features of the intervention and context, since these can help produce better models and explain the heterogeneity of results. The author also highlights the importance of presenting evidence for policymaking and studying how policymakers use evidence from different sources, suggesting they could exhibit some ‘variance neglect’.

Burchett et al. (2011) examine whether health research findings can be applied to settings beyond those originally studied. Through a review of 38 articles describing 25 frameworks for assessing external validity, four key categories emerge, one of which is context. Context encompasses various aspects such as the need for the intervention, the specific characteristics of the setting and target population, and the ease with which the intervention can be implemented in that particular context. For instance, the effectiveness of a public health program developed in one country may not be replicable in another with a different healthcare system or a population with

distinct demographic or cultural features. The authors highlight that many of the frameworks reviewed focus primarily on the intervention itself, often neglecting the importance of the context in which it is applied. Contextual factors, such as healthcare infrastructure, available resources, and cultural practices, can significantly influence the implementation and success of an intervention. Furthermore, Burchett et al. point out that one of the main limitations of the frameworks examined is the lack of empirical data to support the development of context-related criteria and the absence of assessments of their perceived utility. This underscores the need for further empirical research to develop more robust and practical tools for evaluating the applicability and transferability of evidence to new contexts.

Williams (2020) explores the complexity of external validity in evidence-based policy, emphasizing that the effectiveness of a policy in one context does not guarantee its success in a different context. The primary challenge lies in the interaction between a policy's theory of change—namely, the causal logic linking the policy inputs to its outcomes—and the specificities of the local context. For example, the Tamil Nadu Integrated Nutrition Programme (TINP), a successful nutritional program in India, failed in the cultural context of Bangladesh due to fundamental differences in the control of food resources within households. This case illustrates that contextual assumptions, valid in one setting, may not hold elsewhere, leading to entirely different outcomes.

To address this challenge, Williams proposes the method of "mechanism mapping", which allows for comparing a policy's theory of change with the specific conditions of the new context. This method helps identify which elements of the local context might interfere with the policy's functioning, enabling necessary adjustments before implementation. However, Williams cautions that, despite being a powerful diagnostic tool, mechanism mapping is heavily dependent on the subjective judgment of policymakers, which can introduce bias or errors. Despite these limitations, the method represents a significant advance in ensuring that evidence-based policies can be effectively adapted to new contexts, thereby maximizing their relevance and impact.

4.2 Population choice

Population choice is a critical factor in the discussion of external validity and extrapolation. It involves the characteristics of the population studied and how these influence the generalizability of findings to other groups. The demographic composition, socio-economic status, cultural background, and other population-

specific characteristics can significantly impact the outcomes of interventions, making population choice a pivotal aspect of research design.

Findley et al. (2021) discuss the importance of selecting the target population in evaluating external validity. The choice of population is crucial because it largely determines whether the inferences made about a sample can be extended to other groups. The authors introduce the concept of Plausibility of Scope, which requires careful consideration of whether the units analyzed in a study adequately represent the target population. To ensure external validity, it is necessary to compare the characteristics of the sample with those of the broader population, avoiding selection bias that could distort the results. The use of methods such as random or quasi-random sampling is suggested to improve representativeness, while techniques like weighting can be employed when random sampling is not feasible. The authors emphasize that representativeness is a key factor in ensuring that the results are applicable to a broader population, thereby reducing the risk of bias that could compromise external validity.

Rothwell (2005) highlights the critical role of population selection in randomized controlled trials (RCTs) for external validity. However, it is noted that only a small proportion of patients with a given condition actually participate in these studies, which limits the representativeness of the findings. Several factors reduce external validity, such as pre-eligibility selection, where many patients are not even considered for inclusion due to the context in which they are treated or the physician overseeing their care. Additionally, the eligibility criteria used in RCTs often exclude important groups, such as the elderly or patients with comorbidities, further narrowing the generalizability of the results.

Even when eligibility criteria appear appropriate, only a small percentage of patients are actually recruited, and these patients tend to differ significantly from those not included in terms of variables like age, gender, and disease severity. Moreover, run-in periods and enrichment strategies that select patients more likely to respond positively to treatment skew the results, making them less applicable to everyday clinical practice. Finally, Rothwell emphasizes the importance of accurate and comprehensive reporting of eligibility criteria and the patient selection process, which is often insufficient and further limits the ability to assess the external validity of the studies.

Avellar et al. (2017) analyzed how systematic reviews address external validity, with a particular focus on the challenge of replicating intervention outcomes in different contexts. The generalizability of interventions, understood as the ability to extend results to a broader population or context, is often limited in systematic reviews, as

these tend to prioritize internal validity, thereby sacrificing a thorough consideration of contextual variables. For instance, in the HomVEE review, it was found that many studies do not provide sufficient details about the implementation context, such as local resources, the demographics of the population served, or the available infrastructure—elements crucial for determining the applicability and feasibility of a program in different settings.

The applicability of an intervention, which pertains to the relevance of a program to a specific context, is closely tied to understanding the local conditions where the intervention was originally tested. Without detailed information on these aspects, it is difficult to assess whether an intervention that succeeded in a resource-rich environment will be equally effective in a resource-poor setting. Feasibility, or the practicality of implementing the program in a new context, is further complicated when reviews do not provide sufficient data on the logistical and infrastructural support required for the intervention. Essentially, Avellar et al. highlight that the lack of attention to contextual variability significantly reduces the ability of systematic reviews to effectively inform practitioners and policymakers about the potential effectiveness and sustainability of interventions in different environments.

Dekkers et al. (2010) explore the complexity of external validity in clinical trials, which refers to the ability to generalize study findings to different populations. They propose a three-step approach for its evaluation: first, assessing whether the studied population is representative concerning eligibility criteria; second, considering geographic, temporal, and ethnic differences between the study population and the target population; and third, evaluating whether the results can be applied to populations that do not fully meet all eligibility criteria. This approach is essential to ensure that results can be applied to new clinical contexts.

Internal validity, which ensures the accuracy of results within the studied group, is crucial for external validity, yet the latter is often overlooked. The authors distinguish between external validity, which concerns generalizability in identical contexts, and applicability, which deals with the validity of results in different contexts. For example, extending the results of a study on antihypertensive drugs to patients with varying characteristics requires an analysis of ethnic and geographic differences that may influence treatment effectiveness. The authors emphasize that although there is no formal method to establish external validity, it is necessary to assess it to apply findings to new populations, even though repeating studies for every target population is impractical. Therefore, the evaluation of external validity remains a complex issue requiring thorough analysis. Dekkers et al. argue that external validity cannot be formalized in the same way as internal validity and must be considered a complex

reflection that integrates prior knowledge, statistical considerations, and biological plausibility. Thus, assessing external validity is a well-reasoned but fallible judgment on the generalizability of results.

In a similar vein, Esterling et al. (2023) critically examine the concept of external validity in political science, emphasizing the challenges associated with "population choice" in making causal claims and generalizations. They argue that establishing external validity is often impeded by a lack of clarity regarding the conditions that define the studied populations. The authors critique the tendency of researchers to assert causal effects without fully understanding the enabling conditions, which can lead to ambiguity about the true sources of observed outcomes. The paper underscores the necessity of causal specification to support credible generalizations about treatment effects across different political contexts. They illustrate this point using the GSL study (a fictitious example of a laboratory performing quantitative tests to study the causal effect of an intervention), which demonstrated varying results in different settings, highlighting the importance of understanding the specific conditions that influence intervention effectiveness. To enhance the credibility of causal claims in political science, the authors call for rigorous causal specification and a deeper exploration of the conditions that define populations. They emphasize that researchers should improve their reporting practices regarding the contexts and conditions of their studies to bolster the generalizability and applicability of their findings.

4.3 Intervention size

The size of an intervention is a significant factor in the discussion of external validity and extrapolation. It involves the scale and scope of the intervention and how these elements influence the generalizability of findings to other contexts. Larger interventions often face different challenges and produce varying results compared to smaller, more controlled studies.

Cartwright and Hardie (2012) highlight that standard problems such as dilution effects, neighborhood effects, and selection bias can hinder the ability to generalize findings from one context to another. They emphasize that even local randomized controlled trials (RCTs) may not adequately substitute for a comprehensive understanding of how a policy is intended to function within a specific population. The authors argue that while certain programs may meet rigorous criteria for effectiveness, there remains a critical gap in understanding whether the underlying reasons for success in one setting will hold true in another. This concern is particularly relevant when considering the implementation of interventions like Surestart, which may have demonstrated effectiveness elsewhere but lack clear guidance on their applicability in different

contexts. Thus, the authors assert that a framework is necessary to assist policymakers in determining the relevance of evidence to their specific circumstances, particularly in relation to the size and scope of the intervention being considered.

Kern et al. (2016) also explore the challenges associated with the generalization of experimental data, especially when the target population varies in key aspects such as demographics or prior experiences. They use the School Dropout Demonstration Assistance Program (SDDAP) as an example, where the findings from a multi-site experiment may not be directly applicable to all schools or student populations due to differences in local contexts and student needs. The authors emphasize that the lack of comparable individual-level measures across different sites complicates the inference process, making it challenging to determine whether sample ignorability holds. This underscores the importance of employing robust statistical techniques that can adjust for observed differences, thereby enhancing the generalizability of experimental results. Addressing these challenges is crucial for ensuring that research findings are applicable and relevant to diverse real-world contexts.

The issue of intervention size underscores the importance of considering scale when designing and evaluating policies and programs. By understanding the unique challenges associated with larger interventions, researchers and policymakers can better plan for scalability, ensuring that interventions are effective and applicable across different contexts and populations.

4.4. Study design

The design of a study plays a crucial role in the discourse on external validity and extrapolation, as it directly pertains to the structure and methodology of research studies and how these elements influence the generalizability of findings to other settings.

Tipton and Peck (2017) discuss a critical challenge related to the study design employed in evaluations, particularly within social welfare program evaluations that utilize multisite experimental designs to estimate causal treatment impacts. These designs often rely on purposive site selection, leading to samples that are not representative of the broader population. For instance, the Job Search Assistance (JSA) evaluation illustrates how variations in service delivery across states and localities can result in divergent approaches, complicating generalization. This lack of representativeness poses significant limitations for generalizing findings beyond the specific contexts in which the studies are conducted. Researchers often compare the characteristics of study samples to those of the target populations, concluding that they are similar, but this narrative interpretation is insufficient for establishing external

validity. To enhance the generalizability of results, Tipton and Peck argue for adopting more rigorous methodologies that incorporate stratified selection and targeted recruitment plans. By developing a systematic approach to study design that prioritizes representativeness, researchers can better address the challenges associated with external validity, ensuring that selected sites reflect the diversity of the broader population and thereby improving the applicability of findings to various contexts in social welfare policy evaluations.

Similarly, Shadish et al. (2002), in their comprehensive examination of experimental and quasi-experimental designs, discuss the challenges of external validity and extrapolation. These challenges are critical considerations in study design, particularly regarding the generalization of findings beyond the specific conditions of an experiment. Researchers often face the dilemma of determining whether causal relationships identified in controlled settings can be applied to broader populations, settings, or treatments. This concern is heightened by the fact that many studies are conducted in unique contexts that may not accurately reflect the diversity of real-world scenarios. For example, community-based health programs like the Minnesota Heart Health Program were initially tested in controlled environments but later adapted to real-world settings using quasi-experimental designs. These designs helped account for variables that could not be controlled in a purely experimental setup, thereby improving the external validity of the findings. Another example is the National JTPA Experiment, which used quasi-experimental designs to account for selection biases and improve the applicability of findings from training programs to different state contexts. Understanding the extent to which findings can be generalized requires a careful examination of various factors, including the characteristics of the sample and the nature of the intervention. Moreover, the incremental nature of scientific inquiry necessitates that researchers continuously assess the applicability of their findings to untested situations, complicating the design of studies aimed at enhancing external validity.

The study design is crucial for ensuring not only internal validity but also external validity. Chassang and Kapon (2022) discuss several practices that researchers can adopt to enhance external validity through robust design. A central point of their argument is the pre-registration of experiments, which helps prevent bias arising from the ex post selection of results. Pre-registration requires researchers to define the study's objectives and analysis methodologies in advance, thereby reducing the risk that outcomes are influenced by post-hoc methodological decisions that could compromise their generalizability. Furthermore, sharing data and study design with the scientific community allows other researchers to verify the results and test external

validity in different contexts. For example, if a health intervention was effective in a pre-registered study conducted in an urban area, other researchers could use the same design to test the intervention in rural areas, examining whether and how the results vary depending on the context.

Rothwell also discusses the importance of study design in ensuring the external validity of randomized controlled trials (RCTs). He highlights how the context in which these studies are conducted can significantly impact the ability to generalize the results to a broader context. Differences in healthcare systems, national clinical practices, and the selection of participating centers and clinicians can limit the generalizability of findings. For example, a study conducted in a specialized center or by highly experienced clinicians may yield results that are not representative of everyday clinical practice.

Moreover, Rothwell points out that the protocols used in RCTs may differ from routine clinical practice, further compromising external validity. The use of specific diagnostic techniques, experimental or non-standardized treatments, and intensive safety monitoring during trials might not reflect the realities of everyday clinical practice, leading to results that are not easily transferable to other settings. The choice of outcomes measured is another critical aspect; the use of surrogate outcomes or complex, unvalidated scales, as well as overly short follow-up periods, can reduce the clinical relevance of the findings and, therefore, their external validity.

To improve external validity, Rothwell suggests that the design of RCTs should be more closely aligned with everyday clinical practice, with greater transparency in reporting and a particular focus on the details that influence the generalizability of results, such as inclusion criteria, treatment protocols, and outcome measurement.

In the context of external validity, study design is crucial to ensure that findings are generalizable and transferable to other settings. Findley et al. (2021) introduce the conceptual framework M-STOUT, which includes dimensions such as Mechanisms, Settings, Treatments, Outcomes, Units, and Time, to more comprehensively assess external validity. The Study Design dimension emphasizes the need to design studies that not only produce valid causal estimates within a specific context but also consider how these estimates can be applied to other contexts. The authors discuss the importance of having a strong theory and a research design that makes empirical inferences testable, thereby allowing external validity to be evaluated through tests and sensitivity analyses. The significance of adequate sampling and a study plan that accounts for relevant variables is highlighted as a means to ensure that the results can be successfully transported to other populations or settings.

Slough & Tyson (2023) delve into the critical role of study design in meta-analyses to ensure external validity and the generalizability of results. Meta-analysis combines findings from studies conducted in different contexts and times, and to do so reliably, it is essential that the studies are harmonized across two key dimensions: contrast and measurement strategy. The study design in this context implies that the included studies must share a common mechanism and aim at the same empirical objective, which are crucial concepts for achieving comparable results. The authors discuss fixed-effect and random-effect models, explaining that to achieve target equivalence—that is, to ensure that the studies are aiming at the same goal—it is necessary to harmonize both the comparisons between treatments (contrast) and the methods of outcome measurement. Without this harmonization, a meta-analysis risks producing inconsistent or misleading results. Thus, study design is not only about internal validity but also about designing studies in such a way that their results are applicable and comparable within a meta-analysis, requiring careful attention to context, population, and methodology.

Furthermore, the challenge of study design in the context of external validity and extrapolation is a significant concern in consumer behavior research. External validity pertains to the generalizability of research findings to broader populations and settings, which is often compromised when studies rely heavily on specific samples, such as college students. Winer (1999) highlights that much of the consumer behavior literature focuses on theory applications (TA) that prioritize high internal validity through controlled laboratory experiments. However, this focus can lead to questions about the applicability of findings to real-world scenarios. To address these challenges, Winer advocates for integrating secondary data sources, such as scanner panel data, which can provide valuable insights into actual purchasing behaviors that support and validate laboratory findings. This approach allows researchers to bridge the gap between controlled experiments and real-world applications, thereby improving the robustness of their conclusions. Additionally, Winer promotes collaborative efforts between consumer behavior and marketing science researchers. Such joint ventures can facilitate the sharing of methodologies and data, ultimately leading to more comprehensive studies that address external validity concerns. By focusing on these aspects, researchers can better navigate the complexities of study design while ensuring that their findings are applicable beyond the confines of controlled environments. This holistic approach not only enhances the credibility of research outcomes but also contributes to a more nuanced understanding of consumer behavior in diverse contexts.

By understanding the unique challenges associated with study design, researchers and policymakers can better plan for and address issues related to external validity and extrapolation, ensuring that interventions and policies are effective and applicable across different settings and populations.

4.5 Implementation and Scalability

Implementation and scalability are crucial factors in the discussion of external validity and extrapolation, involving the practical aspects of applying an intervention in real-world settings and the challenges of scaling up from small-scale studies to broader applications. These factors are particularly significant as interventions demonstrating efficacy in controlled environments may not yield the same results when implemented in diverse community settings.

Prohaska and Etkin (2010) address the challenges in translating research findings into community programs, highlighting that only a small fraction of scientifically tested interventions are actually implemented on a large scale. The translation of research into practice is often slow and fragmented, particularly in the context of health promotion among older adults. Despite demonstrated efficacy in clinical trials, many programs fail to be adequately disseminated and sustained in real-world settings.

The authors identify four key challenges in the translation process. Internal and external validity pose significant difficulties: while internal validity focuses on attributing outcomes to the program itself, external validity concerns the ability to generalize results to different populations and contexts. Research often emphasizes internal validity, thereby limiting the scalability of programs. A positive example in this context is the Chronic Disease Self-Management Program (CDSMP), which has been successfully replicated across various settings, enhancing its external validity. Another critical issue is the definition of meaningful outcomes. There is a gap between what researchers consider a success and what is truly significant for participants and agency directors. To improve the adoption and dissemination of programs, it is important to include a broader range of outcomes that are relevant to the program's recipients.

Treatment fidelity is crucial to ensure that programs are implemented consistently and according to the intended standards. To maintain effectiveness, it is essential to provide supports such as manuals and checklists that facilitate program consistency. However, even the best-designed program has limited impact if it fails to effectively reach the target population. Therefore, it is vital to consider the characteristics of the population, available resources, and the contexts in which programs are implemented

to ensure effective dissemination. The RE-AIM framework is proposed as a tool for evaluating the impact and dissemination of evidence-based programs in the real world. RE-AIM helps monitor the adoption, effectiveness, implementation, and maintenance of programs, ensuring that they reach the target population and are sustained over time. The authors emphasize the importance of collaboration between researchers and practitioners to adapt programs to operational realities. Documenting essential program elements in detail and developing effective recruitment strategies is crucial. Only through close cooperation and the appropriate use of the RE-AIM framework can the challenges of translation and scalability be addressed, ensuring that health promotion programs for older adults are widely adopted and maintained in community settings.

T. Cook (2014) emphasizes the critical role of external validity in the implementation and scalability of public policies, focusing on the ability to generalize study results to different or broader contexts. He distinguishes between two key functions: representation and extrapolation. The function of representation ensures that samples and treatments are representative of the broader populations or contexts to which results will be applied, but Cook critiques the use of suboptimal methods like opportunistic sampling and propensity score matching, which can undermine generalizability.

Extrapolation is vital for scalability, as it involves applying study findings to new contexts, though Cook notes the difficulty in accurately extrapolating results, given that new contexts may differ significantly from those originally studied. For example, a successful urban educational policy might not work in a rural setting without adjustments. Cook advocates for policy sciences to adopt practices from the natural sciences, such as identifying robust causal mediating processes that can be applied across various contexts, allowing for more reliable predictions in new environments. He suggests methods like response surface modeling for more accurate forecasts and endorses meta-analysis as a tool to enhance external validity, despite its limitations, such as potential biases and representativeness issues.

Acknowledging the complexity of generalizing causally across different dimensions (people, treatments, outcomes, contexts, and time), Cook urges the social sciences to develop explicit methodologies to improve external validity and focus on policy scalability, ensuring that findings are applicable to diverse contexts.

Bold et al. (2013) address the challenge of external validity in the implementation and scaling of educational programs, highlighting how the successful outcomes of small-scale interventions may not be replicated when expanded and managed at a national level. The study examines a contract teacher program in Kenya, initially implemented

successfully by NGOs in limited local contexts, and analyzes its nationwide extension under the Kenyan government's management. While the NGO-led implementation continued to produce significant positive effects on student test scores, confirming previous results observed in trials in Western Kenya and India, these benefits disappeared when the program was managed by the government, with no significant improvement in educational outcomes.

This drastic change is attributed to several factors related to the operational and organizational challenges of the public sector. The weaknesses of government institutions, combined with unfavorable political economy dynamics, such as nepotism and resistance from teacher unions, undermined the program's effectiveness. These constraints weakened the incentives for contract teachers, reducing their motivation and performance quality. Additionally, the lack of effective supervision and delays in salary payments further exacerbated the situation, demonstrating how administrative and political complexities can significantly alter the effectiveness of an intervention when scaled up. This example illustrates that scaling a program from an NGO-managed context to a government-managed one is not a straightforward process; it requires a deep understanding of institutional capacities and local political dynamics. The success of a small-scale program does not guarantee that it can be successfully replicated on a national scale without significant adaptations that consider the operational realities of the public sector.

Chassang and Kapon (2022) emphasize that external validity is inextricably linked to the challenges of implementation and scalability of an intervention. They argue that ensuring a program can be effectively scaled requires an understanding of how implementation dynamics impact outcomes. A key concept introduced by the authors is the "option value" of an intervention, which refers to the flexibility of a program to be adapted or discontinued based on the results observed during its small-scale implementation. This dynamic approach to scalability allows policymakers to start with a pilot phase and use the lessons learned to refine the program before considering broader expansion. For instance, an educational program that succeeds in a small group of schools can be gradually implemented in more schools, with the flexibility to adjust the program according to specific local needs. This strategy reduces the risk of large-scale failures and increases the likelihood that positive outcomes can be replicated across different contexts.

Busetti (2023) highlights the importance of understanding causal mechanisms to enhance the implementation and scalability of public policies. The author argues that knowing how and why a program produces certain outcomes is crucial for replicating

and adapting interventions across different contexts. The proposed strategy is based on a reverse engineering approach that includes four main phases: selecting successful programs, modeling causal mechanisms, assessing the application context, and designing new interventions tailored to the specific context. This approach focuses on identifying the "causal powers," or the intrinsic characteristics of a program that enable it to generate desired effects. These causal powers must be abstracted from their original contexts and adapted to the specifics of new environments where the program is implemented. For example, implementing a program for the digitalization of administrative procedures might require adapting technologies and organizational practices to maintain transparency and efficiency in contexts different from the original.

Busetti emphasizes the importance of considering not only the internal factors of the program but also external elements that may influence its success. A deep understanding of the causal mechanisms allows for the identification of which aspects of a program are essential and which can be modified or replaced without compromising the intervention's effectiveness. This level of understanding is essential for scaling a program on a large scale or transferring it to a different context, ensuring that local conditions are adequately considered and integrated into the program's design.

4.6 Mechanisms and Transferability

Mechanisms and transferability are essential considerations in the discussion of external validity and extrapolation. These factors involve understanding the underlying mechanisms of an intervention and how these can be transferred or adapted to different contexts and populations.

Bareinboim and Pearl (2013) address the critical issues of external validity and extrapolation in causal inference. Understanding the underlying processes that drive causal relationships, known as mechanisms, is essential for ensuring that results are applicable across different settings. For instance, when a treatment is effective in one population, it may not yield the same results in another due to differing mechanisms at play. This underscores the importance of identifying and characterizing these mechanisms to facilitate accurate transportability. Furthermore, the absence of formal frameworks for assessing the transportability of causal effects complicates the generalization of findings. Researchers often struggle to determine whether the causal relationships observed in one context can be reliably extrapolated to another, especially when the populations differ significantly. This challenge highlights the need for rigorous methodologies that account for variations in mechanisms and provide

clearer guidelines for generalizing findings across diverse populations. Addressing these issues is vital for enhancing the robustness of external validity in empirical research. Bareinboim and Pearl provide examples from healthcare, illustrating how understanding the causal mechanisms of treatments can lead to more accurate predictions of effectiveness in different populations. They discuss the importance of considering factors such as disease biomarkers and patient characteristics, which can vary significantly across regions, thus affecting the transportability of causal effects.

Similarly, Pritchett and Sandefur (2014) examine the challenges of external validity and extrapolation in development research. A critical challenge lies in understanding the mechanisms behind treatment effects and their transferability across different contexts. The authors argue that when experimental results are generalized, the underlying mechanisms that produce these effects may not hold in new settings. For example, an intervention that successfully improves educational outcomes in one country may fail in another due to differing social, economic, or institutional factors. This discrepancy highlights the importance of not only assessing the effectiveness of an intervention but also understanding the context-specific mechanisms that drive its success. The authors emphasize that without a thorough examination of these mechanisms, claims of external validity can be misleading, potentially leading to ineffective policy applications. Thus, the transferability of findings from one context to another remains a significant challenge in the pursuit of evidence-based policy in development economics.

In another perspective, Bates and Glennerster (2017) emphasize the importance of focusing on underlying mechanisms rather than merely replicating programs when assessing the transferability of an intervention to new contexts. Understanding the mechanisms that drive behavioral change is crucial for predicting whether a program will be successful elsewhere. A compelling example is the incentive program using lentils to increase child vaccination rates in rural India. The program's success was not primarily due to the specific incentive (lentils) but because it leveraged a general behavioral principle: the difficulty individuals face in maintaining preventive behaviors. The incentives helped overcome this inertia, and this mechanism is applicable in many contexts, even if the type of incentive varies. This case illustrates that programs based on well-understood, general mechanisms are more transferable because they rely on universal behavioral dynamics rather than specific interventions.

Furthermore, Bates and Glennerster propose a framework for evaluating transferability, which begins with identifying the disaggregated theory behind the program, followed by assessing local conditions, the strength of behavioral evidence,

and local implementation capacity. This theoretical approach allows for the extraction of useful lessons from diverse contexts, reducing the risk of program failure when replicated in a new environment. For instance, in the case of vaccinations, the success of the program in India suggests that a similar approach could work in Sierra Leone or Pakistan, provided the incentives are adapted and the reliability of local healthcare services is ensured. This focus on mechanisms makes evaluative research not only more robust but also more valuable for policymakers seeking to adapt effective policies to new contexts.

Chen and Rossi (1987) highlight that external validity is often compromised in traditional research designs, where the overwhelming emphasis on internal validity limits the ability to generalize findings beyond the experimental context. They criticize the approach that overlooks intermediate causal mechanisms— the intervening variables that mediate the effect of treatment on final outcomes. Their critique focuses on the need to understand not only the efficacy of treatment under controlled conditions but also how and why these effects occur, and whether they can be transferred to other contexts or populations. For example, they propose the concept of "explicit generalization," where the study is designed with the specific conditions of the future context in mind, where the results will be applied. An example of this approach would be evaluating a program for prisoners by testing it on a representative sample of recently released prisoners to ensure that the results are applicable to this specific population. This contrasts with "implicit generalization," which, though less precise, attempts to gather useful information across various scenarios, such as evaluating a program on a sample of young, low socioeconomic status males in the hope that they might adequately represent prisoners.

Their analysis underscores the importance of a theory-driven approach that not only identifies relevant variables but also explores their interaction with future contexts to enhance external validity and ensure that the findings are genuinely applicable across a variety of situations.

Bold et al. (2013) explore the critical issue of the transferability of underlying mechanisms in educational programs when implemented in new contexts. A central aspect of the study is the difference in outcomes between contract teachers managed by an NGO versus those managed by the Kenyan government. The research reveals that, despite both groups of teachers having similar qualifications, the outcomes were drastically different, raising questions about the external validity of the behavioral and organizational mechanisms underlying the program. The authors identify three main mechanisms that might explain this performance disparity: teacher selection, monitoring and accountability, and the influence of union dynamics. While the NGO

demonstrated a superior ability to recruit and retain motivated teachers and to effectively monitor their performance, the government struggled to maintain the same level of effectiveness. This was partly due to practices of "local capture," where government hiring processes were influenced by favoritism, and the credibility of government contracts was undermined by union conflicts.

These findings suggest that the mechanisms driving a program's success cannot simply be transferred from one context to another without considering the local institutional and organizational specificities. The effectiveness of a program depends not only on the design of the intervention but also on the institutional context in which it is implemented. In settings where public institutions are weak or subject to political pressures, mechanisms that work in an NGO-managed environment may not have the same impact, necessitating a critical review and adaptation of the program to ensure its transferability and success in new contexts.

Williams (2020) emphasizes the crucial importance of understanding the underlying mechanisms of a policy to ensure its transportability to new contexts. Transportability refers to the ability of a policy, successfully tested in one setting, to be effectively implemented in a different context. A key aspect of transportability is the need to adapt policies according to differences in causal mechanisms that may exist between contexts. For instance, the failure of the large-scale implementation of the Tools of the Mind program in the United States, despite its success in a pilot study, illustrates how mechanisms that work in a small, controlled setting may not function at the national level due to the complexity and variability of school conditions. To enhance transportability, Williams proposes mechanism mapping as a structured approach to identify where a policy's theory of change might falter in a new context. This process helps policymakers anticipate and address potential issues before they arise. For example, in the case of the Bangladesh Integrated Nutrition Program (BINP), mechanism mapping could have revealed that food decisions were not solely made by mothers but involved other family members, suggesting the need to adapt the nutritional counseling component to include husbands and mothers-in-law.

Williams concludes that the transportability of policies requires a balance between using established evidence and adapting to local specificities. Mechanism mapping provides a framework for navigating this balance, helping to determine when and how to modify a policy without compromising its overall effectiveness.

Findley et al. (2021) highlight the challenge of transportability, which is the ability to apply the conclusions of a study to different contexts or populations beyond those originally studied. The Mechanisms dimension in the M-STOUT framework is particularly relevant in this context, as it pertains to the causal processes that link

treatments to outcomes. The authors emphasize that a deep understanding of these underlying mechanisms is crucial for predicting whether a treatment will have the same effect in a new context. Without this understanding, the transportability of results can be seriously compromised.

The concept of Model Utility is explored to assess how well a theoretical or empirical model can be applied to other contexts. The authors suggest that useful models are those that clearly identify causal mechanisms and link them to contextual characteristics, thereby enabling a more accurate assessment of the transportability of results. Transportability requires not only valid causal inference within the original context but also the ability to adapt and apply these mechanisms effectively in new contexts.

The concept of transportability is central to the analysis by Slough & Tyson (2023), particularly when applying the results of a meta-analysis to new contexts. The mechanisms through which a treatment produces effects are crucial in determining whether those effects can be replicated in different settings. The authors distinguish between construct validity, which concerns the alignment between the treatment and theoretical concepts, and external validity, which assesses a mechanism's ability to produce consistent effects across various contexts. They emphasize that without a clear understanding of the underlying mechanisms, the results of a meta-analysis may not be transportable to other contexts. For example, if the mechanisms linking an intervention to its effects vary significantly among the studies included in a meta-analysis, the aggregated results might not accurately reflect the studied phenomenon. Therefore, it is essential for researchers to ensure that mechanisms are consistent across studies to guarantee that the meta-analysis results are valid and transportable to other contexts. The authors also introduce the concept of target-equivalence, which requires that studies included in a meta-analysis harmonize both the contrast and measurement, ensuring that they refer to the same empirical objective. This is crucial to ensure that the transportability of results is not compromised by methodological or contextual differences.

Transportability, then, is not merely about replicating results, but about correctly understanding and applying the mechanisms that produce those effects in new contexts, ensuring that observed differences are not due to inconsistencies in study design or measurement methods.

Busetti (2023) delves into the central role of causal mechanisms in the design and transportability of public policies. Understanding the mechanisms through which a program produces its effects is crucial for adapting it to new contexts and ensuring its success. The author emphasizes that merely replicating a program without a

detailed understanding of the underlying mechanisms can lead to disappointing outcomes, as different contexts can significantly influence the intervention's effectiveness.

The concept of "smart replication" is introduced to describe the approach of adapting programs to new contexts. Successfully replicating a program requires more than copying its superficial features; it necessitates a deep understanding of the causal mechanisms that make it effective and adjusting the design based on the new conditions. For instance, if a nutrition program is effective in a context where mothers control their children's diet, transferring it to a context where this control is exercised by other family members would require rethinking the implementation strategies. The author also discusses the importance of identifying functional equivalents—alternative solutions that trigger the same causal mechanisms as the original program. This is particularly useful when the initial conditions cannot be replicated. A practical example is selecting specific tools or approaches that, although different from those originally used, can produce the same effect by activating the fundamental causal mechanisms.

Busetti proposes an approach that integrates causal mechanism modeling with a careful assessment of the context to ensure that programs can be successfully transported and implemented in different environments while maintaining their effectiveness. This approach requires a balance between fidelity to the original design and adaptation to the specificities of the new context, with a constant focus on understanding the causal processes that drive the desired outcomes.

Khosrowi (2022) explores the importance of causal extrapolation, a crucial process for transferring knowledge gained from a study population to a different target population. Extrapolation is fundamental in evidence-based policy, where mechanisms proven to work in one setting must be adapted and applied to other contexts to ensure their effectiveness. However, this process is complex and involves significant epistemic risks, as differences between populations can affect the success of extrapolation.

The author emphasizes that successful extrapolation requires a deep understanding of the underlying causal mechanisms that determine how and why an intervention produces certain effects. These mechanisms must be analyzed to identify relevant similarities and differences between the study and target populations. For example, in a microfinance intervention, even if the causal pathway between access to microcredit and family welfare appears similar, differences in investment habits between populations could significantly impact the final outcome. According to the author, successful extrapolation requires balancing the use of additional empirical

resources with maintaining the relevance of the original evidence. This process must avoid two main pitfalls: epistemic overreach, which occurs when overly detailed and hard-to-obtain information is demanded, and the "extrapolator's circle," where the additional resources required become so determinant that they render the original evidence irrelevant.

Khosrowi argues that extrapolation strategies must ensure that the original evidence remains central to the decision-making process, supported but not overshadowed by additional resources. This approach is essential for successfully transferring causal mechanisms to new contexts and predicting the effectiveness of interventions in different populations.

4.7 Selection Bias

Selection bias is a critical factor in the discussion of external validity and extrapolation, involving the systematic differences between the participants selected for a study and the population to which the findings are intended to generalize. This bias can significantly limit the applicability of research findings, as results from a non-representative sample may not be applicable to the broader population.

Shadish et al. (2002) examine the intricacies of selection bias in experimental and quasi-experimental research. One of the major challenges identified is the issue of selection bias, which occurs when the sample used in a study does not accurately reflect the broader population. This discrepancy can limit the generalizability of the findings. For instance, the results of a study conducted in urban schools may not be applicable to rural schools due to differing contextual factors. Shadish et al. highlight the critical importance of addressing selection bias to ensure that extrapolations from study findings are valid across different contexts.

They discuss various statistical models designed to correct for selection bias, such as propensity score models and selection bias models. These models adjust for differences between treatment and control groups in studies where random assignment is not possible. It is essential to specify these models accurately and include all relevant covariates to mitigate bias effectively. An illustrative case study on educational interventions is also presented to demonstrate the practical application of propensity score matching. This technique is employed to create equivalent groups, thereby enhancing the validity of causal inferences. By doing so, the results obtained from quasi-experimental designs can be made more comparable to those derived from randomized controlled trials, thus improving the robustness of the study's conclusions.

Findley et al. (2021) delve deeply into the issue of selection bias within the context of external validity, identifying it as a key obstacle to generalizing study results. They distinguish between sample selection bias, which occurs when the sample is not representative of the population, and variable selection bias, which arises when the variables measured in the study do not accurately reflect the theoretical constructs of interest. Both types of bias can severely limit the generalizability of findings, even if the study has high internal validity. This is particularly evident in political science research, where the selection of countries or regions based on convenience rather than representativeness can skew results. For example, studies on democratization often select countries that are accessible or have readily available data, potentially introducing bias that limits the applicability of their conclusions to other contexts.

The authors emphasize that both experimentalists and observationalists often overlook the implications of selection bias, which can significantly distort the applicability of study results. Even large-N studies, which may seem to represent the "real world," can suffer from biases that undermine their external validity. The reliance on pooled or random samples does not inherently guarantee representativeness, as poor indicators for treatments and outcomes can lead to substantial variable selection bias. This underscores the necessity for scholars to rigorously assess and transparently report on external validity, ensuring that the limitations imposed by selection bias are adequately addressed. Findley et al. advocate for more rigorous reporting standards and methodological transparency to enhance the credibility and generalizability of social science research, ultimately contributing to a more nuanced understanding of how results can be extrapolated to broader contexts.

Bareinboim and Pearl (2013) focus on transportability, which involves extending causal effects from one study or setting to another. Their work primarily deals with theoretical aspects of determining when causal effects can be transported across different populations, and they highlight selection bias as a crucial factor in this process. They propose a general algorithm to decide the conditions under which causal effects are transportable, acknowledging that differences in population characteristics can introduce bias. For instance, if a clinical trial is conducted on a specific demographic, the findings may not be applicable to a broader population with different characteristics. Their theoretical framework is critical for researchers looking to generalize findings beyond the original study context, as it underscores the necessity of accounting for selection bias to ensure that transported effects are valid and applicable to the target population. Although their paper does not present a specific case study, it emphasizes the interplay between internal validity and external applicability, and provides a mathematical framework with necessary and sufficient

conditions for assessing when causal relations can be inferred from experimental studies to observational settings. This framework aims to protect researchers from the pitfalls of unwarranted generalization, ensuring that the extrapolation of findings is grounded in a rigorous understanding of the underlying population differences.

Chen and Rossi provide an in-depth analysis of the issue of selection bias, a critical problem when randomization is not feasible. They emphasize that in contexts where randomization cannot be implemented, such as comparative studies between private and public schools, it is essential to use advanced statistical models to specify and control for confounding variables that could distort the results. Their work suggests that, rather than relying solely on randomization to eliminate bias, researchers should incorporate analytical models that account for selection dynamics, thereby reducing the risk of biased estimates. For instance, in cross-sectional surveys, where self-selection bias is common, modeling this bias can enhance the internal validity of the results.

Moreover, Chen and Rossi highlight that even when randomization is possible, it does not completely eliminate the influence of extraneous variables; these variables can still affect the residuals, increasing error variance and thereby reducing the ability to detect the true effects of the treatment. Chen and Rossi's approach, therefore, advocates for a more sophisticated and integrated use of randomization and analytical models to simultaneously address threats to internal validity and minimize selection bias, thereby improving the reliability of conclusions drawn from non-randomized studies.

Degitar and Rose (2023) explore selection bias as a critical obstacle to external validity, highlighting how the representativeness of the study sample relative to the target population influences the ability to generalize results. Selection bias occurs when the characteristics of the sample differ significantly from those of the target population, making it difficult to apply the findings to broader contexts. While internal validity ensures accurate estimates within the specific context of the study, external validity requires that these results be applicable to other populations, which becomes problematic in the presence of selection bias. To address this issue, the authors propose the use of techniques such as stratified sampling and propensity scores, which balance covariates between the study sample and the target population, making the results more representative. They also emphasize the importance of considering effect modifiers—variables that can influence the treatment response differently between the sample and the target population. Techniques like matching and inverse probability weighting can help mitigate these differences.

Degitar and Rose also suggest integrating observational data with data from randomized trials by using synthesis models and calibrated regression. This approach combines the internal validity of randomized trials with the external validity of observational data, enhancing the robustness and generalizability of the estimates. Addressing selection bias is essential to ensuring that study results are applicable to the broader target population, requiring careful study design and the adoption of appropriate analytical methods to balance differences between populations.

5. Methods to Address External Validity and extrapolation

Addressing external validity in research is a multifaceted challenge that has been tackled using various methods and strategies. Below are the primary methods identified in the literature to enhance the external validity of research findings, each expanded with examples and citations.

5.1 Diverse and Representative Samples

One of the most effective methods to enhance external validity is the use of diverse and representative samples. Ensuring that the study population closely mirrors the broader population to which the findings will be applied is crucial for enhancing the generalizability of research results. This approach is particularly important in fields such as public health, education, and social sciences, where population diversity can significantly influence intervention outcomes.

Findley et al. (2021) emphasize the critical importance of external validity in social science research, advocating for rigorous methodologies to enhance the generalizability of findings. They highlight the use of diverse and representative samples as a key approach to improve the accuracy of external validity inferences. By ensuring that samples reflect a wide range of populations, researchers can better assess the applicability of their results to broader contexts. Studies that include participants from various socio-economic backgrounds, geographic locations, and demographic characteristics are more likely to yield findings that can be extrapolated to different settings. This methodological rigor is essential for making credible claims about how results can be generalized to other populations. Random sampling, or "as-if random" sampling in observational setups, is proposed as a benchmark for achieving representativeness. Poststratification, which involves weighting observations to mimic a representative stratification, can enhance representativeness when random sampling is not feasible. Thus, incorporating diverse and representative samples is not only a methodological best practice but also a necessary step toward achieving robust external validity in social science research.

Esterling et al. (2023) also underscore the necessity of construct and external validity in causal inference, highlighting the importance of diverse and representative samples to address these issues. They argue that relying solely on local causal effects limits the generalizability of findings, confining knowledge to specific contexts without providing guidance on broader applicability. External validity is crucial for accumulating causal knowledge across different settings, enabling researchers to make informed claims about the effectiveness of interventions beyond the studied population. Using diverse samples can help identify the conditions under which a treatment may be effective, thereby enhancing the credibility of causal claims. In randomized controlled trials (RCTs), they recommend multisite studies to test for variation in treatment effects across different settings, which requires assumptions about the variation in underlying conditions that enable or disable the cause. This approach mitigates the risks of making causal claims that are only valid in the original study context and ensures that the findings can be extrapolated to broader contexts.

Thomas D. Cook (2014) discusses the critical importance of external validity and extrapolation in policy sciences, emphasizing the need for diverse and representative samples in research. He argues that the generalization of causal relationships relies heavily on the ability to draw conclusions from studies encompassing a wide array of populations, settings, times, and treatment variants. Traditional methods often rely on opportunistic sampling, leading to potential biases such as volunteerism and publication bias, which can skew available data. Propensity score matching can mitigate some of these issues by matching individuals based on observed characteristics, though it requires large samples and well-defined population details. In educational research, a diverse sample of students from various backgrounds can help generalize findings about new teaching methods. In healthcare, including a wide range of patient demographics can better predict treatment effectiveness across different groups. By employing diverse and representative sampling, researchers can produce more robust and applicable results, making their findings more relevant for informing policy decisions in varied contexts. Cook highlights that the strengths of meta-analysis in this regard stem from the potential to replicate cause-effect relationships across different contexts, rather than formal sampling theory. Thus, a methodological focus on diverse and representative samples is vital for improving the external validity of policy research and ensuring that findings can be effectively extrapolated to new situations.

5.2 Contextual Adaptation and Local Tailoring

Contextual adaptation and local tailoring are pivotal strategies in addressing the challenges of external validity and extrapolation in research. Recognizing the

necessity of adjusting interventions to fit the specific characteristics and needs of different contexts enhances the applicability and effectiveness of research findings across diverse settings.

Pritchett and Sandefur (2015) explore the complexities of external validity and extrapolation in development economics, emphasizing the need for contextual adaptation and local tailoring of interventions. They highlight that observational data from the relevant context often provide more reliable estimates than experimental data from different contexts due to the trade-off between internal and external validity. To address this challenge, they propose calculating the root mean squared error (RMSE) for both non-experimental and experimental estimates, measuring reliability by encompassing sampling error and selection bias for non-experimental data, and sampling variance and cross-context parameter heterogeneity for experimental data. Their analysis of microcredit programs illustrates how variations in program design and local conditions can significantly affect results. For example, non-experimental evidence within a specific context can outperform single experimental estimates from other contexts, though this advantage diminishes as more diverse experimental evidence accumulates. By focusing on the heterogeneity of contexts and the specific attributes of interventions, policymakers can better tailor programs to meet local needs, thereby enhancing the relevance and effectiveness of development initiatives.

Williams (2020) introduces the concept of "mechanism mapping" as a key methodological tool to evaluate how well a policy or intervention can be adapted to a new context. This process involves three stages: first, outlining the theory of change, or the logical sequence of steps that lead from an intervention to its intended outcomes. Second, identifying the contextual assumptions that need to be met for the intervention to work effectively. Third, comparing these assumptions with the actual conditions in the new context to detect any discrepancies that might hinder success. A striking example provided by Williams is the failure of the Tamil Nadu nutrition program when applied in Bangladesh, where differing household decision-making dynamics—such as the involvement of husbands and mothers-in-law—undermined the effectiveness of a program designed for a context where only mothers made such decisions. Mechanism mapping, as Williams suggests, enables a structured approach to assessing when and how interventions need to be modified to maintain their effectiveness in a new setting without losing their core elements.

Burchett et al. (2011) similarly stress the importance of evaluating both the new setting and the target population when adapting interventions. They argue that an intervention's success in a new context depends on a detailed understanding of the specific characteristics of the environment, including local financial and human

resources, infrastructure, and policies. They also emphasize the role of sociodemographic and cultural factors, which can influence the acceptability and feasibility of an intervention in its new context. According to Burchett et al., the ability to modify interventions while preserving their core effectiveness is central to successful adaptation. Their approach advocates for a thorough assessment of local capacities and needs to ensure that evidence-based practices can be transferred effectively between different settings. This analysis extends to the need for frameworks that can measure how well an intervention can be flexibly applied to different conditions, ensuring that it aligns with local realities.

Prohaska and Etkin (2010) provide another perspective on contextual adaptation, focusing on the challenges of translating health promotion programs for older adults from controlled research environments to real-world community settings. They highlight the significant gap that often exists between research-based efficacy and practical implementation in local contexts, pointing out that many interventions struggle to achieve the same outcomes when applied in settings with limited resources or different population needs. The authors use the Chronic Disease Self-Management Program (CDSMP) as a case study of successful contextual adaptation, where the program has been replicated in various settings while maintaining the integrity of its core elements. Additionally, Prohaska and Etkin advocate for the use of the RE-AIM framework (Reach, Efficacy, Adoption, Implementation, and Maintenance) as a tool for tracking how well health interventions are adapted to and sustained in local communities. By focusing on aspects like the real-world impact and sustainability of interventions, the framework helps practitioners and researchers ensure that evidence-based programs are not only implemented but are also effective and maintainable over time in diverse community contexts.

Dekkers et al. (2010) explore the challenges of applying clinical trial results to different local contexts, focusing on the complexities of external validity. They emphasize that when translating findings from one population to another, it is necessary to account for geographic, ethnic, and temporal variations that may affect the outcomes. For instance, clinical studies on acute myocardial infarction conducted on Chinese patients may not automatically generalize to other ethnic groups due to differences in treatment responses. Dekkers et al. stress the importance of examining the original eligibility criteria used in trials and determining if they need modification to suit the new target population. This is particularly relevant when strict inclusion criteria in trials—such as the use of "run-in" periods to exclude patients—might lead to an overestimation of benefits and underreporting of adverse effects, thus requiring careful adaptation when applied to more heterogeneous local populations. They argue

that contextual adaptation in clinical settings involves more than simply applying results to a new group; it requires a deep understanding of how local conditions might alter the effectiveness of treatments, making adjustments to ensure that interventions are both relevant and effective in the new context.

Cartwright and Hardie (2012) explore the complexities of external validity and extrapolation in evidence-based policy, underscoring the importance of contextual adaptation and local tailoring. Traditional approaches to external validity often assume that interventions effective in one context will yield the same results elsewhere, an assumption that can lead to failure. To address this, the authors advocate for understanding the specific causal roles and support factors that influence policy effectiveness in different settings. For instance, the Incredible Years parenting program, initially successful in Washington State, required reevaluation and adjustments for implementation in Ireland, Wales, Birmingham, and South London. This reevaluation process involved engaging with local stakeholders, such as community leaders and program originators, to adapt the program while maintaining its core components. By focusing on local conditions, stakeholder perspectives, and the unique characteristics of the target population, policymakers can better assess the relevance of evidence and make informed decisions. This methodology not only enhances the applicability of research findings but also fosters a more nuanced understanding of how policies can be effectively tailored to meet diverse local needs.

This collective exploration by various scholars highlights the multifaceted nature of contextual adaptation, emphasizing that translating interventions across different populations and settings requires a careful balance between preserving core elements and tailoring programs to local needs. Each author underscores that without proper contextual and local adaptations, the external validity of interventions could be compromised, limiting their real-world effectiveness and broader applicability.

5.3 Multi-Site and Cross-Context Studies

Multi-site and cross-context studies are crucial for addressing the challenges of external validity and extrapolation in research. By conducting studies across multiple sites and varied contexts, researchers gather evidence on the effectiveness of interventions in different settings, thereby enhancing the generalizability of their findings. The significance of these studies lies in their ability to test interventions across diverse environments, capturing a range of variables that single-site studies might miss.

Esterling et al. (2023) underscore the necessity of construct and external validity in research, stressing the importance of generalizing causal knowledge across different

contexts. They advocate for Multi-Site and Cross-Context Studies, which involve conducting experiments across various sites to ensure results are not specific to one setting. By comparing outcomes from diverse environments, researchers can identify whether observed effects are consistent or vary significantly, thereby understanding the conditions under which a causal relationship holds. This approach necessitates careful site selection to ensure representativeness and relevance, and employs rigorous statistical methods to collectively analyze data, enhancing the robustness of findings. Ultimately, this strengthens the external validity of research, contributing to more informed policy decisions and practical applications in diverse settings.

Bates and Glennerster (2017) focus on the generalizability of programs through replication in various locations, emphasizing that the success of an intervention depends not only on replicating it in different contexts but also on understanding the underlying mechanisms that drive its effectiveness. For instance, their analysis of the "Sugar Daddies Risk Awareness" program in Kenya, which aimed to reduce HIV transmission among adolescent girls, revealed limitations when transferring the program to Rwanda. Despite certain contextual similarities between the two countries, differences in HIV infection rates and pre-existing knowledge about risks among Rwandan adolescents limited the program's applicability. Similarly, a vaccination incentive program trialed in India was later replicated in Sierra Leone, Pakistan, and Haryana, India. Here, the success of the program depended heavily on local conditions, such as healthcare access and vaccine availability. The program's adaptability across these contexts demonstrated that while geographic and cultural differences posed challenges, leveraging general human behaviors—such as procrastination regarding preventative measures—enabled the program's effectiveness through tailored, context-specific incentives. Bates and Glennerster argue that the multisite approach not only facilitates the replication of interventions across settings but also identifies critical contextual variables that influence the program's success, offering valuable insights into how and when interventions can be generalized.

Similarly, Bold et al. (2013) apply a multisite approach to evaluate the external validity of a contract teacher program in Kenya, extending the intervention to 192 schools across 14 districts. This extensive study tested the program's effectiveness in local contexts characterized by varied economic, geographic, and institutional factors. Their findings highlighted significant discrepancies in outcomes based on whether the program was implemented by a non-governmental organization (NGO) or the government. While the NGO implementation led to improved student performance, the government-run version showed no positive effects, demonstrating how

operational issues such as delayed teacher payments undermined the program's success in public schools. Bold et al. also emphasized the heterogeneity of the intervention's impact based on local conditions, with more pronounced benefits in schools with high student-teacher ratios and lower initial test scores. This multisite methodology allowed for a deeper exploration of local factors that either supported or hindered the program's replication on a larger scale. The study raised important questions about the generalizability of results obtained in smaller or more controlled environments, pointing to the need for context-sensitive implementation.

Rothwell (2005) also underscores the limitations of randomized controlled trials (RCTs) in terms of external validity, focusing on multisite and inter-contextual studies. He discusses how differences in healthcare systems, geographic locations, and clinical settings can significantly influence the generalizability of trial outcomes. One example involves a European trial on carotid endarterectomy, where the time to intervention varied widely across countries, impacting the results and limiting their applicability to other clinical contexts with different practices. Rothwell also refers to the BCG vaccine for tuberculosis, which shows variable efficacy depending on geographic latitude. These examples highlight how contextual factors can constrain the transportability of RCT results. Furthermore, the selection of clinical centers and participating physicians often skews the results toward highly specialized environments, which may not reflect the realities of broader clinical practice. For example, RCTs involving only highly skilled surgeons with exemplary safety records may produce results that are not replicable in more typical clinical settings. Rothwell calls for improvements in the design and reporting of RCTs to ensure that their findings are more applicable to diverse clinical contexts, addressing the variability in healthcare practices across sites and systems.

Khosrowi (2022) approaches multisite and inter-contextual studies through the lens of causal extrapolation, a key process for generalizing findings from one context to another. He highlights the importance of identifying similarities and differences between study populations and target populations to predict whether an intervention will work in new contexts. This is particularly relevant in inter-contextual studies where cultural, economic, and demographic factors may vary significantly. Khosrowi uses the example of a microfinance intervention that succeeded in one population but failed in another due to these contextual differences, illustrating how such factors can dramatically influence outcomes. He introduces the "extrapolator's circle," a methodological challenge in which the need for additional evidence to support inferences becomes so great that the original evidence loses its relevance. This issue underscores the importance of balancing the original evidence with supplemental

resources to ensure valid extrapolation. Khosrowi's analysis highlights the need for careful consideration of contextual factors in multisite and inter-contextual studies, emphasizing the formulation of justified assumptions about population similarities to ensure successful extrapolation.

Together, these works demonstrate that multisite and inter-contextual studies are indispensable for testing the generalizability of interventions across diverse settings. By identifying the critical local conditions that facilitate or hinder the replication of interventions, these studies provide a nuanced understanding of how evidence-based programs can be adapted and applied across varying contexts, ensuring their broader applicability.

5.4 Theoretical and Mechanistic Understanding

Theoretical and mechanistic understanding is essential for addressing the challenges of external validity and extrapolation in research. By delving into the underlying mechanisms and causal pathways that drive outcomes, researchers can develop a deeper comprehension of how interventions work and why they are effective in different contexts. This approach strengthens the ability to generalize findings across diverse settings by providing a robust theoretical foundation.

In her work, Cartwright (2011) explores methodologies to enhance external validity and extrapolation through a theoretical and mechanistic understanding. She emphasizes that addressing external validity involves ensuring findings can generalize across diverse subpopulations or environments by leveraging causal structures to achieve robustness against perturbations. This approach not only enhances replicability but also facilitates accurate predictions and interventions. Furthermore, Cartwright highlights the integration of causal inference methods with perturbation data, such as randomized controlled trials, to strengthen causal claims and assess the stability of causal relationships. This theoretical framework aids in improved external validity, aiding in accurate extrapolation beyond the initial study context. The necessity of horizontal and vertical searches in causal inference is emphasized to identify shared explanatory elements and ensure robust, generalizable conclusions.

Bühlmann (2020) also underscores the importance of causality and external validity, particularly in the context of genome-wide association studies (GWAS). He notes that ensuring findings generalize across different subpopulations or environments involves leveraging causal structures to achieve robustness against perturbations. This theoretical and mechanistic understanding enhances the replicability and robustness of findings, facilitating their application to broader contexts beyond the original study population. Bühlmann further discusses how focusing on causal inference methods

allows researchers to better predict the effects of interventions and assess the generalizability of their results. Integrating perturbation data, such as randomized controlled trials, can further strengthen causal claims and provide insights into the stability of causal relationships. Ultimately, a solid theoretical framework allows for improved external validity and aids in accurate extrapolation, enhancing the robustness of findings across diverse settings.

Chen and Rossi (1987) advocate for a theory-driven approach to address external validity, emphasizing the importance of understanding the causal mechanisms underlying an intervention. Their approach moves beyond traditional randomization by modeling the relationships between treatment, extraneous, and intervening variables, thus avoiding "black-box evaluation" that focuses solely on inputs and outputs without analyzing the mechanisms at play. The framework developed by Chen and Rossi identifies exogenous, intervening, and endogenous variables, enabling researchers to predict how an intervention will function in various contexts. This modeling approach is especially useful when randomization is not feasible, as it allows for the correction of selection bias through statistical models, thereby improving external validity. A practical example provided by the authors involves the selection of students for private versus public schools. Since random assignment is not possible in this context, using models that account for self-selection processes is crucial for generalizing findings to broader contexts. Chen and Rossi's approach balances internal and external validity, offering a more comprehensive understanding of the causal mechanisms needed to generalize findings across diverse contexts.

Similarly, Bold et al. (2013) focus on understanding the mechanisms that underlie the success or failure of interventions. Their study on a contract teacher program in Kenya contrasts the outcomes of the program when implemented by a non-governmental organization (NGO) versus the Kenyan government. They highlight how institutional and political factors—such as local corruption, nepotism, and teacher union pressures—can significantly affect program effectiveness. For instance, while the NGO implementation resulted in higher student test scores, the government-run version did not yield similar improvements due to weaker oversight, delayed teacher payments, and insufficient accountability mechanisms. Bold et al. argue that replicating a program in a new context is insufficient without understanding the institutional mechanisms that influence outcomes. The mechanistic approach here helps to explain why the same intervention can lead to divergent results depending on local governance structures, reinforcing the importance of understanding institutional dynamics for external validity.

Findley et al. (2021) expand on this mechanistic focus through the M-STOUT framework, which adds two dimensions—mechanisms and time—to the traditional UTOS model (Units, Treatments, Outcomes, and Settings). This framework is particularly relevant for understanding how and why interventions work across different contexts. The inclusion of mechanisms emphasizes the importance of identifying the causal processes that link treatments to outcomes, providing a deeper understanding of how an intervention operates. For example, the M-STOUT framework allows researchers to explore whether an intervention that succeeded in one context (e.g., Liberia in 2000) can produce similar outcomes in another (e.g., African countries from 2000-2020) by understanding the underlying mechanisms. The framework demonstrates that changes in one dimension, such as time or setting, may not undermine external validity, but simultaneous shifts across multiple dimensions could affect generalizability. Findley et al. assert that a thorough understanding of causal mechanisms is critical for predicting the transportability of results to new contexts, making their approach an important methodological contribution to the field of external validity.

Busetti (2023) also contributes to the understanding of external validity through a focus on mechanistic insights. He advocates for "reverse engineering" successful programs to model their causal mechanisms and assess their applicability to new contexts. Busetti's approach helps distinguish between essential and non-essential program features during adaptation, ensuring that the core mechanisms responsible for a program's success are preserved. For example, a transparency-based administrative program that reduces processing times cannot be replicated simply by introducing new technology; the key mechanism—transparency—must be understood and maintained. This nuanced understanding of mechanisms allows for more effective adaptation and replication of interventions in different contexts.

Finally, Cook (2014) integrates the concept of UTOSTI—Units, Treatments, Outcomes, Settings, Time, and Interactions—to emphasize the complexity of generalizing causal relationships. Cook argues that a mechanistic understanding of causal processes is essential for predicting how interventions will perform in new contexts, considering the heterogeneity of causal effects across different populations and settings. Tools such as meta-analysis and response surface modeling provide valuable insights into the conditions that affect external validity, aiding in the prediction of how an intervention might function in unstudied contexts. Cook's focus on mechanistic understanding reinforces the importance of identifying causal pathways that enable reliable replication and transportability of results, thus enhancing the robustness of policy applications.

In summary, theoretical and mechanistic understanding provides a robust framework for addressing external validity across diverse contexts. By focusing on the underlying causal mechanisms, these authors contribute to a deeper comprehension of how and why interventions succeed or fail when applied to different populations and settings, thus strengthening the generalizability and applicability of evidence-based programs.

5.5 Iterative Testing and Refinement

Iterative testing and refinement are crucial strategies for enhancing the external validity and extrapolation of research findings. This approach involves a continuous cycle of testing interventions, analyzing outcomes, and making necessary adjustments to improve effectiveness and generalizability. By iteratively refining interventions, researchers can better adapt them to various contexts and ensure their success across different settings.

Shadish et al. (2002) emphasize the importance of external validity and extrapolation in experimental and quasi-experimental designs. To address this, researchers often engage in iterative testing and refinement, conducting multiple studies that vary in persons, settings, treatments, and outcomes to assess the generalizability of findings. By systematically varying key variables and observing outcomes, they can identify patterns and make more accurate generalizations. Understanding the extent to which one can generalize an internally valid finding typically occurs through a gradual process of trial and error across diverse studies. Moreover, by systematically combining results from different research efforts, scientists can build a more comprehensive understanding of the applicability of their findings. This iterative approach not only enhances the robustness of external validity claims but also allows for the identification of conditions under which causal relationships may hold or differ, facilitating more informed extrapolations to new contexts. This methodology is particularly valuable in the social sciences, where context-specific factors can significantly influence outcomes.

Similarly, Pritchett and Sandefur (2014) address the challenges of external validity and extrapolation in development practice, highlighting iterative testing and refinement as a methodological approach to tackle these issues. This process involves conducting multiple rounds of experimentation and analysis to gradually improve the understanding of how findings from one context may apply to another. By systematically testing hypotheses in varied settings, researchers can identify the conditions under which certain interventions are effective or ineffective. This approach is more cost-effective, provides faster feedback, and integrates better into decision-making cycles than traditional independent impact evaluations. Ultimately, iterative

testing and refinement serve as a critical strategy for bridging the gap between internal validity and external applicability in development research, enhancing the robustness of findings and allowing for the adaptation of interventions to better fit local contexts.

Chassang and Kapon (2022) introduce an innovative approach to improving external validity through the concept of iterative testing and refinement. They argue that research should not be viewed as a static process but rather as a continuous cycle of learning. The adoption of a treatment or intervention should not conclude with the publication of a study but evolve through successive implementations that refine predictions and improve the generalizability of results. In this dynamic approach, external validity is enhanced through consistent feedback loops between gathering new evidence and adapting predictive models. This iterative process gradually reduces uncertainty and strengthens the ability to predict how an intervention will perform across different contexts.

The authors emphasize the critical role of collecting data on diverse and relevant covariates, such as demographic or macroeconomic variables, which may influence the effectiveness of an intervention in various settings. Including these covariates in extrapolation models allows for more rigorous testing of external validity and aids in identifying the key factors that could determine the success of a treatment beyond its original context. In addition, Chassang and Kapon introduce the notion of "structured speculation," encouraging researchers to formalize qualitative insights into falsifiable hypotheses that can be tested in subsequent studies. This approach fosters a deeper, more systematic understanding of the mechanisms driving intervention effectiveness.

The concept of "option value" is also highlighted as particularly relevant for policymaking. The authors argue that policymakers should adopt a flexible approach by starting with small-scale interventions that can be adapted or expanded based on initial outcomes. This dynamic strategy enables more efficient resource management and greater adaptability to changing local contexts. Adopting an adaptive learning strategy, where the results of each new implementation are used to refine future predictions, proves especially effective in ensuring more robust and reliable external validity over time. This iterative process, grounded in continuous testing and refinement, offers a methodological advancement for improving the accuracy and applicability of evidence-based interventions across diverse contexts.

5.6 Statistical Techniques and Modeling

Statistical techniques and modeling are fundamental in addressing the challenges of external validity and extrapolation in research. These methods equip researchers with the tools needed to analyze complex data, account for variability across different

contexts, and predict the performance of interventions in new settings. By employing advanced statistical techniques, the robustness and generalizability of research findings can be significantly enhanced.

Bareinboim and Pearl (2013) delve into external validity and the extrapolation of causal effects, emphasizing the importance of statistical techniques and modeling in generalizing experimental findings to different populations. They introduce causal diagrams and graphical models to represent population differences, using an algorithmic framework with do-calculus to establish valid extrapolation conditions. Additionally, selection diagrams capture population differences, detailing how to combine experimental data from source populations with observational data from target populations. This approach ensures bias-free estimates of causal effects and allows for the generalization of empirical results under specific assumptions about population commonalities and differences.

Kern et al. (2016) focus on external validity and extrapolation in experimental research, highlighting the role of statistical techniques in enhancing generalizability. They assess methods such as propensity score approaches and Bayesian Additive Regression Trees (BART) for adjusting observed differences between experimental subjects and target populations. The authors underscore the importance of treatment effect heterogeneity and covariate alignment for accurate estimations, noting that flexible modeling techniques often outperform traditional regression approaches, albeit with strong assumptions. Their findings contribute significantly to the discussion on enhancing external validity through robust statistical frameworks.

Degtiar and Rose (2023) provide a comprehensive overview of statistical techniques designed to address external validity bias and enhance the generalizability of findings from both experimental and observational studies. The paper focuses on the importance of adjusting for differences between study populations and target populations using statistical approaches such as matching, inverse probability of participation weighting (IPPW), and outcome regressions. These methods aim to estimate the Population Average Treatment Effect (PATE), ensuring that the estimates are not skewed by covariate differences between the study sample and the broader population.

One prominent technique discussed is propensity score matching, which balances covariates between the study sample and the target population, thereby minimizing disparities between the two groups. Another key method is inverse probability of participation weighting, which adjusts for bias by balancing the selection probabilities between treated and untreated subjects, allowing for more reliable generalization of the results. However, the authors stress that these techniques should be

accompanied by checks for the positivity of common support (i.e., ensuring a positive selection probability for all subjects), as violating this assumption could undermine the robustness of the results. The paper also explores the use of doubly robust approaches, such as targeted maximum likelihood estimation (TMLE) and augmented inverse probability of participation weighting (A-IPPW). These methods combine modeling for both outcomes and selection probabilities, improving the accuracy of estimates even in cases of model misspecification. Doubly robust methods ensure unbiased estimates as long as at least one of the two models is correctly specified, offering an advanced solution for mitigating external validity bias in statistical analyses.

Finally, Degtiar and Rose emphasize the importance of integrating data from randomized and observational studies to capitalize on the internal validity of the former and the external validity of the latter. Through techniques like cross-design meta-analytic synthesis, researchers can combine information to provide more robust and broadly applicable estimates, significantly improving the scalability of findings across diverse contexts.

5.7 Systematic Reviews and Meta-Analyses

Systematic reviews and meta-analyses are essential methodologies for addressing the challenges of external validity and extrapolation in research. By synthesizing findings from multiple studies, these approaches provide comprehensive evidence on the effectiveness of interventions across diverse contexts, thereby enhancing the generalizability of research conclusions.

Bo and Galiani (2021) explore the concept of external validity and its implications for research findings, proposing a method for evaluating the external validity of randomized controlled trials (RCTs). They emphasize the importance of systematic reviews and meta-analyses as key methodologies for addressing external validity and facilitating extrapolation. By synthesizing findings from multiple studies, these approaches enhance the generalizability of causal estimates across different populations and settings. Aggregating data through systematic reviews and meta-analyses provides a comprehensive assessment of the consistency and robustness of causal relationships, offering a robust framework for understanding the applicability of research outcomes beyond the original study context.

Similarly, Vivalt (2020) addresses the challenges of external validity and extrapolation in impact evaluations, emphasizing the role of systematic reviews and meta-analyses in tackling these issues. By aggregating data from multiple studies, these methodologies provide a comprehensive understanding of treatment effects across

different contexts. This synthesis enables researchers to identify patterns and variations in results, thereby enhancing the generalizability of findings and informing better policy decisions. Using Bayesian hierarchical models, Vivalt demonstrates the effectiveness of these methodologies in systematically analyzing heterogeneity and providing robust estimates of treatment effects across diverse settings.

Avellar et al. (2017) delve into how systematic reviews address external validity, highlighting the challenges of generalizing intervention results to populations and settings that differ from those in the original studies. Traditionally, systematic reviews focus on internal validity—ensuring that an intervention produces effects without interference from other variables—often overlooking external validity, which pertains to whether results can be applied in different contexts. This omission is significant because many end users of systematic reviews, such as policymakers and practitioners, need to know if an intervention will be effective in their specific contexts.

To address this limitation, Avellar et al. emphasize the need to improve the reporting of information related to generalizability, applicability, and feasibility in systematic reviews. Generalizability refers to the extent to which results can be extended to broader populations or settings, while applicability focuses on how relevant an intervention is for a particular context, taking into account local factors such as demographics or political conditions. Feasibility, on the other hand, concerns whether an intervention can be implemented given the available resources.

The authors examined 19 systematic reviews to assess how they handled external validity and found that although many reviews provided information on study contexts and sample characteristics, they often lacked consistency and detail. For example, in the HomVEE review on home visiting programs for at-risk families, it was challenging to assess external validity due to inconsistent reporting on participants and contexts. Not all reviews clearly indicated whether study samples were representative of the target population or if study conditions mirrored real-world settings, making it difficult for practitioners to determine if an intervention would work in their particular environment.

Avellar et al. propose standardizing guidelines to improve external validity reporting in systematic reviews. Among their recommendations is the need to include detailed information about the study context, the demographic characteristics of participants, and subgroup-specific outcomes. A more rigorous and systematic approach to collecting and reporting this information would help decision-makers better assess the transferability and applicability of interventions to their local settings. While systematic reviews are a valuable tool for identifying effective interventions, Avellar et

al. argue that they still need to enhance their consideration of external validity to make the information more useful and relevant for end users.

Slough and Tyson (2023) tackle the issue of external validity in the context of meta-analyses, developing a theoretical framework that highlights the conditions necessary for this method to be effective in ensuring generalizable results. They emphasize that while meta-analysis is a powerful technique for combining findings from multiple studies to draw overarching conclusions, it requires specific conditions to be considered valid. One key aspect is the importance of ensuring that the studies included share a common empirical objective, a condition they refer to as target equivalence.

Achieving target equivalence requires harmonization across two critical elements: contrast (the type of comparison between treatment and control groups) and measurement (how outcomes are assessed). If studies are not harmonized in these aspects, a meta-analysis risks producing inconsistent or misleading results. For instance, Slough and Tyson describe how differences in the timing of information distribution or in methods for measuring voter turnout in a meta-analysis on interventions to increase voter participation can affect final results, thereby compromising the validity of the conclusions.

The authors also propose a distinction between two types of external validity relevant to meta-analyses: projectivism and cross-sectionalism. Projectivism focuses on whether a single study can transport its results to another context, while cross-sectionalism views external validity as a collective feature of a set of studies—central to meta-analyses that combine results from diverse contexts. The latter approach, Slough and Tyson argue, is more appropriate for meta-analyses because it allows for a systematic evaluation of the generalizability of effects across different settings.

Without proper harmonization and target equivalence, Slough and Tyson caution, meta-analyses may be less effective in addressing external validity concerns. They conclude by recommending greater attention to the design of meta-analyses and increased awareness of the potential limitations of these methods, particularly when there is insufficient harmonization between the included studies.

Burchett et al. (2011) examine how systematic reviews and meta-analyses handle external validity, focusing on the applicability and transferability of results to new settings. The authors point out that despite increasing recognition of the importance of external validity in health research, it is still frequently overlooked. To address this gap, Burchett and colleagues identified 25 frameworks used to evaluate external validity, categorizing the criteria into four main areas: context, intervention, outcomes,

and evidence. One of the main goals of their work is to establish how the results of health interventions can be generalized or adapted to different settings.

The methodology used in systematic reviews involves a thorough analysis of the context in which the original studies were conducted and the context where the results are to be transferred. This includes assessing the relevance of the intervention to the needs of the target population and the availability of appropriate resources in the new setting. The authors also highlight the importance of considering the characteristics of the intervention itself, analyzing how it was implemented and whether it is flexible enough to be adapted to different settings. This adaptability is critical to ensuring that an intervention can be customized without losing its effectiveness.

Furthermore, the outcomes of interventions are a critical aspect of assessing external validity. Burchett et al. emphasize that systematic reviews must consider the intervention's effectiveness, the sustainability of its outcomes, and the relevance of the measures used in the new context. Particular attention is given to the possibility that an intervention's effects may vary among subgroups or that unintended adverse effects may arise. Lastly, the consistency of evidence across different studies is considered essential for evaluating whether results can be generalized or transferred to new settings. One framework cited by the authors is RE-AIM, which considers various aspects of implementation and maintenance, demonstrating the importance of evaluating applicability across multiple levels.

Despite the widespread use of frameworks like RE-AIM, the authors acknowledge that there is still insufficient empirical data to demonstrate the effectiveness of these tools among policymakers. As a result, there is a clear need for further empirical research to more thoroughly explore how research findings can be applied and transferred to specific contexts.

6. Conclusion

External validity and extrapolation are fundamental concepts in empirical research, particularly in social sciences and policy evaluation. These concepts ensure that the results of a study can be generalized beyond the specific sample and context in which they were obtained, making them crucial for the applicability and relevance of research findings to broader populations and different settings. The literature identifies three primary models for addressing the challenges associated with external validity and extrapolation: the validity of the original study, statistical adjustments, and the analysis of causal mechanisms. Additionally, several sub-models, including systematic reviews and meta-analyses, multi-site and cross-context studies,

and iterative testing and refinement, provide complementary approaches to enhance the robustness and generalizability of research findings.

Validity of the Original Study

The first model emphasizes the importance of the validity of the original study, focusing on how representative the study is for other populations. This model scrutinizes the internal validity and context-specific factors of the study to determine how well the findings can be generalized. Understanding the contextual factors influencing the implementation and outcomes of interventions allows for better addressing the unique needs and conditions of various populations, thereby enhancing the relevance and impact of findings. This approach involves not only the adaptation of interventions but also the inclusion of local stakeholders in the design and implementation process to ensure cultural and contextual appropriateness. For instance, educational reforms that were successful in small pilot programs often require significant modifications when scaled up to larger, more diverse populations (Burchett et al., 2011; Cook, 2014; Williams, 2020).

Burchett et al. (2011) highlight the necessity of adapting interventions to local conditions to ensure their effectiveness in different settings. They argue that understanding the contextual factors influencing the implementation and outcomes of interventions allows for better addressing the unique needs and conditions of various populations, thereby enhancing the relevance and impact of findings. This approach involves not only the adaptation of interventions but also the inclusion of local stakeholders in the design and implementation process to ensure cultural and contextual appropriateness.

Cook (2014) discusses the critical role of population characteristics in determining the success and transferability of educational policies. Without considering the demographic and socio-economic characteristics of the target population, educational interventions risk being ineffective or even counterproductive. Cook provides examples showing how selection bias in student samples can affect the generalizability of educational interventions, emphasizing the need for studies to reflect the diversity of the broader population.

Williams (2020) also underscores the importance of considering the representativeness of the study population. He points out that interventions often require substantial adjustments when scaled up, and the initial success in controlled settings does not always translate to larger, more diverse populations. For example, educational reforms that worked well in small pilot programs frequently need modifications to address the diverse needs of larger school districts.

6.1 Statistical Adjustments

The second model focuses on using statistical techniques to adjust results, taking into account the characteristics of samples and applying these adjustments to other populations. This quantitative approach involves methods such as propensity score matching to correct data and enhance generalizability. Statistical models are employed to simulate and predict outcomes in different contexts, which involves identifying and adjusting for differences in covariates and context-specific factors that may influence the outcomes. Bareinboim and Pearl (2013) discuss a general algorithm for deciding transportability, which involves using statistical models to determine whether and how findings from one context can be extrapolated to another. This method relies on identifying and adjusting for differences in covariates and context-specific factors that may influence the outcomes.

Degtiar and Rose (2023) emphasize the importance of measuring rich covariates and documenting context-specific variables to facilitate the extrapolation of findings. By incorporating detailed contextual information into statistical models, researchers can better understand how interventions interact with different environments and make more accurate predictions about their performance.

Rothwell (2005) explores the external validity of randomized controlled trials (RCTs), noting that statistical adjustments and modeling are crucial for generalizing findings. By using techniques such as meta-analysis and re-weighting, researchers can account for differences in study populations and contexts. These methods allow for the combination of data from multiple studies to provide a more comprehensive understanding of intervention effects across diverse settings.

Kern et al. (2016) assess methods for generalizing experimental impacts, focusing on the application of statistical models to analyze data from multiple contexts. They argue that modeling interactions between interventions and contextual factors is essential for understanding the variability in outcomes. By using hierarchical models and other advanced statistical techniques, researchers can partition the variance attributable to different sources and identify the key factors that influence intervention success.

Khosrowi (2022) discusses successful extrapolation, emphasizing the importance of robust statistical methods in predicting how interventions will perform in new environments. He argues that statistical techniques such as causal inference models and sensitivity analysis are critical for addressing the uncertainties associated with extrapolation. These methods help to quantify the confidence in predictions and identify potential limitations of the models used.

Chassang and Kapon (2022) also emphasize the role of statistical techniques in designing randomized controlled trials with external validity in mind. They suggest incorporating mechanisms to test the assumptions underlying statistical models and using iterative processes to refine these models based on empirical data. By continuously improving the accuracy and reliability of statistical predictions, researchers can enhance the external validity of their findings.

Analysis of Causal Mechanisms

The third model is based on understanding the causal mechanisms underlying the effectiveness of an intervention. This approach posits that if researchers understand why an intervention works in one context, they can adapt it to other contexts based on these underlying mechanisms. By focusing on the mechanisms that drive outcomes, researchers can identify which aspects of the intervention are crucial for its success and which can be modified to better fit new contexts. A thorough comprehension of the mechanisms through which policies operate allows for more effective customization and adaptation. This understanding helps policymakers identify the critical components of interventions that need to be preserved while allowing flexibility in other aspects to suit local conditions.

Cartwright (2011) emphasizes the importance of understanding the explanatory relevance of evidence and how it can inform the adaptation of interventions to new contexts. Bühlmann (2020) highlights the role of theoretical understanding in improving external validity. By focusing on causality and the underlying mechanisms, researchers can better predict how interventions will perform in different environments. This approach involves developing comprehensive theoretical models that account for various factors influencing the outcomes of interventions, guiding the adaptation and scaling of interventions across diverse settings.

Busetti (2023) and Busetti and Dente (2018) discuss the significance of mechanistic understanding in policy design and implementation. They argue that a thorough comprehension of the mechanisms through which policies operate allows for more effective customization and adaptation. This understanding helps policymakers identify the critical components of interventions that need to be preserved while allowing flexibility in other aspects to suit local conditions.

The Generalizability Framework from the Stanford Social Innovation Review (2017) also underscores the importance of focusing on mechanisms when considering the generalizability of research findings. By understanding the underlying behavioral and contextual mechanisms, researchers can make more informed decisions about whether and how to adapt interventions for different settings.

6.2 Sub-Models and Extensions

While the three primary models provide a robust framework for addressing external validity and extrapolation, several sub-models and extensions complement these approaches. These include systematic reviews and meta-analyses, multi-site and cross-context studies, and iterative testing and refinement.

Systematic Reviews and Meta-Analyses

Systematic reviews and meta-analyses synthesize findings from multiple studies to provide comprehensive evidence on the effectiveness of interventions across diverse contexts. This approach helps identify patterns and variations in outcomes across different settings, providing a broader understanding of intervention impacts. By pooling data from various studies, researchers can achieve greater statistical power and precision in estimating intervention effects. Meta-analyses also facilitate the identification of moderators and mediators that influence the effectiveness of interventions, providing insights into the contextual factors that affect outcomes. These methods help identify the heterogeneity of treatment effects and explore the sources of this variability. By systematically reviewing the literature and using meta-analytic methods, researchers can assess the robustness of evidence and identify gaps in knowledge. Systematic reviews and meta-analyses are crucial for translating research findings into practical applications by providing a clear understanding of what works, for whom, and under what conditions (Avellar et al., 2017; Slough & Tyson, 2023; Bo & Galiani, 2021; Vivalt, 2020; Williams, 2020).

Multi-Site and Cross-Context Studies

Multi-site and cross-context studies involve conducting studies in multiple locations and varied contexts to gather evidence on the effectiveness of interventions in different settings. By comparing results across different sites, researchers can determine the consistency of causal relationships and identify the conditions under which these relationships hold. Conducting experiments in multiple locations helps to understand how different contexts affect the outcomes of interventions. This approach allows researchers to identify context-specific factors and assess whether the intervention's effectiveness can be replicated in various settings. Conducting studies in varied contexts helps in verifying that the constructs being measured are relevant and applicable across different settings, providing a structured approach to integrating evidence from multiple studies conducted in different contexts (Bold et al., 2013; Cook, 2014; Esterling et al., 2023; Stanford Social Innovation Review, 2017).

Iterative Testing and Refinement

Iterative testing and refinement involve a continuous cycle of testing interventions, analyzing outcomes, and making necessary adjustments to improve effectiveness and generalizability. This iterative process of theory development and empirical testing strengthens the external validity of research findings and facilitates their application across diverse contexts. Experimental and quasi-experimental designs allow for systematic variation of conditions and examination of their impact on intervention outcomes. Continuous adaptation and refinement of interventions address the specific needs and characteristics of different contexts, improving the design and delivery of interventions. Interventions should be initially tested in a variety of contexts to identify potential modifications that could enhance their effectiveness, incorporating mechanisms to test and adapt interventions in response to observed outcomes (Shadish et al., 2002; Pritchett & Sandefur, 2014; Tipton & Peck, 2017; Chassang & Kapon, 2022).

In conclusion, the integration of these primary and sub-models provides a comprehensive framework for addressing the challenges of external validity and extrapolation. By employing a combination of detailed contextual understanding, robust statistical techniques, mechanistic insights, and iterative testing, researchers can ensure that their findings are applicable and beneficial across diverse populations and settings. This multifaceted approach not only enhances the reliability and validity of research outcomes but also facilitates the practical implementation of interventions in real-world scenarios.