Mitigating the Building Energy Performance Gap

Outcome-Based Policies and the Potential for Standardization

Faculty of Engineering

Department of Architecture

Dissertation submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Architecture

Khaleel Odeh

December 2025

Supervisors:

Professor Pieter De Wilde

Professor Lori McElroy

Declaration of Authorship

I, Khaleel Odeh, hereby declare that this thesis titled "Mitigating the Building Energy Performance Gap" is the result of my own original work. All sources of information and the work of others have been properly cited and referenced. I further declare that tools such as ChatGPT and Grammarly were used only for the purpose of identifying grammatical issues and enhancing the clarity of writing; they did not contribute to the creation of any original ideas or content.

Date: 8-12-2025

Signature: Khaleel Odeh

Acknowledgements

I would like to begin by expressing my deepest and most heartfelt gratitude to Professor Pieter de Wilde, my primary supervisor, mentor, and role model. Professor de Wilde has guided me not only throughout my PhD journey at the University of Strathclyde but also during my BSc and MSc dissertations at the University of Plymouth. His unwavering support, patience, and wisdom have profoundly shaped me both academically and personally. I truly believe that no words are sufficient to convey how thankful I am for his continuous encouragement and the inspiration he has provided. He has not only been a supervisor but a true example of dedication, intellect, and integrity, qualities I deeply aspire to emulate.

I am also immensely grateful to Professor Lori McElroy, my second supervisor, who joined my supervisory team in my second year. Her involvement was pivotal, especially in helping me build connections with numerous experts whose input significantly enriched my research and allowed me to complete data collection in such a short time. Professor McElroy's kindness, positivity, and constant support have been a major source of motivation, and I deeply appreciate the time, care, and enthusiasm she brought to my work.

My sincere thanks go to Dr. Ombretta Romice, PhD coordinator, whom I truly consider not only a colleague but a friend. Ombretta's warmth and tireless efforts to support every student have made an immeasurable difference to our community. I will truly miss her presence and support.

I would like to express my deepest and most profound gratitude to my parents, whose love, patience, and endless sacrifices have been the foundation of everything I have achieved. Their unwavering belief in me has given me the courage to pursue my goals even when the path seemed uncertain. I am blessed beyond measure to have parents who have never hesitated to stand by my side, who have supported me through every challenge, and who have shown me what unconditional love truly means. Their compassion, generosity, and faith have shaped who I am today, and no words can express how grateful and indebted I am to them. Every step of this journey, every accomplishment, and every moment of perseverance are a reflection of their boundless love and guidance. I owe everything I have become to them.

To my dear friends and colleagues at the Architecture Department, thank you for making this journey memorable. My heartfelt thanks go to Khaled Alassad, my soulmate and companion throughout this PhD. His presence has made this journey not only bearable but joyful, and together we laughed, supported each other, and made the most of every moment. I would also like to thank Farhana Isa and Ahmad Alsaedi for their kindness, support, and the positive spirit they brought into the department, which made the days brighter and more enjoyable.

I would also like to extend my warmest thanks to my dear friends Mohammed Al-Rashdi, Mahmoud Yaghi, Muath Akkad, Nour Ashraf, Abdul Hadi Chibli, Abdulrahman Alhouri, and Anas Al-Meir. Their friendship, encouragement, and constant presence have been a source of strength and comfort throughout this journey. Each of them has contributed in their own special way to making this experience more meaningful, and I am truly grateful for the laughter, conversations, and unwavering support they have shared with me.

Finally, I am profoundly grateful to Dr. Jon Hand for his invaluable insights and feedback, which have significantly shaped and improved many parts of this thesis. I also extend my sincere thanks to Paul Bannister for his efforts in helping me identify a suitable case study, and to the MSc students at LTH, Md Omor Faruk Monsi Parvaz and Jitendra Pavan Jakkula, for their dedicated work in creating the energy simulation models. Their contributions are highly appreciated. I feel truly privileged to have benefited from the wisdom and experience of such esteemed experts and collaborators.

ABSTRACT

Evidence since the mid-1990s has shown a persistent divergence between predicted and measured building energy performance, known as the Energy Performance Gap (EPG). Despite advances in modelling, regulation, and building-performance assessment, the EPG remains unresolved, undermining confidence in sustainable architecture and slowing progress toward zero-carbon buildings. Existing mitigation approaches have largely focused on improving prediction accuracy, while giving less attention to the wider socio-technical, behavioural, operational, and regulatory factors that shape building performance. As a result, the causes and magnitude of the EPG remain inconsistently understood and reported. This thesis aims to critically investigate the EPG as a complex and multifaceted phenomenon. Its objectives are to synthesise evidence on EPG magnitude, identify and evaluate root causes, assess opportunities for mitigation, and develop a more standardised approach to EPG quantification and reporting. A mixed-methods research design was adopted, combining a systematic literature review, root cause analysis, semi-structured interviews, a simulation-based case study using multiple dynamic modelling tools, and a panel discussion. The findings show that the EPG arises from interconnected factors across design, construction, commissioning, and operation. The assessment of NABERS in the UK, an outcome-based rating system originally developed in Australia, indicates that outcome-based approaches can help mitigate design-stage and systemic causes of the EPG by introducing explicit performance targets, post-occupancy evaluation, and stronger accountability. However, their effectiveness depends on sustained implementation and continued operational feedback. The simulation-based case study further demonstrates that reported EPG magnitude is highly sensitive to modelling assumptions and tool selection, with substantial variation occurring even under identical input conditions. This highlights the uncertainty embedded in EPG quantification and shows that model calibration, while valuable, does not by itself ensure predictive validity. The thesis advances a more realistic and accountable understanding of the EPG, supporting improved policy design, more credible performance assessment, and better-informed decision-making. It makes three original contributions to knowledge: first, it provides a comprehensive cross-context synthesis of EPG magnitudes across geographical locations, modelling approaches, and quantification methods; second, it delivers an independent academic assessment of NABERS in the UK; and third, it develops a framework for standardised EPG quantification and reporting that integrates gap definitions and accounts for uncertainty, thereby improving transparency and comparability.

List of Publications

Odeh, K., L. McElroy and P. de Wilde, 2025. Standardizing the quantification of the Energy Performance Gap in office buildings. In: Thomas, P.C., ed. *Building Simulation 2025*, Brisbane, Australia, August 24-27, 2025

Odeh, K., L. McElroy and P. de Wilde, 2025. Mitigating the building energy performance gap through better simulation and control strategies. In: Kumar, B. and Moreno-Rangel, A., eds. *EG-ICE 2025*, Conference on Intelligent Computing in Engineering, Glasgow, UK, 1-3 July 2025

Odeh, K. and P. de Wilde, 2024. The utilization of building simulation results as input data for building level digital shadows and system level digital twins. In: Jenkins et al, eds. *uSIM 2024*, 4th IBPSA Scotland Conference, Edinburgh, UK, 25 November 2024

Odeh, K. and P. de Wilde, 2024. The Building Energy Performance Gap: Potential of Digital Twins and Evidence-Based Policies. In: Reveiro and Arias, eds. *EG-ICE 2024*, Conference on Intelligent Computing in Engineering Vigo, Spain, 3-5 July 2024

Odeh, K. and P. de Wilde, 2023. Exploring the potential of Digital Twins at the District Scale: a framework for investigation. In: Pan and Yan, eds. *Building Simulation 2023*, Shanghai, China, September 4-6, 2023

Manuscripts in preparation for submission:

The Magnitude of the Building Energy Performance Gap: A Systematic Review. **Target journal:** Renewable and Sustainable Energy Reviews.

The Potential of NABERS to Address Cross-Cutting Drivers of the Building Energy Performance Gap. **Target journal:** Energy and Buildings.

NABERS as a Policy Instrument for Influencing Energy Modelling Practices. **Target journal:** Journal of Building Performance Simulation.

The Energy Performance Gap in Buildings: Between Perception and Reality. **Target journal:** Journal of Building Performance Simulation.

Contents

List of Publications **5**

List of tables **10**

List of figures **12**

Abbreviations **14**

1. Introduction **1**

1.1 Research Problem 7

1.2 Aim and Objectives 12

1.3 Thesis Structure 15

2. Methodology Overview **18**

2.1 Systematic Literature Review 19

2.1.1 Root Causes of the EPG 20

2.1.2 Magnitude of the EPG 21

2.2 Mitigation Strategies through Policy 21

2.2.1 Root Cause Analysis 22

2.2.2 Semi-Structured Interviews 23

2.3 Complexity Demonstration 24

2.3.1 Dynamic Simulation Modelling 24

2.3.2 Comparative Analysis 25

2.4 Standardization Efforts 27

2.5 Integration of Methods 28

2.6 Conceptual Clarification and Justification 31

3. Systematic Literature Review **35**

3.1 SLR Methodology 35

3.1.1 SLR Results 37

3.1.2 Quantification and Normalization 40

3.2 Underlying Causes of the EPG 42

3.2.1 Design stage root causes 42

3.2.2 Construction stage root causes 44

3.2.3 Commissioning and operational stages root causes 45

3.2.4 Cross-cutting factors 47

3.3 Magnitude of the EPG 49

3.3.1 Europe 50

3.3.2 Australia ... 78
3.3.3 North America ... 78
3.3.4 South America ... 79
3.3.5 Asia ... 82
3.4 Summary ... 88
4. Mitigation Strategies through Policy ... 94
4.1 Detailed Methodology ... 94
4.1.1 Root Cause Analysis ... 94
4.1.2 Semi-Structured Interviews and Coding ... 96
4.1.3 Ethical Considerations ... 100
4.1.4 Participant Selection ... 100
4.1.5 Analytical Approach ... 101
4.2 Root Cause Analysis ... 101
4.2.1 Design Stage Root Causes ... 102
4.2.2 Construction Stage Root Causes ... 103
4.2.3 Commissioning and Operational Stage Root Causes ... 104
4.2.4 Cross-Cutting Factors ... 105
4.3 Results and Analysis ... 106
4.3.1 Design Stage Root Causes ... 106
4.3.2 Commissioning and Operational Stages Root Causes ... 125
4.3.3 Cross-Cutting Factors ... 140
4.3.4 Combined Synthesis ... 150
4.4 Policy Interventions Towards Mitigating the EPG ... 154
4.4.1 Design Stage Mitigation Strategies ... 154
4.4.2 Commissioning and Operational Stages Mitigation Strategies ... 158
4.4.3 Cross-Cutting Factors ... 160
4.5 Key Takeaways and Summary of Recommendations ... 162
5. Complexity Demonstration ... 167
5.1 Detailed Methodology ... 167
5.1.1 EBD Laboratory ... 168
5.1.2 Simulation Inputs and Assumptions ... 169
5.1.3 Comparative Analysis ... 170
5.2 Results & Analysis ... 172
5.2.1 Inter-Tool Gaps ... 172

5.2.2 Annual Measured Versus Simulated ... 180
5.2.3 Monthly Measured Versus Simulated (21°C) ... 182
5.3 Discussion ... 183
5.3.1 Key Takeaways ... 186
6. Standardization Efforts ... 192
6.1 Methodology ... 192
6.1.1 Participants Selection ... 194
6.2 Establishing the Case for Standardization ... 195
6.2.1 Benchmarks ... 195
6.2.2 Setting Acceptable Deviations ... 198
6.3 Expert Panel ... 201
6.3.1 Modelling Choices and Ambiguity of the Size of the EPG ... 201
6.3.2 Defining Efficiency in the Context of the EPG ... 202
6.3.3 Challenges ... 203
6.3.4 Different Benchmarks Under which the EPG is Quantified ... 204
6.3.5 Anticipated and Unanticipated Performance Gaps ... 208
6.3.6 Classifications of the Energy Performance Gap ... 209
6.4 Summary ... 211
6.4.1 Framework ... 215
7. Synthesis of Knowledge ... 220
7.1 Integration of Key Findings ... 220
7.2 Cross-Cutting Themes ... 225
7.3 Theoretical and Methodological Implications ... 226
7.4 Policy and Practice Implications ... 227
8. Conclusion ... 229
8.1 Contribution to Knowledge ... 230
8.2 Key Lessons ... 232
8.3 Research Limitations ... 235
8.4 Future Research ... 237
References ... 239
Appendices ... 261
Appendix A: PRISMA 2020 Checklist Mapping for the SLR ... 261
Appendix B: Consent Form Template ... 265

Appendix C: Harmonised Input Assumptions and Output Categories for the Simulation Models.. 266

List of tables

Table 2.2: Key governance terms relevant to building energy performance 34
Table 3.1: Analysed records for residential buildings in the UK 52
Table 3.2: Analysed records for commercial buildings in the UK 57
Table 3.3: Analysed records for residential buildings in Switzerland 61
Table 3.4: Analysed records for residential and non-residential buildings in Italy 64
Table 3.5: Analysed records for residential and non-residential buildings in Denmark, Germany, Poland and France 68
Table 3.6: Analysed records for residential and non-residential buildings in Estonia, Cyprus and Ireland 72
Table 3.7: Analysed records for residential and non-residential buildings in other countries in Europe 75
Table 3.8: Analysed records for residential and non-residential buildings in Australia, Canada, Brazil and USA 80
Table 3.9: Analysed records for residential and non-residential buildings in Asia 84
Table 3.10: Conditions influencing the widening or narrowing of EPG magnitude and changes in energy consumption identified through the SLR 92
Table 4.1: Synthesis of perspectives on oversized HVAC components 109
Table 4.2: Synthesis of perspectives on lack of appropriate details 110
Table 4.3: Synthesis of perspectives on poor thermal concepts..111
Table 4.4: Synthesis of perspectives on lack of simplicity 113
Table 4.5: Synthesis of perspectives on lack of buildability...114
Table 4.6: Synthesis of perspectives on lack of sequence of construction process 115
Table 4.7: Synthesis of perspectives on inaccurate performance targets 120
Table 4.8: Synthesis of perspectives on undefined functions and operational requirements...... 122
Table 4.9: Participant-level Distribution of Judgments 123
Table 4.10: Distribution of judgments across themes 124
Table 4.11: Qualitative findings and quantitative judgments occupant behaviour associated with grey area 128
Table 4.12: Qualitative findings and quantitative judgments occupant behaviour associated with psychological reasons 130
Table 4.13: Qualitative findings and quantitative judgments on Overestimation of system efficiency (poor commissioning) 132
Table 4.14: Qualitative findings and quantitative judgments on inaccuracy of sensor network data 134
Table 4.15: Qualitative findings and quantitative judgments on systems installation mismatch (HVAC and lighting) 135
Table 4.16: Synthesis of perspectives on root causes associated with system deterioration 137
Table 4.17: Synthesis of perspectives on root causes associated with functional and operational changes 138

Table 4.18: Participant-level Distribution of Judgments 139
Table 4.19: Distribution of judgments across themes 140
Table 4.20: Participant-level Distribution of Judgments 149
Table 4.21: Distribution of judgments across themes 150
Table 5.1: Annual Measured-to-Simulated Electricity Ratios 181
Table 5.2: Monthly Measured-to-Simulated Electricity Ratios (Setpoint 21 °C) 182
Table 5.3: Annual Statistics of Simulation Gap (Measured/Simulated) 182
Table 5.4: Monthly error metrics for simulated electricity use against measured 2023 electricity consumption at 21°C 183
Table 6.1: Synthesis of perspectives on appropriate benchmarks for calculating the EPG 198
Table 6.2: Comparison of SSI1 and SSI2 perspectives on acceptable deviations in energy performance 200

List of figures

Figure 1.1: Performance Gap Typologies 5
Figure 1.2: Thesis Structure 17
Figure 2.1: Methodological overview of the thesis research design 30
Figure 3.1: PRISMA Diagram 39
Figure 3.2: Geographical coverage of the SLR 49
Figure 4.1: Methodological framework for assessing interpretive saturation within a deductive analytical design 99
Figure 4.2: Design stage root causes 102
Figure 4.3: Construction stage root causes 103
Figure 4.4: Commissioning and operational stages root causes 104
Figure 4.5: Cross-Cutting Factors 105
Figure 4.6: Design Stage interview judgments by all participants 107
Figure 4.7: Participants judgements on inefficient (top three) and efficient (bottom three) design root causes 108
Figure 4.8: Participant judgements on energy modelling related root causes 116
Figure 4.9: Participant judgements on inaccurate performance targets and undefined functions and operational requirements 120
Figure 4.10: Participant judgements on commissioning and operational stages root causes 125
Figure 4.11: Participant judgements on occupant behaviour related causes 126
Figure 4.12: Synthesis of perspectives on deterioration of energy systems 136
Figure 4.13: Synthesis of perspectives on functional and operational changes 137
Figure 4.14: Synthesis of perspectives on cross-cutting causes 141
Figure 4.15: Synthesis of perspectives on poor client's knowledge 142
Figure 4.16: Synthesis of perspectives on lack of post occupancy evaluation 144
Figure 4.17: Synthesis of perspectives on poor communication and collaboration 145
Figure 4.18: Synthesis of perspectives on lack of accountability 146
Figure 4.19: Synthesis of perspectives on over-optimistic prediction 147
Figure 4.20: Mitigation strategies for issues with efficient and inefficient design 155
Figure 4.21: Mitigation strategies to address issues related to uncertainties in building energy modelling 157
Figure 4.22: Mitigation strategies to address inaccuracy of performance target and undefined operational requirement 158
Figure 4.23: Mitigation strategies to address issues in commissioning and operational stages . 160
Figure 4.24: Preventative measures for emergence of cross-cutting factors 161
Figure 5.1: North and south facades of the EBD Laboratory (top left: north facade; top right: south facade), circulation area (bottom right), one of the lighting cells (bottom middle), and one of the thermal cells (bottom left) 169
Figure 5.2: Stages of comparative analysis 171
Figure 5.3: Measured electricity consumption in 2023 and 2024 172

Figure 5.4: Annual simulated electricity consumption under three simulation scenarios 173
Figure 5.5: Monthly simulated electricity consumption at 21°C (kWh) 174
Figure 5.6: Annual simulated end use electricity consumption at 21°C (kWh)......................... 175
Figure 5.7: Monthly simulated fans electricity consumption at 21°C (kWh)............................ 176
Figure 5.8: Annual simulated heating electricity consumption under three simulation scenarios (kWh) 177
Figure 5.9: Monthly simulated heating electricity consumption (kWh).................................... 178
Figure 5.10: Annual simulated cooling electricity consumption under three simulation scenarios (kWh) 179
Figure 5.11: Monthly simulated cooling electricity consumption (kWh).................................. 180
Figure 6.1: Methodologies adopted to achieve the aim of the chapter divided into three stages 194
Figure 6.2: Distinction between different energy benchmarks under which the EPG is calculated and their representativeness of actual EPGs 208
Figure 6.3: Anticipated versus unanticipated gaps 209
Figure 6.4: Classifications of the EPG, their relevance and importance 211
Figure 6.5: Investigation, quantification and reporting framework of the EPG 216

Abbreviations

AAs: Accredited Assessors

AHU: Air Handling Unit

ASHRAE: American Society of Heating, Refrigerating and Air-Conditioning Engineers

BASIX: Building Sustainability Index

BCO: British Council for Offices

BES: Building Energy Simulation

BEMS: Building Energy Management System

BPS: Building Performance Simulation

BREEAM: Building Research Establishment Environmental Assessment Method

CIBSE: Chartered Institution of Building Services Engineers

CPD: Continuing Professional Development

DER: Dwelling Emission Rate

DfP: Design for Performance

DHW: Domestic Hot Water

DOE: Department of Energy

ECON19: Energy Consumption Guide 19

EPBD: Energy Performance of Buildings Directive

EPC: Energy Performance Certificate

EPG: Energy Performance Gap

ESPM: ENERGY STAR Portfolio Manager

EUI: Energy Use Intensity

FM: Facility Manager

HVAC: Heating, Ventilation and Air Conditioning

IDA-ICE: IDA Indoor Climate and Energy

IDRs: Independent Design Reviewers

IES-VE: Integrated Environmental Solutions – Virtual Environment

IBPSA: International Building Performance Simulation Association

LCB: Low Carbon Buildings

LEB: Low Energy Building

LEED: Leadership in Energy and Environmental Design

MKECI: Milton Keynes Energy Cost Index

MINERGIE: The Swiss building standard for comfort, efficiency and climate protection

MINERGIE-A: Active Standard

MINERGIE-P: Passive Standard

MINERGIE-ECO: Ecological Standard

MOPEC: Model Policies for Energy Codes

NABERS: National Australian Built Environment Rating System

NEED: National Energy Efficiency Data Framework

nZEB: Nearly Zero Energy Building

PH: Passivhaus

PHPP: Passive House Planning Package

POE: Post Occupancy Evaluation

PRISMA: Preferred Reporting Items for Systematic Reviews and Meta Analyses

PROBE: Post-Occupancy Review of Buildings and their Engineering

RCA: Root Cause Analysis

RdSAP: Reduced Data Standard Assessment Procedure

SAP: Standard Assessment Procedure

SL: Soft Landings

SLR: Systematic Literature Review

SNECS: Sub-National Energy Consumption Statistics

TM46: Technical Memorandum 46

TM54: Technical Memorandum 54

UHI: Urban Heat Island

UNSDG: United Nations Sustainable Development Goals

VAV: Variable Air Volume

ZEB: Zero Energy Building

1. Introduction

Long before the widespread awareness of building energy efficiency, civilizations across millennia developed sophisticated strategies to moderate indoor environments, thermal massing, solar optimisation, and natural ventilation. As early as 5500 BC, semi subterranean dwellings in the Carpathian region of present-day Romania used the stabilising effect of the earth to maintain constant indoor temperatures throughout the year (Pătrașcu 1984). By 4000 to 3000 BC, builders in Persia and Egypt were already manipulating airflow and thermal mass through wind towers, evaporative cooling, cavity wall systems, thick brick construction, and other methods that reduced indoor overheating (A'zami 2005, Ghaemmaghami and Mahmoudi 2005, El-Shorbagy 2010). Greek and Roman innovations further formalised solar responsive architecture, including the routine south orientation of houses, the development of sun heated rooms, the creation of underfloor heating through the hypocaust system, and the use of solar chimneys to drive natural ventilation (Sear 1982, Sayigh and Marafia 1998, Oxlade 2012). Throughout the first millennium AD, diverse cultures, including communities in Palestine, Native American societies, Icelandic settlers, and builders in southern China, continued to refine thermally effective strategies such as cave dwellings, rock integrated houses, turf construction, and inward facing courtyard buildings that achieved comfort without mechanical systems (Boubekri 2008, Chen 2011, Milek 2012).

Subsequent centuries introduced further advances, including early solar collectors in Switzerland, the discovery of infrared radiation in Germany, the invention of the cast iron radiator in Russia, and the patenting of double layer windows in the United States (Kühnl-Kinel 2000, Jones and Bouamane 2012). The late nineteenth and early twentieth centuries marked the emergence of modern environmental control through inventions such as the electric fan, early highly insulated structures comparable to present passive houses, commercial solar water heaters, mechanical air conditioning, psychrometric analysis, geothermal energy systems, and radiant heating (Hattell and Morrison 1995, Kühnl-Kinel 2000, Carrier 2014). These innovations demonstrate that the principles underpinning contemporary low energy and passive building design did not appear suddenly in response to the energy concerns of the nineteen seventies but instead represent the cumulative outcome of more than seven thousand years of evolving scientific understanding, architectural experimentation, and technological progress.

Chapter One: Introduction

Widespread awareness of building energy efficiency can be traced back to the 1970s oil crisis, a period that catalysed extensive debate on indoor thermal environments, patterns of energy consumption, and optimization strategies for the built environment (Hussein and Alhebsi 2025, Jiang, Zhang et al. 2025). Efforts to improve the energy performance of buildings gained significant momentum during the 1980s, when the concept of Low-Energy Buildings (LEB) began to take shape. Large-scale empirical demonstrations, such as the Linford Low-Energy Houses in Milton Keynes, provided some of the earliest monitored evidence of improved building performance in the UK (Everett, Horton and Doggart 1985). In parallel, the Milton Keynes Energy Cost Index (MKECI) established one of the first measurable benchmarks for assessing energy performance, a contribution widely acknowledged in contemporary literature as a precursor to later rating schemes (Lomas, Steadman and Brown 2019, Allen, Cohen and Hammond 2020, Hammond and Li 2022). These initiatives illustrate how the 1980s laid the foundations for subsequent policies and practices in energy performance assessment and LEBs development.

While empirical demonstrations and early initiatives advanced practical understanding of low-energy design, they also highlighted the need for tools capable of predicting energy performance beyond what could be demonstrated through monitoring alone. This growing methodological demand coincided with the emergence of significant building simulation programs. Early simulation tools developed in the mid-1970s were limited to single-zone dynamic calculations of indoor temperature, humidity, and thermal loads (Kusuda 1976, Bennett and et al. 1977). Throughout the late 1970s and early 1980s, these capabilities expanded to support multi-zone modelling and to incorporate the performance characteristics of HVAC equipment under different operating conditions (Buhl, Curtis et al. 1979, Palmiter, Wheeling and Judkoff 1982, Clarke and McLean 1988). During this same period, modular simulation tools evolved in parallel, beginning with programs for active solar systems (Solar Energy 2000) and HVAC control systems (Kelly, Park et al. 1984). By the 1990s, advances in personal computing enabled the development of graphical user interfaces for simulation programs, greatly improving accessibility (eQUEST 2005). The mid-1990s marked the beginning of the development of DOE's current flagship simulation engine, EnergyPlus (Crawley, Lawrie et al. 2001), arising from the consolidation of two earlier government-sponsored programs (Buhl, Curtis et al. 1979, Office 1992). This period also saw the introduction of the Modelica modelling language for system-level simulation

(Fritzson 2003). As these advances accumulated, the research community formalised its international collaboration through the establishment of the International Building Performance Simulation Association (IBPSA) in 1987, a milestone that contributed substantially to the consolidation, professionalisation, and global coordination of building performance simulation research and tool development (Hensen 2000).

This trajectory paved the way for contemporary sustainable building standards, with a decisive milestone being the emergence of the Passivhaus (PH) concept in the early 1990s, developed through the collaboration of Wolfgang Feist in Germany and Professor Bo Adamson at Lund University in Sweden (Pitts 2017, Moreno-Rangel 2021). Their partnership drew on Scandinavian experience with super-insulated dwellings and German advances in building science, consolidating earlier low-energy lessons into a coherent approach (Moreno-Rangel 2021). The result was a rigorous, testable, and certifiable framework for ultra-low energy demand, defined by clear design criteria and verified through modelling, airtightness testing, and independent certification.

Against this historical backdrop, recent developments in digital technologies have introduced a further shift in the way building performance is predicted and evaluated. While traditional simulation environments remain grounded in white-box, physics-based methods, contemporary research increasingly explores the potential of data-driven and machine-learning approaches to complement or, in certain applications, substitute for conventional modelling. These black-box and hybrid techniques are capable of learning complex input–output relationships directly from empirical data, offering advantages in computational efficiency and the ability to represent phenomena that are difficult to capture explicitly in first-principles simulations (Amasyali and El-Gohary 2018). Machine-learning models have demonstrated strong performance in tasks such as energy-consumption forecasting, load prediction, and indoor-environment estimation (Deb, Zhang et al. 2017). However, their effectiveness depends heavily on the quality and representativeness of available training data, raising important questions concerning model generalisability, interpretability, and the risk of performance degradation when applied beyond their training domain. In response to these challenges, recent work has highlighted the potential of transfer-learning techniques, which enable models trained in one context to be adapted to different building types or climatic regions with substantially reduced data requirements (Wei, Li and Guo 2019, Drgoňa, Arroyo et al. 2020).

Despite these advances, evidence since the mid-1990s has revealed a persistent divergence between predicted and measured building energy performance. First labelled as a "credibility gap" (Norford, Socolow et al. 1994, Bordass, Cohen and Field 2004), this phenomenon has since become broadly recognized in the literature as the Energy Performance Gap (EPG). A pivotal contribution in this field was the series of Post-Occupancy Review of Buildings and their Engineering (PROBE) studies in the UK. These included both detailed occupant surveys (Leaman and Bordass 2001) and broader guidance on feedback and post-occupancy evaluations (POE) (Bordass, Leaman and Eley 2002), which together provided empirical insights into the operational shortcomings of buildings relative to design expectations. Further large-scale evidence was presented in 2016 through Innovate UK's Building Performance Evaluation programme, which assessed 50 non-domestic buildings and revealed that actual CO_2 emissions were, on average, 3.8 times higher than the design predictions stipulated in the UK's building regulations (Petri, Kubicki et al. 2017). Collectively, these investigations demonstrated that numerous buildings underperform in practice, even when explicitly designed to be energy efficient.

Although the EPG has not attracted widespread public attention, it represents a significant financial concern due to its implications for operational energy costs (Gao, Xu et al. 2024). Evidence indicates that the persistence of the EPG constitutes a major barrier to the successful delivery of LEB (Bai, Yu and Pan 2024). Moreover, while the concept of Zero Carbon Buildings (ZCB) has been widely endorsed and has fostered more advanced design approaches, there remains considerable debate regarding the most effective strategies for achieving intended performance outcomes (Tierney and Tennant 2015). The persistence of this EPG not only complicates the management of building energy supply and demand but also risks eroding trust between engineers and clients as well as fostering hesitancy toward the adoption of emerging technologies (de Wilde 2014, Zou, Xiaoxiao et al. 2018, Loukou, Heiselberg et al. 2019). Addressing the EPG is thus pivotal both for enhancing the credibility of sustainable architecture and for advancing the building sector toward energy efficiency in line with the United Nations Sustainable Development Goals (SDGs) (Janser, Hubbuch and Windlinger 2020, Aste, Adhikari et al. 2022).

The term "gap" is generally defined as a "discrepancy" or "difference" (Shi, Si et al. 2019). However, in the context of the EPG, the interpretation of the term and what constitutes this EPG remains contested, particularly with the proliferation of performance gap typologies as depicted in Figure 1.1 below.

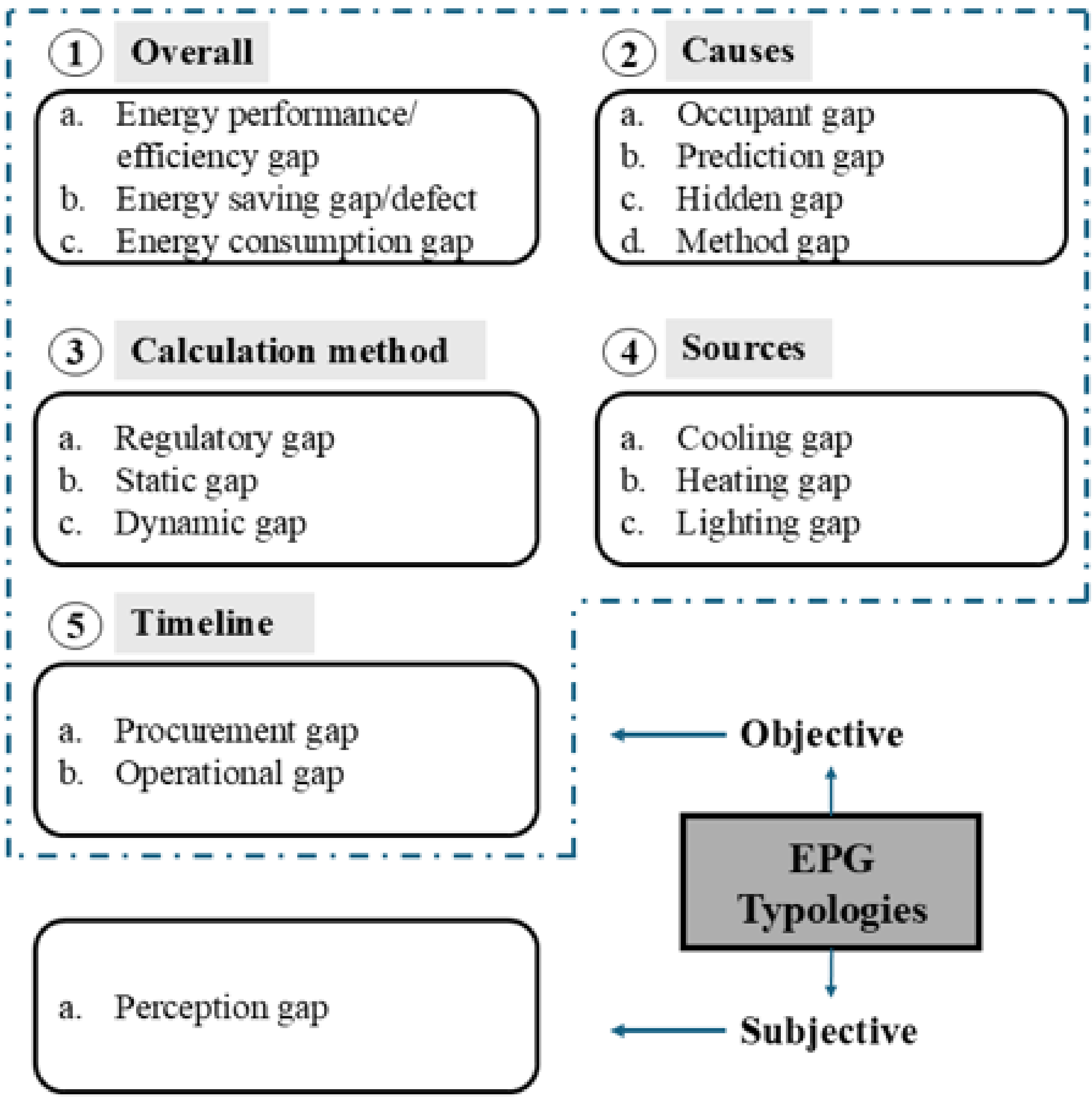


Figure 1.1: Performance Gap Typologies (Bai, Yu and Pan 2024)

As shown in Figure 1.1, EPGs in buildings may be classified as either objective or subjective. The subjective category comprises the perception gap, which relates to occupant satisfaction and describes the divergence between expected and lived experience (Coleman and Robinson

2018). The objective category consists of five interrelated subcategories, each capturing a different dimension of the EPG:

1. **Overall:**

a) **Energy performance (efficiency) gap**: The difference between predicted and actual energy performance of buildings (de Wilde 2014).
b) **Energy saving gap/defect**: The discrepancy between predicted and realised energy savings following interventions (van den Brom, Meijer and Visscher 2020).
c) **Energy consumption gap**: The disparity between expected and actual energy use (Cali, Osterhage et al. 2016).

2. **Causes:**

a) **Occupant gap**: The occupant gap results from user behaviour and is often explained through the notions of rebound and prebound effects which are typically related to interventions such as building retrofit. The rebound effect refers to situations where actual energy use exceeds predicted savings because efficiency gains are offset by occupants raising their comfort expectations, while the prebound effect describes the opposite occurrence, where actual consumption is lower than predicted because occupants deliberately conserve energy (Galvin 2014).
b) **Prediction gap**: Results from inaccuracies in energy modelling, such as simplified assumptions or incorrect input data (Marshall, Fitton et al. 2017).
c) **Hidden gap**: Remains undetected when positive and negative discrepancies offset each other (Salehi, Terim Cavka et al. 2015).
d) **Method gap**: Caused by uncertainties or inconsistencies in calculation and measurement approaches (Wilde and Jones 2014).

3. **Calculation method:**

a) **Regulatory gap**: The discrepancy between compliance-based modelling under standardised conditions and measured energy use (Jradi, Arendt et al. 2018).
b) **Static gap**: The difference between detailed design-stage predictions and actual performance (Jradi, Arendt et al. 2018).

c) **Dynamic gap**: The discrepancy between calibrated models under real operating conditions against measured energy use (Jradi, Arendt et al. 2018).

4. **Sources:**

a) **Cooling gap**: Difference between predicted and actual cooling energy consumption (Palma, Gouveia and Simoes 2019).
b) **Heating gap**: Difference between predicted and actual heating energy demand (Cozza, Chambers and Patel 2020).
c) **Lighting gap**: Difference between predicted and actual lighting energy use (van Someren, Beaman and Shao 2017).

5. **Timeline:**

a) **Procurement gap**: Emerges during construction stages due to procurement decisions and implementation issues (Janser, Hubbuch and Windlinger 2020).
b) **Operational gap**: Occurs during building use due to operation, management, and maintenance practices (Janser, Hubbuch and Windlinger 2020).

This classification brings together several EPG typologies that are often treated separately in the literature. Rather than defining the EPG only by its overall magnitude, source, cause, calculation method, or stage of occurrence, it provides a multidimensional framework for interpreting what a quantified gap represents. For example, a heating-related EPG may have different implications depending on whether it arises from modelling assumptions, operational practices, procurement-stage defects, or the calculation method used. The merit of this classification, compared with more isolated or single-dimensional typologies, is therefore that it supports a more comprehensive and context-sensitive interpretation of the EPG and helps link its measured magnitude to possible mitigation strategies.

1.1 Research Problem

The EPG in buildings persists as an unresolved problem in academia and practice (Chen, Guo et al. 2022). There are significant concerns that the size of this gap cannot be attributed solely to limitations in the accuracy of current energy performance calculation methods (Shi, Si et al. 2019).

Chapter One: Introduction

The current understanding of the EPG, and consequently its broader significance, remains both incomplete and inadequate (Cozza, Chambers et al. 2021). Its implications are far-reaching and detrimental, as the phenomenon embodies substantial complexity in both its magnitude and underlying causes. Advancing knowledge in this area therefore requires more systematic efforts to elucidate the governing principles and intrinsic nature of its magnitude and root causes, with the ultimate goal of establishing generalizable and widely applicable strategies to mitigate the influence of the EPG (Li, Iulo and Poerschke 2023).

One of the pressing issues in the current body of knowledge is that much of the evidence on the magnitude of the EPG is still grounded in static and regulatory perspectives. This is regarded as a key issue in evaluating the EPG and identifying its size and thus its seriousness. Therefore, there is widespread concern regarding the actual magnitude of the EPG. Another layer of complexity stems from the fact such comparisons fail to account for underlying issues in both prediction and measurement. Despite its fundamental importance, there is no definitive answer on the actual magnitude of the EPG at wide scale. There is variability in the ways in which the EPG is defined and reported, and this inconsistency underscores the need for a standardized approach to measuring and presenting the EPG, which would enable clearer comparisons across studies. Furthermore, understanding the scale of EPG and its relationship to variables such as building type, climate, and energy systems remains a central challenge. Progress in this area will depend on adopting uniform methodologies to quantify and report the EPG. It also requires gathering more empirical data and conducting meta-analyses to generate deeper insights (Shi, Si et al. 2019).

The complexity of accurately quantifying the actual extent of the EPG can be realized by observing the number of typologies and only a subset of root causes contributing to its emergence. The adoption of a standardized framework for quantifying and reporting the EPG is therefore widely regarded as essential, since accurate determination of its magnitude is a prerequisite for developing realistic and effective energy policies (Palladino 2023). To date, the extent of the EPG has been shown to be highly contingent upon the baseline selected for its assessment, with accurate identification of its actual size recognized as one of the most critical challenges associated with the phenomenon (Jain, Burman et al. 2020).

Furthermore, the causes that contribute to the emergence of the EPG are acknowledged to be highly complex (de Wilde 2014, Li, Iulo and Poerschke 2023). While accurately identifying the root causes would enable the implementation of effective mitigation measures, research has not yet advanced to the stage where these root causes can be fully identified, managed, or eliminated (Zou, Xiaoxiao et al. 2018). Identifying the factors that contribute to the EPG can yield valuable insights that enhance the reliability of computational tools and strengthen the integrity of modelling practices. Achieving such improvements could generate multiple benefits. First, reducing the magnitude of the EPG would help prevent unrealistic or misleading expectations among stakeholders. Additionally, systematic errors associated with large EPG values may result in suboptimal configuration and sizing of building plants, as these depend heavily on prior load estimations. Furthermore, understanding the common drivers of the EPG can not only improve the accuracy of energy calculation methods and procedures but also inform the design of building features, especially control systems and user interfaces. Finally, deeper investigation of the EPG could incentivize the expansion of knowledge bases and data repositories on occupant behaviour, while also promoting the integration of high-resolution metering infrastructures in both new construction and retrofit projects (Mahdavi and Berger 2024).

On the other hand, research suggests that the precise prediction of a building's performance "may be beyond reach," and that some degree of the EPG should be regarded as an inherent and unavoidable outcome of the complexity of buildings (Mahdavi and Berger 2024). From a single energy criterion, building performance is considered a more complex phenomenon than it is often represented, and thus cannot be meaningfully judged against any simplistic notion of a gap (Patel and Green 2020).

Nonetheless, parts of the literature employ overly optimistic terminology such as "bridging" or "closing" the EPG when proposing solutions centred on the calibration of building energy simulation models. While calibration can improve agreement between simulated and measured data, such approaches often focus on a limited set of adjustable parameters, such as occupancy schedules, infiltration rates, internal gains, or system efficiencies. For example, tuning these parameters may reduce the numerical difference between predicted and measured energy use, but it does not necessarily identify or resolve broader causes of the EPG, such as poor construction quality, commissioning deficiencies, operational practices, or faults in building energy systems (Horrigan, Murphy and Odonnell 2017, Aste, Adhikari et al. 2022, Ozarisoy and Altan 2022). In this sense, claims that calibration can "close" the EPG may overstate what has actually been achieved.

Even where calibration appears to reduce or eliminate the measured gap, research suggests there is a lack of clarity regarding the conditions under which a calibration task can be "terminated" and whether the residual gap may be deemed acceptable (Zhan, Zhu et al. 2025). Similarly, meeting statistical thresholds commonly used under ASHRAE Guideline 14, such as the coefficient of variation of the root mean square error (CVRMSE) and normalised mean bias error (NMBE), does not necessarily ensure predictive reliability across different operating conditions (de Wilde 2014, Sun and Hong 2017). This is important because a model may satisfy calibration criteria while still containing compensating errors, where one incorrect assumption, such as envelope performance, is offset by another adjusted parameter, such as occupancy schedules or internal gains.

These limitations have implications for decision-making. Although numerous studies have claimed to close the EPG through calibration, the reliability of the resulting building energy models may remain uncertain, especially when calibrated models are subsequently used for forecasting, retrofit evaluation, or policy support. For example, a calibrated model that masks compensating errors may lead to inappropriate retrofit decisions, while a model tuned to historical data may perform poorly under future occupancy, weather, or operational conditions (Menezes, Cripps et al. 2012, Fumo 2014). This may create a cascading EPG, in which a model with apparently low predictive error introduces further EPGs in later applications, even when it has been developed using detailed post-occupancy evaluation or monitoring data (Zhan, Zhu et al. 2025). Therefore, calibration should be understood as a useful diagnostic and model-improvement process, rather than as definitive evidence that the EPG has been fully bridged or closed.

The EPG constitutes a regulatory-driven, socio-technical-economic phenomenon in which information, knowledge, technology, behaviour, experience, economic considerations, and policy are deeply interwoven and collectively influential (Shrubsole, Hamilton et al. 2019, Xu and Zou 2020). Therefore, a more rigorous examination of the phenomenon is necessary to highlight both the inherent complexity of the EPG and the conceptual fallacy of reductionist claims.

While technical measures can enhance the accuracy of building performance predictions, effectively managing the EPG remains challenging without mandatory regulatory frameworks and strong communication and coordination among stakeholders (Zheng, Zhou et al. 2024). Energy policy has yet to formally acknowledge the phenomenon, and existing research indicates that the incorporation of strategies to address the phenomenon within policy discourse has received only limited attention (Gram-Hanssen and Georg 2018, McElroy and Rosenow 2019). Therefore, further research is required to deepen understanding of the relationship between existing policies and the EPG, and to explore pathways toward more realistic energy policies and regulations (McElroy and Rosenow 2019). Concurrently, building regulations and policies frequently rely on projected energy performance and are based on assumptions that often do not reflect actual consumption patterns, highlighting the need to mandate transparency in the reporting of building energy performance in operation through outcome-based certification schemes (Cohen and and Bordass 2015). One example of such schemes is the National Australian Built Environment Rating System (NABERS), which is grounded in actual annual energy consumption and is acknowledged to have achieved great success in reducing the EPG between energy targets set at early design stages and measured energy performance (Mirzaie and Menzies 2017).

NABERS was first introduced in Australia as a government-backed initiative to measure and improve the operational performance of buildings, with a particular focus on actual energy consumption. Over time, it became a central tool for driving energy efficiency and transparency in the property sector. Building on its success, NABERS was launched in the UK in 2020 as a performance-based rating framework that assesses the real energy consumption of office buildings. It enables owners to monitor and report their buildings' operational energy performance with accuracy, while also highlighting opportunities for efficiency gains and cost savings. Greater accountability reinforced by mechanisms such as public reporting and contractual obligations in NABERS can serve as a strong driver for narrowing the EPG, exerting a more direct influence on industry practices. NABERS is widely regarded as one of the most effective industry initiatives in aligning predicted and actual building performance; however, discrepancies persist. It is posited that a policy emphasis on actual performance measurement would incentivize the building sector to develop and implement the requisite processes for achieving high operational performance while concurrently enhancing productivity (Tuohy and Murphy 2015). It is recognised that the Soft Landings (SL) framework strengthens collaboration and provides structure throughout procurement and delivery. When paired with performance contracting and performance-based initiatives such as NABERS, it can support meaningful progress toward low-emission outcomes. However, the EPG remains a critical challenge (Mirzaie 2024). Furthermore, it is stated that measurement and verification of in-use energy performance, as outlined in the energy commitment agreement protocol, under NABERS, may contribute to narrowing the EPG (Burman, Kimpian and Mumovic 2018).

NABERS has been widely valued in literature, particularly for its reliance on measured rather than predicted energy performance, which addresses a central limitation of prediction-based approaches that have faced extensive criticism. However, NABERS has not been subject to systematic evaluation in the UK, leaving its impact uncertain, specifically with regard to mitigating the EPG. The evidence available is largely restricted to grey literature. This lack of independent and comprehensive research creates significant uncertainty regarding NABERS's actual impact on mitigating the EPG and underscores the need for critical investigation.

1.2 Aim and Objectives

The discrepancies between as-designed and as-built energy performance are difficult to resolve; yet the gap between as-designed and in-use energy performance represents an even greater challenge, driven by numerous disturbances over the course of the building life cycle (Gupta, Howard and Kotopouleas 2019, Zhan, Zhu et al. 2024). Consequently, addressing the EPG requires not a reductionist view but rather a comprehensive understanding of its nature and the multifaceted causes that contribute to its manifestation across the different stages of a building's lifecycle.

This thesis aims to critically investigate the EPG in buildings, conceptualizing it as a complex and multifaceted phenomenon. It seeks to enhance understanding of its magnitude, assess possibilities for its mitigation, and to advance the development of standardized approaches for its quantification and reporting. Rather than viewing the EPG solely as a discrepancy between simulated and measured energy performance, the thesis frames it as a phenomenon shaped by technical, behavioural, regulatory, and contextual factors. It further considers various manifestations of the EPG that are of relevance to different stakeholder groups, with the objective of deepening understanding of the phenomenon, its severity, and context-specific mitigation strategies aligned with the type of gap which must be mitigated. To achieve this aim, the research pursues the following research objectives (ROs):

- **RO1**: Demonstrate the complexity of the energy performance gap by exploring its root causes and magnitude.

This RO is expected to provide a more systematic identification and structuring of the root causes of the EPG, addressing a gap in the literature where causes remain fragmented and not yet fully resolved. It is further expected to contribute to knowledge by evaluating EPG magnitude as a context-dependent ratio shaped by multiple factors, including those outlined in Figure 1.1 and further explored in Section 3.4. This remains necessary because existing evidence on EPG magnitude is still largely based on static and regulatory perspectives, which may obscure the lifecycle, behavioural, operational, and methodological factors that influence the magnitude of the EPG.

- **RO2**: Explore the potential for mitigating the energy performance gap through approaches that emphasize measured performance outcomes rather than predicted values.

A key expected contribution of this RO lies in repositioning simulation tools as exploratory instruments for evaluating design alternatives and energy-saving measures against predefined energy performance targets, rather than using them primarily to improve prediction accuracy for compliance purposes. The focus on measured performance is motivated by the recognition that the EPG is influenced not only by modelling inaccuracies, but also by factors arising during the design, construction, commissioning, and operational stages. Consequently, improving predictive accuracy alone may not guarantee reliable in-use performance or reduce discrepancies between predicted and actual energy consumption. By emphasising measured performance outcomes, this research promotes an outcome-oriented approach to performance assessment that aligns evaluation with actual building operation and contributes to a broader framework for EPG mitigation beyond model refinement.

- **RO3**: Deepen the understanding of the energy performance gap by clarifying the ambiguity surrounding its magnitude.

The existence of diverse calculation methods for the EPG means that commonly reported EPG ratios are highly sensitive to how the EPG is defined and quantified and therefore cannot be interpreted in isolation. By demonstrating, through the investigation of two key sources of uncertainty in Chapter Five, that even a limited subset of uncertainty sources can significantly influence the magnitude of the EPG, this RO is expected to provide evidence that reported values in the literature may be misleading and can lead to inaccurate conclusions regarding its scale at building, city, or national levels. This establishes a foundation for RO4 by showing why the EPG should be quantified and reported more consistently, enabling more reliable comparisons and a clearer understanding of its trends and implications.

- **RO4**: Assess the potential for standardisation in the quantification and reporting of the energy performance gap, identifying pathways toward greater consistency and comparability.

This RO is anticipated to contribute to knowledge by assessing the potential for greater consistency in the quantification and reporting of the EPG, while acknowledging the inherent challenges of achieving full standardisation across diverse regulatory, climatic, and policy contexts. The value of standardisation lies not in imposing a single universal method, but in improving transparency, comparability, and interpretability across contexts where buildings, regulations, and assessment practices differ. This is important because experiences with building energy assessment, such as EPC implementation across different jurisdictions, demonstrate that shared policy ambitions do not automatically produce harmonised, comparable, or consistent outcomes. By proposing a structured but adaptable framework, this RO supports more reliable cross-study comparisons and opens avenues for future context-sensitive standardisation efforts.

- **RO5**: Develop recommendations for more effective mitigation of the energy performance gap and provide guidance for advancing standardisation efforts.

This RO synthesises the findings from RO1–RO4 to generate practical and research-informed recommendations for mitigating the EPG and improving how it is quantified and reported. Its expected contribution lies in translating the thesis findings into guidance for researchers, designers, policymakers, and building operators, particularly by distinguishing between strategies that improve prediction, strategies that mitigate the EPG, and strategies that improve transparency in reporting. By linking mitigation and standardisation, this RO supports more accountable decision-making and identifies priorities for future research and implementation.

1.3 Thesis Structure

This thesis is organized into eight chapters, which together develop a coherent investigation into the complexity, mitigation, and standardisation of the EPG, as summarised in Figure 1.2. The chapters are sequenced to move from the contextual and methodological foundations of the research, through targeted investigations, to a synthesis of findings and their implications.

1. **Chapter One (Introduction):**

This chapter introduces the EPG in buildings and provides the contextual foundation for the thesis. It outlines the significance of the research problem and presents the research aim, objectives, and overall thesis structure. The chapter also introduces the ROs which collectively guide the investigation.

2. **Chapter Two (Methodology Overview):**

This second chapter sets out the methodological framework underpinning the research. It explains the philosophical stance, research design and methodological choices adopted in the thesis, as well as the approaches to data collection and analysis. The methodological framework presented in this chapter underpins the investigation of all ROs.

3. **Chapter Three (Systematic Literature Review):**

This chapter presents the systematic literature review (SLR), focusing on the root causes of the EPG and the reported magnitude of the gap. It critically examines the technical, behavioural, operational, and systemic factors contributing to the persistence of the EPG, while also highlighting inconsistencies and ambiguities in how its magnitude is reported across previous studies. The chapter directly addresses RO1 by demonstrating the complexity and multidimensional nature of the EPG and provides a foundation for RO3 by identifying ambiguity in reported EPG magnitudes.

4. **Chapter Four (Mitigation Strategies through Policy):**

This chapter investigates how policy interventions can support the mitigation of the EPG. It assesses NABERS in the UK and evaluates its potential role as an outcome-based policy mechanism for reducing the EPG. The chapter examines the strengths and limitations of NABERS within the UK context and considers its broader relevance as a policy-based mitigation strategy. It primarily addresses RO2, while also contributing to RO1 through root cause analysis and to RO5 by informing mitigation recommendations.

5. Chapter five (Complexity Demonstration):

This chapter investigates one subset of the root causes of the EPG to demonstrate why reported values regarding its magnitude should be interpreted with caution. Building on arguments introduced earlier in the thesis, it shows that the complexity of building performance makes precise prediction difficult and suggests that some degree of EPG may be unavoidable. By examining performance against a single energy criterion, the chapter argues that building performance is more complex than is often represented and cannot be meaningfully assessed through simplistic notions of a gap. The chapter directly addresses RO3 by clarifying ambiguities in reported magnitudes and informs RO4 by highlighting the need for a standardised approach to EPG quantification and reporting. It also contributes to RO5 by supporting recommendations for advancing standardisation efforts.

6. Chapter Six (Standardization Effort):

This chapter explores the potential for standardizing the quantification and reporting of the EPG. It builds on the findings of the SLR presented in Chapter Three, which identified inconsistencies in reporting practices, and Chapter Five, which demonstrates the complexity of predicting building performance with precision. Drawing on these insights, the chapter develops a framework for standardized EPG quantification and reporting, which is subsequently refined through expert consultation to improve its relevance and applicability. This chapter directly addresses RO4 and contributes to RO5 by providing guidance for future standardisation efforts.

Each of Chapters Four to Six is structured around a detailed methodology, presentation of results, and focused discussion of findings. This structure ensures that each chapter is analytically rigorous and self-contained, while also contributing to the cumulative argument of the thesis.

7. Chapter Seven (Synthesis of Knowledge):

This chapter synthesizes the findings from the preceding chapters by identifying cross-cutting themes and integrating the main insights generated throughout the thesis. It contributes primarily to RO5 by consolidating knowledge and translating the findings into overarching recommendations for EPG mitigation and standardisation. The chapter also situates the findings within broader theoretical, methodological, and policy contexts, demonstrating how the different strands of the research collectively contribute to an integrated understanding of the EPG.

8. **Chapter Eight (Conclusion):**

This chapter concludes the thesis by summarizing the key findings and consolidating the contributions to knowledge. It reflects on the extent to which RO1–RO5 have been addressed and discusses the theoretical and practical implications of the research. The chapter also identifies the limitations of the thesis and proposes recommendations for future research, thereby providing a foundation for continued investigation of the EPG phenomenon.

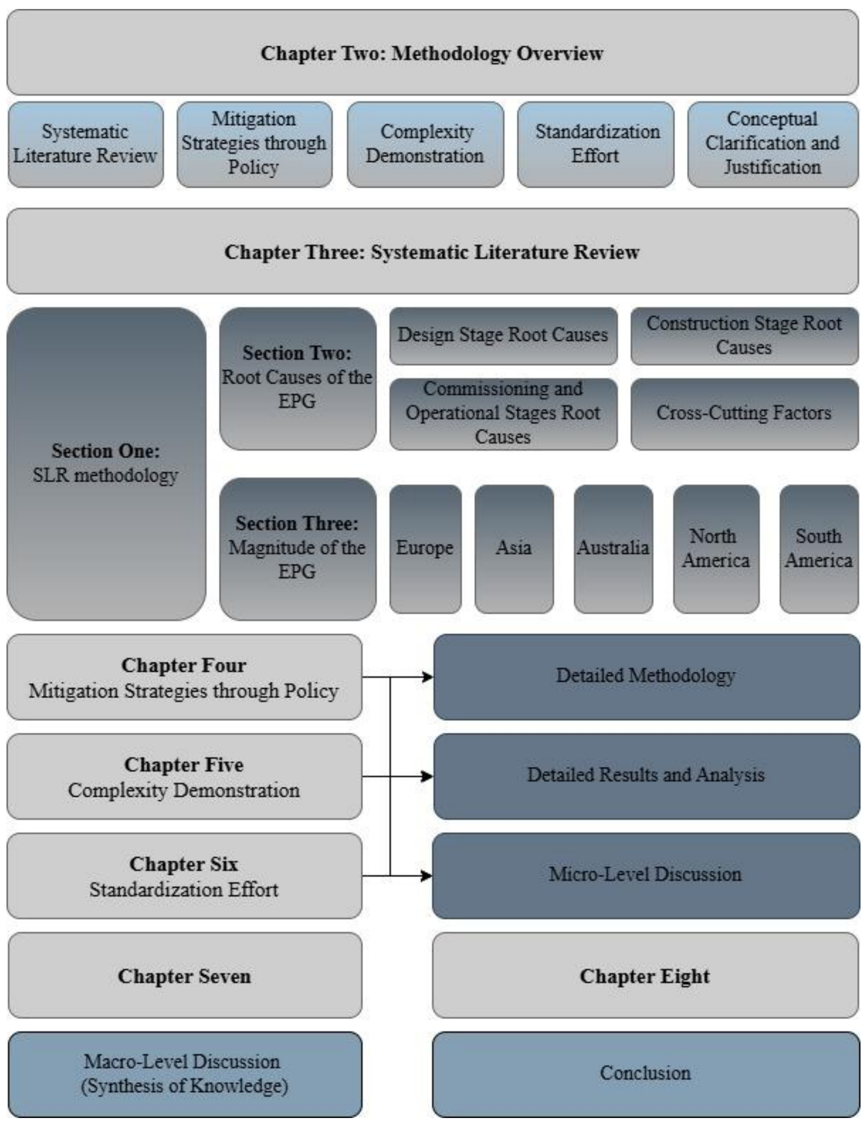


Figure 1.2: Thesis Structure

2. Methodology Overview

The EPG constitutes a regulatory-driven, socio-technical-economic phenomenon in which information, knowledge, technology, behaviour, experience, economic considerations, and policy are deeply interwoven and collectively influential (Shrubsole, Hamilton et al. 2019, Xu and Zou 2020). This thesis critically investigates the EPG in buildings by conceptualising it as a complex and multifaceted phenomenon. It seeks to enhance understanding of its magnitude, assess possibilities for its mitigation, and advance the development of standardized approaches for its quantification and reporting.

A mixed-methods design was therefore adopted to address the inherently complex nature and multidimensional character of the EPG and to support the ROs of the thesis. Mixed-methods research integrates qualitative and quantitative forms of inquiry within a single study, enabling a more comprehensive understanding of complex research problems than either approach can provide alone (Battista and Torre 2023, Sharma, Bidari et al. 2023). The research aim extends to include explanation, interpretation, and policy-relevant insights as the EPG is examined both as a measurable discrepancy between predicted and actual performance and as a phenomenon shaped by design decisions, construction processes, operational practices, policy mechanisms, and reporting conventions. The quantitative components support the analysis of EPG magnitude, modelling uncertainty, and performance variation, while the qualitative components support the interpretation of root causes, policy mechanisms, expert perspectives, and standardisation challenges.

Alternative methodological approaches would have been less suited to the scope of the thesis, as they risked either oversimplifying the phenomenon or limiting the explanatory depth required to address the ROs. A purely quantitative approach, such as one based only on simulation, calibration, or statistical comparison of predicted and measured data, could have supported the analysis of EPG magnitude but would have been insufficient to explain its underlying causes, policy context, and mitigation potential. Conversely, a purely qualitative approach, such as one based only on interviews, document analysis, or expert opinion, could have generated rich explanatory insights but would have lacked the analytical capacity to interrogate variability in reported EPG magnitudes and modelling outcomes. Previous EPG research has often adopted such narrower approaches, including simulation-based and calibration-based studies on the one hand, and qualitative or policy-focused investigations on the other (Zou, Wagle and Alam 2019, Yoo, Shin et al. 2024, Zhan, Zhu et al. 2024). In contrast, a mixed-methods design is more appropriate for complex interdisciplinary ROs because it allows triangulation across evidence sources and supports the integration of measurable, interpretive, and policy-relevant findings (Hongling, Yishu et al. 2025).

The chapter begins by presenting the structure and strategy of the Systematic Literature Review (SLR), which establishes the foundation for the thesis by synthesizing evidence on both the causes of the EPG and its magnitude. This is followed by detailed accounts of the methodological approaches applied across the three following chapters: (1) policy mitigation strategies investigated through semi-structured interviews, (2) complexity demonstration through dynamic simulation modelling, and (3) development of a framework for standardization of quantification and reporting, informed by semi-structured interviews and refined through an expert panel discussion.

2.1 Systematic Literature Review

The SLR presented in the thesis, covering the period 2014–2024, is structured around two interconnected strands that address the root causes and the magnitude of the EPG. An SLR is defined as a review of an existing body of literature that follows a transparent and reproducible methodology for searching, quality assessment and synthesis, with a high level of objectivity (Kraus, Breier and Dasí-Rodríguez 2020). The SLR is suited to address the first RO, which aims to demonstrate the complexity of the EPG by exploring both its root causes and magnitude.

The EPG literature is dispersed across design, construction, commissioning, operation, policy, modelling and post-occupancy evaluation, while the thesis requires a structured basis for clarifying both the underlying causes and the magnitude of the EPG before progressing to subsequent stages in which root cause analysis, semi-structured interviews, comparative analysis associated with simulation and standardisation efforts are conducted. The SLR was therefore not used merely as a background review, but as the evidence base that informs the subsequent investigations of the thesis.

Alternative literature review approaches were considered but were less preferable specifically to the investigation of the magnitude of the EPG. A narrative review can provide a broad qualitative overview of a subject, but it is less structured than an SLR and is not required to follow stringent parameters such as explicit sources, inclusion and exclusion criteria or a defined analytical framework, making it more vulnerable to selection bias (Juntunen and Lehenkari 2019). A scoping review would have been useful for mapping the breadth of available literature, but it is less suited to critical assessment and synthesis of findings, where SLRs have an advantage when the aim is to draw conclusions and implications rather than only describe the scope of evidence (Munn, Peters et al. 2018, Khatri, Dabas et al. 2025). Finally, a meta-analysis was not adopted because it requires comparable quantitative measures across studies, whereas the EPG literature uses heterogeneous definitions, baselines, building types, modelling assumptions and reporting metrics; such variability limits the usefulness of statistical aggregation (West and Martin 2023). Accordingly, the SLR offered the most appropriate balance between breadth, methodological rigour, transparency, and analytical depth for addressing RO1 while also supporting the later ROs.

The SLR was sourced from three major academic databases widely recognized in engineering and built environment research: *Web of Science, Scopus, and Engineering Village*. Predefined search strings were applied, and inclusion criteria limited the review to peer-reviewed journal articles, conference papers and selected technical reports published between 2014 and 2024. Only studies directly addressing either the causes or the magnitude of the EPG were included. In addition to identifying root causes and synthesizing magnitudes, the review systematically maps the typologies of the EPG, as outlined in Chapter One (Figure 1.1). This ensures that subsequent empirical work is grounded in a nuanced understanding of how the EPG is variably defined, reported, and contested in the literature. The proposed approach under sections 2.1.1 and 2.1.2 seeks to balance exhaustive and saturation-based review methods with practical efficiency, recognising that, depending on the aims of the SLR, extending the search beyond what is methodologically necessary may offer limited additional value.

2.1.1 Root Causes of the EPG

The first part of the SLR identifies and categorises the diverse drivers of the EPG across design, construction, commissioning, and operational stages. To ensure comprehensive coverage without redundancy, a saturation approach was applied (Saunders, Sim et al. 2018). Records were analysed year by year (2014–2025), with key and seminal contributions retained until no new categories of causes emerged. This process resulted in 15 records, with 2 additional studies identified through forward citation chasing snowballing, bringing the total to 17 cause-focused records. The resulting taxonomy establishes the foundation for Chapter 4, where these causes are interrogated through a 5 Whys RCA and subsequently discussed in semi-structured interviews.

2.1.2 Magnitude of the EPG

The second strand of the SLR synthesises evidence on the magnitude of the EPG across different geographical, climatic, and building contexts and across different interpretations of the EPG consistent with the typologies presented under Figure 1.1. In contrast to the root causes strand, this component adopts an exhaustive review strategy also known as exhaustive coverage or comprehensive searching. This approach aims to identify and include all relevant studies that meet the defined eligibility criteria, thereby ensuring comprehensive coverage and minimising the risk of omission (Templier and Paré 2015). Such completeness is important for enabling robust comparison, aggregation and understanding of the magnitude of the EPG across different contexts.

The magnitude of the EPG varies substantially across studies depending on building type, geographical context, modelling approach, performance baseline, energy-use category, and method of calculation. A selective or narrative sampling strategy would overrepresent particular contexts or definitions of the EPG and therefore produce a distorted understanding of its magnitude.

This review highlights inconsistencies in reporting and challenges in comparability, providing the foundation for Chapter Five, where the complexity of quantifying the EPG is demonstrated through a case study using four dynamic simulation tools. Insights from this review also contribute directly to Chapter Six, which develops a preliminary framework for standardization by drawing on observed inconsistencies in reporting and the complexity revealed through simulation.

2.2 Mitigation Strategies through Policy

The chapter investigates the potential of outcome-based policies, with a focus on NABERS, to mitigate the EPG. To achieve this, the chapter adopts two methodological approaches: RCA and semi-structured interviews with key stakeholders.

2.2.1 Root Cause Analysis

The taxonomy of causes identified in Chapter Three is further interrogated using the "5 Whys" root cause analysis (RCA) technique to examine how causes emerge and interact across building life-cycle stages. RCA is a structured problem-solving approach intended to move beyond visible symptoms to identify the mechanisms that generate or sustain a problem (Al-Rifai 2025). In this thesis, RCA is applied following the SLR, which identifies and categorises factors contributing to the EPG according to the stage at which they emerge. This additional analytical step is necessary because some identified factors may represent intermediate symptoms rather than underlying causes. Accordingly, the "5 Whys" technique is employed as a structured analytical aid to support the identification of deeper underlying EPG drivers, rather than assuming that every EPG issue can be reduced to a single linear root cause.

The "5 Whys" technique was selected because it provides a transparent and systematic means of tracing causal relationships through iterative questioning (Al-Rifai 2025). This supports RO1 and RO2 by enabling investigation into why the EPG persists, rather than merely cataloguing contributing factors, while also supporting the exploration of mitigation strategies targeted at identified root causes. The method is also suited to the subsequent semi-structured interviews, because participants can explain why specific EPG drivers arise and propose solutions aimed at preventing underlying causes rather than merely addressing symptoms.

Observable discrepancies in building performance frequently arise from complex chains of interrelated factors spanning design, construction, commissioning, and operation. By applying the "5 Whys" technique, this research systematically interrogates deeper drivers of the EPG to support the development of more effective and targeted mitigation strategies. This is important because interventions focused solely on intermediate symptoms may fail to address the structural and systemic conditions that give rise to the EPG. Alternative analytical approaches, such as cause categorisation and visual mapping techniques, are useful for organising contributory factors and illustrating relationships between causes. However, previous research suggests that some RCA tools, including cause-and-effect diagrams and interrelationship diagrams, may have limited ability to identify causal interdependencies or rigorously interrogate deeper causal structures (Doggett 2005). In contrast, the "5 Whys" RCA supports deeper exploration of causal pathways and investigation into how and why EPG drivers emerge and interact, thereby aligning with of RO1, RO2, and RO5.

2.2.2 Semi-Structured Interviews

Interviews were conducted with two stakeholder groups: NABERS Independent Design Reviewers (IDRs) and Accredited Assessors (AAs). The selection of these two groups provided targeted insights into how NABERS mechanisms, such as commitment agreements, independent design reviews, and accredited assessments address EPG root causes identified through the "5 Whys" RCA.

Semi-structured interviews are used as a qualitative method in which the researcher follows a prepared guide or framework of themes while retaining flexibility to ask follow-up questions, probe explanations, and explore issues that emerge during the conversation (Ruslin, Mashuri et al. 2022). This approach enables the exploration of expert perspectives while maintaining comparability across respondents. This stage of the research required expert interpretation rather than broad measurement as the aim was not simply to confirm whether NABERS mechanisms address predefined root causes, but to understand how IDRs and AAs perceive the capacity of NABERS to address design-stage, commissioning, operational, and cross-cutting root causes of the EPG. It also enabled participants to identify practical measures for addressing root causes that may not be fully addressed by NABERS.

The semi-structured format was preferred over structured interviews or questionnaires because it allowed interviewees' viewpoints to be expressed in a more open and context-sensitive manner than standardised instruments, while still maintaining focus on the predefined themes (Ruslin, Mashuri et al. 2022). It was also preferred over unstructured interviews because the research required consistency and comparability across participants while remaining flexible enough to probe causal reasoning and practical experience.

The interviews were analysed using a colour-coded matrix to systematically organise and code responses, enabling the identification of patterns and relationships across stakeholder groups. This approach supported transparent interpretation of the qualitative data and facilitates comparison with findings from other components of the research. A colour-coded matrix was considered appropriate because the interviews were targeted and theme-driven, and the purpose was to compare expert interpretations against predefined RCA categories and NABERS mechanisms rather than to generate an entirely new grounded theory from the data.

The analysis of NABERS is situated within the broader policy–practice gap identified in Chapter One, where reliance on predicted performance in regulation has proven inadequate. By integrating insights from RCA and semi-structured interviews, the research connects underlying causes with real world policy mechanisms, thereby supporting the development of more comprehensive and practically grounded mitigation strategies. This directly supports RO2 and contributes to RO5 by informing recommendations for more effective mitigation strategies.

2.3 Complexity Demonstration

This chapter critiques the reliability of reported EPG magnitudes in literature by demonstrating the inherent complexity of predicting building performance. This responds directly to the ambiguities identified in the SLR (Chapter Three), where reported values of the EPG varied widely due to differences in defining the EPG, prediction and measurement approaches. This strand of the research directly addresses RO3, which seeks to deepen the understanding of the EPG by clarifying the ambiguity surrounding its magnitude. While these inconsistencies highlight the difficulty of comparison across studies, they also raise a deeper concern whether accurate prediction of building performance is, in fact, achievable. The following methods are adopted:

2.3.1 Dynamic Simulation Modelling

Dynamic building energy simulation models are developed for the Energy and Building Design Laboratory at Lund University. Four widely used simulation tools are employed (IES-VE, ClimateStudio, DesignBuilder, and IDA-ICE). Using the same building as a case study across all tools with identical input data and geometry enables a controlled investigation into how simulation scenarios and tool-specific features contribute to variability in predicted outcomes.

The requirements of RO3 influenced the selection of dynamic building energy simulation tools because the aim is to examine how modelling assumptions, simulation scenarios, and tool-specific calculation procedures influence the estimated magnitude of the EPG, rather than simply to optimise predictive accuracy. Dynamic simulation represents the time-varying interactions between building fabric, systems, internal gains, occupancy, weather conditions, and operational schedules. This directly supports RO3, which focuses on clarifying the ambiguity surrounding

EPG magnitude and demonstrating why reported gap values may vary depending on how prediction and comparison are undertaken.

White-box models are appropriate for this purpose because they are based on physical principles, including heat transfer thermodynamics, and building-system operation. Their model structure allows the relationship between inputs, assumptions, and outputs to be examined more explicitly than in data-driven approaches (Hamdaoui, Benzaama et al. 2021). They also do not require historical operational data, which makes them suitable for controlled scenario testing and for examining how alternative modelling assumptions affect predicted performance before or independently of calibration (Yu, Chang and Dong 2022). This transparency is essential for a multi-tool comparison across the four tools employed, where the same building geometry and input assumptions are applied across different dynamic simulation environments to investigate how modelling choices contribute to uncertainty in EPG quantification. This is particularly relevant because even established white-box tools can differ in their calculation procedures and predictive performance for specific buildings (Mazzeo, Matera et al. 2020).

Data-driven or black-box models are less suitable for addressing RO3 because they derive statistical relationships from historical datasets. While they can be useful where detailed building information is unavailable, they require sufficient, high-quality and representative data, missing data, data errors, data variability, and non-standardised data types can reduce accuracy and limit transferability. They are also less appropriate for causal interpretation because their internal relationships are not explicitly explained through physical heat-transfer processes (Manfren, James and Tronchin 2022). Grey-box models are also considered less preferable for this stage because, although they combine simplified physical structure with empirical or data-driven elements and can improve computational efficiency, their development methods remain less standardised, and they lack the widely adopted software environments associated with established white-box simulation tools (Yu, Chang and Dong 2022).

2.3.2 Comparative Analysis

Inter-model comparison has long been used in building performance simulation as a way of examining how different simulation programs represent the same physical problem, most notably through the Building Energy Simulation Test (BESTEST) methodology and its later incorporation into the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE)

Standard 140. ASHRAE Standard 140 includes comparative tests in which simulation results are compared under carefully specified test conditions; these exercises are intended to increase confidence in simulation tools and to reveal differences arising from algorithms, input interpretation, modelling assumptions, and implementation choices (ASHRAE 2020). However, BESTEST and ASHRAE Standard 140 are based on standardised test cases rather than measured real-building case studies.

In Chapter Five, this comparative logic is adopted for a different purpose. The case study does not constitute a formal BESTEST or ASHRAE Standard 140 validation exercise, nor are the models calibrated using measured performance data. Instead, the case study building is represented in four dynamic simulation tools using architectural drawings, building-services documentation, harmonised input assumptions, and common weather data. The resulting outputs are first compared against each other through an inter-tool gap investigation and are then benchmarked against measured electricity consumption from the case study building to examine the measured-versus-simulated EPG. This dual comparison provides empirical evidence of methodological variability in model outputs and illustrates how different interpretations of the EPG for the same building can emerge depending on the simulation tool, modelling workflow, and scenario assumptions. Where comparable monthly outputs are available, additional diagnostic indicators such as MBE and CVRMSE are used to examine whether apparent annual agreement masks temporal discrepancies. However, the analysis remains an applied comparative case study rather than a software validation exercise, and the results are therefore interpreted as evidence of modelling uncertainty in this specific case rather than as a general assessment of any tool's validity.

This strand of the thesis deepens the understanding gained from the SLR on EPG magnitude conducted in Chapter Three. While the SLR identified inconsistencies in how the EPG is defined, measured, and reported across the literature, Chapter Five provides empirical evidence explaining why such inconsistencies can arise and why simplistic gap figures may be unreliable. By demonstrating how two uncertainty dimensions, namely inter-tool variability and measured-to-simulated variability, can substantially affect the resulting gap, the investigation shows that apparently simple EPG figures can be highly sensitive to the procedures through which they are produced. The purpose is therefore not to establish which simulation tool is most accurate, but to demonstrate how methodological variability affects the quantification and reporting of the EPG.

Building on insights from Chapters Three and Five, the findings strengthen the case for standardisation by exposing both the variability of reported values and the uncertainties embedded in modelling practices. In doing so, Chapter Five provides critical evidence and justification for Chapter Six, where the potential for standardising the quantification and reporting of the EPG is explored through the development of a framework and its refinement through discussions with key experts in the field.

2.4 Standardization Efforts

This chapter develops and refines a framework for standardizing the quantification and reporting of the EPG. This strand moves beyond diagnosis and critique to construct a solution-oriented response to the inconsistencies identified in the literature and the methodological challenges exposed through simulation modelling. The objective is to create a structured, transparent, and transferable framework that can improve both academic comparability and practical reliability in reporting the EPG.

- **Preliminary Framework Development:** The framework is initially constructed by synthesizing insights from the SLR (Chapter Three) and the complexity demonstration (Chapter 5). The literature review highlighted major inconsistencies in the way the EPG has been defined and reported globally, while Chapter 5 demonstrated empirically why modelling-based EPG figures are inherently variable. Together, these findings establish the conceptual basis for a standardized approach, ensuring that the framework directly responds to inconsistencies and uncertainties.

- **Expert Engagement through Semi-Structured Interviews:** To test and refine the preliminary framework, semi-structured interviews are conducted with three building energy performance experts, selected for their diverse perspectives across research, practice, and policy. These interviews provide critical insights into the practical viability of the proposed framework, highlighting potential limitations, contextual challenges, and areas where greater clarity or flexibility is needed.

- **Expert Panel:** Following the interviews, the framework is further examined through a panel discussion with internationally recognised simulation specialists. This step subjects the framework to rigorous collective scrutiny by leading figures in the field and facilitates dialogue on its transferability across contexts and its capacity to address a long-standing issue in building performance assessment. While the use of an expert panel introduces a potential risk of convergence toward shared assumptions or what might be termed group thinking, this risk is mitigated by the diverse professional backgrounds and methodological perspectives represented among the participants. As a result, the panel functions not as a mechanism for consensus for its own sake, but as a means of testing the robustness, clarity, and applicability of the proposed framework.

The outcome of this chapter is a final framework for standardized quantification and reporting of the EPG. The framework addresses academic needs by providing a consistent basis for comparison across studies, while also responding to industry and policy priorities by offering a transparent and outcome-oriented structure that can be applied in practice. By building on the inconsistencies identified in Chapter Three and the modelling complexities demonstrated in Chapter 5, and refining these insights through expert engagement, Chapter Six establishes a constructive pathway toward greater reliability, comparability, and accountability in addressing the EPG.

2.5 Integration of Methods

Although each strand employs distinct methods, they are deliberately interconnected and sequenced to form a cumulative research design. The SLR establishes the foundation by clarifying the contested causes and magnitudes of the EPG. Building on this, the policy-based analysis tests the potential of outcome-based frameworks to mitigate these causes, integrating both policy mechanisms and practitioner insights. The simulation-based modelling then interrogates the reliability of reported magnitudes, exposing the inherent complexity of predicting building performance and reinforcing the need for standardized approaches. Finally, the standardization framework synthesizes these insights into constructive solutions, refined and validated through expert engagement.

Together, these methods ensure both depth within each strand and coherence across the thesis, enabling the research to progress logically from diagnosis, through critique, to solution-building. Importantly, the integration of methods allows the thesis to address the EPG not only as a technical issue, but as a socio-technical-economic phenomenon. Thus, the research provides a holistic investigation of the EPG and advances knowledge in a way that is academically rigorous and practically relevant.

The methodology presented in this chapter provides a high-level overview of the research design and the integration of its constituent strands. Figure 2.1 maps the ROs against the methodological approaches, specific focus, and chapters, illustrating how each strand contributes to the achievement of the overall aim of the thesis. Detailed methodological accounts are then provided in the respective chapters. Chapter Three explains the SLR strategy in full, including database selection, search protocols, and inclusion/exclusion criteria. Chapter 4 sets out the procedures for RCA and semi-structured interviews, clarifying sampling, interview design, and analytical approach. Chapter 5 outlines the development of the simulation models, the case study selection, and the comparative analysis of inter-tool and measured–simulated EPGs. Finally, Chapter Six details the process of developing, refining, and validating the standardization framework, including semi-structured interviews and the expert panel discussion. Together, these chapters provide the necessary methodological depth to support the research design summarized here.

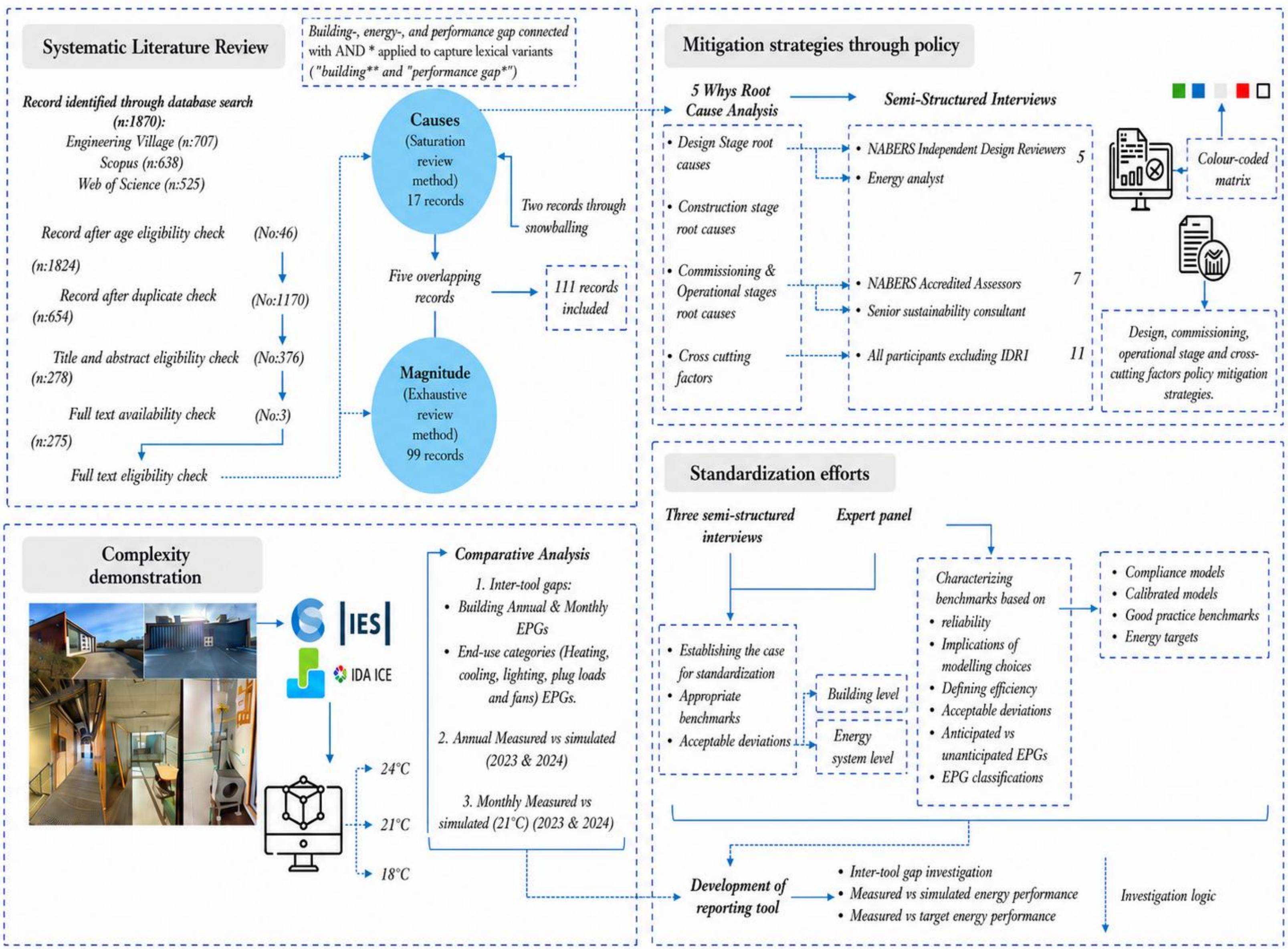


Figure 2.1: Methodological overview of the thesis research design

2.6 Conceptual Clarification and Justification

NABERS was developed in the late 1990s in Australia in response to growing recognition that improvements in building design were not consistently translating into improved operational performance. Initially introduced as the Building Greenhouse Rating in 1999, the scheme evolved into a nationally standardised framework for assessing the in-use environmental performance of buildings, with a particular emphasis on energy efficiency. Over time, NABERS expanded in both scope and geographic reach, incorporating additional performance domains such as water, waste, and indoor environmental quality, and extending beyond Australia to contexts such as New Zealand and the United Kingdom. Today, NABERS represents one of the most mature performance-based rating systems globally, with extensive uptake across the commercial building sector and demonstrated long-term reductions in energy use and emissions associated with repeated ratings (Bannister and Flores 2024).

Given the centrality of policy in this thesis, it is important to clarify the terminology used throughout, as well as the role of NABERS within this landscape. NABERS is formally understood as an environmental rating system rather than a policy in itself. However, it becomes policy-relevant when it is embedded within, or used to support, wider governance mechanisms such as disclosure requirements, benchmarking, contractual commitments, performance targets, and market-based incentives. This thesis therefore treats NABERS as a standardised outcome-based rating system through which the potential of outcome-based policy can be examined, rather than as a standalone policy instrument.

NABERS evaluates buildings using a standardised star rating scale, ranging from one to six stars, based on measured operational performance derived from independently verified metered data (Bannister and Flores 2024). This enables comparison of building performance across a national dataset, supporting the identification of inefficiencies and opportunities for improvement. The system assesses multiple dimensions, including energy use and associated greenhouse gas emissions. Crucially, these data are normalised to account for contextual factors such as climate, occupancy, and building use, ensuring that ratings reflect efficiency rather than external conditions. The requirement for annual reassessment creates a continuous performance feedback loop, while the visibility of performance data enhances transparency, which may influence market behaviour and rental values. In this sense, NABERS operates not only as an outcome-based certification system but also as a market-enabling mechanism linking performance outcomes to economic incentives.

Within the broader landscape of building environmental assessment initiatives, NABERS differs from systems such as Leadership in Energy and Environmental Design (LEED), Building Research Establishment Environmental Assessment Method (BREEAM), and Green Star in the type of performance evidence it prioritises. LEED, BREEAM and Green Star are points-based green building rating systems that assess multiple sustainability categories, including energy, indoor environmental quality, materials, water and emissions, although their structures, weighting systems and certification pathways differ across schemes (Doan, Ghaffarianhoseini et al. 2017). For the purposes of this methodology, the key distinction is that many conventional certification pathways are substantially based on design-stage documentation, predicted performance or asset-based assessment. NABERS, by contrast, is selected because it is centred on post-occupancy performance and uses measured operational data to rate actual building outcomes. This makes it particularly relevant to the investigation of the EPG, as the method adopted in this thesis requires an assessment based on measured outcomes. Accordingly, NABERS is adopted in this thesis as an outcome-based methodological case because its assessment structure aligns directly with the thesis focus on measured operational performance, while prior comparative work distinguishes NABERS from other major rating systems by its reliance on post-occupancy performance data (Gui and Gou 2020).

While alternative operational rating systems exist, such as the U.S. ENERGY STAR Portfolio Manager (ESPM), NABERS was selected as the primary analytical lens in this research due to its stronger alignment with verified operational performance assessment, market integration, and explicit focus on ongoing building operation. Both ESPM and NABERS use measured energy data; however, NABERS is structured as a formal annual rating scheme centred on assessed post-occupancy performance, whereas ESPM functions primarily as a benchmarking and tracking platform. This distinction makes NABERS more suitable for this thesis because the research is concerned not only with measured outcomes, but also with how operational performance is assessed, communicated, and linked to management practices in ways that can support mitigation of the EPG. Although ESPM provides a valuable operational benchmarking framework, peer-reviewed research has highlighted limitations in the accuracy and explanatory capacity of its underlying calculation method, leading to proposals for enhanced approaches (Arjunan, Poolla and Miller 2020).

This distinction between NABERS as a rating system and its role within policy is therefore critical. Policy design literature recognises benchmarking, disclosure mechanisms, and performance measurement systems as key tools through which policy objectives are operationalised (Howlett 2019). Accordingly, NABERS is used in this research not as a prescriptive solution, but as an evaluative lens through which the potential of outcome-based policy approaches can be explored. Root causes of the EPG are first identified through the SLR and structured through RCA, and then held constant as a predefined analytical framework. Semi-structured interviews are subsequently conducted to evaluate the extent to which NABERS addresses each identified root cause, and where gaps are identified, participants are asked to propose potential interventions based on their professional experience.

The resulting data were analysed through a combined qualitative and quantitative approach, whereby explanatory narratives were captured through manual coding, and evaluative judgements represented through a colour coded matrix. The overall analytical process linking predefined root causes, participant evaluations, and the assessment of interpretive saturation within the predefined analytical framework is synthesised in Figure 4.1. In this way, the thesis moves beyond descriptive assessment to generate evidence-based and context-sensitive mitigation strategies, grounded in both established literature and the experiential knowledge of practitioners directly involved in NABERS and related projects.

To ensure conceptual clarity and consistency in terminology throughout the thesis, Table 2.2 provides definitions and examples of key terms, including policy, regulation, rating systems, standards, and norms. These distinctions are maintained consistently in the analysis to avoid conflating different types of governance mechanisms.

Table 2.1: Key governance terms relevant to building energy performance

Term	Definition	Examples
Policy	A system of principles, decisions, and actions adopted by governments or organisations to achieve specific objectives or to address identified societal problems (Howlett 2019).	• Climate Change Act in the United Kingdom, which establishes legally binding carbon reduction targets (Rosenow, Cowart and Thomas 2017). • The UK Clean Growth Strategy, which sets out policy ambitions for improving energy efficiency and reducing emissions across the built environment (Department for Business and Industrial 2017).
Regulation	Legally binding rules established by governmental authorities to control behaviour and ensure compliance with policy objectives (Baldwin, Cave and Lodge 2012).	• Part L of the Building Regulations in England, which sets legally binding requirements for the conservation of fuel and power in buildings and requires energy performance calculations to demonstrate compliance (Pan and Garmston 2012). • Energy Performance of Buildings Directive in the European Union, which sets a framework for improving building energy efficiency (European 2018).
Standards	Documented and consensus-based technical specifications or guidelines that provide agreed methods, criteria, or best practices and are typically developed by recognised organisations (Brunsson and Jacobsson 2000).	• ASHRAE standards for ventilation and thermal comfort (ASHRAE 2021). • CIBSE TM54 which provides guidance for evaluating operational energy performance in buildings (CIBSE 2013).
Rating systems/ certification schemes	Third-party, often voluntary assessment frameworks that evaluate and communicate building performance against defined criteria, using standardised methods, benchmarks, scoring systems or rating scales (Marchi, Antonini and Politi 2021).	• NABERS, which rates buildings based on measured operational performance using a standardised star-rating scale. • BREEAM and LEED, which assess broader sustainability performance through points-based certification systems. • ENERGY STAR Portfolio Manager, which benchmarks operational energy performance.
Norms	Informal and socially shared expectations that influence behaviour within professional or cultural contexts without being legally enforced (Bicchieri 2006).	• Professional expectations that building energy modellers should use transparent and defensible assumptions for occupancy, internal gains, HVAC schedules, weather files and system operation when producing simulation outputs. • Industry norms around post-occupancy evaluation and operational energy monitoring, where design teams are increasingly expected to compare predicted and measured performance even where no legal requirement exists.

3. Systematic Literature Review

This chapter presents the systematic literature review (SLR) undertaken to investigate the causes of the EPG and its magnitude across global contexts. The review is organized into five sections. Section 3.1 outlines the SLR methodology, including the search strategy, databases, inclusion and exclusion criteria, and analytical approach. Section 3.2 explores the causes of the EPG, structured around the building lifecycle stages: design (3.2.1), construction (3.2.2), and commissioning and operation (3.2.3). Cross-cutting factors are also explored as systemic drivers that influence multiple stages under (3.2.4). Section 3.3 reviews the magnitude of the EPG across different regions: Europe (3.3.1), Australia (3.3.2), North America (3.3.3), South America (3.3.4) and Asia (3.3.5). Section 3.4 provides a summary of the SLR findings and key takeaways.

3.1 SLR Methodology

The SLR was undertaken to investigate the causes that contribute to the emergence of the EPG and its magnitude across global contexts. In line with best practice, the process adhered to the Preferred Reporting Items for Systematic Reviews and Meta Analyses (PRISMA) guidelines to enhance transparency (Page, McKenzie et al. 2021). PRISMA was selected as the reporting framework because it provides a recognised structure for documenting the stages through which records are identified, screened, assessed for eligibility and included in the review. PRISMA is not a single fixed review method, nor does it prescribe one type of evidence base (Page, McKenzie et al. 2021). Rather, it supports transparent reporting of systematic reviews with different scopes and source strategies, including reviews based only on academic databases and reviews that incorporate grey literature where this is relevant to the research question. PRISMA-S further emphasises the importance of reporting search strategies, information sources and search decisions in sufficient detail to support transparency and reproducibility (Rethlefsen, Kirtley et al. 2021).

The review proceeded primarily through peer-reviewed academic databases because the aim was to establish a robust academic evidence base on the causes of the EPG and how it has been conceptualised, measured and explained. This was important because the thesis uses the SLR to construct the root-cause taxonomy and to synthesise reported magnitudes before progressing to the later methodological stages. However, a strictly academic-only review would have been too restrictive because the EPG is also shaped by policy programmes, industry initiatives, technical guidance and international energy-efficiency organisations. Grey literature can improve the comprehensiveness of systematic reviews and reduce the risk of publication bias, particularly where relevant evidence is produced outside peer-reviewed journals, although it also introduces challenges relating to source quality, search reproducibility and consistency of appraisal (Mahood, Van Eerd and Irvin 2014, Paez 2017). For this reason, the review adopted a controlled and selective version of PRISMA rather than an unrestricted grey-literature search.

Accordingly, the SLR was based mainly on peer-reviewed academic sources, complemented by limited grey literature where such material was directly relevant, methodologically transparent and capable of providing consolidated evidence not readily available in journal articles. This decision reflects a pragmatic balance between academic rigour and thematic completeness. For example, one of the two additional records identified through snowballing was included because it synthesised evidence on the causes that contribute to the emergence of the EPG and classified them according to the stage at which they emerge (Hinge 2019). The inclusion of this source strengthened the review while avoiding the reduced reproducibility and potentially uneven quality that could result from a broad, open-ended grey-literature search. Therefore, the version of PRISMA adopted in this thesis is best understood as a systematic academic review with targeted grey-literature supplementation, selected to support RO1 while maintaining transparency.

It is acknowledged that several approaches are thematically relevant to the EPG because they also address the relationship between intended and actual building performance. These include NABERS, PH, SL and Post-Occupancy Evaluation (POE). NABERS is particularly relevant to operational energy performance because it is an outcome-based rating scheme grounded in measured building performance. PH seeks to minimise performance shortfalls through rigorous design, construction quality assurance and verification. SL and POE focus on handover, operational feedback, occupant use and the learning processes needed to close gaps between design intent and in-use performance.

Although these initiatives were not treated as separate review domains within the SLR, their relevance was not excluded from the thesis. In particular, evidence relating to SL, PH and NABERS was considered where it appeared within the EPG literature and was further explored through the investigation of reported EPG magnitudes in Section 3.3. This enabled relevant studies and projects captured by the EPG-centred search strategy to be examined as part of the wider assessment of how the EPG is quantified and reported across different contexts. The decision not to treat these initiatives as separate search strands was therefore intentional rather than an omission. The purpose of the SLR was narrower: to identify and synthesise literature that explicitly explains the causes contributing to the emergence of the EPG and investigates its reported magnitude, rather than to conduct a wider review of all performance-assurance and building-evaluation frameworks.

Including full literatures on NABERS, PH, SL and POE would have broadened the review substantially and risked reducing the focus and reproducibility of the SLR. Instead, the review adopted a focused EPG-centred search strategy, based mainly on peer-reviewed academic literature and complemented by targeted grey literature where directly relevant. This approach allowed the review to maintain academic rigour while acknowledging the practical and industry-based contexts in which the EPG is often investigated. Nevertheless, not treating these initiatives as separate review domains is recognised as a methodological boundary at the point of method selection.

3.1.1 SLR Results

The PRISMA 2020 checklist was used as a reporting aid to ensure that the review process was described transparently and that the reader can follow how records were identified, screened, assessed and included. The PRISMA 2020 statement comprises a checklist of items and sub-items, supported by an expanded checklist that provides more detailed reporting recommendations for each item (Page, McKenzie et al. 2021, PRISMA 2024). The checklist was not treated as a rigid methodological protocol, but as a structure for verifying that the relevant elements of the review were reported clearly. The rationale and objectives of the SLR are introduced at the beginning of this chapter; the eligibility criteria, databases, search string, date range and language restrictions are reported within this section; the identification, screening and inclusion process is summarised in Figure 3.1; and the approach to extracting, recalculating and normalising EPG values is explained in Section 3.1.2.

Some PRISMA checklist items were not directly applicable because the review did not conduct a statistical meta-analysis or assess intervention effects. Accordingly, items relating to effect measures, certainty assessment, synthesis of effect estimates and intervention-specific risk of bias were not applied. Instead, rigour was pursued through transparent search reporting, duplicate removal, explicit inclusion criteria, documented screening stages, consistent recalculation of EPG ratios, and the use of two review strands suited to the thesis objectives: saturation-based synthesis for root causes and exhaustive extraction for EPG magnitude. A summary mapping of the PRISMA 2020 checklist items to the corresponding thesis sections is provided in Appendix A to make explicit how the relevant checklist items were addressed, and where non-applicable items were identified and justified.

The SLR covered the period from 2014 to 2024 and was conducted across three major academic databases: *Web of Science (n=525), Scopus (n=638), and Engineering Village (n=707),* selected for their complementary coverage of engineering, energy and built environment research (Mongeon and Paul-Hus 2016, Gusenbauer and Haddaway 2020, Schultz and McAllister 2024). Using multiple complementary databases improves recall and reduces the risk of disciplinary or regional coverage bias (Bramer, Rethlefsen et al. 2017, Martín-Martín, Orduna-Malea et al. 2018).

A deliberately simple but targeted search string was used, combining the terms *building*, *energy*, and *performance gap* connected with AND. Asterisks were applied to capture lexical variants ("*building*" and *"performance gap*"). A language filter limited results to English, and where records were flagged as *"other languages",* but the full text was in English, they were included. The initial search retrieved 1,870 records in total.

All records were imported into EndNote for management and screening, following the PRISMA stages of identification, screening, eligibility and inclusion as depicted in Figure 3.1. The PRISMA flow diagram summarises this process. Two distinct review methods were then applied to reflect the dual focus of the SLR:

- Causes of the EPG (Saturation review method): Many records discussed causes contributing to the EPG. To ensure comprehensive coverage without redundancy, a saturation approach was applied (Saunders, Sim et al. 2018). Records were analysed year by year from 2014 to 2024, with key and seminal contributions retained until no new categories of causes emerged. This process resulted in 15 records, with two additional studies identified through forward citation chasing, also referred to as snowballing, bringing the total to 17 cause-focused records.
- Magnitude of the EPG: Reported magnitudes varied widely and inconsistently across studies. To capture the full breadth of evidence, an exhaustive review was conducted, including all eligible records that provided quantitative data on the magnitude of the EPG. Out of the 275

records assessed for this strand, 99 records met these criteria, allowing a global synthesis across geographical and regulatory contexts.

An overlap of five records between the two strands of the SLR was identified, leaving a final total of 111 unique records included in the synthesis.

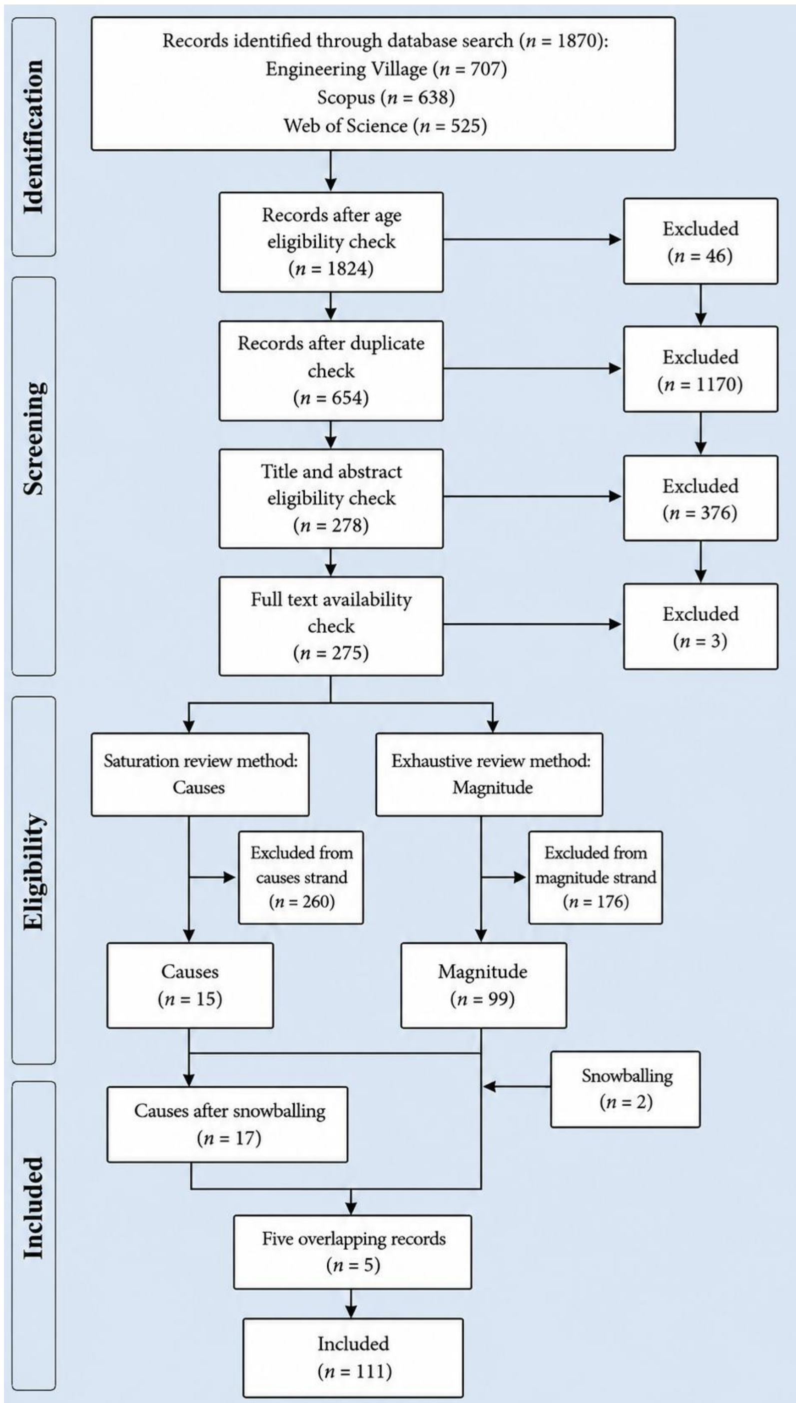


Figure 3.1: PRISMA Diagram (Page, McKenzie et al. 2021)

3.1.2 Quantification and Normalization

For the second strand of the SLR, EPG values were not extracted in their original form because the literature uses many different definitions and calculation methods. Instead, only the measured and simulated energy data were collected, and a consistent framework was applied to recalculate the EPG for all studies. In this framework, the factor, which is the ratio of measured energy performance to simulated, predicted or target performance, was calculated for each EPG using *equations* 3.1, 3.2 and 3.3. Where a factor is greater than one, this indicates measured energy exceeds simulated or target values wherein *equation 3.2* was used. For over performance (factor less than one), where measured consumption falls below predictions, *equation 3.3* was used.

The measured-to-simulated factor was first calculated to determine whether the measured energy use was higher or lower than the simulated value:

$$\text{Factor} = \frac{Measured}{Simulated} \qquad \text{(Eq. 3.1)}$$

where M is the measured energy use and S is the simulated energy use.

Where the factor was greater than one, the EPG was calculated as the percentage by which measured energy use exceeded simulated energy use:

$$\text{EPG} = \frac{Measured\text{ - }Simulated}{Simulated} \times 100 \qquad \text{(Eq. 3.2)}$$

Where the factor was below one, the result was expressed as the measured energy use as a percentage of the simulated energy use:

$$\text{Measured-to-simulated percentage} = \frac{Meaured}{Simulated} \times 100 \qquad \text{(Eq. 3.3)}$$

For illustration, a factor of 1.2 indicates that a building consumed 20% more energy than predicted, an unfavourable outcome, whereas a factor of 0.80 indicates that a building consumed 80% of predicted energy, indicating a favourable outcome of 20% below expected demand. To address methodological variation across studies, the following rules were applied:

1. Where measured outcomes were compared against policy or standard benchmarks such as PH thresholds, the benchmark value was treated as simulated.
2. Where simulated data is only available for compliance model, calibrated or in use models with no measured data, the simulated values from calibrated or in-use models are treated as measured and the value from compliance models as simulated to demonstrate the potential of calibration on mitigating discrepancies.
3. For consistency, all factors reported in tables within *section 3.3* were rounded to one decimal place, and percentages derived from *equation 3.2* and *equation 3.3* and used in the narrative were also rounded to one decimal place.

The second strand of the SLR proved especially demanding because each record required close reading to determine how the EPG had been defined, what was being compared, and whether the reported value could be recalculated consistently using the framework adopted in this thesis. In many cases, studies reported only final percentage values, ratios or performance comparisons, while using definitions or equations that differed from those presented above. This meant that reported values could not simply be extracted and compared directly; instead, each ratio had to be identified where possible and re-expressed using a consistent basis.

For example, studies relating to PH were not always directly comparable because the EPG could be based on different reference points, including SAP calculations, typical or good-practice benchmarks, PH requirements, or Passive House Planning Package (PHPP) outputs. As a result, the review had to determine whether EACH comparison represented a regulatory, a PH benchmark-based gap, or another form of benchmark-based comparison. In addition, in some records, measured operational data were not available, and the comparison was instead made between compliance models and calibrated or in-use models. In such cases, the calibrated or in-use model output was treated, as a pragmatic classification rule, as the closer representation of measured performance, while the compliance model was treated as the simulated or target value.

Furthermore, the process required careful examination of the parameters underlying both sides of each comparison and, where possible, recalculation of values to improve consistency across Section 3.3. The process was therefore time-intensive not because of data volume alone, but because apparently similar EPG values often referred to different baselines, modelling assumptions, benchmarks and calculation procedures.

3.2 Underlying Causes of the EPG

This section synthesises the causes of the EPG identified in the reviewed literature and organises them according to the stage at which they most commonly emerge: design (Section 3.2.1), construction (Section 3.2.2), commissioning and operation (Section 3.2.3), together with cross-cutting factors (Section 3.2.4) that affect multiple stages. The purpose of this section is not to conduct the RCA developed later in Chapter Four, but to establish the literature-derived cause groups that inform that subsequent analysis. To improve clarity, a concise discussion of how these causes contribute to the EPG is presented followed by a summary table for each section that consolidates the main causes reported in the literature.

3.2.1 Design stage root causes

Design-stage causes relate to the way performance intentions are defined, modelled and translated into specifications before construction begins. A first group of causes concerns communication and target-setting. Several studies identify unclear or poorly communicated performance targets as a source of later divergence between predicted and actual performance (de Wilde 2014, Jradi, Arendt et al. 2018, Liang, Qiu and Hu 2019, Eon, Breadsell et al. 2020). These communication problems are important because they shape modelling assumptions, design priorities and the extent to which energy performance is integrated with comfort, indoor air quality and wider project objectives.

A second group of causes concerns the design itself. The literature indicates that both inefficient design and poorly buildable efficient design can contribute to the EPG. Poor thermal concepts, oversized HVAC systems and inadequate detailing can embed performance weaknesses directly, while excessive complexity, weak buildability, poor sequencing and insufficient simplicity can make otherwise efficient designs difficult to deliver in practice (de Wilde 2014, van Dronkelaar, Dowson et al. 2016, Zou, Xiaoxiao et al. 2018, Zou, Wagle and Alam 2019). In this sense, design-stage causes are not limited to poor energy concepts; they also include the practical ability of the design to be understood, constructed, commissioned and operated. Evidence from Eon, Breadsell et al. (2020) further shows that inadequate construction experience among design teams, unclear documentation, unaddressed site constraints and missing details can force on-site improvisations, increasing the risk that design assumptions are not realised in the constructed building.

A third group relates to uncertainty in future use and system performance. The reviewed studies show that the EPG can emerge when future functions, operational requirements, occupancy, plug loads, internal gains, construction materials and stochastic factors are inaccurately represented during design (de Wilde 2014, Jradi, Arendt et al. 2018, Zou, Alam and Hebert 2019). Over-optimistic assumptions about system efficiency and technological performance are also repeatedly identified, particularly where robustness, end-user requirements or functional changes are insufficiently considered (Zou, Xiaoxiao et al. 2018). Padey, Goulouti et al. (2021) further illustrate this issue through the overestimation of solar and internal gains and the use of normative values that do not reflect real building conditions.

Energy modelling and simulation form a fourth design-stage cause group. The literature shows that prediction uncertainty is not only a matter of tool accuracy, but also of modelling practice, input assumptions, analyst expertise and the scenarios selected for testing. Reported issues include tool misuse, inappropriate methods, limited analyst knowledge, lack of formal error testing, weak audit trails and limited robustness testing under alternative scenarios of use and occupancy (de Wilde 2014). van Dronkelaar, Dowson et al. (2016) provide a useful taxonomy of uncertainty, distinguishing specification uncertainty, modelling uncertainty, numerical uncertainty, scenario uncertainty and heuristic uncertainty. These uncertainties relate to incomplete or inaccurate input data, modelling simplifications, computational choices, assumptions about weather or occupancy, and human judgement. Other studies reinforce these issues by identifying inappropriate assumptions, inaccurate design parameters, tool uncertainty, oversimplified fabric assumptions, limited simulation expertise, inaccuracies in occupant-behaviour modelling, improper façade modelling and mistaken inputs or settings (Jradi, Arendt et al. 2018, Zou, Xiaoxiao et al. 2018, Zou, Alam and Hebert 2019, Zou, Wagle and Alam 2019, Cozza, Chambers et al. 2021, Zare, Saryazdi et al. 2023, Mahdavi and Berger 2024).

Several studies stress the role of climate and micro-environmental conditions. Shi, Si et al. (2019) shows that reliance on standard climate files fails to reflect local conditions such as urban heat islands, which significantly increase energy consumption. Mahdavi and Berger (2024) extends this to contextual settings, including adjacent buildings, airflow patterns, and greenery, all of which can diverge from assumptions. Miller, Sodagar et al. (2021) criticizes the persistent use of historical climate files and average conditions, highlighting again the role of urban context. Cozza, Chambers et al. (2021) and Zare, Saryazdi et al. (2023) both emphasize the use of unrepresentative climate files, often derived from airport weather stations distant from urban centres, while Zheng, Zhou et al. (2024) explicitly lists weather data as a weak link at the design stage.

Finally, incomplete information and weak integration also contribute to design-stage EPGs. Missing details, incomplete datasets, poor model inputs, unspecified operational assumptions, unclear instructions for operation and inaccurate material, geometry or system data can reduce the reliability of predictions (de Wilde 2014, van Dronkelaar, Dowson et al. 2016, Zou, Xiaoxiao et al. 2018, Hinge 2019). Eon, Breadsell et al. (2020) further shows that inadequate information and unclear integration of fabric and services can result in incorrect specifications, design errors and later construction-stage faults. These issues demonstrate that design-stage causes are not isolated technical problems, but are closely connected to communication, documentation and the transfer of information into construction and operation.

3.2.2 Construction stage root causes

Construction-stage causes concern the translation of design intent into the as-built building. The literature shows that the EPG can emerge when design intent is compromised by construction uncertainties, malpractice, procurement decisions, value engineering, change orders, cost-first procurement, supply constraints, substitutions or on-the-job solutions (de Wilde 2014, Shi, Si et al. 2019, Eon, Breadsell et al. 2020, Mahdavi and Berger 2024). Poor workmanship and construction quality form a second major cause group. Studies repeatedly identify poor workmanship, cut corners, improper construction techniques and inadequate construction quality as contributors to discrepancies between predicted and actual energy performance (Zou, Xiaoxiao et al. 2018, Zou, Alam and Hebert 2019). These problems often materialise through specific defects, including incorrect or missing installation in external walls, partitions, openings, floors and roofs (Eon, Breadsell et al. 2020). Envelope and fabric defects are important because poor envelope assembly, improper envelope construction, gaps in insulation, thermal bridges, air leakage and vapour barrier failures directly undermine thermal performance (de Wilde 2014, Jradi, Arendt et al. 2018, Eon, Breadsell et al. 2020, Zare, Saryazdi et al. 2023, Zheng, Zhou et al. 2024).

Material, component and equipment quality also affect construction-stage performance. The literature identifies poor-quality components, inappropriate or substandard materials and economically driven substitutions as causes of underperformance (Jradi, Arendt et al. 2018, Zou, Xiaoxiao et al. 2018, Zou, Alam and Hebert 2019). Eon, Breadsell et al. (2020) adds that low-quality insulating materials may degrade over time, creating a temporal dimension in which construction-stage decisions continue to affect operational performance after handover.

A further group of construction-stage causes concerns knowledge, skills and management practices. Limited experience and knowledge, poor site management, deficient communication, insufficient inspection, limited contractor skills and inadequate attention to execution capability all reduce the likelihood that design intent will be realised on site (Zou, Xiaoxiao et al. 2018, Eon, Breadsell et al. 2020, Miller, Sodagar et al. 2021, Zare, Saryazdi et al. 2023). Finally, time and budget pressures can intensify construction-stage risks. Time pressure is associated with incomplete work and rushed commissioning, while economically motivated decisions and cost- or programme-driven changes can lead to substitutions that negatively affect energy outcomes (Jradi, Arendt et al. 2018, Zou, Alam and Hebert 2019, Eon, Breadsell et al. 2020).

3.2.3 Commissioning and operational stages root causes

de Wilde (2014) highlights the weaknesses of handover processes, arguing for continuous commissioning, while noting that buildings are often completed without appropriate tests. Zou, Alam and Hebert (2019) reinforces this by identifying incomplete commissioning as a recurrent shortcoming. Similarly, van Dronkelaar, Dowson et al. (2016) identifies poor commissioning as a distinct category of failure, while Zou, Wagle and Alam (2019) points to the lack of post-construction testing as a persistent issue.

More detailed evidence is offered by Eon, Breadsell et al. (2020), which estimates that poor installation and commissioning can increase building energy use by up to 20%. Low-carbon technologies such as heat pumps, solar thermal systems, and condensing boilers are found to be particularly vulnerable to poor installation practices, incorrect parameter settings, and oversizing. Skill gaps among installers and sequencing problems, where building services are added after fabric construction leading to thermal bridges and air leakage, exacerbate these problems. The same study also stresses the rarity and unreliability of post-construction verification, citing inconsistent testing protocols, poorly calibrated equipment, and low-quality tools as reasons why results cannot be trusted. In parallel, Zou, Xiaoxiao et al. (2018) and emphasize that full performance testing is often curtailed due to time and budget constraints, while the latter also reports that designers are frequently disengaged from commissioning stage.

Once buildings enter operation, occupant behaviour is repeatedly identified as a dominant driver of the EPG. de Wilde (2014) attributes mismatches in plug loads and internal gains to unanticipated occupant activity, compounded by higher-than-expected IT-related loads and operational practices such as thermostat settings, operation hours, and building energy management system (BEMS) configurations. A broad review in Zou, Xiaoxiao et al. (2018) details how personal and social characteristics, interaction among occupants, attitudes, experience, and comfort expectations influence energy outcomes, with rebound effects and limited knowledge of system controls further widening the EPG. Zou, Alam and Hebert (2019) echoes this through evidence of complex user behaviour, while Jradi, Arendt et al. (2018) cites occupant behaviour, activities, and knowledge as sources of underperformance. Similarly, Zare, Saryazdi et al. (2023) shows how window openings, shadings, setpoint adjustments, and cultural or lifestyle preferences alter actual energy consumption, while Liang, Qiu and Hu (2019) identifies both higher-than-designed occupancy levels and higher-than-expected consumption by occupants as common factors. Hinge (2019) consolidates these findings, pointing to a range of occupant-related drivers including attitudes, social interactions, comfort, rebound effects, and the inability to control systems.

Beyond occupants, technical performance and control issues contribute significantly to operational gaps. Zou, Xiaoxiao et al. (2018) records that controls often do not work as predicted, a finding echoed by Zare, Saryazdi et al. (2023), which highlights malfunctioning equipment, failures of energy systems, poorly tuned controls, and strategies that fail in practice. van Dronkelaar, Dowson et al. (2016) and Cozza, Chambers et al. (2021) identify malfunctioning equipment as a major driver, alongside poor operational practices. Liang, Qiu and Hu (2019) cites technology failure, particularly of energy-efficient systems, as a common cause of underperformance. Evidence from Jradi, Arendt et al. (2018) adds detail, noting faulty systems and components, variation in operational modes, and inappropriate building management and control strategies, while Zheng, Zhou et al. (2024) also flags general system performance failures.

The skills and practices of building operators are another source of divergence. Zou, Alam and Hebert (2019) highlights the shortage of knowledgeable and skilled facility managers (FMs) and the absence of ongoing monitoring, while Jradi, Arendt et al. (2018) documents inappropriate management and control strategies, poor maintenance practices, and weak service regimes. Zheng, Zhou et al. (2024) explicitly names poor management as an operational cause. Hinge (2019) reinforces this, pointing to lack of fine-tuning in the first years of operation, inadequate training of occupants, and failures in standardized monitoring practices.

Weaknesses in monitoring and measurement infrastructure also play a substantial role. de Wilde (2014) highlights uncertainty in sensor data, gaps in measurement science and the absence of standardized monitoring procedures throughout the building life cycle. Mahdavi and Berger (2024) expands this, pointing to the limited coverage and resolution of both occupant monitoring and metering systems, as well as errors in documenting and interpreting collected data. Padey, Goulouti et al. (2021) similarly identifies errors in measurement systems and incorrect configurations of technical systems. Faulty, poor or poorly installed sensors and meters are reported in Jradi, Arendt et al. (2018) as one of the sources of discrepancies.

Several sources emphasize that building performance is not static but evolves over time. van Dronkelaar, Dowson et al. (2016) refers to longitudinal variability in operation, where changes in use, equipment, or occupant behaviour lead to energy consumption drifting away from original predictions. Zou, Alam and Hebert (2019) similarly highlights changes in space usage patterns and varying indoor and outdoor conditions as contributors to operational discrepancies. This temporal variability complicates attempts to measure or close the EPG, as performance is constantly in flux.

3.2.4 Cross-cutting factors

Cross-cutting factors are not confined to a single life-cycle stage but influence how causes arise, persist and interact across design, construction, commissioning and operation. A first cross-cutting factor is industry culture. The literature describes a construction culture resistant to change, where traditional practices persist and problems of quality, integrity and responsibility recur (de Wilde 2014). The same paper further cautions that regulatory pressure to demonstrate ambitious efficiency outcomes can encourage over-optimistic predictions, while the absence of routine POE processes removes a key feedback mechanism for learning and improvement.

A second cross-cutting factor is lack of accountability. The literature repeatedly shows that when responsibility for performance outcomes is fragmented across clients, designers, contractors, operators and occupants, no single actor is clearly accountable for whole-building energy performance (Beu 2014, Zou, Alam and Hebert 2019, Zou, Wagle and Alam 2019, Eon, Breadsell et al. 2020, Zheng, Zhou et al. 2024). illustrates this problem by showing how responsibility may be divided into separate packages of work, meaning that overall building quality becomes no single actor's responsibility. Without a main contractor or project manager focused on energy outcomes, faults may be overlooked, quality control may become episodic, and responsibility can ultimately shift to occupants through higher-than-expected energy bills.

Zou, Alam and Hebert (2019) highlights poor communication and collaboration between stakeholders and a lack of feedback loops. While Zheng, Zhou et al. (2024) records similar communication failures and adds that conflicts of interest further impede collaborative practices. Eon, Breadsell et al. (2020) provides detailed evidence, showing that decisions made during construction are often not communicated back to design teams, project documentation is fragmented at the team rather than project level, and collaboration is obstructed by a lack of shared interests or integrated delivery methods. Zou, Wagle and Alam (2019) identifies insufficient information transfer as a recurring challenge. Detailed evidence is provided by Eon, Breadsell et al. (2020), which shows that information is inconsistently recorded, inadequately shared, or misinterpreted between stakeholders, causing performance drift between design intent and as-built reality.The role of misaligned incentives is highlighted in Zheng, Zhou et al. (2024), which reports conflicts of interest as a barrier to mitigating the EPG. Similarly, Eon, Breadsell et al. (2020) shows that commercial and contractual arrangements often prioritize cost over energy performance, weakening the incentives to resolve issues that contribute to the EPG.

Knowledge deficits and skills shortages run across the delivery chain. Zou, Wagle and Alam (2019) lists lack of knowledge and experience as central contributors to the gap, while de Wilde (2014) also links poor client knowledge to underperformance. Eon, Breadsell et al. (2020) reinforces this view, noting that insufficient experience with energy-efficient technologies amplifies the effects of poor communication, weak accountability, and fragmented project delivery. Finally, knowledge and skills deficits affect multiple actors and stages. Limited expertise among clients, designers, contractors, installers, operators and occupants can amplify performance risks across the life cycle (de Wilde 2014; Zou, Wagle and Alam 2019; Eon, Breadsell et al. 2020). These deficits are cross-cutting because they influence how performance targets are set, how designs are developed, how construction is managed, how systems are commissioned and how buildings are operated.

3.3 Magnitude of the EPG

This section examines the magnitude of the EPG across different geographical contexts, with the aim of uncovering patterns in how it is defined, calculated, and reported. The geographical distribution of the eligible records included in the analysis is illustrated in Figure 3.2, which highlights the global scope of the evidence base, spanning five continents and 27 countries. While the regional structure of this section follows broad geographical groupings, the analysis is not intended as a purely geographical comparison. Instead, it focuses on identifying how methodological choices and contextual factors influence reported EPG magnitudes. To support this objective, the analysis draws on a dataset comprising 714 calculated EPG factors derived from the reviewed studies. These factors capture how EPG magnitudes vary according to construction period, building rating, calculation method, ventilation strategy, prediction method, simulation scenario, and occupancy patterns.

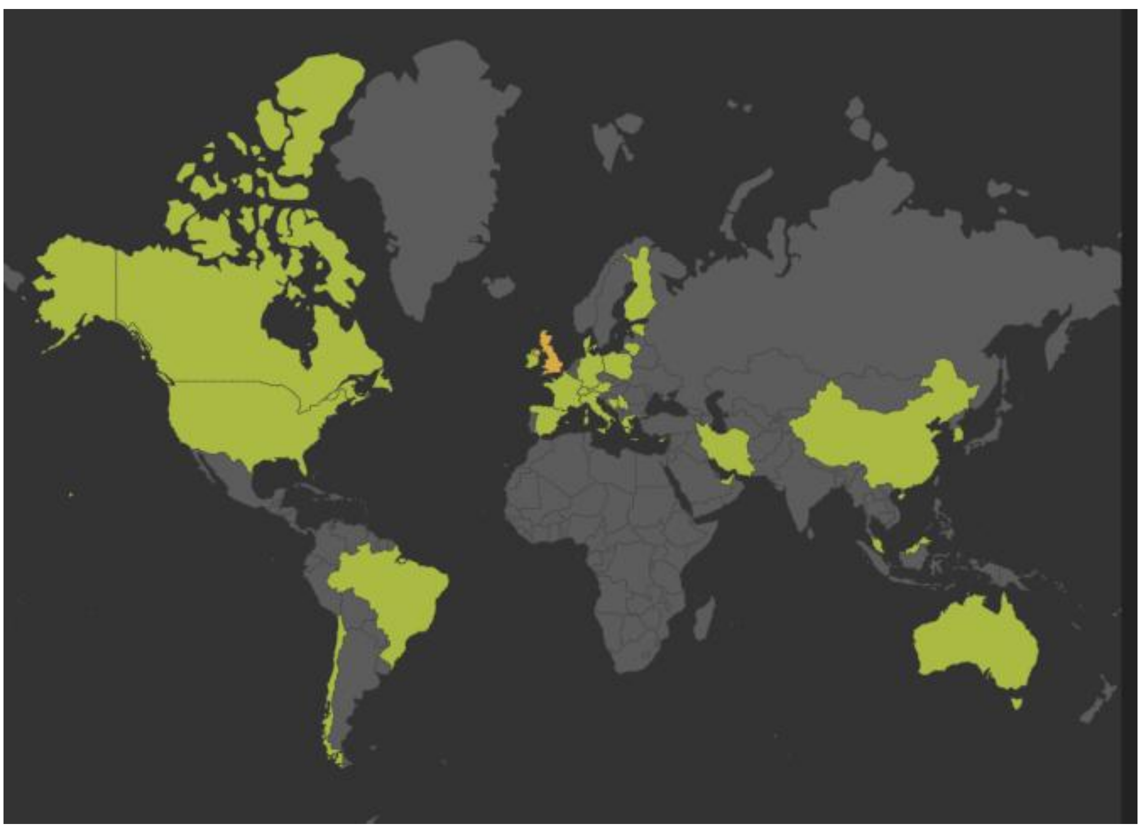


Figure 3.2: Geographical coverage of the SLR, showing studies from 27 countries across five continents

3.3.1 Europe

Several studies have highlighted significant EPGs in residential buildings in the UK, as summarized in Table 3.1 . Jones, Fuertes and De Wilde (2015) analysed six flats, reporting the measured gas consumption was 59% of SAP predictions, while electricity consumption was 104% higher than predicted, indicating a substantial mismatch. The same sample was subsequently compared with simulated energy performance, where gas consumption represented 66% of the simulated values and electricity consumption exceeded simulations by 25%. In as similar study, Kane, Firth et al. (2015) examined ten homes and observed mean EPGs equivalent to 58% and 92% of SAP predictions for gas and electricity, respectively. Bros-Williamson, Stinson and Currie (2015) compared a conventional dwelling with a PH dwelling. The conventional building consumed 42% more gas for hot water and heating than predicted by SAP, while its electricity consumption for lighting, fans, pumps, and appliances was only 57% of a typical benchmark for semi-detached houses. The PH dwelling, by contrast, consumed 37% more gas than SAP predictions and 6% more electricity than the benchmark.

Gupta and Kapsali (2016) examined two dwellings and found electricity consumption to be 80% higher than expected when measured against an extended SAP methodology that incorporated unregulated loads. Expanding on this work, Gupta, Howard and Kotopouleas (2019) investigated 12 PH and 50 non-PH dwellings, reporting mean heating EPGs of 161% and 91% above SAP predictions for PH and non-PH samples respectively. Jones, Xiaojun et al. (2020) identified an initial heating EPG of 170% in a SOLCER house during the heating season, which was reduced to 18% following calibration. For a single week in December, the heating EPG remained 9% above the simulated value. Similarly, Jain, Burman et al. (2020) observed extreme EPGs in an apartment block, with electricity and gas consumption exceeding simulated values by 472% and 177% respectively. However, when compared to good practice energy benchmarks, electricity and gas consumption were lower by 20% and 55%, indicating more favourable outcomes. Mitchell and Natarajan (2020) studied 97 PH dwellings and reported an average heating EPG of 92% of predicted energy performance. Nevertheless, the results varied considerably depending on the benchmark used. When measured energy was compared with PH targets, most buildings exhibited more favourable EPGs than when measured against predictions from the PH simulation tool.

Miller, Sodagar et al. (2021) analyzed four single-family homes and reported water heating EPGs ranging from 13% below predictions to 216% above SAP calculations. For heating, the discrepancies ranged from 28% below to 80% above predicted values. Bahadori-Jahromi, Salem et al. (2022) investigated seven family houses and demonstrated that adjusting heating set points reduced heating EPGs from an initial range of 27–32% above predictions to 19–22%. Similarly, modifying heating schedules reduced initial heating EPGs from 27–36% to 18–20%.

At a larger scale, Few, Manouseli et al. (2023) reported that dwellings rated A or B consumed more energy than predicted by SAP, whereas those rated C to G consumed less. For example, C-rated dwellings consumed 92% of predicted energy, while F- and G-rated dwellings consumed only 52%. Duncan, Hayles and Littlewood (2024) examined affordable housing across six units monitored from November 2021 to November 2022. All units showed substantial deviations from DER predictions, with performance gap factors ranging between 1.14 and 3.15, underscoring inconsistencies in energy modelling for low-income housing. Finally, Firth, Allinson and Watson (2024) conducted a national-scale study of residential buildings in England and Wales, comparing Reduced Data Standard Assessment Procedure (RdSAP)-predicted gas usage with actual meter readings. The ratio of predictions to measurements ranged from 0.44 in Blaenau Gwent to 1.24 in the City of London, demonstrating clear regional disparities in prediction accuracy.

Table 3.1: Analysed records for residential buildings in the UK

References	Building Type	Period	Baseline & Unit	Details		Factor	Measured / predicted
Jones, Fuertes and De Wilde (2015)	Apartment / flat (Six flats)	Annual Nov 2013–Oct 2014	SAP (kWh)	Gas	Measured mean for six flats against predicted	0.59	2,170 / 3,701
				Electricity	Measured mean for six flats against predicted	2.04	2,332 / 1,142
			DesignBuilder (kWh)	Gas	Measured mean against mean of seven models	0.66	2,170 / 3,278
				Electricity	Measured mean against mean of seven models	1.25	2,332 / 1,873
Kane, Firth et al. (2015)	Detached and semi-detached houses (10 dwellings)	Monthly Mar-2014	SAP (kWh)	Gas (Space and hot-water heating)	Mean	0.58	1,833 / 3,182
					Min	0.43	702 / 1,603
					Max	0.64	3,202 / 5,001
				Electricity (Lighting and appliances)	Mean	0.92	367 / 401
					Min	0.77	511 / 658
					Max	1.25	244 / 195
Bros-Williamson, Stinson and Currie (2015)	Conventional semi-detached house built to 2010 Scottish building regulations (One dwelling)	Annual	SAP ($kWh \cdot m^{-2} \cdot a^{-1}$)	Gas (DHW, heating)	Measured against SAP design value	1.42	85 / 60
			Typical UK benchmark for semi-detached dwellings ($kWh \cdot m^{-2} \cdot a^{-1}$)	Electricity (Lighting, fans, pumps and appliances)	Measured against UK semi-detached benchmark	0.57	27.3 / 48
	Semi-detached PH (One dwelling)		SAP ($kWh \cdot m^{-2} \cdot a^{-1}$)	Gas (DHW, heating)	Measured against SAP design value	1.37	63 / 46
			Typical UK benchmark for semi-detached dwellings ($kWh \cdot m^{-2} \cdot a^{-1}$)	Electricity (Lighting, fans, pumps and appliances)	Measured against UK semi-detached benchmark	1.06	51 / 48
Gupta and Kapsali (2016)	End-terrace and mid-terrace low-carbon social houses (Two dwellings)	Annual Jan–Dec 2013	Extended SAP (includes unregulated loads) ($kWh \cdot m^{-2} \cdot a^{-1}$)	Electricity	Actual annual energy use against design prediction	1.8	— / —
Gupta, Howard and Kotopouleas (2019)	PH: Flats, houses and bungalows (12 dwellings)	Annual 2010–2014	SAP ($kWh \cdot m^{-2} \cdot a^{-1}$)	Space heating	Mean	2.61	23.0 / 8.8
					Min	0.91	4.8 / 5.3
					Max	2.73	50.3 / 18.4
	Non-PH low-energy dwellings: Fats and houses (50 dwellings)				Mean	1.91	58.3 / 30.5
					Min	0.48	3.3 / 6.8
					Max	2.72	175.0 / 64.2
Jones, Xiaojun et al. (2020)	Energy-positive detached SOLCER house (1 dwelling)	Heating season Oct-Mar 2016	HTB2 simulation tool (kWh)	Heating	Baseline model	2.71	4,732 / 1,749
					Calibrated (Accounting for loft leakage and secondary causes)	1.18	4,732 / 3,991
		20–26 Dec 2015			Calibrated	1.10	189 / 172

Table continues next page

Continued from previous page

References	Building Type		Period	Baseline & Unit	Details		Factor	Measured / Predicted
Jain, Burman et al. (2020)	Apartment block (98 flats across two buildings)		Annual	DesignBuilder / compliance modelling ($kWh \cdot m^{-2} \cdot a^{-1}$)	Electricity	Actual against simulated (Equipment energy excluded)	5.7	35.5 / 6.2
					Gas	Actual against simulated	2.77	113.4 / 40.8
				CIBSE Guide F good-practice benchmark ($kWh \cdot m^{-2} \cdot a^{-1}$)	Electricity	Actual against good-practice benchmark	0.81	35.5 / 44.0
					Gas		0.46	113.4 / 247.0
Mitchell and Natarajan (2020)	97 PH dwellings	Flats and houses (97 dwellings across 13 sites)	1–3 years, depending on site	PHPP prediction ($kWh \cdot m^{-2} \cdot a^{-1}$)	Space heating	Mean measured against mean PHPP prediction	0.92	10.8 / 11.7
				PH target (15 $kWh \cdot m^{-2} \cdot a^{-1}$)		Mean measured against PH target	0.72	10.8 / 15
		Flats (21 dwellings)		PHPP prediction ($kWh \cdot m^{-2} \cdot a^{-1}$)		Mean measured against mean PHPP prediction	1.65	12.7 / 7.7
				PH target (15 $kWh \cdot m^{-2} \cdot a^{-1}$)		Mean measured against PH target	0.85	12.7 / 15
		PH 76 dwellings		PHPP prediction ($kWh \cdot m^{-2} \cdot a^{-1}$)		Mean measured against mean PHPP prediction	0.86	11.1 / 12.9
				PH target (15 $kWh \cdot m^{-2} \cdot a^{-1}$)		Mean measured against PH target	0.74	11.1 / 15
		Flats and houses, Category A (5 flats and 27 houses)	1–3 years; varied frequency	PHPP prediction ($kWh \cdot m^{-2} \cdot a^{-1}$)		Mean measured against mean PHPP prediction	0.88	12.2 / 13.9
				PH target (15 $kWh \cdot m^{-2} \cdot a^{-1}$)		Mean measured against PH target	0.81	12.2 / 15
		Flats and houses, Category B (4 flats and 35 houses	Three years; monthly frequency	PHPP prediction ($kWh \cdot m^{-2} \cdot a^{-1}$)		Mean measured against mean PHPP prediction	0.90	11.0 / 12.2
				PH target (15 $kWh \cdot m^{-2} \cdot a^{-1}$)		Mean measured against PH target	0.73	11.0 / 15
		Flats and houses, Category C (12 flats and 13 houses)	Two years; biannual frequency	PHPP prediction ($kWh \cdot m^{-2} \cdot a^{-1}$)		Mean measured against mean PHPP prediction	1.41	11.3 / 8.0
				PH target (15 $kWh \cdot m^{-2} \cdot a^{-1}$)		Mean measured against PH target	0.75	11.3 / 15
Miller, Sodagar et al. (2021)	Attached single-family end-terrace house (One dwelling)		Annual Jul 2012–Sep 2014	SAP Heating ($kWh \cdot m^{-2} \cdot a^{-1}$) Water heating (kWh $person^{-1}$ $year^{-1}$)	Heating	Actual against predicted	1.80	50 / 28
					Water heating		0.87	976 / 1,122
	Attached single-family mid-terrace house (One dwelling)				Heating		0.73	24/ 33
					Water heating		3.17	2,439 / 770
	Attached single-family mid-terrace house (One dwelling)				Heating		0.75	23 / 30.4
					Water heating		1.53	1,131 / 738
	Attached single-family end-terrace house (One dwelling)				Heating		1.24	36/ 29
					Water heating		1.20	1,028 / 854

Table continues next page

Continued from previous page

References	Building Type		Period	Baseline & Unit	Details		Factor	Measured / Predicted
Bahadori-Jahromi, Salem et al. (2022)	Detached, semi-detached, terraced and end-/mid-terrace single-family houses (Seven dwellings)		Annual	TAS ($kWh\cdot m^{-2}\cdot a^{-1}$)	Total energy demand	Baseline model: range across houses	1.27-1.32	— / —
					Heating energy		1.27-1.36	— / —
					Total energy demand	After heating set-point adjustment	1.19 -1.22	— / —
					Heating energy	After heating-schedule adjustment	1.18- 1.20	— / —
Few, Manouseli et al. (2023)	Dwellings (1,374 households)	EPC bands A and B	Annual (2021)	EPC / SAP 2012 ($kWh\cdot m^{-2}\cdot a^{-1}$)	Total energy demand	Metered PEUI against EPC-modelled PEUI	Measured more than predicted	
		EPC band C					0.92	— / —
		EPC band D & E					Measured less than predicted	
		EPC bands F and G					0.52	298 / 573
Duncan, Hayles and Littlewood (2024)	Affordable housing dwellings (Six dwellings)	(Unit A)	Annual 2 Nov 2021 – 2 Nov 2022	DER calculations total energy use ($kWh\cdot m^{-2}\cdot a^{-1}$)	Gas and electricity	Measured against DER prediction	2.47	129 / 52.21
		(Unit B)					2.55	149 / 58.48
		(Unit C)					3.15	202 / 64.15
		(Unit D)					1.62	103 / 63.68
		(Unit E)					1.14	71 / 62
		(Unit F)					2.7	126 / 46.6
Firth, Allinson and Watson (2024)	Existing residential housing stock in England and Wales		Annual Mid-May 2019 - mid-May 2020	EPC / RdSAP gas prediction ($kWh\cdot a^{-1}$)	Gas	National average gas measured Sub-National Energy Consumption Statistics (SNECS) against RdSAP predicted	0.82	— / —
	Existing residential housing stock in England and Wales					National average gas measured (NEED) against RdSAP predicted	0.83	— / —
	Existing residential housing stock in England and Wales					England average measured gas against RdSAP predicted	0.82	— / —
	Existing residential housing stock in Greater London					Greater London measured gas against RdSAP predicted	0.77	— / —
	Existing residential housing stock in City of London					Local authority with lowest performance gap; RdSAP under-predicted	1.24	— / —
	Existing residential housing stock in Blaenau Gwent					Local authority with highest performance gap; RdSAP over-predicted	0.44	— / —
	Existing residential housing stock in Birmingham					Local authority measured gas against RdSAP predicted (14,701- 17.862 $kWh\ yr^{-1}$)	0.82	14,701 / 17,862

In the commercial sector in the UK as summarized in *Table 3.2*, Burman, Mumovic and Kimpian (2014) calibrated an energy model of an educational building under the Energy Performance of Buildings Directive (EPBD), reporting relatively small gaps of 1.2%, 3.4%, and 3.4% above simulated values for gas, electricity, and total energy consumption respectively. By comparison, de Wilde (2014) employed an EnergyPlus model combined with an uncertainty workbench. In this case, total electricity consumption in 2011 was 98% of predicted values and measured and simulated values in 2012 were nearly identical, suggesting no gap. However, electricity use for HVAC was considerably lower, amounting to only 32% and 34% of predictions in 2011 and 2012, while gas consumption was substantially higher, at 281% and 289% above simulated performance. Brady and Hanmer-Dwight (2016) further observed gaps as high as 99% above CIBSE TM46 for both gas and electricity in an educational building, although this reduced to 38% when assessed against a TM54-based simulation.

Brady and Abdellatif (2017) analysed an educational building using measured data from Display Energy Certificates (2008–2015) and simulated outcomes of IES-VE model extended with TM54. The resulting gaps ranged from 47% to 114% above predictions for gas and from 75% to 174% for electricity. Gupta, Kapsali and Gregg (2017) quantified the gap in a library and community centre. The library consumed 80% more energy than Part L predictions (extended with TM54), with electricity use exceeding typical and good practice benchmarks by 23% and 70% respectively. By contrast, the community centre consumed only 30% and 50% of typical and good practice benchmarks, respectively. Mohareb, Hashemi et al. (2017) reported electricity and gas consumption to be 30% and 184% higher than simulated values, while Pritchard and Kelly (2017) identified gaps in an office building of 10–29% above typical benchmarks and 93–127% above good practice benchmarks. They further found that the difference between measured performance and CIBSE TM46 benchmarks ranged from 9% to 24%, while performance exceeded logbook estimates for core building functions by 40%.

In the hospitality sector, Rotimi, Bahadori-Jahromi et al. (2017) reported gaps of up to 44% in a hotel when comparing measured consumption with regulatory benchmarks. When compared with simulated performance, however, the gap narrowed to 28%, and further to 12% when catering loads were incorporated into the model. Burman, Kimpian and Mumovic (2018) examined five schools, documenting wide variations in electricity use ranging from 37% to 191% above typical benchmarks and from 66% to 254% above good practice benchmarks. Jain, Burman et al. (2020) extended this line of research, exploring an office, school, and hospital, and reported significant variations depending on whether measured energy use was compared with simulated predictions or good practice benchmarks. A subsequent study by Jain, Burman et al. (2021) on an office building found that measured energy use was 36% greater than predicted by a TM54-based simulation.

More recent work has expanded the analysis to larger samples. Thompson, Burman et al. (2022) studied 39 non-domestic buildings, demonstrating how procurement methods, ventilation strategies, and the use of SL can both widen and reduce the gap. Reguis, Tunzi et al. (2023) examined 121 non-domestic buildings in Edinburgh, focusing on heating performance. The study reported that conventional buildings consumed between 208% and 648% more than the PH retrofit target and revealed that older buildings tended to exhibit narrower gaps than newer ones. Benedict, Regnier et al. (2024) conducted a calibrated simulation of a 2,636 m² hotel, where measured electricity use (137.30 MWh) closely matched the simulated value (144.58 MWh), resulting in a factor of 0.95. Similarly, Lizana, Wheeler et al. (2024) analyzed a 1,154 m² educational building and found that energy consumption exceeded the European nZEB target range by 45%–82% and the PH classic standard by a 142%.

Table 3.2: Analysed records for commercial buildings in the UK

References	Building Type	Period	Baseline	Unit	Details		Factor	Measured / Predicted
Benedict, Regnier et al. (2024)	Hotel	Annual 2022	IESVE (calibrated using hourly data)	MWh	Electricity		0.95	137 / 144
Burman, Mumovic and Kimpian (2014)	Educational (secondary school)	Annual 2011-2013	IESVE (calibrated)	$kWh·m^{-2}·a^{-1}$	Gas		1.01	87/ 86
					Electricity		1.04	140 / 135
de Wilde (2014)	Educational (mixed-use academic building)	2011	EneryPlus Model combined with uncertainty workbench $kWh·a^{-1}$	$kWh·a^{-1}$	Total electricity		0.98	1.19×10^6 / 1.21×10^6
		2012					1	1.21×10^6 / 1.21×10^6
		2011			HVAC electricity		0.33	1.20×10^5 / 3.69×10^5
		2012					0.35	1.29×10^5 / 3.69×10^5
		2011			Gas		3.81	8.81×10^5 / 2.31×10^5
		2012					2.89	6.68×10^5 / 2.31×10^5
Brady and Hanmer-Dwight (2016)	Educational	Annual 2013-2014	TM46	$kWh·a^{-1}$	Gas and electricity	Median	1.99	— / —
			TM54				1.38	— / —
Brady and Abdellatif (2017)	Educational	Annual DEC data 2008–2015	IES-VE compliant CIBSE TM54	$kWh·m^{-2}·a^{-1}$	Space heating, cooling, fans and pumps, with longhand calculations for other energy streams	Gas range (Lowest-highest)	1.47-2.14	219–232 / 149
						Electricity range (Lowest-highest)	1.18-2.74	112–260 / 95
Gupta, Kapsali and Gregg (2017)	Library	Annual	BRUKL / UK Part L design prediction extended using TM54 to include additional end uses	$kWh·m^{-2}·a^{-1}$	Total energy use		1.8	— / —
			CIBSE TM46		Electricity		1.23	— / —
			CIBSE Guide F good-practice benchmark				1.70	— / —
	Community centre		BRUKL / UK Part L design prediction extended using TM54 to include additional end uses	$kWh·m^{-2}·a^{-1}$	Total energy use		1	— / —
			CIBSE TM46 benchmark				0.30	— / —
			CIBSE Guide F good-practice benchmark				0.50	— / —
Mohareb, Hashemi et al. (2017)	Educational	—	Design estimate for regulated energy use	—	Electricity (regulated loads)		1.30	— / —
					Gas (regulated loads)		2.84	— / —

Table continues on next page.

Table continued from previous page.

References	Building Type	Period	Baseline		Unit	Details		Factor	Measured / Predicted
Pritchard and Kelly (2017)	Office	Annual 2010 - Jun 2014 (4.5 years)	Range acceptable according to energy consumption guide 19 for office (ECON19) excluding the labs within the building	Typical	$kWh \cdot m^{-2} \cdot a^{-1}$	Office energy, excluding laboratory loads		1.10-1.29	— / —
				Good practice				1.93-2.27	— / —
			ND-NEED statistics for offices and the CarbonBuzz metric for university buildings			Total metered energy		1.49-1.69	— / —
			CIBSE TM46			Total metered energy, excluding laboratory loads		1.09-1.24	— / —
			Measured against estimated by building logbook (building core functions)			Core regulated energy functions		2.40	— / —
Rotimi, Bahadori-Jahromi et al. (2017)	Hotel	—	EDSL TAS v9.3.3 UK Building Regulation model		$kWh \cdot a^{-1}$	Actual building consumption against Building Regulation estimate		1.44	5,504,655 / 3,811,393
			EDSL TAS systems model			Actual building consumption against system simulation estimate		1.28	5,504,655 / 4,287,790
			EDSL TAS systems model + CIBSE TM50 catering benchmark			Actual building consumption against system simulation plus catering estimate		1.12	5,504,655 / 4,923,991
Burman, Kimpian and Mumovic (2018)	Educational (Five schools)	Annual	DEC secondary-school dataset, typical benchmark		$kWh \cdot m^{-2} \cdot a^{-1}$	Electricity	Measured range in the five schools against typical benchmark	1.37- 2.91	— / —
			DEC secondary-school dataset, good-practice benchmark			Electricity	Measured range in the five schools against good practice benchmark	1.66- 3.54	— / —
			DEC secondary-school dataset, good-practice benchmark			Gas and biomass	School one School two	0.84 0.57	— / —
			DEC secondary-school dataset, typical benchmark			Gas and biomass	School three	1.29	— / —
			CIBSE TM46					1.04	— / —
			DEC secondary-school dataset, typical benchmark			Electricity	School one & four	1.35	— / —
			CIBSE TM46					1.73	— / —
Lizana, Wheeler et al. (2024)	Educational	Annual Oct 2022 - Mar 2023	Recommended nZEB threshold defined by the European guidelines (80 to 100)		$kWh \cdot m^{-2} \cdot a^{-1}$	Primary energy consumption		1.45-1.82	145 / 100; 145 / 80
			PH (Classic) standard (60)					2.42	145 / 60

Table continues on next page.

Table continued from previous page.

References	Building Type	Period	Baseline	Unit	Details			Factor	Measured / Predicted
Jain, Burman et al. (2020)	Office	Annual	DesignBuilder (Performance modelling)	$kWh \cdot m^{-2} \cdot a^{-1}$	Electricity			1.18	68 / 57
					Gas			1.52	29 / 19
			CIBSE guide F		Electricity			1.25	68 / 54
					Gas			0.28	29 / 79
	School		DesignBuilder (Compliance modelling)		Electricity			3.07	61 / 20
					Gas			8.67	168 / 19
			CIBSE guide F		Electricity			2.46	61 / 25
					Gas			1.55	168 / 108
	Hospital		DesignBuilder (Compliance modelling)		Electricity			3.25	114 / 35
					Gas			3.62	303 / 83
			CIBSE guide F		Electricity			1.52	114 / 74
					Gas			0.71	303 / 422
Jain, Burman et al. (2021)	Office		DesignBuilder (Compliant with CIBSE TM54)	$kWh \cdot m^{-2} \cdot a^{-1}$	Total building energy			approx. 1.36	approx. 95 / 70
Thompson, Burman et al. (2022)	Non-domestic buildings (39 buildings)	Annual	TM46/Good practice	$kWh \cdot m^{-2} \cdot a^{-1}$	Total building energy	Soft Landing		0.74/0.90	— / —
						No Soft Landing		0.98/1.23	— / —
						Procurement (Design and build)		0.97/1.15	— / —
						Procurement (Traditional)		0.77/1.16	— / —
						Mechanical Ventilation strategy		1.06/1.31	— / —
						Natural Ventilation strategy		0.77/1.01	— / —
						Seasonal buildings		1.07/1.31	— / —
						Non-seasonal buildings		0.80/1.01	— / —
						Mechanical ventilation- seasonal buildings		1.19/1.45	— / —
						Natural ventilation-non seasonal buildings		0.63/0.86	— / —
Reguis, Tunzi et al. (2023)		Annual	PH for existing retrofit	$kWh \cdot m^{-2} \cdot a^{-1}$	Heating	Average EUI of 178 $kWh \cdot m^{-2} \cdot a^{-1}$ for Edinburgh sample		7.12	178 / 25
						Average EUI for a national of 187 $kWh \cdot m^{-2} \cdot a^{-1}$		7.48	187 / 25
						Average for two buildings of 77 $kWh \cdot m^{-2} \cdot a^{-1}$ built after 2011		3.08	77 / 25
			EPC-calculated energy use			Measured EUI against predicted based on build date	*Average from pre-1919-1979	1.20	— / —
							1980-1990	1.32	— / —
							1991-2002	1.48	— / —
							2003-2010	1.57	— / —

In Switzerland, several studies have examined the EPG in residential buildings, as summarized in Table 3.3. Lehmann, Khoury and Patel (2017) quantified the EPG in eight MINERGIE-P-ECO standard student apartments using measured energy performance data from July 2010 to June 2012. The study found that overall electricity consumption was only 71% of the standard value set by the Swiss Society of Engineers and Architects (SIA). By contrast, hot water and space heating demand were 100% higher than the planned value under Minergie standards and 27% higher than the SIA limit value. For hot water alone, measured consumption exceeded simulated values by 55%. Cozza, Chambers et al. (2019) examined 56 energy-efficient residential buildings and reported that normalized primary energy use during operation was 94% higher than predicted during the design phase. At the same time, measured energy performance achieved only 86% of retrofit targets under the various Minergie standards.

Flourentzos, Ivanov Antonov and Pantet (2019) analysed a refurbished residential building, finding that heating demand in 2012–2013 was 97% of the mean value of a comparable statistical sample. Following optimisation measures introduced between 2013 and 2017, heating consumption in 2017 was reduced to 70% of the sample mean, highlighting the impact of targeted interventions. Mojic, Lehmann et al. (2019) assessed 65 apartments and observed that heating demand was 44% higher than energy performance certificate (EPC) calculations, depending on the EPC version applied. The study further demonstrated how the magnitude of the EPG varies with the choice of simulation tool and with heating set-point assumptions. Cozza, Chambers et al. (2020) with a larger sample, of 1,172 buildings, reported median pre-retrofit and post-retrofit EPGs of 73% below theoretical values and 11% above theoretical values respectively. Expanding to a national sample of 34,816 buildings, Cozza, Chambers and Patel (2020) found median actual consumption to be 88% of theoretical predictions, representing a favourable EPG of 12%. Their study further showed that favourable gaps were more common in lower-rated buildings, with the exception of A-rated dwellings, and that buildings constructed prior to 2000 tended to consume less energy than theoretical calculations, in contrast to post-2000 constructions which tended to consume more. Padey, Goulouti et al. (2021) investigated a Minergie-Eco multi-family building and found heating demand to be 59% higher than predicted in 2014. This discrepancy grew steadily, reaching 113% by 2017.

Table 3.3: Analysed records for residential buildings in Switzerland

Reference	Building Type	Period	Baseline, Unit & Details			Factor	Measured / Predicted
Lehmann, Khoury and Patel (2017)	Residential Student housing 8 apartments (MINERGIE-P-ECO standard)	Annual Two years Jul 2010 – Jun 2012	Measured overall electricity against standard value of set by the Swiss society of engineers and architects $MJ\cdot m^{-2}\cdot a^{-1}$			0.71	71 / 100
			Measured hot water and space heating ($38\ kWh\cdot m^{-2}\cdot a^{-1}$)	Planned value according to Minerge		2	38 / 19
				Limit value according to SIA		1.27	38 / 30
			Measured hot water against predicted			1.55	— / —
			Measured thermal energy consumption against: $kWh\cdot m^{-2}\cdot a^{-1}$	Mean of two relevant residences		0.97	224 / 232
				Average of 35 residences		0.45	224 / 502
				Average multifamily buildings		0.46	224 / 483
			Measured average energy use per capita against average for multifamily buildings $kWh\ a^{-1}$			0.54	580 / 1081
Cozza, Chambers et al. (2019)	Residential High-performance buildings (56 buildings)	Annual	Normalized net primary energy in the operation phase against normalized primary energy consumption in the design phase $kWh\cdot m^{-2}\cdot a^{-1}$			1.94	— / —
			Measured against retrofit target under various Minergie standards $kWh\cdot m^{-2}\cdot a^{-1}$			0.86	— / —
			Actual final energy consumption against target final $kWh\cdot m^{-2}\cdot a^{-1}$	28 new buildings *a few buildings consumed over their targets		0.96	— / —
				Five new buildings against (Model Policies for Energy Codes MoPEC) 2014 target		1.07	38 / 35.5
				Five new buildings against Minergie		1.15	41 / 35.7
				14 new buildings against Minergie-P		0.89	29 / 33
				Four new buildings against Minergie-A		0.85	30 / 35
				Retrofitted buildings (28 buildings)		0.88	— / —
				15 retrofitted buildings against MoPEC 2014 target		0.93	— / —
				Three retrofitted buildings against Minergie		0.96	— / —
				Seven retrofitted buildings against Minergie-P		0.71	— / —
				Three retrofitted buildings against Minergie-A		0.93	— / —
Flourentzos, Ivanov Antonov and Pantet (2019)	Residential (Refurbished nZEB apartment building)	Annual 2012-2013	Heating $kWh\cdot m^{-2}\cdot a^{-1}$	After refurbishment		0.97	97 / 100
		Annual 2017		After energy performance optimization in the period from 2013-2017		0.70	70 / 100
Mojic, Lehmann et al. (2019)	Residential multi-family apartment buildings (65 apartment buildings)	Annual	Measured against EPC utilizing various versions of Swiss standard Heating $kWh\cdot m^{-2}\cdot a^{-1}$	Average for the sample		1.44	— / —
				Range for four buildings		2-2.15	— / —
				In three homes		0.94	— / —
				Average for buildings without separate DHW measurement		1.51	— / —
				Average for buildings with separate DHW measurement		1.4	— / —
			Measured against simulated Heating $MJ\cdot m^{-2}$	Lesosai	22ºC	1.12	— / —
					23ºC	1.24	— / —
					24ºC	1.37	— / —
				IDA ICE	22ºC	1.37	— / —
					23ºC	1.28	— / —
					24ºC	1.43	— / —

Table continues on next page.

Table continued from previous page.

Reference	Building Type	Period		Baseline, Unit & Details			Factor	Measured / Predicted
Cozza, Chambers et al. (2020)	Residential Retrofitted single- and multi-dwelling buildings (1172 buildings)	—		CECB theoretical final thermal energy Heating + DHW $kWh \cdot m^{-2} \cdot a^{-1}$	Before retrofit	Minimum	0.94	14.7 / 15.6
						Maximum	0.67	321 / 478
						Mean	0.71	153 / 216
						Median	0.73	146 / 199
					After retrofit	Minimum	0.66	4.24 / 6.39
						Maximum	0.86	226 / 264
						Mean	1.01	66.3 / 65.8
						Median	1.11	59.3 / 53.2
Cozza, Chambers and Patel (2020)	Residential (34,816 buildings)	—		CECB theoretical final thermal energy heating & DHW, $kWh \cdot m^{-2} \cdot a^{-1}$ Median	Building rating	Entire sample (34816 buildings)	0.88	113 / 128
						A (156 buildings)	0.94	37 / 39
						B (2554 buildings)	1.20	50 / 42
						C (7395 buildings)	1.07	84 / 79
						D (9067 buildings)	0.96	116 / 121
						E (6564 buildings)	0.84	137 / 164
						F (4039 buildings)	0.75	151 / 202
						G (5041 buildings)	0.56	174 / 308
					Construction period	Pre 1919	0.83	— / —
						1919-1945	0.84	— / —
						1945-1960	0.84	— / —
						1960-1970	0.88	— / —
						1970-1980	0.87	— / —
						1980-1990	0.94	— / —
						1990-2000	0.97	— / —
						2000-2010	1.10	— / —
						Post 2010	1.11	— / —
Padey, Goulouti et al. (2021)	Residential Multi-family apartment building (Minergie-Eco)	Annual	2014	Lesosai Minergie deterministic normative model Heating kWh_useful·a^{-1}	Measured against deterministic normative demand		1.59	71,273 / 44,899
			2015				1.97	88,477 / 44,899
			2016				2.10	94,509 / 44,899
			2017				2.14	95,999 / 44,899

In Italy, several studies have investigated the EPG across different building types, as summarized in Table 4.4. Kampelis, Gobakis et al. (2017) quantified the EPG in both industrial and residential buildings, reporting that measured consumption was 31% and 177% above EPC predictions, respectively. Tagliabue, Maltese et al. (2018) assessed two schools using IESVE; the first exhibited an EPG of 42% above predicted values, whereas the second consumed only 88% of the simulated energy demand. Filippi and Sirombo (2019) evaluated 152 A-rated flats over two heating seasons, finding that the EPG increased from 41% above the estimated standard value in 2016–2017 to 59% in 2017–2018. Kampelis, Papayiannis et al. (2020) analysed the HVAC system of an industrial nZEB, where baseline models indicated EPGs of 29% in 2017 and 25% in 2018. Following calibration, the EPGs were reduced to 5% and 2%, respectively.

Salvia, Morello et al. (2020) examined 154 retrofitted apartments during the 2018–2019 heating period, reporting that actual energy use was on average 80% higher than calculated values. Aste, Adhikari et al. (2022) studied an nZEB residential complex monitored from 2014–2015, initially identifying an EPG of 44% (1.44) above simulations. After calibration, this was narrowed to 93% of predicted values (0.93). Vivian, Carnieletto et al. (2023) quantified the EPG in 11 refurbished apartments across a heating season by comparing measured performance before and after modifying indoor environmental conditions. For the first group of six flats, increasing the air change rate led to an average EPG of 12% above baseline conditions, while for the second group of five flats, reducing the air change rate resulted in an average EPG of just 2% above baseline. Maracchini, Latini et al. (2024) investigated a four-storey multifamily residential building using DesignBuilder. Both unoccupied and occupied apartments were modelled, with pre-calibration gap factors ranging from 0.85 to 1.05. After calibration, these narrowed to between 0.90 and 1.02, reflecting closer alignment between predicted and measured performance.

Table 3.4: Analysed records for residential and non-residential buildings in Italy

Reference	Building Type	Period	Baseline, Unit & Details		Factor	Measured / Predicted
Kampelis, Gobakis et al. (2017)	Industrial	Annual	EPC kWh·m^{-2}·a^{-1}	Operational normalised net primary energy against design-stage normalised primary energy	1.31	35.4 / 26.9
	Residential Apartment building				2.77	54.4 / 19.6
Tagliabue, Maltese et al. (2018)	Educational (School one)	Annual 2014-2016	IESVE kWh·m^{-2}·a^{-1}	Before alterations or making improvements (whether on building or simulation model)	1.42	174.38 / 123.07
				2014	1.06	243.86 / 230.36
				2015	1.29	297.19 / 230.36
				2016	1.25	286.92 / 230.36
	Educational (School two)		IESVE kWh·m^{-2}·a^{-1}	2014 (before making alterations)	0.88	141.97 / 161.68
				2015	1.10	148.86 / 134.74
				2016	0.88	118.91 / 134.74
Filippi and Sirombo (2019)	Residential High-performance multi-family social housing (A rated 152 certified flats)	Monthly Oct 15th-Apr 15th 2016-2017	Average actual heating consumption against estimated standard value Heating kWh·m^{-2}·a^{-1}		1.41	63.3 / 44.9
		Oct 15th-Apr 15th 2017-2018			1.60	71.8 / 44.9
Kampelis, Papayiannis et al. (2020)	Industrial (nZEB)	Annual 2017	EnergyPlus HVAC MWh·a^{-1}	Annual measured against baseline model electricity consumption	1.29	232.67 / 180.53
				Annual measured against calibrated electricity consumption	1.05	232.67 / 221.95
		Annual 2018		Annual measured against baseline model electricity consumption	1.25	250.05 / 200.46
				Annual measured against calibrated electricity consumption	1.02	250.05 / 244.02
Salvia, Morello et al. (2020)	Residential Retrofitted social housing apartments (154 retrofitted apartments)	Monthly Heating period 2018-2019	Average actual (63 kWh·m^{-2}·a^{-1}) energy use against average calculated (energy certificates) of 35 kWh·m^{-2}·a^{-1})		1.80	63 / 35
Aste, Adhikari et al. (2022)	Residential (nZEB complex)	Annual Oct 2014 - Sep 2015	EnergyPlus Net primary energy kWh·m^{-2}·a^{-1}	Baseline model	1.44	25.5 / 17.7
				Calibrated 1: Weather data	1.50	— / —
				Calibrated 2: U-values	1.33	— / —
				Calibrated 3: Use profile	1.30	— / —
				Calibrated 4: Setpoints	1.29	— / —
				Calibrated 5: HVAC features	0.93	— / —
				Calibrated 6: PV production	0.93	— / —
			EnergyPlus Heating kWh·m^{-2}·a^{-1}	Calibrated 1: Weather data	1.49	— / —
				Calibrated 2: U-values	1.13	— / —
				Calibrated 3: Use profile	1.03	— / —
				Calibrated 4: Setpoints	1.03	— / —
			EnergyPlus Cooling kWh·m^{-2}·a^{-1}	Calibrated 1: Weather data	1.34	— / —
				Calibrated 2: U-values	1.27	— / —
				Calibrated 3: Use profile	1.12	— / —
				Calibrated 4: Setpoints	1.12	— / —

Table continues on next page.

Table continued from previous page.

Reference	Building Type	Period	Baseline, Unit & Details				Factor	Measured / Predicted
Vivian, Carnieletto et al. (2023)	Residential 11 retrofitted apartments	Monthly Heating period 14 Nov–21 Dec	EnergyPlus Heating demand $kWh \cdot m^{-2}$	Representative flat 1			1.04	— / —
				Representative flat 2			1.10	— / —
				Group one (six flats) (increased air change rate)		Range	0.93-1.32	— / —
						Average	1.12	43.4 / 39.6
				Group two (five flats) (reduced air change rate)		Range	0.82-1.24	— / —
						Average	1.02	33.2 / 33.3
Maracchini, Latini et al. (2024)	Residential Apartment in a four-storey multi-family building	2^{nd}- 6^{th} Sep 2021	DesignBuilder Space cooling electricity	Unoccupied apartments *Factors are approximate	A	Pre-calibration	0.96	— / —
						Post-calibration	0.98	— / —
					B	Pre-calibration	0.97	— / —
						Post-calibration	0.99	— / —
					C	Pre-calibration	0.97	— / —
						Post-calibration	0.98	— / —
				Occupied apartments *Factors are approxim. Figure-derived	D No occupancy data	Pre-calibration	0.85	— / —
						Post-calibration	0.90	— / —
					E Occupancy data	Pre-calibration	1.05	— / —
						Post-calibration	1.02	— / —

In Denmark (Table 3.5), Jradi, Arendt et al. (2018) investigated regulatory EPGs in an educational building in 2016 using the Danish compliance tool (BE10). They reported that lighting and heating systems consumed 64% and 27% more than predicted, respectively. A calibrated dynamic model was then developed, showing that heating, overall electricity, and lighting electricity were 97%, 97%, and 99% of simulated values. For ventilation units, the EPG ranged from 98% of predicted consumption to 4% above predictions. Loukou, Heiselberg et al. (2019) assessed a residential building against the European recommended range for new single-family houses and found performance to vary between 16% above the highest recommended value and 89% of the minimum recommended value. Using the Be.18 tool, they reported that the building consumed 89% more energy than predicted for heating, ventilation, DHW, tanks, and heat pumps, while total primary energy use was 86% of predicted values. Jradi, Liu et al. (2020) analysed an educational building over two weeks in May 2018 using a calibrated OpenStudio model. They reported EPGs of 8% and 25% above predicted values for total electricity and heating, respectively. At the floor level, lighting discrepancies ranged from 9% above predictions to only 69% of predicted values, while ventilation performance ranged from 2% above predictions to 79% of predictions.

Carpino, Loukou et al. (2022) studied an nZEB single-family house (December 2017–May 2018) with DesignBuilder, showing that heating demand was 97% of predicted values and electricity use only 0.03% above predictions. At a larger scale, Hansen and Gram-Hanssen (2023) examined 114,022 Danish households. They found that A-rated buildings classified under a newer regulation consumed 4% less than those under the earlier version, while the magnitude of the EPG steadily decreased from B to G rated dwellings. Importantly, A and B rated buildings consumed significantly less energy than lower-rated ones.

In Germany (*Table 3.5*), Galvin (2014) analysed heating EPGs in three retrofitted residential blocks of 30 apartments each. Results showed minimal gaps in the first block (2% above retrofit targets for building with district heating and radiators), larger discrepancies in the second (57% above targets for building with underfloor heating), and the greatest gap in the third (273% above targets for building with heat pumps). Cali, Osterhage et al. (2016) compared three residential buildings, finding that measured primary energy consumption represented only 49% of predicted values, suggesting considerable overestimation. Grossmann, Galvin et al. (2016) studied four public buildings before and after retrofitting. Pre-retrofit, the music academy consumed 2% more than predicted, while the police headquarters and two customs offices consumed 86%, 84%, and 96% of predictions, respectively, indicating mostly favourable performance. Post-retrofit, all four buildings consumed less than predicted, with the music academy, for example, utilising only 56% of predicted demand. Moeller, Weber et al. (2020) investigated six residential buildings with 40 apartments, finding average heating gaps of 15% at the building level and 25% at the apartment level, but with wide individual variation from 318% above predictions to 41% of predicted values. More recently, Moeller (2024) analyzed two smart-home equipped multi-apartment buildings with 137 households, reporting average space heating consumption (53.5 $kWh \cdot m^{-2} \cdot a^{-1}$) closely aligned with predictions (51.4 $kWh \cdot m^{-2} \cdot a^{-1}$, factor 1.04), though household variation ranged from 0.39 to 1.52.

In Poland (Table 3.5), Cholewa, Balaras et al. (2020) assessed heating EPGs in 11 multifamily buildings before and after retrofit. Prior to retrofit, eight buildings exhibited favourable performance, consuming less than predicted. Post-retrofit, however, 10 buildings consumed more than predicted, while only one consumed 93% of predicted values. Szulgowska-Zgrzywa, Stefanowicz et al. (2020) analysed heating and DHW use in 15 apartments with varied heating systems. Most showed favourable gaps, except for one solid-fuel heated apartment (70% above predictions) and two natural-gas heated apartments (13% above predictions). In a later study, Szulgowska-Zgrzywa, Stefanowicz et al. (2023) monitored six apartments between January and March. Five out of six consumed less than predicted for heating, while DHW results varied: two apartments consumed 47% and 27% more than predicted, and one consumed 58% of predicted values. In France (Table 3.5), Guyot, Giraud et al. (2020) analyzed an office building, finding heating and cooling EPGs of 12% and 28% above calibrated simulations, respectively. Charlier (2021) conducted a large-scale study of 1,853 dwellings using 2012 energy data. On average, measured consumption was only 65% of calculated values, while a subsample of 517 gas-heated dwellings consumed 71%, indicating broadly favourable EPGs.

Table 3.5: Analysed records for residential and non-residential buildings in Denmark, Germany, Poland and France

Reference	Building Type	Location	Period	Baseline, Unit & Details			Factor	Measured / Predicted
Jradi, Arendt et al. (2018)	Educational University building	Denmark	Annual 2016	Danish official compliance tool (BE10) Regulatory EPG	Lighting kWh		1.64	— / —
					Heating MWh		1.27	— / —
			Monthly Feb- mid Mar 2017	EnergyPlus Dynamic EPG	Heating MWh		0.97	71 / 73
					Overall electricity MWh		0.97	24 / 25
					Lighting electricity kWh		0.99	5972 / 6047
					Ventilation units' electricity kWh	Unit 1	0.98	1213 / 1241
						Unit 2	1.05	1165 / 1105
						Unit 3	0.99	1151 / 1166
						Unit 4	1.04	1332 / 1281
			Monthly Sep-Dec 2016	Calibrated model created on EnergyPlus (OpenStudio) (Dynamic EPG)	Heating MWh	Sep	0.95	2.4 / 2.7
						Oct	1.09	25 / 23
						Nov	0.97	45 / 46
						Dec	0.97	47 / 48
					Ventilation electricity kWh	Sep	0.96	4471 / 4650.3
						Oct	0.98	3212 / 3269
						Nov	1.07	4224 / 3932
						Dec	1.02	3258 / 3193
					Lighting electricity kWh	Sep	0.96	4841 / 5025
						Oct	1.03	6095 / 5894
						Nov	0.96	5873 / 6106
						Dec	0.97	5800 / 5966
Loukou, Heiselberg et al. (2019)	Detached single-family nZEB house		Annual 2018	European recommended yearly primary energy range for new single-family house Heating, ventilation, DHW, Tank & Heat Pump $kWh \cdot m^{-2} \cdot a^{-1}$			0.90 - 1.16	58 / 50 - 65
				Be.18 tool $kWh \cdot m^{-2} \cdot a^{-1}$	Heating, ventilation, DHW, Tank & Heat Pump		1.89	58 / 31
					Total primary energy use		0.86	74 / 85.6
Jradi, Liu et al. (2020)	Educational		Weekly (1-14 May 2018)	OpenStudio Calibrated	Total Electricity consumption		0.92	
					Total Heating consumption		0.75	
					Lighting based on floor level	Parterre Ground First	0.91 1.31 1.09	— / —
					Ventilation units 1/2/3/4		1.21 0.98 1.07 0.93	— / —
Carpino, Loukou et al. (2022)	Detached single-family nZEB house		1st Dec- 31st May 2018	DesignBuilder Calibrated kWh	Heating		0.98	7213 / ≈ 7360
					Electricity		1.003	1587 / ≈ 1582

Table continues on next page.

Table continued from previous page.

Reference	Building Type	Location	Period	Baseline, Unit & Details			Factor	Measured / Predicted
Hansen and Gram-Hanssen (2023)	Residential dwellings (114,022 households)	Denmark	Annual 2019-2021	Danish EPC calculated Mean Heating & DHW MWh·a⁻¹ per household	A2015/2020(N:7416)		1.78	10.3 / 6
					A2010 (N:6483)		1.82	14 / 8
					B (N:2915)		1.48	18 / 12
					C (N:31957)		1.24	21 / 17
					D (N:38216)		1.03	24 / 23
					E (N:16343)		0.83	24 / 28
					F (N:7063)		0.68	23 / 33
					G (N:3637)		0.47	22 / 47
Galvin (2014)	Residential apartment block (30 apartments)	Germany	Annual 2012	Retrofit targets Heating kWh·m⁻²·a⁻¹	Metered against target (District heating and conventional radiators)		1.02	51 / 50
	Residential apartment block (30 apartments)				Metered against target (District heating and underfloor heating)		1.57	58 / 37
	Residential apartment block (30 apartments)				Metered against target (Heat pumps, third of apartments used under floor heating, third used ventilation heating and third used ceiling radiant heating)		3.73	88 / 23.6
Cali, Osterhage et al. (2016)	Residential Three multi-apartment buildings		Annual Average 2005-2007	Primary energy kWh·m⁻²·a⁻¹			0.49	171 / 347
Grossmann, Galvin et al. (2016)	Non-residential public service building		Annual Multi-year pre-/post-retrofit observations	Actual space heating energy consumption against calculated kWh·m⁻²·a⁻¹	Pre-retrofit	Music Academy	1.02	143 / 140
						Police HQ	0.86	106 / 123
						Customs office 1	0.84	204 / 244
						Customs office 2	0.96	206 / 215
					Post-retrofit	Music Academy	0.57	47 / 83
						Police HQ	0.85	52 / 61
						Customs office 1	0.86	117 / 136
						Customs office 2	0.92	88 / 96
Moeller, Weber et al. (2020)	Residential Six energy-efficient multi-apartment buildings (40 Apartments)		Monthly Heating periods 2011-2017	Heating energy demand of each apartment was determined under assumptions used in the German Energy Saving Ordinance (EnEV)	Average EPG on apartment building level MWh		1.15	52 / 45
					Average individual apartment level		1.25	— / —
					Individual apartment level range		0.41- 4.18	— / —
Moeller (2024)	Residential Two energy efficient multi apartment buildings (137 households)		Heating period 2021-2022	Heating kWh·m⁻²·a⁻¹	Measured average for sample against average demand		1.04	54 / 51
					Household one		1.52	54 / 35
					Household two		0.39	21 / 55
					Household three		1.49	73 / 49
					Household four		1.44	72 / 50
					Measured heating against reference energy efficient multi apartment buildings with thermostatic radiator valves		0.97	52 / 54
					Measured heating against conventional heating control system energy efficient multi apartment buildings		0.40	52 / 129

Table continues on next page.

Table continued from previous page.

Reference	Building Type	Location	Period	Baseline, Unit & Details			Factor	Measured / Predicted
Cholewa, Balaras et al. (2020)	Residential 11 comparable multifamily buildings each consists of a number of flats ranging from 19 to 60 flats	Poland	Monthly Eight heating seasons	Heating Pre-retrofit GJ·m⁻²	Similar characteristics	H1	0.85	0.54 / 0.63
						H2	1.01	0.71 / 0.70
						H3	0.79	0.61 / 0.77
						H4	1.07	0.66 / 0.62
					H5		1.16	0.64 / 0.55
					Similar characteristics	H6	0.79	0.50 / 0.63
						H7	0.82	0.56 / 0.68
						H8	0.88	0.68 / 0.76
					Similar characteristics	H9	0.92	0.77 / 0.83
						H10	0.80	0.61 / 0.76
						H11	0.92	0.68 / 0.73
				Heating Post-retrofit GJ·m⁻²	Similar characteristics	H1	1.13	0.42 / 0.38
						H2	1.22	0.54 / 0.44
						H3	0.93	0.45 / 0.48
						H4	1.49	0.49 / 0.33
					H5		1.63	0.45 / 0.28
					Similar characteristics	H6	1.12	0.46 / 0.41
						H7	1.20	0.49 / 0.41
						H8	1.31	0.62 / 0.47
					Similar characteristics	H9	1.07	0.54 / 0.50
						H10	1.04	0.49 / 0.43
						H11	1.08	0.49 / 0.45
Szulgowska-Zgrzywa, Stefanowicz et al. (2020)	Residential 15 apartments		14 Jan–9 Mar 2020	Polish regulations and ISO Standard for energy performance Heating and DHW kWh·m⁻²·a⁻¹	Solid fuel heated apartments	A1	1.30	69 / 53
						Average (A2, A3, A4)	0.43	31 / 72
					Electricity heating apartments	Average (A5, A6, A7)	0.30	15 / 50
					District heating apartments	Average (A8, A9, A10, A11)	0.51	25 / 50
					Natural gas apartments	A12	0.87	33 / 38
						A14	0.87	32 / 37
						Average (A12-A15)	0.79	32 / 41
Szulgowska-Zgrzywa, Stefanowicz et al. (2023)	Residential Six apartments		14 Jan- 9 Mar 2020	Calculations models for DHW and heating were created using Polish regulations and ISO Standard for energy performance	Actual heating demand against calculated kWh	A1	1.29	2770 / 2150
						A2	0.30	1095 / 3671
						A3	0.25	360 / 1460
						A4	0.38	869 / 2302
						A5	0.58	1902 / 3298
						A6	0.81	1445 / 1781
					Actual hot water demand against calculated kWh	A1	1.47	281 / 191
						A2	0.58	102 / 176
						A6	1.27	348 / 275
Guyot, Giraud et al. (2020)	Office	France	Annual	EnergyPlus (Calibrated) Heating kWh·m⁻²·a⁻¹	Measured against simulated heating needs		1.12	31 / 28
					Measured against simulated cooling needs		1.28	14 / 11
Charlier (2021)	Residential 1853 dwellings		Annual 2012	EPC (3CL) theoretical consumption Heating, hot water, cooling, lighting and appliances kWh·m⁻²·a⁻¹	Mean actual for the sample against mean theoretical		0.65	168 / 265
					Mean actual against mean theoretical in 517 buildings that use gas for hot water and heating		0.85	177 / 209

In Estonia *(Table 3.6)*, Hamburg, Kuusk et al. (2020) evaluated the performance of an nZEB residential building consisting of 80 apartments, quantifying the gap both before and after renovation. Measured energy consumption exceeded predictions by 217% prior to renovation and by 55% after renovation. Post-renovation measured consumption was also 47% above the nZEB target. The study further quantified EPGs across other energy aspects. Kuivjogi, Uutar et al. (2021) analysed 33 trade buildings (11 hardware stores and 22 grocery stores) and six office buildings. For around 21 of the trade buildings, measured energy consumption exceeded the range of primary energy limits for major renovation set by Estonian EPC regulations by between 13% and 86%. For the six office buildings, measured energy use exceeded EPC primary energy limits by between 11% and 146%.

In Cyprus *(Table 3.6)*, Kampelis, Gobakis et al. (2017) studied the energy performance of an educational building, finding that normalized primary energy consumption during operation was 98% of the normalised design-phase value. Ozarisoy and Altan (2022) examined 36 residential tower blocks, analysing electricity consumption in three representative flats at different floor levels. The first-floor flat consumed 12% more than predicted, the intermediate flat 18% more, while the upper-floor flat consumed only 10% of the simulated value, illustrating significant variability across vertical building levels. Nigumann, Kalamees et al. (2024) evaluated the performance of an apartment building before and after renovation. Pre-renovation, the building exhibited a performance factor of 0.79, improving to 0.93 post-renovation. However, specific system components varied considerably: heating and ventilation, for example, showed a post-renovation factor of 1.26, whereas hot water demand aligned more closely with predictions.

In Ireland (Table 3.6), Coyne and Denny (2021) analysed a large sample of residential properties using observed energy performance data from November 2014 to June 2017. Comparing observed and theoretical annual gas consumption across building ratings and types, they found that AB-rated buildings consumed 19% more gas than predicted, while buildings rated C to G consumed less than theoretical estimates. When analysed by building type, measured consumption was consistently lower than theoretical predictions. Colclough, Hegarty et al. (2022) studied eight one-bedroom homes retrofitted to nZEB standards (A-rated) from initial classes F and G. Using DEAP, Ireland's national energy rating tool, they reported significant gaps in space heating and DHW consumption, with discrepancies ranging from 29% to 138% above predicted values.

Table 3.6: Analysed records for residential and non-residential buildings in Estonia, Cyprus and Ireland

Reference	Building Type	Location	Period	Baseline, Unit & Details			Factor	Measured / Predicted
Hamburg, Kuusk et al. (2020)	Residential (nZEB building consists of 80 apartments)	Estonia	Annual	Before renovation $kWh \cdot m^{-2} \cdot a^{-1}$	Heating, ventilation, hot water and electricity		3.18	302 / 95
					Heating and ventilation		6.33	152 / 24
					Hot water		3.78	53 / 14
					Electricity		1.70	97 / 57
				After renovation $kWh \cdot m^{-2} \cdot a^{-1}$	Measured against nZEB target (Heating, ventilation, hot water and electricity)		1.47	147 / 100
					Heating, ventilation, hot water and electricity		1.55	147 / 95
					Heating and ventilation		1.67	40 / 24
					Hot water		2.57	36 / 14
					Electricity		1.25	71 / 57
Kuivjogi, Uutar et al. (2021)	Trade buildings (11 Hardware stores and 22 grocery stores) *EPG is for 65% of the sample		Annual Hardware stores Average 2015-2017 Grocery stores Average 2017-2018	Primary energy limit for trade buildings according to EPC in Estonia for major renovation $kWh \cdot m^{-2} \cdot a^{-1}$	Measured range against primary energy limit for trade buildings according to EPC in Estonia for major renovation		1.13-1.86	260-427 / 230
				Primary energy limit for trade buildings under EPBD scope (without appliances) according to EPC in Estonia for major renovation $kWh \cdot m^{-2} \cdot a^{-1}$	Measured range against primary energy limit for trade buildings under EPBD scope (without appliances) according to EPC in Estonia for major renovation		1.33-1.88	254-421 / 224
	Office (Six offices)		Annual Average 2005-2017	Primary energy limit for office buildings according to EPC in Estonia for major renovation $kWh \cdot m^{-2} \cdot a^{-1}$	Measured range against primary energy limit for office buildings according to EPC in Estonia for major renovation		1.11-2.46	177-394 / 160
				Primary energy limit for office buildings under EPBD scope (without appliances) according to EPC in Estonia for major renovation $kWh \cdot m^{-2} \cdot a^{-1}$	Measured range of against primary energy limit for office buildings under EPBD scope (without appliances) according to EPC in Estonia for major renovation		1.14-2.92	139-356 / 122
Nigumann, Kalamees et al. (2024)	Apartment building			Expected use before renovation $kWh \cdot m^{-2} \cdot a^{-1}$	Total	Measured energy performance value before renovation against expected use before renovation	0.79	173-220
					Heating + ventilation		0.77	98-127
					Hot water		0.48	10-21
					Electricity		0.9	65-72
				Calculated use after renovation $kWh \cdot m^{-2} \cdot a^{-1}$	Total	Measured energy performance value after renovation against calculated use after renovation	0.93	92-99
					Heating + ventilation		1.26	34-27
					Hot water		0.95	21-22
					Electricity		0.76	38-50
				Expected use before renovation $kWh \cdot m^{-2} \cdot a^{-1}$	Total	Measured energy performance value before renovation against expected use before renovation	0.75	199-264
					Heating + ventilation		0.77	151-195
					Hot water		0.48	16-33
					Electricity		0.89	32-36
				Measured delivered after renovation against calculated $kWh \cdot m^{-2} \cdot a^{-1}$	Total	Measured energy performance value after renovation against calculated use after renovation	1.03	103-100
					Heating + ventilation		1.24	52-42
					Hot water		0.97	32-33
					Electricity		0.76	— / —

Table continues on next page.

Table continued from previous page.

Reference	Building Type	Location	Period	Baseline, Unit & Details			Factor	Measured / Predicted
Kampelis, Gobakis et al. (2017)	Educational (Tertiary research laboratory)	Cyprus	Annual	Normalized net primary energy consumption in the operation phase against normalized primary energy in the design phase kWh·m^{-2}·a^{-1}			0.98	62.9 / 64.4
Ozarisoy and Altan (2022)	Residential 36 tower blocks		Annual 2016	IES Calibrated (Electricity) kWh	First-floor representative flat (low occupancy)		3.08	3,079 / 999
					Intermediate floor representative flat (moderate occupancy)		1.91	5,259 / 2,755
					Upper floor representative flat (high occupancy)		0.63	1,000 / 1,591
Coyne and Denny (2021)	Residential Housing sample observation for same sample for two years making total observations 19251	Ireland	Annual (observed from Nove 2014 - Jun 2017, sixteen bimonthly 8 periods)	Energy use kWh·a^{-1} Mean observed against mean theoretical EPC energy use	Building rating	AB (n:2601)/C (n:8269)/ D (n:4835)/ E (n:2051)/ FG (n:1495)	1.54/1.11/ 0.84/0.68/ 0.49	— / —
					Building type	Apartment/ Detached/ Semi-detached/Terrace	0.74/0.81/ 0.91/1.02	— / —
				Gas use kWh·a^{-1} Mean observed against mean theoretical EPC energy use	Building rating	AB (n:2601)/C (n:8269)/ D (n:4835)/E(n:2051)/ FG (n:1495)	1.19/0.86/ 0.65/0.55/ 0.40	— / —
					Building type	Apartment (n:1674)/Detached (n:2316)/ Semi-detached (n:6905)/Terrace (n:8356)	0.60/0.64/ 0.71/0.79	— / —
Colclough, Hegarty et al. (2022)	8 one-bedroom houses transformed from F and G to A nZEB		Annual Sep 2019-Aug 2020	The national energy rating software DEAP kWh	Recorded space heating and DHW primary energy consumption against estimated performance	H1/H2/ H3/H4/ H5/H6/ H7/H8	1.72/1.41/ 1.44/1.31/ 2.38/1.29/ 1.74/1.51	— / —

In Serbia (Table 3.7), Anelkovi, Kljaji et al. (2021) analysed 16 residential buildings. For the eight with district heating, heating consumption ranged from 19% below to 70% above EPC estimates. Buildings with individual gas boilers showed wide variation, from as little as 7% of EPC predictions to as much as 355% above, indicating significant inconsistency in estimation accuracy. In Lithuania (Table 3.7), Motuzien, Lapinskien et al. (2021) studied 62 multi-storey residential buildings. A-rated dwellings exhibited gaps between 26% above and 58% below EPC predictions, A+ buildings ranged from 73% above to 46% below, while A++ buildings showed gaps of 56% to 144% above predictions. More recently, Motuziene, Lapinskiene and Rynkun (2024) examined 13 energy-efficient offices, reporting measured-to-simulated ratios varying widely from 0.89 to 7.83. In Bosnia and Herzegovina (Table 3.7), Katic, Krstic and Marenjak (2021) analysed 185 educational buildings and found average consumption of 145.6 $kWh \cdot m^{-2} \cdot a^{-1}$, compared to predicted use of 242.9 $kWh \cdot m^{-2} \cdot a^{-1}$, corresponding to a favourable EPG of 40% below estimates. In Finland (Table 3.7), Ahmed, Hasu and Kurnitski (2022) examined five nZEB schools and daycare centres, identifying heating and electricity gaps of 110% to 160% above EPC predictions when unregulated loads were included. In Greece (Table 3.7), Kotarela, Kyritsis et al. (2023) compared dynamic and quasi–steady-state tools for single-family homes before and after retrofit. Before retrofit, dynamic simulations exceeded quasi–steady-state estimates by 9% for heating. In most other cases, quasi–steady-state predictions were higher, highlighting methodological sensitivity in estimating heating, cooling, DHW, and total energy use. In Spain (Table 3.7), Hernandez-Cruz, Giraldo-Soto et al. (2023) assessed 481 dwellings across six buildings in the Basque region. Lower-rated dwellings consumed 17%–30% less than predicted, while higher-rated dwellings consumed 90%–128% more. Heating demand was 60%–68% below estimates, while DHW averaged 38% above. Arbulu, Perez-Bezos et al. (2024) studied 126 social rental dwellings in northern Spain, showing discrepancies between normative and calibrated models. Incorporating occupant behaviour into the calibrated model improved accuracy (e.g., factor 0.97 vs. 0.82), underscoring the importance of behavioural modelling. In Austria (Table 3.7), Casian, Berger et al. (2023) conducted three simulated scenarios on five flats using. Results showed both widening and narrowing of the EPG depending on assumptions, with year-to-year variations further highlighting the impact of simulation scenario on the size of the EPG.

Table 3.7: Analysed records for residential and non-residential buildings in other countries in Europe

Reference	Building Type	Location	Period	Baseline, Unit &Details			Factor	Measured / Predicted
Anelkovi, Kljaji et al. (2021)	Apartment buildings (16 five-storey residential buildings)	Serbia	Annual Average 2018-2020	EPC Heating $kWh \cdot m^{-2} \cdot a^{-1}$	Buildings with District Heating Systems	B1	0.81	41.0 / 50.92
						B2	0.88	46.0 / 52.07
						B3	1.30	49.7 / 38.22
						B4	1.15	34.0 / 29.52
						B5	1.70	40.0 / 23.50
						B6	0.99	44.3 / 44.84
						B7	1.61	42.7 / 26.42
						B8	1.05	40.7 / 38.55
					Buildings with Individual Gas Boilers	B9	1.74	74.0 / 42.53
						B10	1.07	58.7 / 54.81
						B11	1.45	79.7 / 54.81
						B12	2.20	78.7 / 35.77
						B13	2.07	100.3 / 48.38
						B14	1.76	91.7 / 52.14
						B15	1.91	63.0 / 33.00
						B16	4.55	100.0 / 21.98
Motuzien, Lapinskien et al. (2021)	Apartment buildings	Lithuania	Annual	EPC Heating $kWh \cdot m^{-2} \cdot a^{-1}$	A rated building (43 buildings)		0.42-1.26	— / —
					A+ rated building (13 buildings)		0.46-1.73	— / —
					A++ rated buildings (6 buildings)		1.56-2.44	— / —
Motuziene, Lapinskiene and Rynkun (2024)	Offices		Jan 5 – Nov 27 2021	Theoretical calculations Heating $kWh \cdot m^{-2} \cdot a^{-1}$	Office one		1.37	59 / 43
			Jul 15- 23 Dec 2019		Office two		2.95	56 / 19
			Apr 7- Apr 28 2022		Office three		0.89	114 / 128
			May 5- Jul 30 2022		Office four		6.69	174 / 26
			—		Office five		6.56	59 / 9
			—		Office six		2.14	62 / 29
			—		Office seven		2.59	75 / 29
			—		Office eight		7.83	47 / 6
			—		Office nine		0.91	89 / 98
			—		Office 10		3.31	53 / 16
			—		Office 11		4.54	59 / 13
			—		Office 12		7.05	155 / 22
			—		Office 13		1.36	34 / 25
Katic, Krstic and Marenjak (2021)	Educational (185 Schools)	Federation of Bosnia and Herzegovina	Annual	Detailed energy audit calculation Heating $kWh \cdot m^{-2} \cdot a^{-1}$	Average actual energy for sample against average estimated		0.60	146 / 243
					Range of actual energy consumption against estimated range for 50% of the sample		0.53 0.63	82 / 155 193/ 304

Table continues on next page.

Table continued from previous page.

Reference	Building Type	Location	Period	Baseline, Unit &Details			Factor	Measured / Predicted
Ahmed, Hasu and Kurnitski (2022)	Educational buildings (Six nZEB schools and daycare centres)	Finland	Annual	EPC $kWh \cdot m^{-2} \cdot a^{-1}$	Measured (primary energy factor) ranges against predicted in 3 schools and 2 daycare centres Delivered heat, delivered electricity and primary energy Includes appliances and lighting		2.1-2.6	— / —
				IDA ICE Calibrated $kWh \cdot m^{-2} \cdot a^{-1}$	One daycare centre measured against simulated	Delivered heat	0.99	230 / 233
						Delivered electricity	1.02	112 / 110
						Primary energy	1.00	364 / 364
					One school measured against simulated	Delivered heat	1.00	— / —
						Delivered electricity	0.99	— / —
						Primary energy	0.99	— / —
Kotarela, Kyritsis et al. (2023)	Residential (Detached single-family dwelling)	Greece	Annual	EnergyPlus against TEE-KENAK quasi-steady-state calculation $kWh_e \cdot m^{-2} \cdot a^{-1}$	Before retrofit (G-rated)	Heating	1.09	106.6 / 97.8
						Cooling	0.86	6.1 / 7.1
						DHW	0.55	17.3 / 31.3
						Total	0.95	130.0 / 136.2
					After retrofit A+ rated based on dynamic simulation and B+ rated based on quasi-steady state	Heating	0.74	14.2 / 19.3
						Cooling	0.69	5.3 / 7.7
						DHW	0.26	9.1 / 35.4
						Total	0.46	29 / 62
Hernandez-Cruz, Giraldo-Soto et al. (2023)	Social housing apartments in multifamily residential blocks (481 dwellings distributed in six buildings)	Spain	Annual	EPC Total energy $kWh \cdot m^{-2} \cdot a^{-1}$	Energy consumption for sample		1.19	— / —
					EPG in 3 out of the six buildings (buildings with low EPCs)		0.70-0.83	— / —
					EPG in 3 out of the six buildings (buildings with high EPCs)		1.90-2.28	— / —
				EPC Heating $kWh \cdot m^{-2} \cdot a^{-1}$	Range for the sample		0.32-0.40	— / —
					Consumption in one building against predicted value		0.33	15.9 / 47.7
				EPC DHW $kWh \cdot m^{-2} \cdot a^{-1}$	Average for the sample		1.38	— / —
					Range in five out of the six buildings		0.71 -1.72	— / —
Arbulu, Perez-Bezos et al. (2024)	Social rental housing apartments in two multi-storey residential blocks (126 dwellings)		Monthly Dec 2020-31 Mar 2021	DesignBuilder Heating $Wh \cdot m^{-2}$	Normative model	Cluster one	0.82	11.24 / 13.68
						Cluster two	0.40	5.51 / 13.68
						Cluster three	0.19	2.55 / 13.68
						Cluster four	0.75	10.28 / 13.68
						Cluster five	0.36	4.88 / 13.68
						Cluster six	0.57	7.80 / 13.68
					Calibrated model using occupant-behaviour data	Cluster one	0.97	11.24 / 11.54
						Cluster two	0.46	5.51 / 12.37
						Cluster three	0.17	2.55 / 15.20
						Cluster four	0.78	10.28 / 13.17
						Cluster five	0.41	4.88 / 12.03
						Cluster six	0.67	7.80 / 11.58

Table continues on next page.

Table continued from previous page.

Reference	Building Type	Location	Period	Baseline, Unit &Details			Factor	Measured / Predicted
Casian, Berger et al. (2023)	Residential (three flats in a multi-unit building)	Austria	2017	EnergyPlus (OpenStudio) Space heating and DHW kWh·m⁻²	Simulation scenario 1	Flat 1/ Flat 3/ Flat 4	1.52/1.12/1.73	— / —
					Simulation Scenario 2		1.35/0.89/1.60	— / —
					Simulation Scenario 3		1.17/0.66/0.51	— / —
			2018		Simulation Scenario 1	Flat 1/ Flat 3/ Flat 4	1.60/1.16/1.71	— / —
					Simulation Scenario 2		1.47/0.94/1.56	— / —
					Simulation Scenario 3		1.32/0.72/1.41	— / —

3.3.2 Australia

The EPG in Australia has been examined across both residential and non-domestic contexts. Evidence presented in Table 3.8 illustrates the considerable variability observed across different building types and end-uses. Munsami, Prasad and Lan (2017) monitored three homes during the cooling season and reported discrepancies ranging from 60% above predictions to as high as 2,992%. When focusing on January, one dwelling aligned closely with predictions, another consumed only 14% of the predicted value, while the third showed demand 563% higher than expected. These findings, reflected in Table 3.8, underline the magnitude of household-level variation. Zou and Alam (2020) assessed a NABERS-certified office building and observed annual electricity consumption to be 11% above predictions, with monthly deviations spanning from 34% above to 7% below modelled values. Alam and Devjani (2021) reported even larger gaps in an educational building, with measured electricity and gas consumption exceeding predictions by 140% and 210%, respectively. Miller, Sodagar et al. (2021) reinforced the problem of inconsistent outcomes, finding heating electricity use to be 81% lower than predicted in one home yet 172% higher in another, with cooling demand exceeding predictions by 480%.

The role of calibration is evident in Kang, Chang et al. (2021) and Kang, Ma et al. (2022) who studied PH student accommodation. Initial results showed space heating demand at only 12% of predicted values and plug loads more than 18 times higher than assumed. However, through progressive calibration, heating demand narrowed to within 10% of predictions, plug loads converged to near parity, and lighting performance shifted from under- to over-estimation before finally settling below predicted values. As Table 3.8 highlights, these results demonstrate how iterative model refinement can substantially reduce the scale of apparent performance gaps.

3.3.3 North America

In North America, EPG research has covered a broad mix of residential, commercial, and public buildings, with the diversity of findings summarized in Table 3.8. In Canada, Fedoruk, Cole et al. (2015) observed electricity use in an educational building to be 29% higher than design values and 85% above performance objectives, yet 60% lower when compared to typical office benchmarks. Salehi, Terim Cavka et al. (2015) reported similar discrepancies in other educational buildings, where consumption exceeded IES and eQuest predictions by around 60%. Rouleau, Gosselin and Blanchet (2018) documented large positive gaps in 40 social housing units, while Cavka, Cavka and Salehi (2022) reported mixed outcomes: mechanical systems performed marginally better than predicted, but overall electricity use was 23% higher, with monthly deviations ranging from 7% to 62% above predicted. Mikhail, Mather et al. (2023) further disaggregated results for an office building, identifying HVAC loads as the principal driver of excess demand. Favourable performance was reported in the USA by Lin and Chen (2022), who found an nZEB office to consume slightly less than predicted (ratios of 0.92 and 0.97 across two operational years).

3.3.4 South America

In Brazil, Ing Liang, Menoncin Loper et al. (2020) showed that measured energy use in bank branches varied by 25% to 187% above EPC estimates, although results fell within the benchmark range for typical banks. This benchmark-dependence, also visible in Table 3.8, illustrates how the interpretation of EPGs changes depending on the chosen reference point. Similarly, Seguel-Vargas, Rubio-Bellido et al. (2024) identified overestimation of heating demand in Chilean residential buildings, with measured consumption amounting to only 62% of simulated (DesignBuilder) predictions.

Table 3.8: Analysed records for residential and non-residential buildings in Australia, Canada, Brazil and USA

Reference	Location	Building Type	Period	Baseline & Unit	Details		Factor	Measured / predicted
Munsami, Prasad and Lan (2017)	Australia	Residential (3 dwellings)	Cooling season October- March	BASIX predicted kWh	Cooling	Actual air-conditioning energy use against BASIX-predicted cooling load; House 8 / House 15 / House 25	4.38 30.92 1.60	759 / 173 1,608 / 52 857 / 598
			January				0.99 6.63 0.14	172 / 173 345 / 52 54 / 398
Zou and Alam (2020)		Office (NABERS certified)	Annual Jul 2018- Jun 2019	IES VE MWh	Measured electricity against predicted value		1.11	632.54 / 568.23
				Heating degree-days K·day	Actual heating degree-days against predicted heating degree-days		0.96	1,322 / 1,383
				Cooling degree-days K·day	Actual cooling degree-days against predicted cooling degree-days		1.42	62.35 / 44
			Monthly Jul 2018–Jun 2019	IES VE kWh	Variations in monthly EPG for electricity consumption *Approximate	Highest EPG	1.34	— / —
				IES VE kWh		Lowest EPG	0.93	— / —
Alam and Devjani (2021)		Educational	September 2018 to August 2019	Design-stage predicted consumption from Green Star energy efficiency report	Electricity MWh		2.4	— / —
					Gas GJ		3.1	— / —
Miller, Sodagar et al. (2021)		Detached single-family houses (2 houses)	Annual Jul 2014–Dec 2016	NatHERS / AccuRate $kWh{\cdot}m^{-2}{\cdot}a^{-1}$	Heating	House 2; House 3	0.24; 2.21	11.52 / 48.89; 52.38 / 23.72
					Cooling	House 2	6.49	17.13 / 2.64
					Heating and cooling	House 2; House 3	0.56; 1.90	28.65 / 51.53; 52.38 / 27.55
Kang, Chang et al. (2021)		Residential (PH student accommodation)	Annual Aug 2019–Jul 2020	PHPP $kWh{\cdot}m^{-2}{\cdot}a^{-1}$	Heating	Measured against PHPP-predicted value	0.12	0.9 / 7.3
				PH threshold $kWh{\cdot}m^{-2}{\cdot}a^{-1}$		Measured against PH threshold	0.06	0.9 / 15
Kang, Ma et al. (2022)		Residential (PH student accommodation)	Annual	PHPP $kWh{\cdot}m^{-2}{\cdot}a^{-1}$	Heating	Measured against three PHPP model configurations	0.12; 0.45; 0.90	0.9 / 7.3; 0.9 / 2.0; 0.9 / 1.0
					Plug loads		1.53; 1.16; 0.70	25.1 / 16.4; 25.1 / 21.6; 25.1 / 25.9
					Lighting		19.22; 1.25; 0.96	17.3 / 0.9; 17.3 / 13.8; 17.3 / 18.1

Table continues on next page.

Table continued from previous page.

Reference	Location	Building Type	Period	Baseline & Unit	Details		Factor	Measured / predicted
Fedoruk, Cole et al. (2015)	Canada	Educational	Annual Apr 2012- Mar 2013	Design estimate $MWh\cdot a^{-1}$	Electricity	Measured against design estimate	1.29	755 / 585
				Design target $kWh\cdot m^{-2}\cdot a^{-1}$		Measured against design target	1.86	130 / 70
Salehi, Terim Cavka et al. (2015)		Educational	Annual Apr 2012–Mar 2013	IES-VE $kWh\cdot m^{-2}\cdot a^{-1}$	Electricity	Measured against predicted	1.59	125.5 / 78.9
				eQuest $kWh\cdot m^{-2}\cdot a^{-1}$			1.61	125.5 / 78.0
Rouleau, Gosselin and Blanchet (2018)		Multi-residential social housing (40 apartments)	Annual 2016	PHPP model $kWh\cdot m^{-2}\cdot a^{-1}$	Total energy	Measured against PHPP predicted	1.74	129.1 / 74.3
					Heating		2.22	36.9 / 16.6
Cavka, Cavka and Salehi (2022)		Educational	Annual Mar 2012-2013	2010 BC Hydro $MWh\cdot a^{-1}$	Electricity	Actual electricity consumption in the first year of operation against intended	1.23	— / —
					Mechanical systems	Mechanical-system energy consumption against predicted	0.96	612 / —
					HVAC	Monthly HVAC range	1.07-1.62	— / —
						Average monthly HVAC	1.42	— / —
Mikhail, Mather et al. (2023)		Office	Annual 2019	EnergyPlus $kWh\cdot m^{-2}\cdot a^{-1}$	Total energy		1.21	96.9 / 79.8
					Heating, cooling, fans and pumps		1.58	43.6 / 27.6
					Interior equipment		0.85	29.3 / 34.5
Ing Liang, Menoncin Loper et al. (2020)	Brazil	11 Banks	Annual 2014-2016	INI-C calculation, $kWh\cdot m^{-2}\cdot a^{-1}$	Total energy consumption Electricity	Actual energy against calculated INI-C total; range across 11 bank branches	1.25-2.87	96.73–183.50 / 56.62–86.04
Seguel-Vargas, Rubio-Bellido et al. (2024)	Chile	Residential	Annual	DesignBuilder $kWh\cdot m^{-2}\cdot a^{-1}$	Heating	Measured against simulated	0.62	125.7 / 202.0
Lin and Chen (2022)	USA	Office (Net zero energy)	Annual First year of operation Annual Fifth year of operation	EnergyPlus Validated with first- and fifth-year measured energy data $kWh\cdot m^{-2}\cdot a^{-1}$	Electricity	Total measured building electricity-use intensity in first year against simulated electricity-use intensity	0.93	43.93 / 47.37

3.3.5 Asia

Research into the magnitude of the EPG in Asia and the Middle East reveals both favourable and unfavourable outcomes across a diverse range of building types, as summarised in *Table 3.9* In China, Geng, Ji et al. (2018) reported efficient outcomes in an office building, where actual baseload consumption was 12% lower than predicted. A more extensive analysis by Wang, Pang et al. (2020) across four offices and one hotel highlighted the variability of EPGs depending on system type and modelling method. Static EPGs for the offices ranged from 34% to 93% below predictions, while the hotel exhibited a small positive gap of 6%. By contrast, dynamic modelling showed EPGs between 3% and 53% above predicted values for the offices and 24% for the hotel. HVAC systems were consistently problematic, with offices ranging from 7% to 114% above predicted and the hotel recording an 88% gap. Non-HVAC dynamic EPGs were favourable, ranging from 1% to 5% below predictions. These findings, reflected in Table 3.9, indicate that the choice of modelling approach and system boundary substantially affects reported gaps.

System and season-specific variability is further illustrated in Bai, Yu and Pan (2023) , who reported an annual EPG of 4% in an office building, with cooling demand up to 49% above predictions and lighting demand up to 79% below. Bao and Meng (2023) reinforced the importance of calibration through a study of a cooling plant: under two simulation scenarios, chiller energy consumption ranged from 17% below to 0.08% above predicted values, with measured performance improving to 8% below predictions once the model was refined. Bai, Pan and Yu (2024) extended this work to high-rise residential buildings, showing annual measured energy use of 175 kWh/m² compared to 145 kWh/m² in the as-designed model, yielding a factor of 1.21. After calibration incorporating occupancy data, the factor narrowed to 1.01, demonstrating how passive design measures and accurate representation of occupant behaviour can reduce discrepancies.

In South Korea, Ahn, Kim et al. (2016) recorded gas use 61% higher than predicted and electricity use 6% lower in a non-residential building. Ji, Hong et al. (2022) broadened the scope to 364 non-residential buildings, finding regional variations but little correlation with building ratings: in the southern region, EPG factors ranged from 0.88 to 1.83, while in the central region they ranged from 1.45 to 1.98. Yoo, Shin et al. (2023) examined 302 buildings and found strong correlations between ratings and heating EPGs, but none for cooling or total energy use. A follow-up study

(Yoo, Shin et al. 2024) analysed 158 certified buildings under the national energy rating system, highlighting that total energy use in the highest-rated buildings (1+++) was nearly three times higher than predicted (factor 2.79). Heating gaps were particularly pronounced (factor 5.81), while cooling demand was lower than expected (factor 0.66). Differences by building type were also marked, with hotels showing the largest total EPG (factor 3.77). As reflected in *Table 3.9*, these studies reveal the complexity of translating ambitious ratings into consistent performance outcomes.

In Malaysia, Zaid and Kiani (2016) found that a green office consumed 26% less energy than planned, but still 61% more than comparable conventional buildings. A subsequent study by (Zaid, Rad and Zainon 2017) showed the same office consumed 18% less energy than a selected code of practice, highlighting the tension between design targets and benchmarking choices.

Mixed findings have also been reported in the United Arab Emirates (UAE). Kim, Bande et al. (2019) analysed an educational building and found that in-use energy consumption was only 2% higher than Abu Dhabi code requirements but 33% greater than the as-designed model, while still 5% below a baseline model. Disaggregation showed EPGs of 30% for lighting, 39% for fans, and 24% for cooling. Belpoliti (2021) reported even larger initial discrepancies in a prototype dwelling, where cooling and lighting demand exceeded predictions by 116% during a 10-day monitoring period. After refining the energy model, the gaps narrowed substantially, with cooling and lighting demand measured at 6% below predictions, and appliance demand at 3%–4% below. In Iran, Zare, Saryazdi et al. (2023) studied four flats on different floor levels and highlighted how performance varied significantly depending on location within the building. By incorporating actual weather data, boiler efficiency, and survey-based occupancy information across three calibration stages, the researchers demonstrated the crucial role of accurate input data in narrowing performance gaps. Their results, also summarized in *Table 3.9*, underscore how context-specific and data-driven modelling is essential to aligning predictions with real-world outcomes.

Table 3.9: Analysed records for residential and non-residential buildings in Asia

Reference	Location	Building Type	Period	Baseline & Unit	Details		Factors	Measured / Predicted
Geng, Ji et al. (2018)	China	Office	Annual (2016)	Regression result from simulated reference/design-stage energy model $kWh\cdot day^{-1}$	Lighting, plug and auxiliary systems	Actual bill-weather regression baseload against simulated reference-model baseload	0.88	822 / 934
Wang, Pang et al. (2020)		Office Three-star green label	Annual	EnergyPlus static & dynamic reconciled models $kWh\cdot m^{-2}\cdot a^{-1}$	Total static Total dynamic HVAC dynamic non-HVAC dynamic		0.35 1.03 1.08 0.98	23.9 / 68.6 23.9 / 23.2 12.3 / 11 11.6 / 11.8
		Hotel Two-star green label					1.06 1.24 1.88 0.99	160 / 151 160 / 128 68 / 36 92 / 92
		Office Two-star green label					0.14 1.11 1.46 0.95	10.8 / 75.4 10.8 / 9.7 4.4 / 3.0 6.4 / 6.7
		Office Three-star green label					0.67 1.33 2.14 0.98	50.8 / 75.5 50.8 / 38.0 24.9 / 11.6 25.9 / 26.4
		Office Two-star green label					0.93 1.53 1.78 0.98	66.3 / 71.1 66.3 / 43.2 52.1 / 29.2 14.2 / 14.0
Bai, Yu and Pan (2023)		Office	1 Jun 2018 – 30 Apr 2019	Whole building annual EPG: Measured against simulated MWH			1.04	3,584 / 3,485
				Whole building range of monthly EPG MWh/month			0.89-1.35	— / —
				Dynamic EPG: Average hourly measured against simulated $kWh\cdot h^{-1}$			1.04	448 / 430
				Annual lighting EPG/ April/ July MWh			0.25/0.27/0.21	— / —
				Annual cooling EPG/ April/ July MWh			1.49/1.98/1.31	— / —
				Whole EPG in July			1.01	— / —
Bao and Meng (2023)		Cooling plant (Chillers)	July 20 - August 17	EnergyPlus $MWh\cdot day^{-1}$	Operation mode one: Low load	Measured against simulated average daily chiller energy consumption	0.83	4.9 / 5.9
					Operation mode two: Medium load		0.92	9.3 / 10.1
					Operation mode three: High load		1.01	12.7 / 12.6
Bai, Pan and Yu (2024)		Residential	Annual 2018	DesignBuilder $kWh\cdot m^{-2}\cdot a^{-1}$	Measured against as designed model		1.21	175 / 145
					Measured against as occupied model (calibrated)		1.01	175 / 173

Table continues on next page.

Table continued from previous page.

Reference	Location	Building Type	Period	Baseline & Unit	Details		Factors	Measured / Predicted
Ahn, Kim et al. (2016)	South Korea	Office	Annual	DesignBuilder	Gas $Nm^3 \cdot a^{-1}$	Uncalibrated	1.62	426,132 / 262,383
						Calibrated	0.99	426,132 / 432,663
					Electricity $kWh \cdot a^{-1}$		0.95	10,257,547 / 10,831,690
Ji, Hong et al. (2022)		Non-residential (364 certified buildings)	Annual	ECO2 tool $kWh \cdot m^{-2} \cdot a^{-1}$	BEEC level 1+++	Southern region	1.20	92.0 / 76.4
					BEEC level 1++	Central region Southern region	1.99	236.2 / 118.7
							1.22	135.9 / 111.8
					BEEC level 1+		1.73	289.5 / 167.5
							1.39	220.3 / 158.7
					BEEC level 1		1.46	334.4 / 229.4
							1.25	288.3 / 230.0
					BEEC level 2		1.48	422.3 / 284.7
							0.89	250.8 / 282.5
Yoo, Shin et al. (2023)		Non-residential buildings: office, educational, hotel and multi-type buildings (302 Buildings)	Annual	ECO2 tool $kWh \cdot m^{-2} \cdot a^{-1}$	Total Heating Cooling	1+++	1.61	116.7 / 72.7
							3.48	54.3 / 15.6
							0.52	9.1 / 17.5
						1++	1.92	280.6 145.8
							2.10	96.8 / 45.9
							1.44	30.5 / 21.1
						1+	2.16	400.0 / 184.9
							1.70	97.5 / 57.1
							1.85	53.6 / 28.9
						1	1.63	404.7 / 246.9
							1.35	89.9 / 66.3
							0.98	56.3 / 57.4
						2	1.95	616.2 / 315.1
							0.99	113.1 / 114.2
							1.23	86.3 / 70.1

Table continues on next page.

Table continued from previous page.

Reference	Location	Building Type	Period	Baseline & Unit	Details		Factors	Measured / Predicted
Yoo, Shin et al. (2024)	South Korea	158 non-residential certified buildings	Annual	ECO2 tool kWh·m⁻²·a⁻¹	Total	1+++	2.79	192.3 / 68.9
						1++	2.50	299.4 / 120
						1+	2.67	440 / 164.8
						1	1.98	457.1 / 231.3
						2	2.18	672.5 / 287.9
					Heating	1+++	5.81	118 / 20.3
						1++	2.19	96.7 / 44.1
						1+	1.88	106.8 / 56.7
						1	1.39	98.1 / 70.5
						2	0.97	113 / 116.6
					Cooling	1+++	0.66	10.3 / 15.5
						1++	1.48	31.6 / 21.4
						1+	1.84	57.8 / 31.4
						1	1.11	63.8 / 57.6
						2	1.09	70.2 / 64.5
					Based on chiller type	Electric	1.81	56 / 30.9
						Gas	1.25	43 / 34.4
						Mixed	1.07	48.5 / 45.4
					Based on building type	Office	1.92	353.1 / 184.4
							1.38	78.6 / 57.1
							1.07	45.9 / 42.7
						Education	1.99	288.8 / 144.9
							1.63	84.5 / 51.8
							1.51	43.6 / 28.9
						Hotel	3.77	568.3 / 151
							2.91	167.0 / 57.4
							2.23	66.0 / 29.6
						Multi-apartment	2.98	539.0 / 180.7
							1.69	112.7 / 66.7
							1.74	52.0 / 29.9
Zaid and Kiani (2016)	Malaysia	Office	Annual 2013–2014	Intended design performance kWh·m⁻²·a⁻¹	Electricity		0.74	0.74
				Conventional office building kWh·m⁻²·a⁻¹			1.61	111 / 69
Zaid, Rad and Zainon (2017)		Office	Annual 2013–2014	Intended design performance kWh·m⁻²·a⁻¹	Electricity		0.74	111 / 150
				MS 1525:2007 recommended benchmark for non-residential buildings kWh·m⁻²·a⁻¹			0.82	111 / 135
				Conventional office building kWh·m⁻²·a⁻¹			1.61	111 / 69

Table continues on next page.

Table continued from previous page.

Reference	Location	Building Type	Period	Baseline & Unit	Details		Factors	Measured / Predicted
Kim, Bande et al. (2019)	UAE	Educational	Annual 3 Jun– 1 Jul	Trace 700 kWh·a⁻¹	Base utilities, Lights, Fans, cooling, heating	In use model against: Abu Dhabi Code/As designed model/Baseline model (ASHRAE 90.1 Appendix G technique)	1.02 / 1.33/ 0.95	— / —
					Lighting	In use model against: As designed model	1.30	— / —
					Fans		1.39	— / —
					Cooling		1.24	— / —
Belpoliti (2021)		House prototype	18-27 Nov	DesignBuilder kWh	Cooling and lighting		2.17	231 / 106
			21-26 Nov		Cooling and lighting		0.94	60 / 64
			18-27 Nov		Appliances		0.97	127 / 130
			21-26 Nov		Appliances		0.96	75 / 78
Zare, Saryazdi et al. (2023)	Iran	Residential apartment block (four flats)	Annual Oct 2022– Mar 2023	ASHRAE design-stage assumptions, airport EPW file, boiler catalogue efficiency kWh·m⁻²	Heating (gas)	First floor Second floor Third floor Fourth Floor	3.79 4.43 2.03 2.59	194.2 / 51.3 149.4 / 33.7 68.5 / 33.7 202.8 / 78.3
				ASHRAE assumptions with actual boiler efficiency of 56% kWh·m⁻²			2.12 2.49 1.14 1.45	194.2 / 91.4 149.4 / 60.1 68.5 / 60.1 202.8 / 139.5
				ASHRAE assumptions with real EPW file and actual boiler efficiency kWh·m⁻²			22.0 2.33 1.07 1.38	194.2 / 94.8 149.4 / 64.0 68.5 / 64.0 202.8 / 146.5
				Questionnaire-based calibrated simulation with real EPW file and measured boiler efficiency kWh·m⁻²			1.42 1.26 1.06 1.15	194.2 / 137.1 149.4 / 118.3 68.5 / 64.5 202.8 / 176.2

3.4 Summary

The first part of the SLR (*Section 3.2*) makes clear that the EPG phenomenon cannot be understood simply as a difference between simulated and measured energy performance. Rather, it emerges from a complex interplay of technical, organizational, and behavioural factors that extend across all stages of the building lifecycle. At the design stage, recurring problems include miscommunication of performance targets, issues with design itself whether its efficient or inefficient (complex) and issues that stem from uncertainties when conducting building energy modelling. At the construction stage, persistent discrepancies between design intent and as-built reality are widely documented. These stem from poor workmanship, inappropriate material substitutions, and lapses in quality assurance, often shaped by procurement practices that prioritize cost and time over performance integrity.

Commissioning and operational stages introduce additional layers of complexity. Handover processes are frequently incomplete, with insufficient commissioning, limited verification, and inconsistent testing protocols constraining performance outcomes. Once in use, buildings are strongly shaped by occupant behaviour, with cultural practices, comfort expectations, and limited system knowledge producing energy use patterns that diverge from predictions. These effects are compounded by equipment malfunctions, poorly tuned controls, inadequate maintenance regimes, and shortages of skilled FMs. Furthermore, building performance evolves dynamically over time, reflecting changes in use, occupancy, and environmental conditions that complicate the maintenance of efficiency. Beyond stage-specific causes, a set of cross-cutting factors intensifies the persistence of the EPG. Among these, lack of accountability for energy outcomes, lack of communication and collaboration between stakeholders, and prioritizing cost minimization over long-term performance. Skills shortages and knowledge deficits across the delivery chain, coupled with an entrenched construction culture resistant to change, further restricts the capacity to address these challenges systematically.

Some of the causes commonly cited as contributing to the EPG may in fact reflect limitations in modelling assumptions or changes in operational conditions rather than clearly demonstrated deficiencies in design choices, construction quality or energy systems. This underscores the need for a more explicit distinction between factors that should be recognised as genuine causes of the EPG, meaning those that directly influence energy use, and factors that are merely perceived or inferred as a result of imperfect modelling and no standardised measurement practices. Establishing such a distinction, together with the development of standardised procedures for simulation and measurement, would enable more accurate and rigorous diagnosis of both the origins and the magnitude of the EPG.

In summary, Section 3.2 shows that the EPG is produced by a combination of stage-specific and cross-cutting causes. Design-stage causes shape predicted performance through assumptions, targets, modelling choices and specifications. Construction-stage causes determine whether design intent is realised in the as-built building. Commissioning and operational causes influence whether systems perform as intended once the building is occupied. Cross-cutting factors, particularly accountability, communication, incentives, industry culture, documentation and skills, affect all stages and help explain why the EPG remains persistent. This synthesis provides the foundation for Chapter Four, where these literature-derived causes are examined more deeply through the 5 Whys RCA and then used as the analytical basis for assessing the potential of NABERS and related policy interventions.

On the other hand, the SLR on the magnitude of the EPG (Section 3.3) highlights a series of critical lessons for research and practice. Reported values of the EPG must be interpreted with caution, as they are highly sensitive to the choice of benchmark and modelling approach, whether compliance (compliance gap), performance (static gap), or calibrated (dynamic gap) models are employed. Without methodological consistency and explicit treatment of uncertainty, numerical estimates risk misrepresenting actual magnitude of the EPG and being misleading rather than informative. The challenge of comparability across studies further complicates attempts to establish reliable conclusions. Correlations can only be drawn when the EPG is quantified against the same benchmark and for the same end uses, yet such consistency is rare. The widespread variation in parameters and reporting practices combined represents a fundamental barrier to building a coherent evidence base.

The literature also demonstrates that the EPG is not a static value but a dynamic phenomenon that shifts over time and in response to interventions. Longitudinal studies provide evidence that gaps frequently widen, although optimization measures, improved calibration of models, or adjustments such as refined heating setpoints have been shown to reduce discrepancies. For instance, the narrowing of performance factors from 0.97 to 0.70 following targeted optimization illustrates the potential of technical interventions to deliver tangible improvements (Flourentzos, Ivanov Antonov and Pantet 2019).

Technical and operational choices further influence outcomes. Buildings connected to district heating systems exhibited smaller gaps than those relying on individual gas boilers (Anelkovi, Kljaji et al. 2021), while delivery strategies such as SL and natural ventilation supported closer alignment between predicted and measured performance (Thompson, Burman et al. 2022). These findings emphasize that energy outcomes are determined as much by procurement and operational practices as by design intent.

Evidence on building ratings and regulatory frameworks reveals complex relationships with the EPG. In some cases, highly rated buildings exhibit wider gaps than conventional buildings, despite consuming less energy in absolute terms. Case studies from Switzerland and Denmark indicate that newer rating schemes may achieve narrower gaps than earlier versions, while research from South Korea suggests that links between ratings and the EPG are more apparent for heating than for overall energy consumption. This paradox of high ratings combined with persistent gaps signals a need for more nuanced understanding of performance certification.

Construction period has also emerged as a relevant factor (Cozza, Chambers and Patel 2020). In several contexts, newer buildings displayed wider EPGs than older ones, raising concerns that technological and regulatory advancements are not consistently translating into better real-world outcomes. At the same time, comparisons with stringent performance standards such as PH underscore the seriousness of the problem, showing how far conventional buildings fall short of ambitious energy efficiency targets.

On the other hand, the SLR shows that the magnitude of the EPG is highly dependent on how its defined, calculated, and reported. Across the reviewed studies, there is no single standardised method for quantifying the EPG. In some cases, the gap is calculated by comparing measured energy performance with simulated performance, although the simulation baseline may vary between compliance model (regulatory gap), performance model (static gap), calibrated model (dynamic gap). In other cases, measured performance is compared with PH targets, simulated using PHPP, nZEB targets, regulatory benchmarks, good-practice benchmarks, typical-practice benchmarks, or statistical datasets representing comparable buildings within a local, regional, or national context.

This variation in reference points has a direct influence on the magnitude and interpretation of the EPG. Studies such as Mitchell and Natarajan (2020) which report the EPG against multiple benchmarks, illustrate that the same building can appear to have different EPG values depending on the baseline used. A building may perform favourably when compared with typical energy benchmark, but less favourably when assessed against more ambitious benchmarks such as PH or nZEB targets. Conversely, a building may meet a regulatory or good-practice benchmark while still falling short of its original design-stage simulated energy performance on PHPP or PH performance expectation.

Within the PH context, this distinction is particularly important because comparing measured energy use against PH targets indicates the extent to which a building meets the specific performance requirements associated with the PH standard, rather than simply whether it performs better or worse than broader regulatory or stock-based benchmarks. However, when measured performance is compared with broader regulatory benchmarks, the EPG reflects a different form of performance assessment. It indicates how far a building is from meeting external policy, market, or comparative performance expectations rather than its specific design intent. Therefore, evidence on the EPG in PH buildings may appear inconsistent where studies use different baselines, but such variation often reflects methodological differences rather than direct contradiction.

The reviewed literature therefore suggests that the EPG should not be interpreted as a fixed or absolute value. Instead, its meaning depends on the selected benchmark and the purpose of the assessment. Quantifying the EPG against different reference points can be useful where the aim is to understand how far a building or a group of buildings at a neighbourhood, city, or national stock are from achieving particular performance targets. This approach can also support the identification of appropriate interventions, ranging from minor operational adjustments to more substantial improvements in airtightness, building fabric, HVAC systems, lighting, commissioning, or occupant-related practices.

Overall, the SLR indicates that the absence of a standardised method for calculating and reporting the EPG remains a key methodological limitation in the field. Nevertheless, the use of multiple benchmarks can provide a more nuanced understanding of building performance, particularly where different stakeholders, regulatory frameworks, or client requirements define success in different ways.

The SLR covers the full range of objective EPG typologies presented in Figure 1.1, excluding the subjective/perception-related dimension, which falls outside the scope of the thesis. However, the analysis also extends beyond these typologies by identifying additional conditions under which EPG magnitude may widen or narrow and energy consumption may increase or decrease as summarised in Table 3.10 below. Section 3.3 provides extensive evidence on EPG magnitudes across 27 countries and five continents, but direct cross-study correlations remain difficult because studies use different baselines, energy-use categories, calculation methods, modelling assumptions, and reporting conventions. The importance of this section therefore lies not only in demonstrating coverage of established EPG typologies and the magnitude of the EPG calculated for each typology, but also in showing that EPG magnitude is shaped by a broader set of conditions, including whether they are associated with both the widening or narrowing of EPG magnitude and increases or decreases in energy consumption.

Table 3.10: Conditions influencing the widening or narrowing of EPG magnitude and changes in energy consumption identified through the SLR

Condition		**Influence**	**References**
Air change rate	Increased air change rate	Wider EPG	Vivian, Carnieletto et al. (2023)
	Decreased air change rate	Narrower EPG	
Calibration in relation to occupancy data	No occupancy data used	Wider EPG	Maracchini, Latini et al. (2024)
	Occupancy data	Narrower EPG	
Soft Landings	SL adopted	Lower energy consumption	Thompson, Burman et al. (2022)
	SL not adopted	Higher energy consumption	
Procurement	Design and build	Narrower EPG and higher energy consumption	
	Traditional	Wider EPG and lower energy consumption	
Ventilation strategy	Mechanical	Higher energy consumption and wider EPG	
	Natural	Lower energy consumption and narrower EPG	
Build date	Older buildings	Narrower EPG and higher energy consumption	Reguis, Tunzi et al. (2023)
	Newer buildings	Wider EPG and lower energy consumption	
Energy system	District Heating Systems	Narrower EPG and lower energy consumption	Anelkovi, Kljaji et al. (2021)
	Individual Gas Boilers	Wider EPG and higher energy consumption	

A final observation concerns the geographical distribution of research, which remains heavily skewed toward the western world, with the UK standing out as the dominant context in which the phenomenon has been theorized and empirically examined. This emphasis reflects the historical origins of the EPG concept, which was first articulated in the UK, as well as biases in building performance research, where countries with established regulatory frameworks often shape methodological norms and influence the direction of scholarly inquiry. Technological and contextual factors also contribute to this pattern. In regions such as Scandinavia, where buildings are generally more energy efficient, the EPG has attracted comparatively limited research interest, reducing its visibility as a policy or scientific issue.

Beyond Europe, the geographical coverage becomes significantly thinner. Notably, no EPG studies were identified from the African continent, and only one study was found from the Middle East, highlighting substantial regional blind spots. These gaps limit the extent to which observations through literature can be generalized globally and obscure how differing climatic conditions, construction practices, cultural expectations, and regulatory maturities shape the magnitude and evolution of the EPG. Addressing this lack of diversity is therefore essential for developing a more robust and contextually sensitive evidence base.

4. Mitigation Strategies through Policy

This chapter introduces the analysis and synthesis of findings in a structured way, focusing on the investigation of mitigation strategies through policy. Section 4.1 outlines the research design, participant selection, and analytical approach. Section 4.2 presents the approach followed to conduct the root cause analysis using the "5 Whys" technique to trace the underlying causes of the EPG identified in Chapter Three. Section 4.3 reports the results of the semi-structured interviews. The analysis is organized into four parts: high level quantitative analysis of interview results (4.3.1) followed by detailed qualitative and quantitative design stage insights from NABERS Independent Design Reviewers (IDRs) (4.3.2); commissioning and operational stage insights from NABERS Accredited Assessors (AAs) (4.3.3); cross-cutting factors identified through discussions with both IDRs and AAs (4.3.4). Section 4.4 provides a micro-level discussion and develops a framework for the mitigation of the EPG informed by expert's insights. It interprets the results, maps NABERS' potential to mitigate the EPG and identifies limitations, gaps, and areas for further development.

4.1 Detailed Methodology

This section sets out the methodological framework adopted to explore the potential for mitigating the EPG through approaches that emphasize measured performance outcomes rather than predicted values.

4.1.1 Root Cause Analysis

The methodological foundation for this chapter is the RCA of EPG factors, conducted using the "5 Whys" technique. This simple and often powerful approach to solving problems encourages deep thinking through repeated questioning, and it can be adapted quickly for use with most issues. Most importantly, the method is tied to the idea of systematic problem solving. Without that guiding intent, the method becomes only an empty outline of the real process. Effective use of the technique depends on three key factors: first, stating problems accurately and completely; answering each question with full honesty; and having the determination to reach the root of the problem and resolve it (Serrat 2017).

The “5 Whys” technique, originally developed by Toyota as part of its lean problem-solving approach (Ohno 1988). Its iterative questioning structure allows surface-level issues to be traced back to systemic drivers, making it valuable in contexts where problems emerge across multiple stages. Prior studies show that “5 Whys” technique helps move beyond immediate technical faults to uncover organizational and behavioural factors, leading to more durable solutions (Murugaiah, Benjamin et al. 2010). Applied to the EPG, the method offers a practical and systematic way to integrate insights across lifecycle stages and expose the interdependencies that sustain performance shortfalls.

Here, the 5 Whys technique is applied to the causes of the EPG identified in Chapter Three. Each cause identified in prior literature is examined for how and why it arises across the building lifecycle, from design, construction to commissioning and operation and cross-cutting factors. The purpose of this analysis is not to generate a deterministic map of causes, but rather to reveal the interdependencies that produce the EPG. This structured process provides the conceptual grounding for evaluating how NABERS may intervene in various stages.

While the “5 Whys” technique traditionally continues through five iterative layers of questioning, in this analysis the number of “whys” varied according to the point at which further questioning no longer provided new explanatory insight. This reflected an analytical saturation point within the RCA, whereby additional iterations did not generate new root causes but instead reiterated previously identified explanations, consistent with the principle of diminishing analytical returns in qualitative inquiry. For several causes identified in the literature, the root cause became clear after the second or third iteration, at which point additional questioning would simply repeat or reinforce the same explanation rather than uncover a deeper causal factor.

This was evidenced through cross-comparison of the 17 reviewed records, where recurring causal pathways converged at similar levels of explanation, indicating that further iterations were unlikely to yield substantively new insights. In these instances, the point of stopping was consistently supported across multiple sources rather than by extending the questioning to an arbitrary fifth level. Accordingly, saturation within the “5 Whys” RCA was defined as the point at which no new causal layers emerged, rather than as adherence to a fixed number of iterations. The stopping point was determined by when the inquiry reached what could reasonably be considered an underlying cause supported across the reviewed studies.

In this way, the analysis remained consistent with the intention of the "5 Whys" technique, which is to reveal fundamental drivers, while avoiding unnecessary extension of the questioning beyond the point of meaningful analytical value. It is also important to recognise that several causes identified within Section 4.2 emerge within interconnected causal structures, where a given factor may simultaneously act as both a root cause and a consequence of other conditions. For example, lack of accountability may be rooted in broader issues related to overall industry culture, while lack of communication and collaboration can emerge as a downstream effect associated with lack of accountability.

4.1.2 Semi-Structured Interviews and Coding

Following the RCA, semi-structured interviews were conducted with two stakeholder groups: NABERS IDRs and AAs. Semi-structured interviews were selected because they balance consistency across participants with flexibility to pursue emergent themes (Gill, Stewart et al. 2008, Kallio, Pietilä et al. 2016). This approach is suited to exploring the contextual and technical complexity of the EPG, as it allows participants to elaborate on their experiences while still addressing predefined themes linked to the RCA.

The semi-structured interviews were conducted online via Zoom over a twelve-month period. Each interview lasted between thirty minutes and over one hour depending on the depth of discussion. All interviews were recorded, transcribed using an online transcription tool, and then manually coded against the causes identified in Chapter Three. Manual coding of interviews fosters immersion, reflexivity, and closeness to the data. Mattimoe, Hayden et al. (2021) highlight that manually working through transcripts promotes a deeper engagement with the data and enables researchers to notice nuances that might otherwise be overlooked when relying solely on software. Similarly, Skjott Linneberg and Korsgaard (2019) argue that manual coding is especially valuable for smaller projects, as it enhances transparency and analytical rigour by keeping the researcher closely connected to participants' words. In total, twelve interviews were conducted: five with IDRs and seven with AAs. To protect anonymity, participants are referred to as IDR1–IDR5 and AA1–AA7 throughout the thesis.

Semi-structured interviews were analysed using a deductive coding approach, in which predefined root causes of the EPG based on the SLR and RCA were held constant throughout the analysis.

Each interview transcript was manually coded against this fixed analytical framework, with participants asked to evaluate the extent to which NABERS addresses each predefined root cause. Coding was conducted in two complementary ways. First, a quantitative layer of analysis was implemented through a colour-coded matrix, in which each root cause was assigned one of five evaluative categories: addressed by NABERS (green), policy intervention required (blue), uncertain (grey), elaboration without assigning a category (uncategorised), or remains out of control (red). This enabled a structured comparison of evaluative judgments across participants and the calculation of distributions by root cause and stage. Second, a qualitative layer of analysis was conducted through manual coding of interview transcripts to capture the reasoning, explanations, and contextual nuances underlying participants' judgments. This dual approach allowed not only the aggregation of evaluative positions but also the interpretation of differences in perspective, consistent with mixed qualitative–quantitative coding strategies in applied qualitative research (Miles, Huberman and Saldaña 2014).

The decision to conduct twelve interviews is justified on the basis of interpretive or meaning saturation, rather than thematic saturation. Meaning saturation is defined as the point at which additional interviews do not yield new evaluative positions or explanatory mechanisms regarding the predefined root causes indicating that the data were sufficient to address RO2 and the part of RO5 associated with this chapter. Given the deductive design of this part of the thesis, where root causes were predefined through the SLR and RCA, saturation was not assessed in terms of the emergence of new themes. Interpretive saturation, aligned with the concept of meaning saturation, refers to the point at which issues are fully understood, and no new insights emerge regarding the meaning, explanatory mechanisms, or relationships associated with existing codes or categories. In contrast to code or thematic saturation, which focuses on the repetition of themes, meaning saturation emphasises the depth, nuance, and explanatory richness of interpretations, requiring that no additional interviews yield new interpretive dimensions or mechanisms (Hennink and Kaiser 2022, Rahimi and Khatooni 2024).

Empirical evidence supports the adequacy of sample sizes in the range of approximately 9-17 interviews for achieving saturation in studies with relatively homogeneous participants and narrowly defined research objectives (Hennink and Kaiser 2022). Within this range, earlier interviews typically contribute the majority of new insights, while later interviews predominantly reinforce existing interpretations. In this research, the participant group consisted of professionals with closely related expertise in NABERS and building performance analysis, and the analytical focus was constrained to a predefined set of root causes, thereby aligning with conditions under which saturation is expected to be achieved within a limited sample.

Evidence of saturation was assessed through two complementary indicators. First, a cumulative analysis of interview coding demonstrated diminishing returns in interpretive insights, whereby early interviews contributed diverse evaluative positions, while subsequent interviews largely confirmed and elaborated existing interpretations without introducing substantively new explanatory mechanisms. Second, the colour-coded matrices showed convergence in evaluative patterns across participants, where variation in responses, such as policy intervention required (blue) or elaboration without assigning a category (uncategorized) reflected differences in interpretation and category selection rather than the identification of new issues. This pattern is consistent with the transition from code saturation to meaning saturation described in the literature, where interpretive depth stabilises even as minor variation in expression persists (Hennink, Kaiser and Marconi 2017).

The colour-coded matrices were used as the main evidence for assessing convergence and interpretive saturation. For each predefined root cause, participant responses were coded to determine whether later interviews introduced new evaluative positions or mainly reinforced existing patterns of interpretation.

The evidence suggests that interpretive saturation was broadly reached within the predefined analytical framework, particularly for design-stage causes, where participants repeatedly converged around similar evaluative positions. However, saturation was not uniform across all analytical categories. Commissioning and operational causes produced a higher number of uncategorised responses, indicating that participants often elaborated on these issues without assigning a definitive evaluative category. These areas are therefore treated as more exploratory and are acknowledged as a methodological boundary of the analysis rather than as fully saturated themes.

Saturation was therefore considered to have been reached for the predefined analytical framework when additional interviews did not yield new evaluative positions or explanatory mechanisms regarding the predefined root causes, indicating that the interpretive scope of the analytical framework had been sufficiently explored. Where evaluative disagreement remained, this was interpreted as evidence of difference in expert judgement based on their experience rather than the emergence of new analytical categories. This process, including the interaction between predefined causes, participant evaluations, and convergence in interpretation, is illustrated in Figure 4.1.

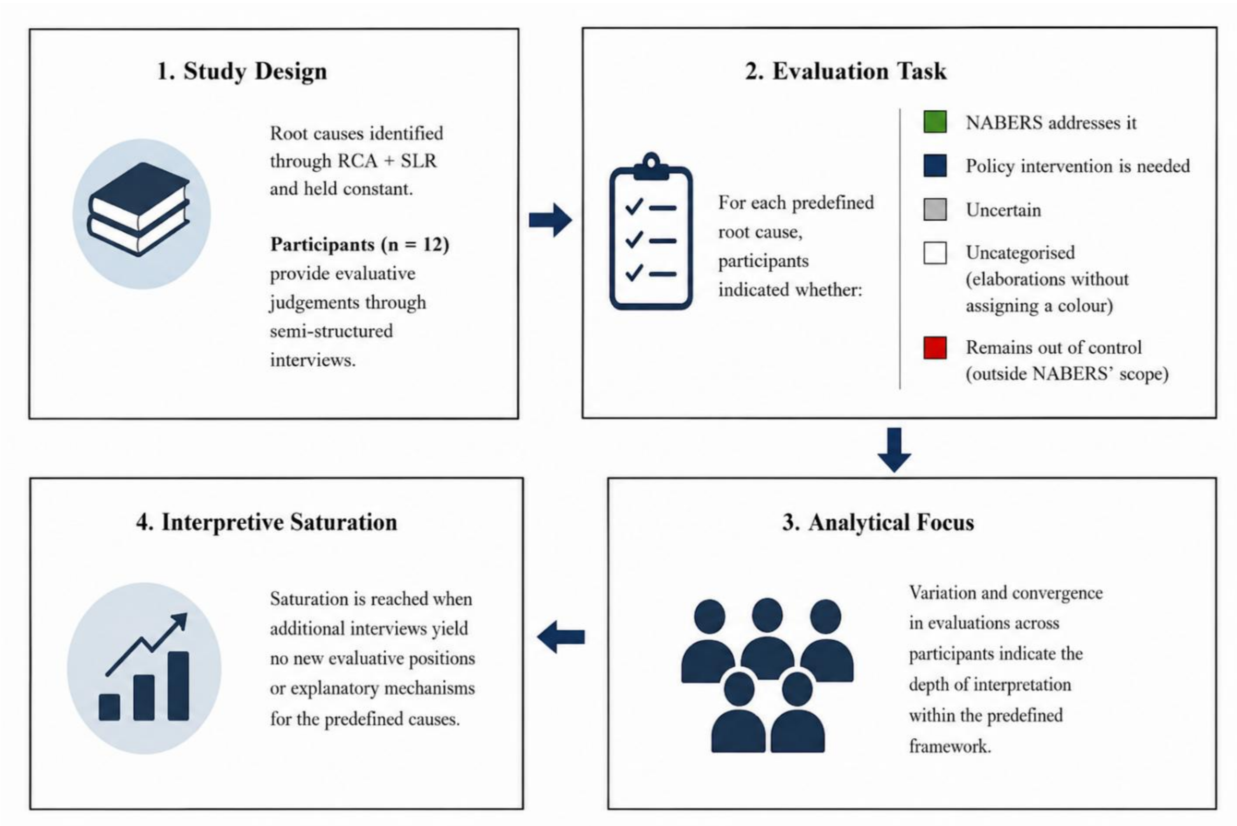


Figure 4.1: Methodological framework for assessing interpretive saturation within a deductive analytical design

It is important to note that two additional interviews were conducted with NABERS IDRs, one based in the UK and one in Australia. These interviews were not included in the results and analysis. The interview with the Australian participant could not be incorporated because of a recording failure, which limited the depth of analysis that could be performed. The interview with the participant in the UK was excluded because the participant did not return the required consent form, which prevented the use of the material in accordance with ethical requirements. Two further interviews were conducted with construction managers to explore causes of the EPG arising during the construction phase. These interviews were also excluded because both participants were based in Malaysia, a context that differs substantially from that of the United Kingdom, and because the very small number of participants did not provide a sufficiently robust representation of construction sector perspectives. In addition, maintaining a focus on outcome-oriented policies required a consistent framing across the interview sample.

4.1.3 Ethical Considerations

Ethical approval for the research was obtained prior to data collection. All participants were provided with a participant information sheet and consent form in advance of the interviews. Consent was secured for recording and transcription; the consent form template is provided in Appendix B. Given the professional roles of participants, particular care was taken to anonymize details of employers, clients, and specific projects. During the interviews, certain insights raised by one expert were explicitly shared with others as a form of respondent validation. This iterative process helped ensure the robustness and credibility of the findings.

4.1.4 Participant Selection

IDRs were selected because they hold a unique role in NABERS, conducting detailed reviews of building designs and associated energy models. They assess whether the design and its simulation make a plausible case for achieving the target NABERS rating, while also identifying practical risks and opportunities that could affect operational performance (Foo and Bannister 2022). Participants had extensive experience in NABERS, including exposure to mature schemes, providing valuable insights for understanding its adoption in the UK context.

Focusing on commissioning and operational stages, seven interviews were conducted, of which six participants were NABERS accredited assessors (AAs). AAs were selected for their formal training, examination, and supervised assessment process, which equips them to perform NABERS ratings (Nabers 2023). Their expertise allowed them to reflect on the scheme's practical application and influence on building performance. One participant was included due to their senior leadership role in sustainability and direct involvement in NABERS-related projects, providing additional insights into operational practices, landlord–tenant dynamics, and post-occupancy behaviour.

4.1.5 Analytical Approach

The interview transcripts were thematically coded in relation to the causes identified in Chapter Three. In analysing and presenting the findings, care was taken to avoid making categorical claims about whether NABERS definitively does or does not address a given cause. This was intended to reflect the complexity and context-dependence of the EPG, rather than reducing it to overly simplistic conclusions.

To complement the qualitative synthesis, a descriptive quantitative layer was introduced. Participant judgements were aggregated into counts and percentages and subsequently visualized through color-coded matrices to illustrate areas of convergence and divergence across experts and themes. Importantly, the colour categories (Green, Blue, Grey, Red, and Uncategorized) were assigned directly by participants themselves during interviews; the researcher's role was limited to tallying and visualizing these inputs. In this scheme, green signifies that NABERS addresses a given cause; blue indicates that further policy intervention is required; grey denotes causes where participants were uncertain or offered limited elaboration; red reflected causes considered beyond control (e.g., occupant behaviour in certain cases); and white represented uncategorized cases, where participants elaborated on causes but decided not to assign a dedicated colour.

This mixed methods design aligns with guidance that integrating qualitative and quantitative methods enables researchers to capitalize on the strengths of each while offsetting their weaknesses. Schoonenboom and Johnson (2017) argue that careful integration enhances the coherence and robustness of inferences, while Eisman and et al. (2022) emphasise that combining complementary epistemological perspectives increases explanatory power and better captures the complexity of practice and implementation contexts. To enhance reliability, points raised by one participant were sometimes introduced in subsequent interviews to test resonance with others, and this respondent validation step complements the quantitative distributions by surfacing stable agreements, contested areas, and genuine uncertainty that warrants further investigation.

4.2 Root Cause Analysis

The identification of root causes was conducted through the analysis of 17 records reviewed in Chapter Three. Using the "5 Whys" technique, each stage of the building lifecycle was interrogated to trace the underlying causes through which the EPG emerges. This structured approach enabled the causes to be unpacked progressively, moving from higher-level issues to specific, actionable root causes. The process was visualized in stage-specific diagrams, each illustrating the branching pathways from the first "why" to deeper levels of causation.

4.2.1 Design Stage Root Causes

The first "Why" in the design stage was framed as design-stage root causes. From this, a second layer was identified as "issues with the design itself." Two distinct pathways were then mapped: "inefficient design" and "efficient but complex design". Inefficient design was further broken down into three root causes: "oversized HVAC components", "lack of appropriate details", and "poor thermal concept". The structure of this analysis is illustrated in Figure 4.2.

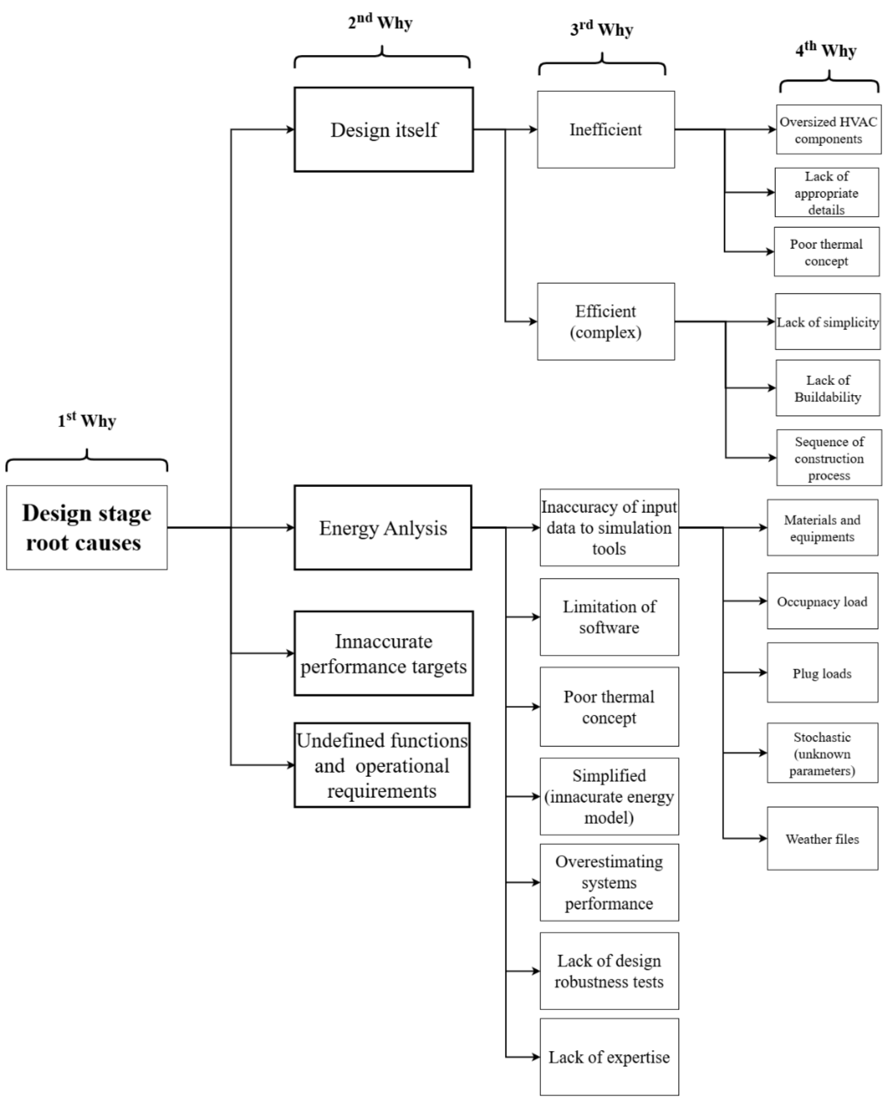


Figure 4.2: Design stage root causes

4.2.2 Construction Stage Root Causes

For the construction stage, the first "why" was framed as "construction stage root causes". The second level included issues such as "poor workmanship", "design changes", "time pressure" and "cut corners" are consistent with findings reported in analyzed records. These were then unpacked into more specific root causes. The analysis revealed that errors in construction often cascade directly into operational inefficiencies, highlighting the need for rigorous quality assurance throughout the delivery process. This causal pathway is presented in Figure 4.3.

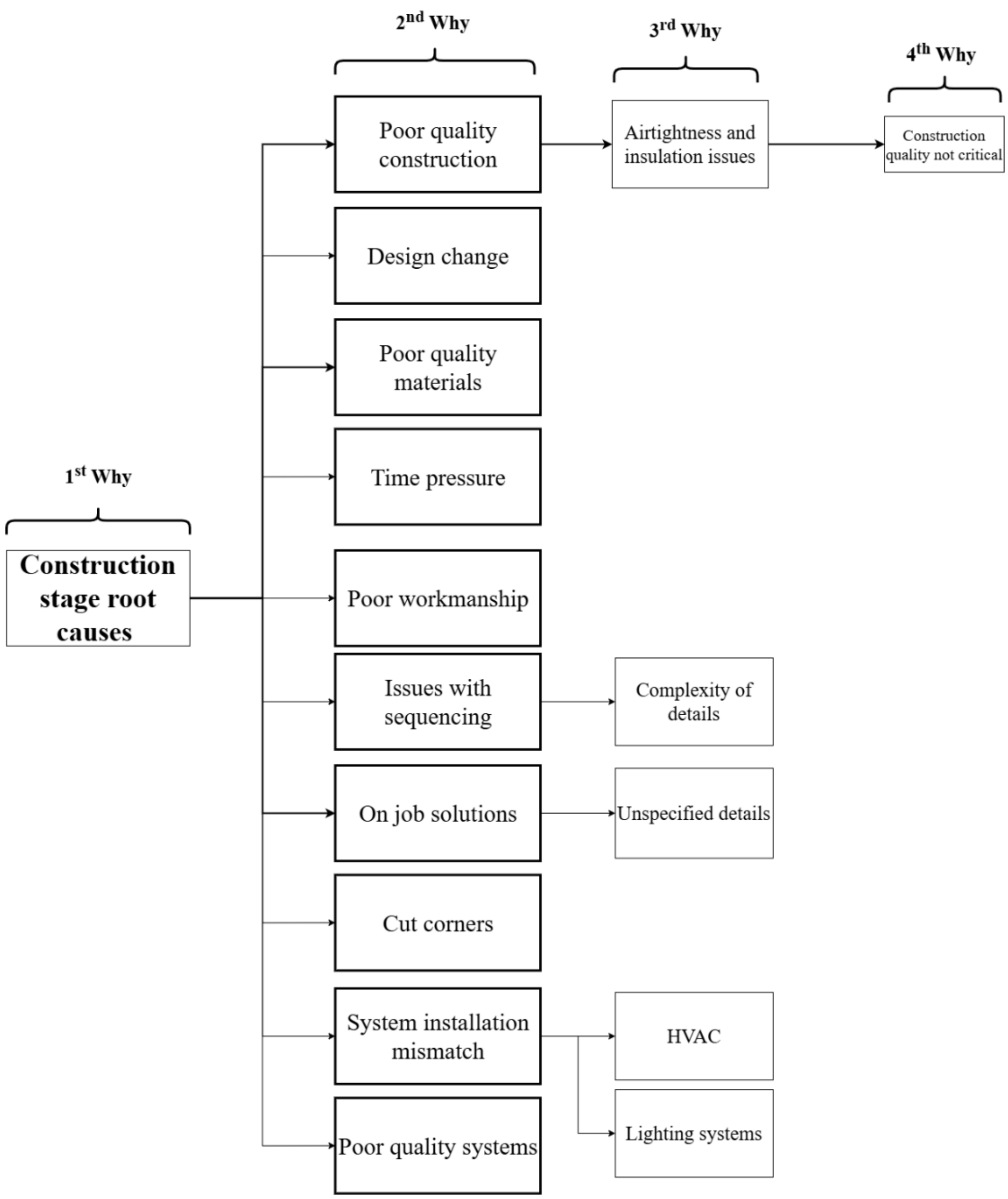


Figure 4.3: Construction stage root causes

4.2.3 Commissioning and Operational Stage Root Causes

In the commissioning and operational stage, the initial "why" was defined as "commissioning and operational stage root causes". This was expanded in the second and final levels to include six and 16 factors, respectively. Figure 4.4 presents this causal structure.

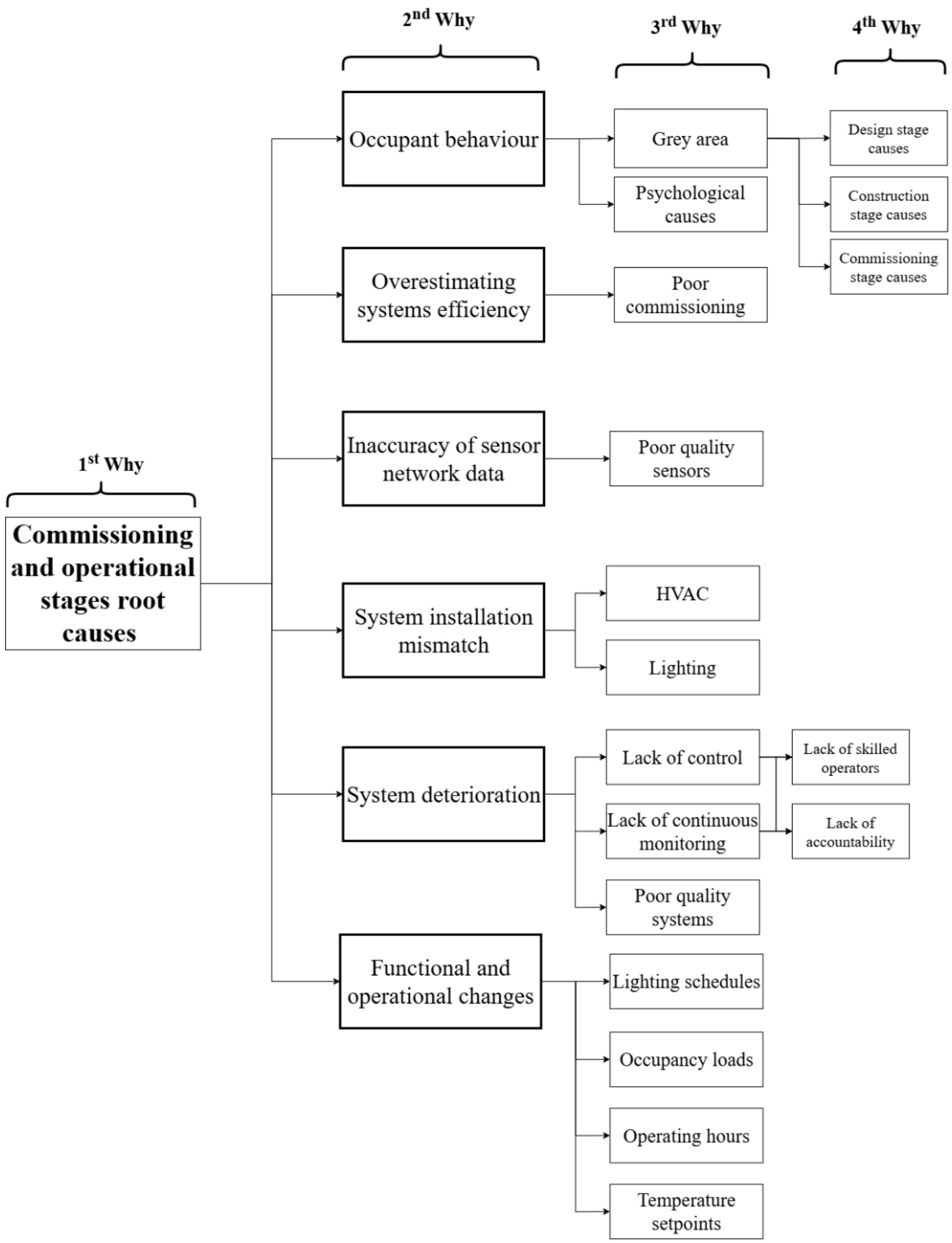


Figure 4.4: Commissioning and operational stages root causes

4.2.4 Cross-Cutting Factors

Cross-cutting factors were also analysed, as these causes transcend individual lifecycle stages and influence performance outcomes throughout the process. At the second "why", six broad categories were identified: "overall industry culture", "poor client knowledge", "lack of post-occupancy evaluation", "poor communication and collaboration", "lack of accountability", and "regulatory pressure". A third layer was developed only for two categories. For "industry culture", recurring themes of "quality", "integrity", and "responsibility" were identified. For "regulatory pressure", the pathway led to "over-optimistic predictions", a finding consistent with analyzed records. These factors illustrate that systemic cultural and regulatory influences act as enablers of the EPG, reinforcing issues that manifest at individual project stages. The causal structure of cross-cutting factors is shown in Figure 4.5.

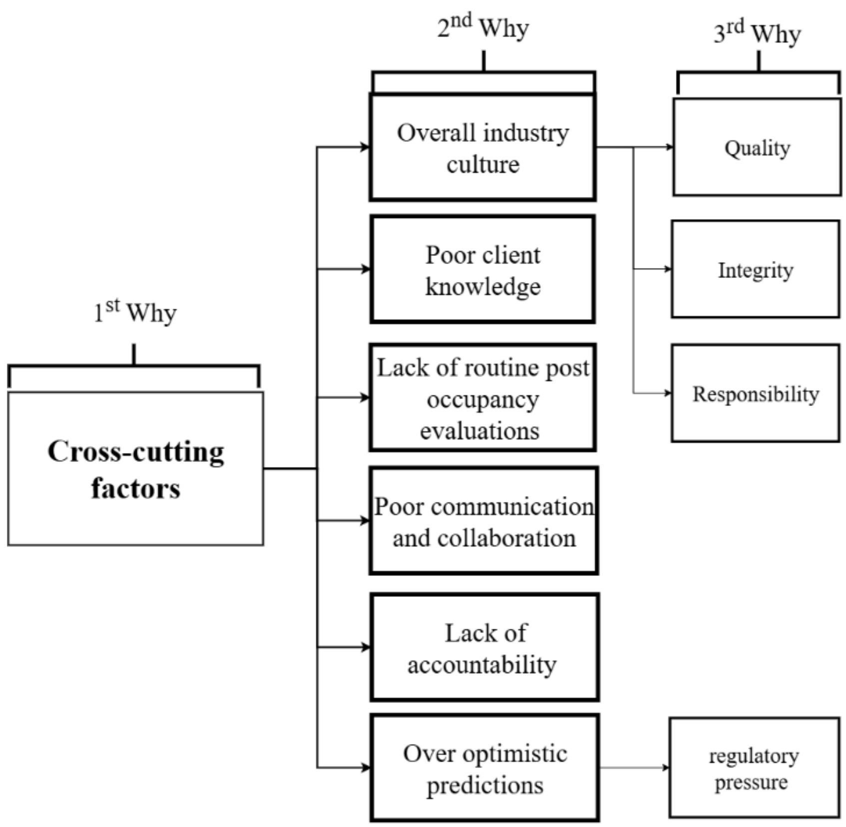


Figure 4.5: Cross-Cutting Factors

The RCA, supported by the 17 reviewed records, therefore provides a comprehensive framework for understanding the emergence of the EPG and the diagrams (Figures 4.2.1–4.2.4) form the analytical foundation for the interview findings presented in Section 4.3.

4.3 Results and Analysis

Building upon the RCA, a total of 12 interviews were conducted with two stakeholder groups: five NABERS IDRs and six NABERS AAs and a sustainability director with extensive NABERS experience. In addition, discussions of cross-cutting factors were explored with all participants, apart from one IDR (IDR5). Sections (4.3.1–4.3.3) present a structured qualitative and quantitative analysis of interview findings across design stage, commissioning and operational stages combined and cross-cutting factors. Each of those sections conclude with thematic findings and co-occurrence analysis. Section 4.3.4 is combined synthesis of all interview findings, followed by a comprehensive policy framework (section 4.5) for the mitigation of the EPG informed by experts' insights. The final section (4.6) is where key takeaways of this chapter are summarized.

4.3.1 Design Stage Root Causes

During the five interviews conducted, participants evaluated 19 root causes of the EPG that emerge in the design stage, producing a total of 95 judgments. Across all judgments, 61 (64%) were coded green, indicating high potential of NABERS in addressing certain causes. On the other hand, 14 (15%) judgements were coded with blue indicating the need for policy intervention to address such causes. Furthermore, three (3%) judgements were coded with grey, indicating uncertainty and limited elaborations on certain causes while 17 (18%) judgements were uncategorized, indicating causes where participants elaborated on them but preferred not to assign a dedicated colour. However, when the analysis is structured by themes, differences become evident as presented under the three themes within sections (4.3.1.1-4.3.1.3). Figure 4.6 below presents a color-coded matrix of the 95 judgements on the 19-design stage root causes discussed.

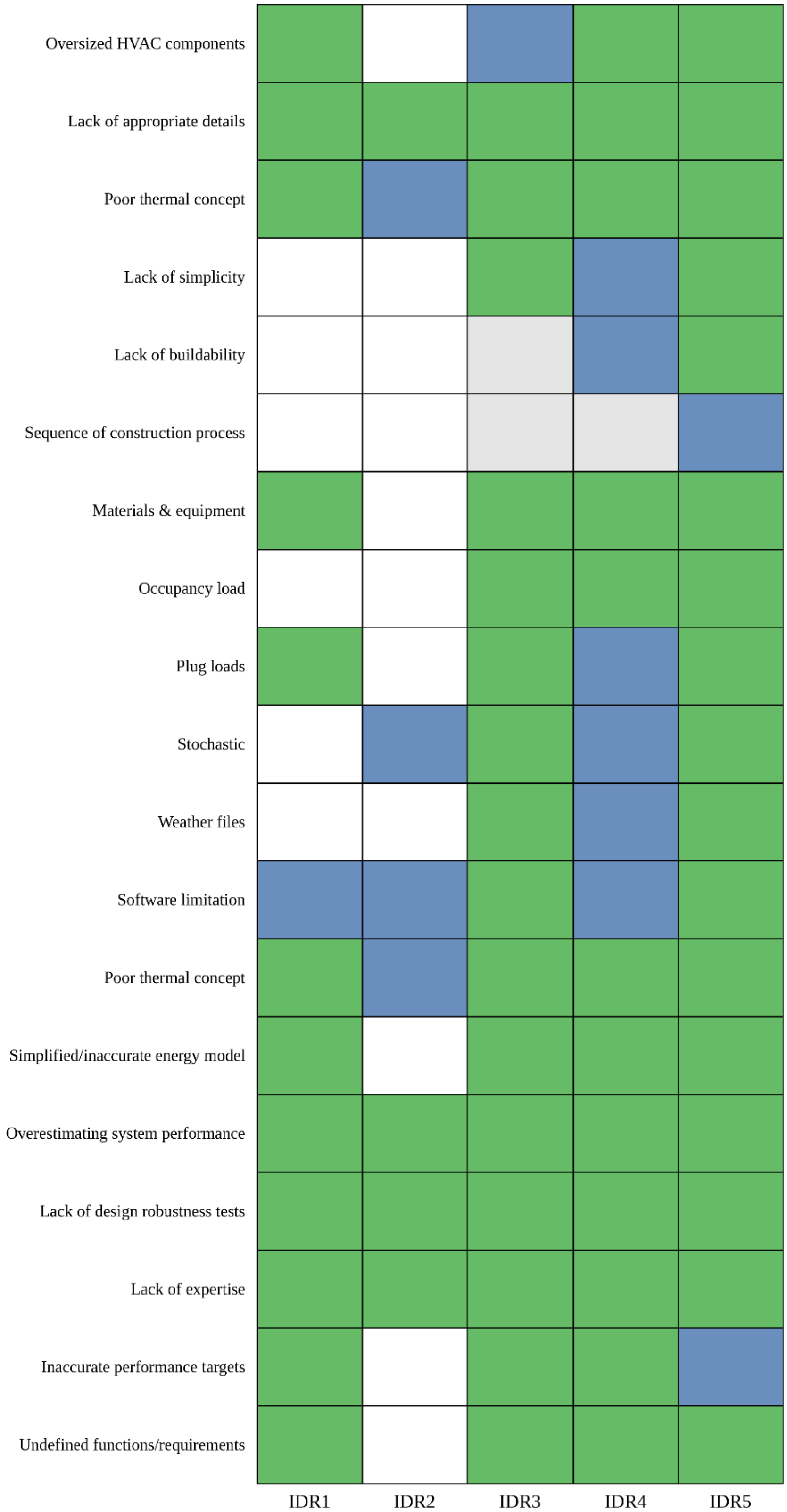


Figure 4.6: Design Stage interview judgments by all participants

4.3.1.1 Issues with the Design Itself

This theme comprises six causes and 30 judgments on inefficient and efficient designs and was divided between 15 (50%) green, 5 (17%) blue, 3 (10.0%) grey, and 7 (23%). Figure 4.7 below shows the color-coded matrix.

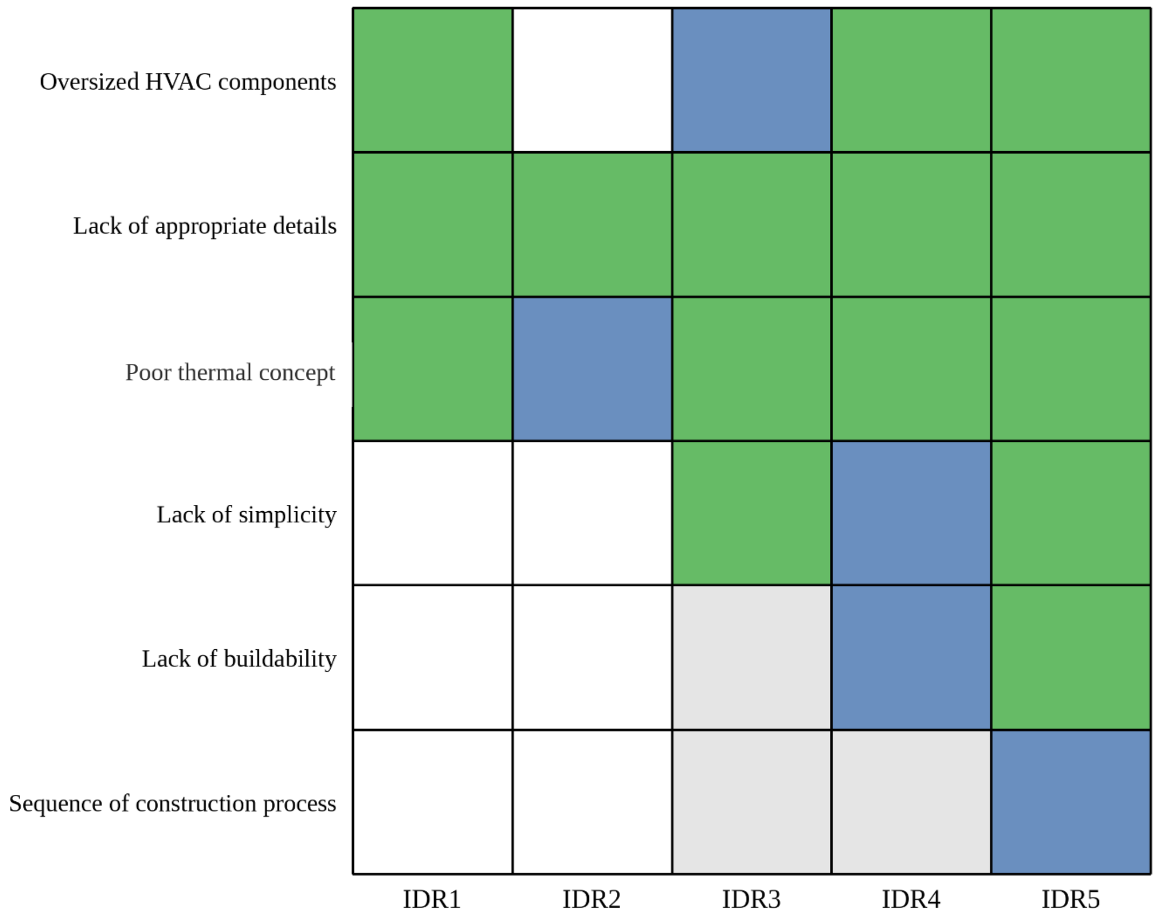


Figure 4.7: Participants judgements on inefficient (top three) and efficient (bottom three) design root causes

Inefficient Design: This subtheme was relatively clear-cut, with 12 of 15 judgments (80%) coded green, 2 (13%) blue, and one (7%) uncategorized cause. This subtheme contains three root causes: oversized HVAC component, poor thermal concept and lack of appropriate details.

On oversized HVAC components, IDR5 explained that "every project has got oversized HVAC systems as engineers add a percentage for safety" before stressing that the crucial test under NABERS is whether systems are controlled so they "deliver just the amount of required energy at the right time." He gave an example where a chiller plant is large enough to serve two buildings, and still the energy target was achieved because the sequencing logic switched units off when not needed. IDR4 also observed that oversizing occurs “reasonably frequently” not only for main plant but also for terminal units such as fan coils and radiators, often because plant availability issues that force designers to select equipment larger than necessary. IDR2 noted that traditional design practices are “obsessed about systems being big enough to cope with the worst possible day” whereas NABERS shifts the focus to “what happens day in, day out all the time.” For IDR2, efficiency at part-load is more critical. IDR3 described how simulations reveal when a system is “running at really low speeds” in which case design teams must “do another simulation with a smaller size” to improve efficiency. Synthesis of participant perspectives is summarized in Table 4.1 which maps quantitative and qualitative judgements.

Table 4.1: Synthesis of perspectives on oversized HVAC components

Oversized HVAC	**Quantitative**	**Qualitative**
IDR1		• Teams ensures systems are of good standard in terms of efficiency
IDR2		• Traditional design practices are “obsessed about systems being big enough to cope with the worst possible scenarios • NABERS shifts the focus to “what happens day in, day out all the time.”
IDR3		• Simulations reveal when a system is “running at really low speeds” in which case another simulation with a smaller size” to improve efficiency is needed.
IDR4		• Oversizing occurs frequently for main plant and terminal units because of plant availability issues that force designers to select equipment larger than necessary.
IDR5		• Every project has got oversized HVAC systems as engineers add a percentage for safety” • Crucial test under NABERS is whether systems are controlled so they “deliver just the amount of required energy at the right time.” • A chiller plant in accredited building was large enough to serve two buildings, and still the energy target was achieved because the sequencing logic switched units off when not needed

Consensus is observed amongst participants on the potential of NABERS to mitigate or prevent the emergence of issues associated with lack of appropriate details, especially regarding controls and metering, which were consistently highlighted as a weakness in conventional projects. According to IDR2, they started to “always chase a description of operations,” because at early stages “mechanical engineers never produced this” yet NABERS requires it, compelling design teams to define control logic far earlier than before. Same participant added, Design for Performance (DfP) also demands lots of details on metering, sub-metering and validation, including sign-off that systems are properly installed and set up. IDR4 confirmed that as IDR she often “force people to provide a level of controls detail to review,” which she described as “one of the big changes” compared with non-NABERS projects. However, she also clarified that

“NABERS itself doesn’t solve that problem” if applied only at the end; rather, “the DfP process does,” and most UK NABERS projects are currently on that pathway in her perspective. IDR5 highlighted the importance of testing chiller staging and realistic pump curves. These insights confirm that NABERS has materially changed practice by requiring explicit, early, and practical control and metering strategies. Unlike other schemes that allow such detail to be deferred, NABERS makes it central to the design review process, thereby reducing the risk that vague or absent specifications undermine performance. Synthesis of participant perspectives is summarized in Table 4.2 which maps quantitative and qualitative judgements.

Table 4.2: Synthesis of perspectives on lack of appropriate details

Lack of appropriate details	Quantitative	Qualitative
IDR1		
IDR2		• Chasing a description of operations because, at early design stages, mechanical engineers have not traditionally produced this level of documentation; however, NABERS requires it, compelling design teams to define control logic much earlier than before. • DfP demands extensive detail on metering, sub-metering, and validation, including formal sign-off confirming that systems have been properly installed and configured.
IDR3		
IDR4		• Forcing design teams to provide a greater level of controls detail for review, which was described as one of the major changes compared with non-NABERS projects. • NABERS itself does not resolve this issue if applied only at the end of the project; rather, the DfP process addresses it, with most UK NABERS projects currently following the DfP pathway.
IDR5		

Poor thermal concepts, particularly façade design and glazing ratios, were another concern. IDR2 stated that IDRS are “quite often appointed at Stage 3,” by which point “the façade concept, how much glass, what kind of build-up” are already defined and “all those opportunities have gone.” To address this limitation, he advocates for a pre-IDR workshop “about two months earlier,” which allows reviewers to influence envelope performance at the “sweet spot” in the design process. IDR3 gave a practical example of a heavily glazed office where IDRS pressed to “drop g-values or reducing the windows,” which reduced cooling loads and made the target achievable. According to IDR4, NABERS has increased collaboration with architects, explaining that “most of the time when we’re designing any NABERS building, there will be a lot of dialogue and increasingly architects asking about criteria to achieve for the envelope.” She stressed that NABERS has made such conversations more routine. IDR5 also pointed out that “the designer may comply with building regulations, but still risk not achieving the NABERS target," which often forces teams to go beyond the minimum requirements, for example by improving glazing performance or insulation. Finally, IDR1 added that system controls and fabric must be reviewed together because "both really can influence efficiency." These accounts suggest that NABERS can drive better thermal concepts, but its effectiveness is highly dependent on the timing of reviewer engagement and the willingness of architects and clients to accept changes. Synthesis of participant perspectives is summarized in Table 4.3 which maps quantitative and qualitative judgements.

Table 4.3: Synthesis of perspectives on poor thermal concepts

Poor thermal concepts	**Quantitative**	**Qualitative**
IDR1		• System controls and building fabric must be reviewed together, as both have a direct influence on operational efficiency.
IDR2		• IDRs are often appointed at Stage 3, by which point the façade concept, glazing extent, and build-up have already been defined, meaning that many opportunities to influence envelope performance have been lost • Advocates for a pre-IDR workshop approximately two months earlier, enabling reviewers to influence envelope performance at the most effective point in the design process.
IDR3		• In a heavily glazed office, the IDR encouraged the team to reduce g-values or window areas, which lowered cooling loads and helped make the NABERS target achievable.
IDR4		• NABERS has increased collaboration with architects, with envelope performance now forming a more routine part of early design dialogue. Although some resistance remains, NABERS has made conversations around envelope criteria and performance expectations more common.
IDR5		• Compliance with building regulations does not necessarily ensure achievement of the NABERS target, often requiring teams to exceed minimum standards through measures such as improved glazing performance or enhanced insulation.

Overall, the IDRs confirmed that NABERS functions as a disciplining framework. Oversizing remains endemic, but its impacts are mitigated by attention to part-load performance and control strategies. Lack of appropriate detailing is much more strongly addressed, since NABERS requires realistic control and metering logic to be defined and reviewed early. Poor thermal concepts are the most conditional, with real influence possible only when IDRs are engaged before design decisions are locked in. In sum, NABERS alters the treatment of inefficient design by embedding operational realism, demanding greater design detail, and promoting earlier attention to thermal performance. While it cannot fully overcome entrenched practices such as oversizing, it nonetheless mitigates their consequences and increases the likelihood that design intent is delivered in use.

Efficient Design (Complex): The efficient but complex design subtheme was more fragmented: only 3 of 15 judgments (20%) were green, 3 (13%) were blue, 3 (20%) were grey, and 6 (40%) were uncategorized, reflecting hesitation and a higher level of uncertainty when dealing with complexity of design. Interviewees agreed that efficiency can introduce its own risks.

The IDRs' accounts show where NABERS surfaces these risks and where it struggles to control them. On lack of simplicity, IDR2 warned that excessive complexity can be detrimental arguing that documents like the description of operations must be written in clear, operator-oriented language rather than "controls jargon" that facilities staff do not understand. He invoked the "keep it simple" principle and cautioned that investigations of buildings repeatedly show complex arrangements underperform because "people do not understand how they work." He also gave practical examples of sophisticated building management systems left switched off and teams who "have no ability to control that BMS," as well as a wider point that training, SL and seasonal commissioning are often prerequisites for complex strategies to succeed. He added a counter-nuance: we should not assume that "simplicity always saves energy," because there is need for more qualified and experienced engineers to operate buildings, otherwise the technology is not used well. IDR3 came to the same conclusion from the modelling side. He noted that a building management system can be "designed to do a thousand things" but "half of them is never set up properly or get overridden." Where geometry is very complex, he tends to address this with zoning or sampling rather than rejecting a design, and "rarely" tells teams "We cannot build that in the software." IDR5 highlighted system configurations that are efficient in principle but fragile in practice. He pointed to "series counterflow chillers" as "hard to model," and when that happens, he advises teams to model a more conservative scenario so that the performance prediction remains robust even if the idealized sequence cannot be represented exactly.

IDR4 added a policy-design tension: stringent targets can unintentionally "encourage complexity," as teams layer on optimization rather than choosing simpler, more robust solutions. IDR1 framed complexity as mainly a controls and domestic hot water issue on all-electric office typologies, where the industry is "still learning" between local electric, high-temperature heat pumps, or cascade arrangements; her view was that design review helps "stop people doing silly things," but real learning will come from seeing how completed projects perform. Reviewers stressed that NABERS itself is the framework that enables such scrutiny. By setting measurable outcomes, NABERS makes conventional, minimal approaches impossible. As one IDR put it, once a project targets NABERS, "convention is no longer possible," and the scheme becomes the baseline against which design logic is tested. Synthesis of participant perspectives is summarized in Table 4.4 which maps quantitative and qualitative judgements.

Table 4.4: Synthesis of perspectives on lack of simplicity

Lack of simplicity	Quantitative	Qualitative
IDR1		• The industry is still learning • IDRs help prevent inappropriate design decisions, but more substantive learning will emerge once completed projects are assessed against their actual operational performance.
IDR2		• The description of operations should be written in clear, operator-oriented language rather than technical controls terminology that facilities staff may not understand. • Operators do not always have a sufficient understanding of how building systems are intended to function. • Sophisticated building management systems are sometimes left switched off or underused in practice • Training, SL, and seasonal commissioning are prerequisites for complex operational strategies to perform successfully. • It should not be assumed that simplicity will always reduce energy use, as complex buildings require suitably qualified and experienced engineers to operate them effectively; otherwise, the intended technology may not be used as designed
IDR3		• A building management system may be designed with extensive functionality, but many features are often not properly configured or are later overridden. • Where building geometry is highly complex, zoning or sampling can be used rather than rejecting the design or assuming it cannot be represented in the modelling software. • Particular attention is required where complexity may threaten performance outcomes, such as in cases involving overheating risks or micro-loads.
IDR4		• Stringent energy targets can unintentionally encourage additional system complexity, as teams may layer optimisation strategies rather than prioritising simpler and more robust solutions.
IDR5		• Some system configurations may be efficient in principle but fragile in practice. For example, where series counterflow chillers are difficult to model accurately, a more conservative surrogate may be used to maintain a robust performance prediction when the idealised sequence cannot be represented exactly.

Lack of buildability was also recognized as a recurring concern. IDR5 gave the clearest example of an efficient concept that becomes risky in practice: "passive chilled beams in a humid climate," where humidity control is challenging and "if you lose humidity control you will get condensation in the space." He tells teams that in such climates "meeting your NABERS target now has a higher risk" because the system is "less flexible," and asks them to decide whether to proceed or change the design. IDR4 said design teams do look at workmanship-sensitive items such as airtightness detailing and sometimes use prefabrication or modular construction to improve consistency. But "there is nothing to compel the contractor to look at buildability," and "NABERS does not compel them to," so she would rate buildability as an area beyond what NABERS currently requires. She added, however, that the increased scrutiny of performance under NABERS does have indirect influence, making design teams more attentive to these details than they might otherwise be. IDR1 pointed to governance instruments around NABERS that track delivery rather than guarantee it, such as rating achievement plans and quarterly reporting, and the recommendation to maintain an as-built model that sets subsystem targets for comparison with meters. Table 4.5 which maps quantitative and qualitative judgements.

Table 4.5: Synthesis of perspectives on lack of buildability

Lack of buildability	Quantitative	Qualitative
IDR1		• NABERS governance mechanisms track delivery rather than guarantee it, including rating achievement plans, quarterly reporting, and the recommendation to maintain an as-built model with subsystem targets for comparison against metered performance. • These mechanisms help maintain visibility of delivery changes and system tuning through to the first NABERS rating.
IDR2		
IDR3		• Buildability is largely considered the responsibility of the contractor rather than the design reviewers. • While IDRs may identify risks such as overheating or micro-loads arising from a complex design concept, construction practicality remains primarily within the contractor's remit.
IDR4		• Design teams consider workmanship-sensitive aspects, such as airtightness detailing, and may use prefabrication or modular construction to improve buildability. • NABERS does not directly compel contractors to assess buildability. • The increased performance scrutiny under NABERS can indirectly influence buildability by making design teams more attentive to construction-sensitive details than they might be on non-NABERS projects.
IDR5		• Efficient design concepts can become risky in practice, such as passive chilled beams in humid climates, where achieving the NABERS target may carry greater risk because the system is less flexible.

On sequencing, the reviewers agreed that NABERS has little formal impact. IDR4 said plainly that sequence "is governed by programme and cost," and that structural engineers tend to lead it "with no awareness of NABERS," so designers have little opportunity to shape overall sequencing in service of operational energy aims. She again noted partial mitigation at the level of the mechanical and electrical disciplines, where prefabrication can reduce workmanship variability on ductwork and fittings that otherwise raise fan power. IDR5 confirmed that reviewers "are commenting on the design, not the build method or sequence." He also linked sequencing to procurement: efficient chillers often have long lead times, and if they "do not come in time you may buy a cheaper chiller because you just cannot get it," with obvious implications for energy outcomes. Rising costs and material scarcity reinforce the risk of later substitutions that depart from the performance intent. IDR1 returned to the role of rating achievement plans and quarterly reports, which keep milestones visible to the team but do not direct the order of works. These processes illustrate NABERS' broader effect: it may not control sequencing, but by requiring reporting and outcome targets it enforces a level of transparency and accountability that is absent in conventional approaches. Table 4.6 which maps quantitative and qualitative judgements.

Table 4.6: Synthesis of perspectives on lack of sequence of construction process

Sequence of construction process	**Quantitative**	**Qualitative**
IDR1		• Rating achievement plans and quarterly reports help keep project milestones visible to the team but do not directly determine the order of works. • NABERS may not control construction sequencing; however, by requiring reporting and outcome-based targets, it introduces a level of transparency and accountability that is often absent in conventional approaches.
IDR2		
IDR3		
IDR4		• Construction sequencing is primarily governed by programme and cost considerations. • Structural engineers often lead sequencing decisions with limited awareness of NABERS requirements, leaving designers with limited opportunity to shape sequencing around operational energy objectives. • Partial mitigation is possible within mechanical and electrical disciplines, where prefabrication can reduce workmanship variability in ductwork and fittings that may otherwise increase fan power.
IDR5		• IDRs typically comment on design quality and performance risks rather than construction methods or sequencing. • Rising costs and material scarcity increase the risk of later substitutions that may diverge from the original performance intent.

For complex designs, the line between success and shortfall is less about missing detail and more about fragility and deliverability. On simplicity, the reviewers warned that complex control logic, novel plant configurations and domestic hot water strategies can outstrip what facilities teams can operate and can be difficult to model faithfully. NABERS reviews expose these issues and encourage robust surrogates in modelling, yet the scheme also sets stringent targets that some teams respond to by adding layers of optimisation. On buildability, the decisive factors sit in construction practice, climate suitability and supply chain realities which NABERS does not regulate, although it raises awareness and drives teams to consider risks more carefully. On sequence, programme and cost dominate decisions, and procurement delays for high-efficiency technologies can derail delivery even when design intent is sound. NABERS therefore strengthens visibility and realism, and by embedding reporting and accountability mechanisms it increases the pressure on contractors and facilities managers, but its influence on these three risks is constrained by responsibilities and incentives that lie outside the design review.

4.3.1.2 Issues in building energy modelling

Uncertainty in building energy modelling was identified through the RCA, and the IDRs were asked to comment on the issues that had been highlighted. Their reflections confirmed the persistence of these challenges but also showed how NABERS has reshaped the way they are managed. While conventional compliance modelling was described as "a very constrained, fictitious view of the world" (IDR2), NABERS was seen as a more rigorous process in which "almost everything is accounted for" (IDR1), with the crucial implication that predictions are later reconciled against actual outcomes and there is "nowhere to hide" (IDR2). This theme spanning 11 causes and 55 judgments, presented 39 (70.1%) green, 8 (14.5%) blue, and 8 (14.5%) uncategorized. Although NABERS coverage is perceived as relatively strong here, this theme contained the highest concentration of blue responses, highlighting policy and methodological gaps in areas such as stochastic processes, plug loads, weather files, and software tools as shown in Figure 4.8 below.

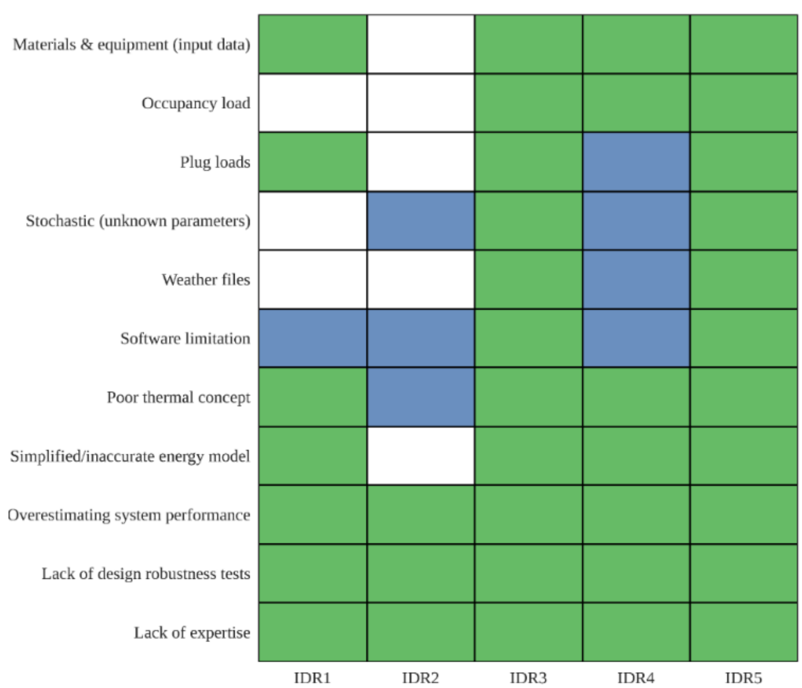


Figure 4.8: Participant judgements on energy modelling related root causes

When prompted about the accuracy of inputs to simulation tools, IDR1 explained that part of the reviewer's role is to "make sure that things like U values are as design" though he regarded such parameters as relatively minor compared with other uncertainties. IDR2 indicated confidence in their understanding and control of materials and systems, but emphasized that models often neglect internal thermal mass, so that "buildings are not as reactive as the model suspects." IDR4 confirmed that "on NABERS projects we are 100% attentive to fabric and system performance," while recognizing that these assumptions remain simplifications.

The greatest uncertainty was seen in occupancy. IDR1 described it as "the big unknown," noting that although tenant lighting and power are excluded from the base-building rating, "they still drive loads, and they need to be represented." He explained that "we generally use defaults for occupancy hours," adjusting them where extended use is expected, with NABERS' rated-hours normalization providing a safeguard. IDR2 considered occupancy to be manageable "quite well, provided the assumptions are realistic," while IDR3 described running workshops so that they are not just making assumptions, and using NABERS defaults only when unavoidable.

Plug loads were another source of variability. IDR1 reported reliance on NABERS defaults with diversity allowances, while IDR2 considered them "captured quite well, provided they are metered correctly." IDR4 suggested "there could be a lot better information in that space," since tenants "nearly always over-state what they will need." IDR3 warned that if contractors "bundle tenant and landlord power together, you can't separate it, and you can't verify it for the rating."

In terms of stochastic or unknown factors, IDR4 admitted candidly that "you are literally just making things up," and warned that such assumptions "can have a huge impact" on outcomes. IDR2 reformed this issue as a structural weakness of models, highlighting that the omission of thermal mass renders simulated buildings far more reactive than their real counterparts. IDR3 emphasised transparency and accountability: "we document everything so someone can pick up the report and find out what we assumed," and they regularly run off-axis scenarios to test sensitivities.

Weather files were also described as imperfect. IDR1 stated that his practice adopts CIBSE 2020 files with a future climate scenario, knowing that "the weather we get will always be different."

IDR4 agreed that "the files are fairly granular, but they're not fully in step with climate change," and suggested that NABERS could provide clearer guidance by location.

The limitations of software provoked strong comments. IDR1 noted weaknesses "particularly on the water side" and in the representation of heat pump controls, though he acknowledged that requiring full HVAC and controls modelling is "already a big step forward." IDR2 was more critical, insisting that "IES is pretty poor at modelling part-load performance of heat pumps, and it doesn't do PV properly." He explained that his team now calculates part-load and PV performance externally and re-imports the results, adding that "NABERS is driving us towards EnergyPlus via DesignBuilder." IDR4 reported that her team "almost always model heat pumps in Excel, not IES," and are increasingly trialling EnergyPlus for complex systems. IDR3 offered a different perspective, noting that "IES has the advanced plant modelling we need for NABERS," though still acknowledging limitations.

Reviewers also stressed that modelling cannot compensate for poor thermal concepts introduced earlier in design. IDR1 stated plainly that "if the model says the building won't meet the target, I'm not signing it off." IDR2 argued that "by the time you're doing detailed modelling, it's too late" to correct façade mistakes such as unshaded west-facing glass. IDR4 confirmed that under NABERS, reviewers often tell architects directly, "you will need this level of airtightness or insulation," which can influence choices at a stage when change is still possible.

On the question of model fidelity, IDR1 insisted that "you need to model fan curves and pump curves" and justify any simplifications. IDR3 echoed that "we strive for the best accuracy," coordinating closely to resolve uncertainties. By contrast, IDR5 stressed that "sometimes a simpler model is more reliable," warning that excessive complexity can produce fragile results. For him, NABERS is driving them to do modelling for "risk-management simulation, not prediction."

The problem of overestimating system performance was a recurring concern. IDR1 cautioned that modelers must not be "enthusiastic" in ways that exaggerate efficiency. IDR3 recounted that "IES was giving us pump energy way lower than reality," which led his team to export flows and recalculate hourly values externally. IDR4 described routinely derating manufacturer claims, "maybe 80% of the figure" and requiring factory acceptance testing of part-load performance curves. IDR2 stressed that "there's nowhere to hide" under NABERS, since models are compared with metered results and must be accurate within roughly 10%. IDR5 added that "a proper chiller map needs at least twenty points" and that a simple fan curve is inadequate unless ducts are included. He also provided a broader example from Australia: "all our central air conditioning systems have got a damper on every floor, we can shut off an entire floor" when vacant.

IDR5 also reflected on the industry learning curve in Australia, arguing that NABERS gradually raised professional capability by making performance expectations explicit and unavoidable. In his account, ratings that were initially difficult to achieve became routine as engineering teams adapted their practices and accumulated experience. This progression eventually required the introduction of a higher six-star category, illustrating how the scheme not only measured performance but also shifted the market's understanding of what was achievable. IDR5 suggested that the UK may follow a similar trajectory as expertise develops under NABERS.

Overall, the IDR accounts indicate that NABERS has reframed energy modelling practice. The scheme does not remove uncertainty or eliminate software limitations, but it requires these issues to be exposed, documented and tested for resilience. It also strengthens accountability by linking design-stage assumptions to operational outcomes. More broadly, participants portrayed NABERS as a mechanism for long-term cultural change: raising expectations of professional expertise, challenging over-reliance on automation, and encouraging design practices that embed energy-conscious controls into the built environment.

4.3.1.3 Inaccurate Performance Targets and Undefined Operational Requirements

This final theme concerns two linked problems: inaccurate performance targets that are aspirational, and vague or shifting operational requirements. Interviewees affirmed both issues as persistent contributors to the EPG, while also explaining how NABERS partly reframes and manages them. This theme included 2 causes and 10 judgments. Of these, 7 (70%) were green and 2 (20%) uncategorized, and (10%) blue. This distribution shows that participants believe NABERS to be most effective at handling targets and requirements, which were considered adequately addressed across most responses as shown in Figure 4.9 below.

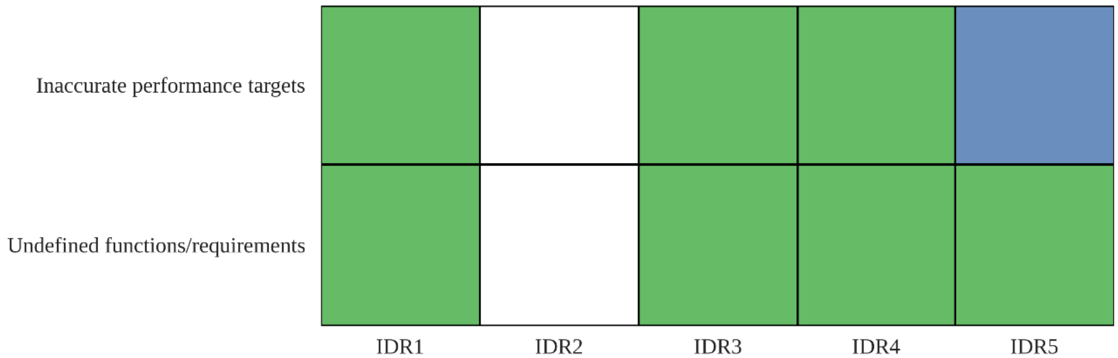


Figure 4.9: Participant judgements on inaccurate performance targets and undefined functions and operational requirements

On performance targets, IDR1 emphasized that under NABERS "the owner commits to achieving a certain number of stars and the stars can be converted into an energy target," making the aim explicit from the outset. IDR4 described the discipline this creates: "we are given an energy use intensity… which is a metric that we cannot alter," in contrast with EPC-style assessments that are "very difficult to… match [to] in-use performance." Yet the IDRs cautioned that targets can be declared before their implications are understood. As IDR3 put it, clients sometimes arrive saying "we want a six stars NABERS," only to discover once the model shows the interventions required are "outside of the budget," and projects then "drop from six to five." His conclusion: "we need to be upfront and accurate at each stage." Table 4.7 which maps quantitative and qualitative judgements.

Table 4.7: Synthesis of perspectives on inaccurate performance targets

Inaccurate performance targets	**Quantitative**	**Qualitative**
IDR1		• Under NABERS, the owner commits to achieving a defined star rating, which can then be translated into an explicit energy target from the outset.
IDR2		
IDR3		• Clients may initially seek a six-star rating, but once modelling identifies the interventions required, the associated costs can exceed the available budget, leading some projects to reduce the target from six to five stars. • Clear and accurate communication is required at each project stage to ensure that the implications of the NABERS target are understood.
IDR4		• The NABERS commitment agreement introduces greater discipline by establishing an energy use intensity target that cannot be altered, in contrast with EPC-style assessments, which are more difficult to align with in-use performance.
IDR5		

Turning to undefined functions and operational requirements, IDR1 judged the risk lower for offices in the current UK scope because developers must commit to a target and “it’s only offices,” with atypical uses (e.g., data centres) treated by specific provisions rather than being buried in the base-building rating. IDR4, however, warned that clarity about use is still hard to obtain: “NABERS has pushed us to engage with that a lot more… we are not very often successful,” and “the industry is not prepared, particularly outside London, to fully address how tenants are going to use an office space.”

In response, IDR3 described a procedural discipline: where operational data are missing, teams “apply defaults,” document the assumptions, and critically embed energy budgets into the BMS so that monthly metered consumption can be compared with the modelled splits; this helps pinpoint whether “a tenant is exceeding the limit” or whether “a plant is spending too much heating,” revealing control or storage issues. IDR2 added a subtlety about design-stage set-points and guidance: standard office specifications (e.g., heating sized at 20 °C) are not how spaces are actually operated (“people would turn the heating up to 21 or 22”), which means targets based on design performance can be mis-aligned with reality unless NABERS’ operational frame is applied.

IDR5, drawing on the Australian context, linked both problems to risk management and sector scope. Outside offices, “a hospital is very difficult to set a NABERS target for,” because multiple subsystems and highly variable operations undermine confidence in target-setting; “maybe after five or 10 years, we will learn more.” Where functions remain uncertain, IDR5 was explicit that the NABERS remedy is scenario-based risk management: begin with a most-probable base case, then stress-test “what happens if the weather changes” or if “25% of your building is not occupied.” Table 4.8 which maps quantitative and qualitative judgements.

Table 4.8: Synthesis of perspectives on undefined functions and operational requirements

Undefined functions and operational requirements	Quantitative	Qualitative
IDR1		• The risk is lower for offices within the current UK scope because developers must commit to a defined target, and atypical uses, such as data centres, are addressed through specific provisions rather than being hidden within the base-building rating.
IDR2		• Standard office specifications, such as heating sized at 20 °C, do not always reflect how spaces are actually operated, as occupants may increase heating setpoints to 21 °C or 22 °C. This can misalign design-based targets with operational reality unless NABERS' in-use performance framework is applied.
IDR3		• Where operational data are unavailable, teams apply default assumptions, document them, and embed energy budgets into the BMS so that monthly metered consumption can be compared against modelled energy splits. This helps identify whether a tenant is exceeding their allowance or whether central plant is consuming excessive heating energy, indicating potential control or storage issues.
IDR4		• Clarity around actual building use remains difficult to obtain. • NABERS has encouraged greater engagement with tenant-use assumptions, although this is not always successful. The industry, particularly outside London, is not yet fully prepared to address how tenants will use office spaces in practice.
IDR5		• Drawing on the Australian context, uncertainty around non-office sectors is linked to risk management and sector scope. For example, hospitals are difficult to set NABERS targets for because their multiple subsystems and highly variable operations reduce confidence in target-setting, suggesting that further sector learning may be needed over time. • Where building functions remain uncertain, NABERS supports scenario-based risk management by starting with a most probable base case and then stress-testing alternative conditions, such as weather variation or partial occupancy, through additional simulations.

In synthesis, the IDRs depict NABERS as shifting the culture from aspirational targets plus vague usage to explicit, audited targets plus declared operational assumptions that are tested for robustness and wired into BMS monitoring. This mitigates, but does not eliminate, the risks. Targets can still be set for marketing then revised for budget, and definitions of use can drift as tenants arrive. Where UK benchmarks and normalization are less developed (beyond offices), inaccurate targets and undefined functions remain a structural source of uncertainty, even within the NABERS frame.

4.3.1.4 Thematic Findings and Co-occurrence Analysis

Patterns at the level of individual causes reflect thematic findings. Stochastic and software limitation attracted the largest number of blue responses, showing that participants repeatedly identified them as requiring further intervention. By contrast, causes such as lack of appropriate details, undefined functions or operational requirements, overestimation of system performance, lack of design robustness tests and lack of expertise were coded green almost by all participants, indicating widespread confidence in NABERS' ability to address them. At the level of individual participants, differences in coding style were evident. IDR5 gave 17 of 19 judgments (89.5%) as green and 2 (10.5%) as blue, while IDR3 gave 16 of 19 green (84%), 1 blue (5%), and 2 grey (10.5%). IDR4 presented a more mixed distribution with 12 of 19 green (63%), 6 blue (32%), and 1 grey (5%). IDR1 made 12 of 19 green (63%), one blue (5%), and 6 uncategorized (32%). Finally, IDR2 departed significantly from the group, coding only 4 of 19 green (21%), 4 blue (21%), and leaving 11 (58%) uncategorized. The analysis show that while most participants leaned towards green, IDR4 placed particular emphasis on policy gaps and IDR2 preferred to elaborate without assigning a colour, leaving more than half of their responses uncategorized. *Table 4.9* participant level distribution of judgements.

Table 4.9: Participant-level Distribution of Judgments

Participant	Addressed by NABERS (Green)	Policy Intervention required (Blue)	Uncertain (Grey)	Elaboration without a category (Uncategorized)
IDR1	12	1	-	6
IDR2	4	4	-	11
IDR3	16	1	2	-
IDR4	12	6	1	-
IDR5 (Australia)	17	2	-	-

Co-occurrence analysis enriches this interpretation by identifying where pairs of causes were categorized together by the same participant. In green responses, consistent clustering was observed. For example, lack of appropriate details, overestimating system performance, lack of design robustness tests, and lack of expertise by all 5 participants, producing co-occurrence values of 5 in each case. This suggests that NABERS' coverage is not seen as isolated, but rather as systemic across groups. Blue responses, although fewer, also showed clustering. Stochastic input data to simulation tools for unknown parameters and software limitation were coded blue together two times. Grey and uncategorized co-occurrences were rare and occurred primarily within the efficient but complex design subtheme, where participants were more hesitant or uncertain. No Red judgments occurred, meaning that none of the 19 root causes was perceived as entirely beyond practical control.

Overall, high-level quantitative analysis demonstrates that green judgments dominate, affirming NABERS' current or indirect coverage of the identified causes. However, 14 responses (14.7%) signaled the need for further policy action, particularly in relation to analytical and modelling issues, while 17.9% remained uncategorized, reflecting either uncertainty in applying the coding system or reluctance to commit to a definitive category. The three grey judgments (3.2%) underscore the areas where participants expressed genuine uncertainty. The co-occurrence analysis shows that NABERS' perceived strengths are cohesive and systemic across clusters of design-related causes, while the perceived weaknesses requiring policy support are similarly clustered, especially around analytical tools and assumptions. This implies that interventions and reforms should be designed to act on interconnected sets of problems, rather than addressing each root cause in isolation. Table 4.10 illustrates the quantitative results. Rows one to six present the first theme, while rows seven to 17 represent the second theme and the last two rows represent the final theme.

Table 4.10: Distribution of judgments across themes

Category	Total Judgments	Addressed by NABERS (Green)	Policy Intervention required (Blue)	Uncertain (Grey)	Elaboration without a category (Uncategorized)
Overall (All Themes)	95	61 (64%)	14 (15%)	3 (3%)	17 (19%)
Theme 1: Design itself	30	15 (50%)	5 (17%)	3 (10%)	7 (23%)
Theme 1a: Inefficient design	15	12 (80%)	2 (13%)	0 (0.0%)	1 (7%)
Theme 1b: Efficient but complex design	15	3 (20%)	3 (20%)	3 (20%)	6 (40%)
Theme 2: Energy analysis	55	39 (71%)	8 (14.5%)	0 (0.0%)	8 (14.5%)
Theme 3: Performance targets & operational requirements	10	7 (70.0%)	1 (10.0%)	0 (0.0%)	2 (20%)

4.3.2 Commissioning and Operational Stages Root Causes

The RCA identified 16 root causes of the EPG emerging during the commissioning and operational stages, grouped into six thematic categories. Seven participants reflected on these causes, and their responses were color coded. Across all 112 judgements (16 causes × 7 participants), the distribution was: green 25 (22%), blue 17 (15%), grey 1 (1%), uncategorized 68 (61%), and red 1 (1%). The dominance of uncategorized categories shows that participants frequently elaborated on causes without assigning a code, reflecting hesitation or a reluctance to judge NABERS' coverage in this stage. Nevertheless, over one-fifth of responses were green, affirming NABERS' ability to address specific causes, while more than 15% were blue, signalling consistent calls for further policy or procedural interventions The color-coded matrix in Figure 4.10 below illustrates how judgements vary by theme, cause, and participant.

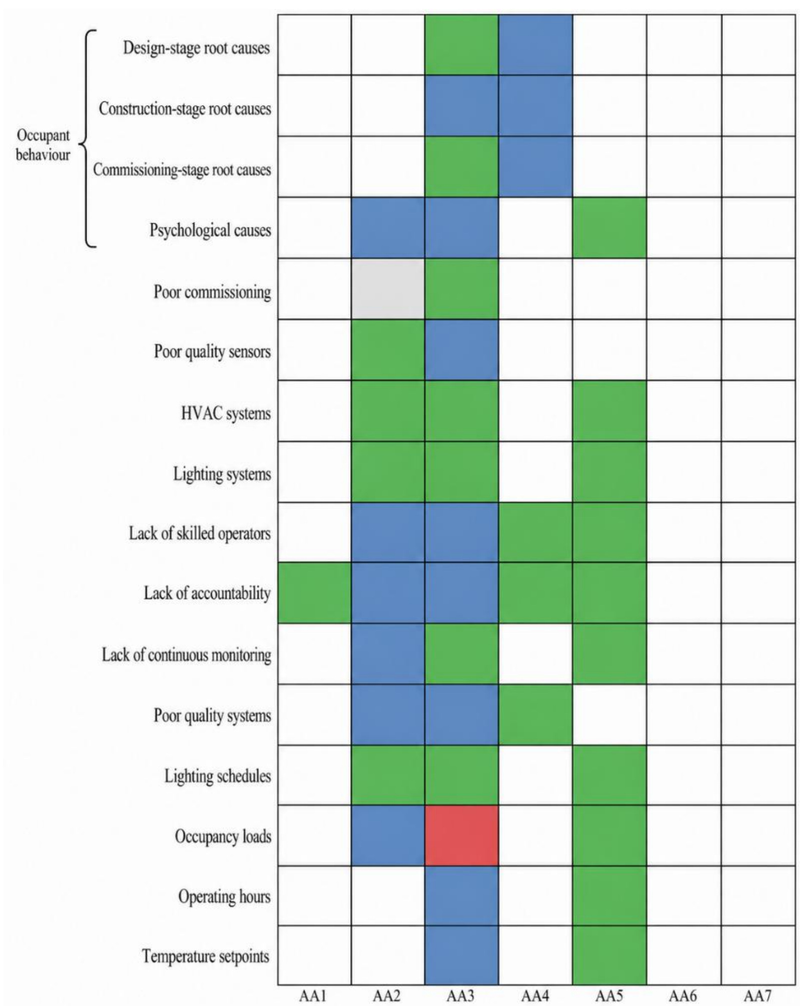


Figure 4.10: Participant judgements on commissioning and operational stages root causes

4.3.2.1 Occupant Behaviour

This theme comprises 28 judgements on occupant behaviour related causes, divided into grey areas issues wherein issues that emerged in previous stages might have led occupants into changing their behaviour and another aspect associated with psychological causes. Of these, 19 (68%) were uncategorized, 3 (11%) green, and 6 (21%) blue. This theme therefore reflects both hesitation in categorization and need for policy intervention as depicted in *Figure 4.11* below.

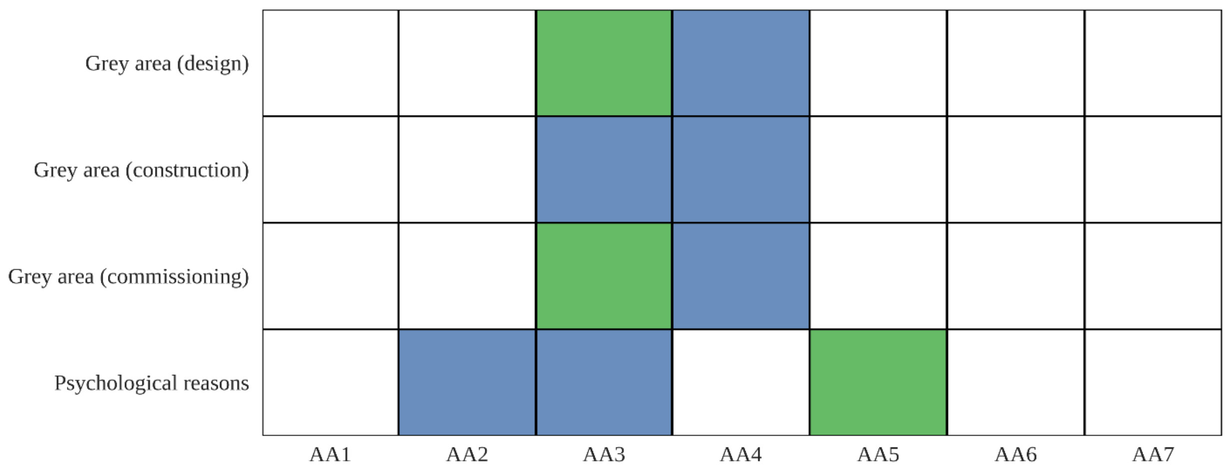


Figure 4.11: Participant judgements on occupant behaviour related causes

Grey Area: AA1 argued that NABERS cannot remove all modelling uncertainty because tenants are often unknown until after design and construction, meaning "there is always a degree of difference between the model and the reality." For him, attempting to perfect the design model is "super expensive, super time consuming" and offers limited benefit compared with investing in making the occupied building perform well. Other participants, however, stressed that early engagement can still improve modelling assumptions. AA2 explained that NABERS modelling seeks to capture likely operating hours, densities, controls and metering logic more accurately than Part L compliance, where many inputs are fixed by the NCM. In this sense, NABERS encourages a shift from compliance modelling toward performance-oriented modelling.

Tenant-side evidence further showed how landlord–tenant interfaces shape outcomes. AA3 described a mixed-mode headquarters fit-out where NABERS 4.5–5 star aspirations informed assumptions about window use and daily operation. After occupation, BMS commissioning delays threatened to delay the 12-month monitoring period and embed poor habits, but the NABERS target created pressure to resolve the issue rather than allow it to drift. Similarly, AA4 argued that

NABERS' monitoring and metering requirements make behavioural effects more visible. Since design-stage scenarios are often developed without named tenants, teams test extended hours, higher densities and occupant control over set points, servicing hours and window operation. Unlike BREEAM, which was described as focusing largely on design parameters, NABERS relies on actual meter data and therefore exposes operational performance.

Participants also highlighted the importance of governance, handover and facilities management. AA5 described the NABERS rating achievement plan as a soft-landings-style process involving designers, FM teams and tenant representatives, supported by monthly in-use end-use reporting. On some projects, modelled energy is replaced with actual data and BMS trends are reviewed remotely, with out-of-hours use logged where tenants extend operation. AA6 provided a more striking example: two similar office buildings diverged sharply in energy use, with the single-tenant, internally managed building approaching NABERS five stars, while the multi-tenant, third-party managed building used roughly twice as much energy. The difference was attributed to poor commissioning, uncalibrated meters, FM teams lacking system knowledge, underfloor heating running in summer and lighting left at full output overnight. The remedy was not further design refinement, but training, communication and regular monitoring against targets.

AA7 added that lease structures can permit energy-intensive behaviours that overwhelm design intent, such as a single trader triggering whole-building HVAC at 2 a.m. NABERS can reveal and address these patterns, for example through local cooling or designated extended-hours zones, but only where metering resolution is sufficient. Base building ratings with richer sub-metering are more likely to detect such behaviour than whole-building ratings based only on utility data. He also noted that scheme effectiveness depends on leasing norms, contrasting UK tenant control of fan coil units with stronger landlord control in Australia.

Overall, NABERS does not eliminate grey-area issues, but it redirects attention from perfecting design-stage predictions toward managing actual performance. Its value lies in clear targets, prescriptive metering, independent review and feedback routines that expose irregular operation and support corrective action. The interviews suggest that FM capability, tenant governance, structured handover and operational rules are often more decisive than marginal improvements in model precision. Table 4.11 presents the associated qualitative findings and quantitative judgments.

Table 4.11: Qualitative findings and quantitative judgments occupant behaviour associated with grey area

Occupant behaviour associated with grey area	Quantitative	Qualitative
AA1	Design stage Construction stage Commissioning stage	• Greater value is achieved by directing resources toward ensuring that the occupied building operates effectively with its users.
AA2	Design stage Construction stage Commissioning stage	
AA3	Design stage Construction stage Commissioning stage	• Landlord–tenant interfaces have a significant influence on occupant behaviour and operational energy outcomes. • The presence of a NABERS performance target helps ensure that commissioning issues are actively addressed rather than allowed to remain unresolved.
AA4	Design stage Construction stage Commissioning stage	• NABERS monitoring requirements and prescriptive metering make behavioural effects more visible in operation. • As design-stage central scenarios are often developed without confirmed tenant information, teams undertake off-axis testing, such as extended operating hours or higher occupancy densities, while also considering the degree of occupant control over setpoints, servicing hours, and window operation.
AA5	Design stage Construction stage Commissioning stage	• The Rating Achievement Plan supports a SL-style handover, involving meetings between designers, facilities management teams, and tenant representatives to explain intended building operation, followed by monthly in-use end-use reporting. • Although these practices are initiated within NABERS projects and enforced by the Independent Design Reviewer, they are beginning to transfer into non-NABERS sectors through the adoption of EUI targets and similar feedback loops.
AA6	Design stage Construction stage Commissioning stage	• A single-tenant, internally managed site was close to achieving a NABERS five-star rating, whereas a multi-tenant, third-party managed site consumed approximately twice as much energy. • Training and regular monitoring against defined targets were identified as key remedies for poor operational performance.
AA7	Design stage Construction stage Commissioning stage	• Lease arrangements can encourage energy-intensive behaviours that undermine design intent, such as out-of-hours cooling requests that activate whole-building HVAC systems despite only serving a small area. • The NABERS process can identify and advise on irregular operational patterns, for example by recommending local cooling for out-of-hours areas or designating a small extended-hours zone; however, effective detection depends on sufficient metering resolution.

Psychological: Interviewees stressed that while NABERS can address many technical and operational causes of the EPG, psychological dimensions of occupant behaviour remain harder to manage. These include situations where spaces are objectively well lit, ventilated and thermally conditioned, yet occupants still switch on lights, adjust heating or open windows in energy-intensive ways. AA1 noted that NABERS does not directly account for such behaviour at design stage, explaining that "lighting does not impact NABERS rating really." However, he argued that good design and the IDR process can reduce risks by limiting unnecessary occupant control, particularly over heating and cooling systems. Design can therefore constrain, but not eliminate, behavioural effects. AA2 similarly acknowledged that misuse or misunderstanding of systems can increase energy use, but suggested that scenario modelling, education, user guides and SL can mitigate this. He also noted that annual NABERS re-rating makes such issues visible and gives clients an incentive to respond.

AA3, a tenant-side practitioner, argued that design can indirectly influence psychology. For example, openable windows may reassure occupants because "psychologically, they know they can open the windows." In his office, the ambition to achieve a NABERS five-star rating became a shared cultural focus, giving staff "a single point" around which to coordinate operational behaviour. He added that occupiers have more control over behaviour in their own buildings than designers do in client commissions. AA4 emphasized the subjectivity of thermal comfort, noting that acceptable set points vary between individuals. He cautioned that NABERS should not compromise comfort for efficiency and argued that dialogue with occupiers is essential, since users should not be surprised by a building they perceive as uncomfortable.

Beyond technical design, organizational factors can also affect energy demand. AA5 noted that formal dress codes may lead staff to demand cooler environments, while mixed-mode buildings require careful control strategies. His team uses indicator panels and automatic disabling of fan coil units when windows are opened, alongside off-axis simulations to test misuse scenarios such as windows being left open overnight. AA6 observed a growing tendency to reduce direct occupant control to protect performance, arguing that achieving a good NABERS rating requires limiting occupant influence "as much as possible." He praised NABERS for requiring off-axis misuse simulations and allowing reviewers to challenge designs where occupant risks are too high. By contrast, AA7 was sceptical that NABERS UK currently addresses psychological or comfort factors systematically. Drawing on Australian experience, he suggested this would require an expanded scheme similar to NABERS Indoor Environment. Table 4.12 presents the related qualitative findings and quantitative judgments.

Table 4.12: Qualitative findings and quantitative judgments occupant behaviour associated with psychological reasons

Occupant behaviour associated with psychological reasons	**Quantitative**	**Qualitative**
AA1		• Good design and the IDR process can limit operational problems by restricting unnecessary controls. • Design can constrain occupant behaviour, but it cannot fully eliminate behavioural variation in use.
AA2		• Off-axis simulations are used to test how alternative operating conditions or user behaviours may affect performance. • Education, user guides, and SL are important mechanisms for supporting effective building operation. • NABERS annual re-rating makes operational issues visible and provides clients with an incentive to address them.
AA3		• Design choices can indirectly shape occupant psychology; for example, providing openable windows can reassure occupants by giving them a sense of control. • The NABERS rating can become a cultural focus, encouraging ongoing collaboration with colleagues around how the building should be operated.
AA4		• Thermal comfort remains subjective, with occupants varying widely in their preferred setpoints. • NABERS is not intended to compromise comfort for energy efficiency, and good design processes should capture acceptable setpoint ranges, such as the British Council for Offices guidance allowing 26 °C in summer. • Dialogue with occupiers is crucial, as occupants should not be surprised by a building they may perceive as uncomfortable. • Organisational factors, such as formal workplace dress codes, can indirectly increase energy demand where staff expect cooler internal conditions.
AA5		• Off-axis simulations can test the consequences of misuse, such as windows being left open overnight, with the outcomes reported to clients.
AA6		• Occupant influence should be excluded from the base-building assessment as far as possible. • Off-axis simulations provide a method for exploring behavioural and operational risks beyond the central design case.
AA7		• An expanded scheme similar to Australia's NABERS Indoor Environment rating could help capture psychological and comfort-related factors more systematically.

Through restricted controls, off axis scenario modelling, user education, and annual ratings, the scheme reframes occupant behaviour as a design and governance issue rather than a post hoc excuse. Success depends on cultural alignment and FM leadership (AA3), clear communication of comfort expectations (AA4), and automation or control restriction (AA6). However, the UK scheme remains limited compared to Australia, where indoor environment ratings explicitly integrate well-being with energy outcomes.

4.3.2.2 Overestimation of Systems Efficiency

This theme comprises seven judgments, of which 5 (71%) were uncategorized, one (14%) green, and one (14%) grey. The dominance of uncategorized causes demonstrates reluctance to categorize overestimating systems efficiency when paired with the root cause of poor commissioning, while the presence of grey indicates uncertainty about NABERS' ability to cover commissioning practices. The participants agreed that NABERS plays a limited but important role in dealing with the recurring problem of overestimated system efficiency. AA3 stated that *"the beauty of NABERS is that it shows what is actually happening, not what the model assumed."* This view highlight that while NABERS does not prevent errors at the design stage, it acts as a reality check once buildings are occupied.

AA2 reinforced this by adding, "NABERS is effective in identifying performance shortfalls, but the overestimation problem begins at the design stage and is outside NABERS' reach." AA5 also agreed that the system works retrospectively: "You only know systems are not as efficient as expected when NABERS results come in." Some experts saw an indirect deterrent effect. AA6 noted, "Designers know their buildings will be rated, so NABERS discourages them from exaggerating system efficiency." Similarly, AA7 remarked that "the continuous accountability mechanism reduces the likelihood of inflated efficiency claims." At the same time, concerns were raised about NABERS' limitations. AA4 emphasised: "It is not NABERS' role to validate efficiency claims at design; it is simply a tool to measure outcomes." This highlights that NABERS is reactive, not preventative, in nature. Overall, the consensus is that NABERS cannot prevent efficiency overestimation during design, but it exposes inaccuracies once the building is operating and exerts accountability pressure that discourages unrealistic claims. Table 4.13 below presents qualitative findings and quantitative judgments.

Table 4.13: Qualitative findings and quantitative judgments on Overestimation of system efficiency (poor commissioning)

Overestimation of system efficiency (poor commissioning)	Quantitative	Qualitative
AA1		• NABERS cannot correct inaccurate assumptions directly, but it exposes the gap once a building is in operation.
AA2		• NABERS is effective in identifying performance shortfalls through measured operational outcomes.
AA3		• NABERS reveals what is actually happening in use, rather than what was assumed in the design model.
AA4		
AA5		
AA6		• NABERS discourages overstatement of system efficiency by making performance claims subject to operational verification.
AA7		• The continuous accountability mechanism reduces the likelihood of inflated efficiency claims.

4.3.2.3 Inaccuracy of Sensor Network Data

This theme also contained 7 judgments, distributed as 5 (71%) uncategorized, one (14%) green, and one (14%) blue. This suggests participants frequently elaborated without categorization but still highlighting issues related to reliability of data and importance of correct placement of sensors to prevent inaccuracies. While NABERS is designed to evaluate overall energy consumption, it does not regulate the specification or functioning of individual sensors.

The problem was characterized by AA2 as "one of the biggest problems with operational energy". He elaborated on the frequency and consequences of meter and sensor failures, noting that "usually when we get data back… it's dropped out. It's reading a dead value or it's just zero… meters can just fail or have errors … you end up having to send multiple people to site to kind of figure this out." While AA2 acknowledged that NABERS projects often require stronger metering strategies compared to conventional buildings, emphasizing that this does not necessarily guarantee better quality or fewer failures. Thus, the issue remains systemic and persistent across projects. As AA1 explained, "NABERS doesn't really care about what you have for your sensors or how they operate. It wants you to monitor your energy consumption. If your energy is bad and that's because loads of your sensors aren't working, then it might make you look at them a bit better to figure out what that is." This indicates that although NABERS may indirectly incentivize stakeholders to investigate malfunctioning sensors when poor performance is observed, it does not provide a mechanism to ensure the quality, calibration, or long-term reliability of the installed sensor infrastructure.

Another challenge identified relates to the fragmented ownership and sourcing of sensors. AA3 highlighted the interpretive difficulties created by data originating from multiple stakeholders, observing that “because the sensors are going to be coming from lots of different places and some of them are related to the landlord, some is our plant, some will be related to how we're using the buildings… there's a question about whether what's the data, what's the sensor information that's going to tell our NABERS assessor to include.” This reflects a broader ambiguity in the operational stage, where the integration of diverse data sources complicates both interpretation and accountability in the assessment process. Accountability was further emphasized by AA4, who underscored the limitations of design professionals in influencing the quality of installed equipment. As AA4 explained, “as a designer, I have zero control over poor quality of equipment that is actually installed.”

This highlights the shift in responsibility to FMs and equipment suppliers during commissioning and operation, with NABERS offering little capacity to address failures arising from poor equipment quality or inadequate BMS oversight. The significance of proper commissioning and sensor placement was emphasized by AA5, who pointed out that “if the sensors are in the back of the fan coil unit, that all needs to be set up correctly… if you’re not doing that and people are too cold or too hot, then it’s going to affect the energy at the end of the day.” AA5 also highlighted practical examples of malfunctioning systems, explaining that “the lighting sensors just don’t switch off … or you might have a lighting system that just stays on fully all the time because it was that complicated to actually go to the back end and actually change the parameters.” Such insights illustrate how even technologically advanced systems can fail to deliver expected outcomes if not correctly commissioned or if their complexity hinders operational fine-tuning.

Finally, AA6 captured the multidimensional nature of the problem, attributing it to the interaction of multiple factors: “it’s poor-quality sensors and meters as well, not being commissioned properly.” This reinforces the argument that sensor inaccuracy arises not from a single cause but from the convergence of poor equipment quality, inadequate commissioning, and faulty management systems. Table 4.14 below presents qualitative findings and quantitative judgments.

Table 4.14: Qualitative findings and quantitative judgments on inaccuracy of sensor network data

Inaccuracy of sensor network data (poor quality sensors)	Quantitative	Qualitative
AA1		• NABERS may indirectly incentivise stakeholders to investigate malfunctioning sensors when poor operational performance is identified. • NABERS does not provide a direct mechanism to ensure the quality or long-term reliability of the installed sensor infrastructure. • Fragmented ownership and sourcing of sensors present an additional challenge for operational energy management.
AA2	(green)	• Sensor reliability and data quality represent major challenges for operational energy performance. • NABERS projects often require more robust metering strategies than conventional buildings; however, this does not necessarily guarantee higher sensor quality or fewer failures.
AA3	(blue)	• Interpretive difficulties can arise when operational data are generated, owned, or managed by multiple stakeholders.

4.3.2.4 Systems Installation Mismatch (HVAC and lighting)

Consisting of two root causes related to HVAC and lighting systems, this theme comprises 14 judgements, of which 8 (43%) were uncategorized and 6 (57%) green. AA2 argued that such mismatches would be "very apparent in the model" and stressed that NABERS is "very sensitive" to system mismatches ensuring that any changes are captured through updated modelling exercises. AA3 reinforced this by highlighting the role of predictive energy modelling conducted at RIBA stage four, where detailed specifications for all MEP plants were tied directly to achieving the NABERS rating. In this context, NABERS was described as a "strong reason to choose to really narrow down that gap between what you expect and what we want."

AA4 described the process as iterative, noting that once technical specifications are finalized at stage four, the model is updated to reflect them, and further adjustments are made at the as-built stage using actual installation drawings. However, AA4 also acknowledged the realities of value engineering, whereby cost constraints might drive substitutions: "you might specify something, but because of cost, it might not really be installed." Nevertheless, NABERS was seen to maintain a level of discipline, since any such changes would still be evaluated for their impact on the targeted rating, even if the intended outcome is achieved later in operation rather than immediately.

AA5 stressed that system changes, particularly major ones such as air handling units or lighting systems, would necessarily trigger re-modelling to ensure compliance with NABERS targets.

While minor adjustments might not require the same treatment, more significant substitutions could erode the "modelling margin," which provides a buffer to maintain a five-star rating despite small variations. This margin, however, was not seen as a license for negligence, but rather as a tool for making "everybody aware" of the operational implications of design or installation changes.

AA6 highlighted that NABERS provides a structured mechanism to address installation mismatches, both during design and construction. Changes are formally captured in the rating achievement plan, and "if during the construction stage, anything changes, it's highly encouraged to rerun the simulation and see how it affects the rating." AA6 further pointed out the contractual aspect, where tenants are often bound by fit-out specifications aligned with NABERS modelling assumptions (e.g., maximum lighting and plug load intensities). In this respect, NABERS was considered to exert "a greater level of control compared to other schemes," ensuring that deviations in installation do not undermine long-term performance. The insights suggest that NABERS has considerable potential to limit risks associated with system installation mismatches, though its effectiveness is partly contingent upon enforcement, contractual rigor, and the willingness of stakeholders to engage in re-modelling when deviations occur. Table 4.15 below presents qualitative findings and quantitative judgments.

Table 4.15: Qualitative findings and quantitative judgments on systems installation mismatch (HVAC and lighting)

Systems installation mismatch (HVAC and lighting)	Quantitative	Qualitative
AA2		• Mismatches are very apparent in models and NABERS is very sensitive to that.
AA3		• Predictive modelling at RIBA stage four ties detailed MEP plant specifications to achieving the NABERS rating.
AA4		• Models are updated iteratively at stage four and as built, but value engineering may alter installed systems.
AA5		• Major changes like air handling units or lighting require re-modelling. • Modelling margin helps maintain target but is not a licence for negligence.
AA6		• Changes during construction should trigger re simulation. Fit out specifications contractually bound to NABERS assumptions.

4.3.2.5 System Deterioration

This theme, which included four root causes, produced 28 judgments. Of these, eight were green, seven categories coloured blue, and 13 causes were left uncategorized. This theme therefore contained the highest proportion of blue judgments, highlighting areas in need for policy intervention as shown in colour coded matrix in Figure 4.12.

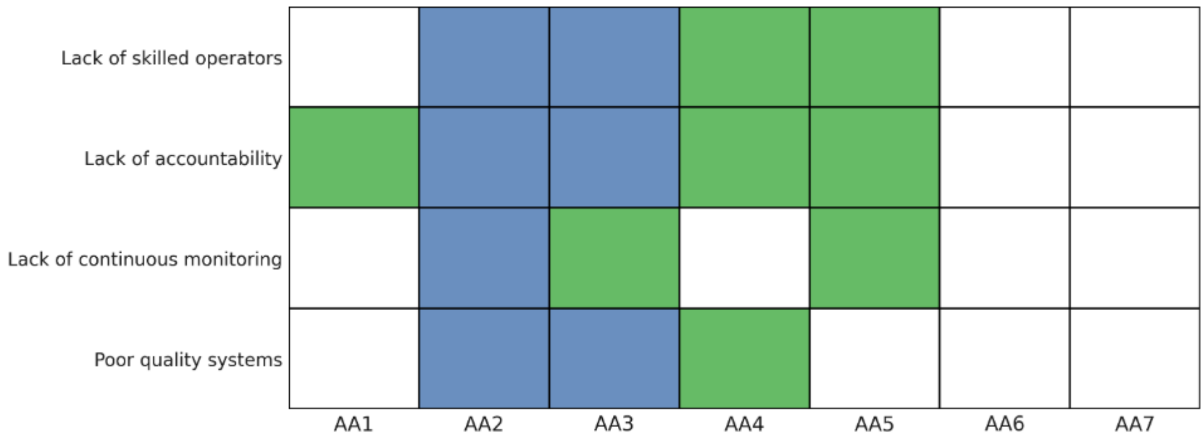


Figure 4.12: Synthesis of perspectives on deterioration of energy systems

Several participants emphasized the role of skilled operators in maintaining long-term system performance. AA1 and AA2 argued that, without operator expertise, building systems can drift from design intent and accumulate inefficiencies over time. AA3 similarly noted that “if operators are not properly trained, systems become dependent on automation that can easily fall out of calibration,” highlighting the fragility of performance without knowledge and oversight.

Participants also identified weak accountability as a barrier to continuous monitoring. AA4 explained that responsibility is often “fragmented across contractors, consultants, and building owners,” allowing issues to persist without clear ownership. AA5 suggested that NABERS could strengthen accountability through more transparent reporting of operational outcomes, while AA6 linked the issue to organizational culture, arguing that effective continuous monitoring requires a broader management shift, not just better operators.

AA7 added that monitoring and operator skill cannot compensate for poor system quality once weak systems are installed. He argued that NABERS could help by setting clearer expectations for system quality, rather than focusing only on outcomes. Overall, participants acknowledged that NABERS cannot directly solve structural problems such as operator shortages or weak procurement but saw its outcome-based framework as capable of mitigating system deterioration through ongoing verification, transparent reporting, stronger accountability, and incentives for better training and equipment quality. Table 4.16 presents participants' perspectives on root causes associated with systems deterioration.

Table 4.16: Synthesis of perspectives on root causes associated with system deterioration

Root cause	AAs perspectives
Lack of skilled operators	• Without operator expertise, systems drift from design intent and inefficiencies accumulate. • If operators are not trained, systems rely on automation that can fall out of calibration.
Lack of accountability	• Accountability fragmented across contractors, consultants and owners. • NABERS could strengthen accountability by requiring transparent reporting.
Installation of poor-quality systems	• Once poor systems are installed, no amount of monitoring or operator skill can deliver high performance. • NABERS could set clearer standards on system quality.

4.3.2.6 Functional and Operational Changes

This theme, encompassing four root causes (lighting schedules, occupancy loads, operating hours, and temperature setpoints), generated 28 judgments. These were divided into six (21.4%) green, 3 (10.7%) blue, 18 (64.3%) uncategorized, and 1 (3.6%) red. This indicates that while most responses were left uncategorized, a notable share highlighted gaps in operational practices, and one judgment indicated causes viewed as beyond practical control as shown in the color-coded matrix in Figure 4.13.

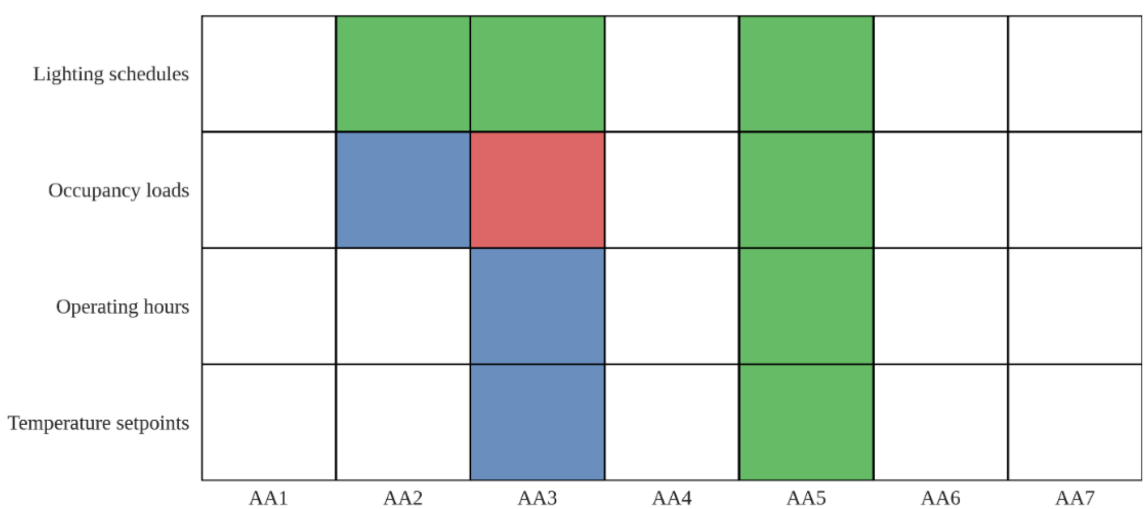


Figure 4.13: Synthesis of perspectives on functional and operational changes

The responses demonstrated a consensus that these operational dynamics exert a strong influence on energy outcomes and consequently on the reliability of NABERS ratings. AA1 highlighted that "NABERS is capable of capturing the effect of changes in operational schedules, such as lighting or operating hours, because the rating is based on actual utility data." This was echoed by AA2, who emphasized that the reliance on measured consumption ensures that operational shifts are reflected in the rating outcome. At the same time, AA3 and AA4 raised concerns about the limits of NABERS in anticipating or fully accounting for changes in occupant load or set-point adjustments. AA3 explained that while NABERS captures the "outcome of increased loads or altered temperatures," it does not prevent such changes, leaving responsibility with FMs. Similarly, AA4 stressed that the system "reflects operational reality but cannot regulate it," underlining that the framework benchmarks performance impacts but does not control day-to-day choices.

AA5 and AA6 further emphasized that NABERS, while valuable, is reactive rather than proactive. According to AA5, the system's role is "to measure and report," which ensures accountability but does not directly influence operational decision-making. AA6 reiterated that NABERS can reveal the impacts of functional and operational changes, yet its effectiveness depends on whether building operators use the feedback to adjust practices. Overall, the discussion revealed that NABERS has clear potential in recognizing the implications of functional and operational changes since these are reflected in its energy performance ratings. However, its effectiveness as a regulatory mechanism is limited, as it primarily reports and benchmarks outcomes rather than shaping day-to-day decisions. Table 4.17 below presents participants perspectives on root causes associated with functional and operational changes.

Table 4.17: Synthesis of perspectives on root causes associated with functional and operational changes

Root cause	AAs perspectives
Functional and operational changes	• Reliance on measured consumption ensures operational shifts are reflected in outcomes. • NABERS captures outcomes of increased loads or altered temperatures but does not prevent them. • NABERS reflects operational reality but cannot regulate day to day choices. • NABERS role is to measure and report, ensuring accountability but not shaping decisions. • NABERS reveals impacts of changes, but effectiveness depends on operators responding to feedback.

4.3.2.7 Thematic Findings and Co-occurrence Analysis

At the level of individual causes, psychological causes associated with occupant behaviour, lack of skilled operators, lack of accountability, and poor-quality systems each attracted 29% blue, underscoring persistent concerns about human and organizational capacity. By contrast, system installation mismatches, and lighting schedules were coded green in 43% of cases, reflecting confidence in NABERS' coverage of installation and scheduling issues. Meanwhile, Poor commissioning, Poor quality sensors, Operating hours, and Temperature setpoints were uncategorized in 71% of judgments, highlighting uncertainty and reluctance to categorise causes that are often context dependent.

Differences were also apparent across participants. AA5 was green leaning, with 10 of 16 judgments. AA3 presented a more balanced distribution, with 7 of 16 green and 8 of 16 (50.0%) blue, highlighting both confidence in NABERS' coverage and repeated identification of needy intervention. AA2 also produced a mixed profile, with 4 of 16 green and 4 of 16 blue, leaving the remainder uncategorized. AA4 contributed 3 of 16 (19%) green and 1 of 16 (6%) blue, with most responses uncategorized. In contrast, AA1, AA6, and AA7 coded almost all of their responses uncategorized, showing a preference for elaboration without categorization. Table 4.18 presents distribution of judgments.

Table 4.18: Participant-level Distribution of Judgments

Participant	Addressed by NABERS (Green)	Policy Intervention required (Blue)	Uncertain position (Grey)	Beyond control (Red)	Elaboration without a category (Uncategorized)
AA1	1	–	–	–	15
AA2	4	6	1	–	5
AA3	7	8	–	1	–
AA4	3	3	–	–	10
AA5	10	–	–	–	6
AA6	–	–	–	–	16
AA7	–	–	–	–	16

Co-occurrence analysis confirmed systemic patterns. In green responses, HVAC systems and Lighting systems were consistently paired, reflecting perceptions that NABERS' coverage extends across installation-related causes. In Blue responses, co-occurrence was strongest among Lack of skilled operators, Lack of accountability, and Poor-quality systems, revealing that participants viewed these as interconnected organisational gaps requiring bundled interventions. Grey and uncategorized co-occurrences were rare but occurred in commissioning and operational setpoint causes, highlighting areas of hesitation. Table 4.19 show distribution of judgments across themes.

Table 4.19: Distribution of judgments across themes

Category	Total Judgments	Addressed by NABERS (Green)	Policy Intervention required (Blue)	Uncertain position (Grey)	Beyond control (Red)	Elaboration without a category (Uncategorized)
Occupant Behaviour (Grey Area + Psychological)	28	3 (11%)	6 (21%)	–	–	19 (68%)
Overestimation of System Efficiency	7	1 (14%)	–	1 (14%)	–	5 (71%)
Inaccuracy of Sensor Network Data	7	1 (14%)	1 (14%)	–	–	5 (71%)
System Installation Mismatch (HVAC & lighting)	16	6 (37.5%)	–	–	–	10 (62.5%)
System deterioration	28	8 (29%)	7 (25%)	–	–	13 (46%)
Functional and operational Changes	28	6 (21%)	3 (11%)	–	1 (4%)	18 (64%)
Overall (All themes)	112	25 (22%)	17 (15%)	1 (1%)	1 (0.9%)	68 (61%)

Overall, the analysis shows that commissioning and operational causes were uncategorized (61%), reflecting uncertainty and reluctance to categorize. Within the categorized responses, green judgments (22%) highlight NABERS' perceived strength in installation mismatch and some operational scheduling, while blue judgments (15%) underscore gaps in human capacity, accountability, and data quality. The clustering of blue causes demonstrates that weaknesses are not isolated but systemic, requiring integrated reforms that address operator skills, accountability mechanisms, commissioning practices, and data reliability together, rather than isolation.

4.3.3 Cross-Cutting Factors

During the interviews, both IDRs and AAs (apart from IDR5) evaluated eight cross-cutting root causes, producing a total of 88 individual judgments. Each judgment was categorized using the same colour code as presented in Figure 4.14 below. Across all judgments, 67 (76%) were green, 14 (16%) were blue, six (7%) were uncategorized, and one (1%) was grey. The predominance of green demonstrates that participants generally believe NABERS has strong potential to mitigate or prevent cross-cutting factors. At the same time, 14 judgments highlight areas where additional policy or procedural intervention is considered necessary, while six judgments remained uncategorized and one reflected uncertainty. However, when analysed by theme, different emphases emerged as in the following sections.

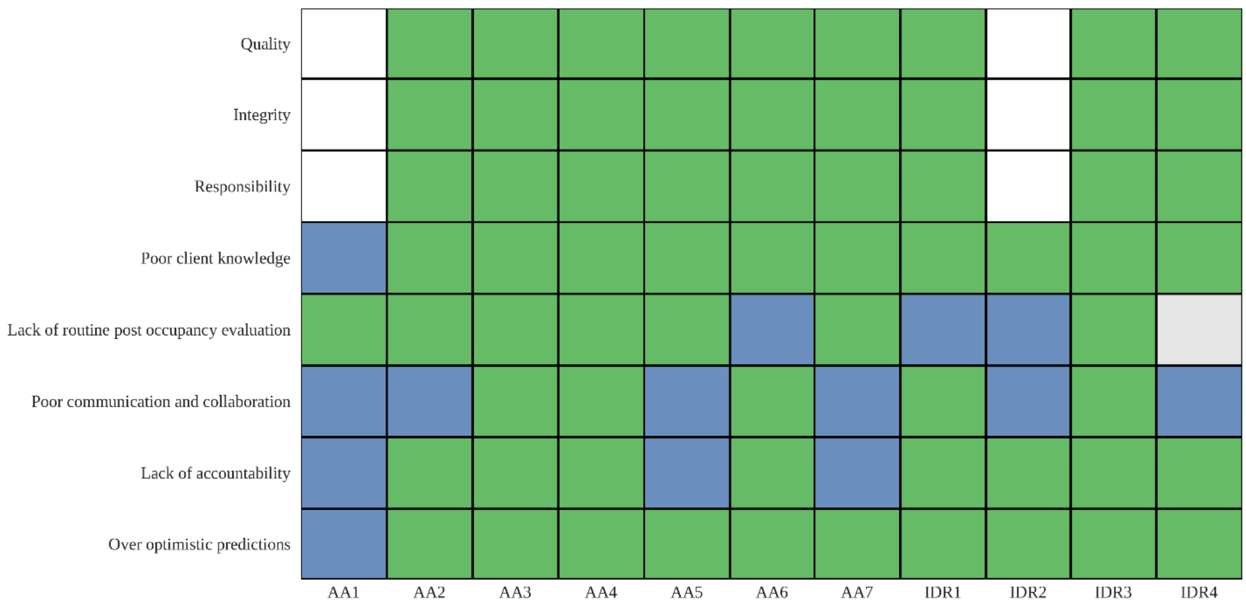


Figure 4.14: Synthesis of perspectives on cross-cutting causes

4.3.3.1 Overall industry culture (quality, integrity, responsibility)

This theme produced 33 judgments, of which 27 (82%) were green and six (18%) uncategorized causes. The results show strong confidence that NABERS can help shape cultural drivers when categories were applied, though some participants elaborated without assigning a definitive code. Both IDRs and AAs agreed that NABERS is shifting norms, yet the effect is uneven and still focused on offices. IDR1 argued that NABERS is “upskilling modellers and making people aware of the things that make buildings not perform,” while cautioning that this impact is strongest among larger, more capable clients. IDR2 called the change a “slow burn,” limited by the current office focus, noting that many other sectors remain driven by compliance ratings that bear little relation to real energy use. By contrast, IDR3 saw clear potential across the triad of quality, integrity and responsibility, emphasizing that designs are reviewed from many points of view and that contractors can be held accountable for what they deliver. IDR4 reported that clients are beginning to realize they cannot hide behind compliance tricks, describing the “shock” when those used to achieve BREEAM or EPC targets and learn that with NABERS “there is nothing we can really do,” she judged that the cultural shift has already started.

AA echoed these experiences. AA1 was cautious about claims of transformation in the UK, judging it too early to see a step change, although he accepted that expansion beyond offices could move the needle. AA2 agreed that change is real but presently concentrated in high end offices with strong net zero commitments. AA3 was unequivocal, responding “yes, all three” when asked whether NABERS lifts quality, integrity and responsibility, and described it as a clear framework that shows best practice. AA4 linked culture change to market incentives and corporate strategies, observing that a developer who pursues a rating signals higher integrity and responsibility and is tied to published targets through competition.

AA5 pointed to concrete practice shifts: “the quality of modelling has improved as teams adopt more advanced tools and stay involved through to in use so that a five-star model must be matched by a five-star outcome”. AA6 described the steep learning curve for the next three to five years, with many workshops becoming lectures on what to do, but also found people keen to learn. AA7 added that the star system simplifies efficiency for clients and that evidence of rental premium for higher ratings reinforces demand and cultural change.

4.3.3.2 Poor client knowledge

Under this theme, 11 judgments are produced, with 10 (91%) green and 1 (9%) blue, indicating widespread agreement on NABERS’ potential but recognition of the need for additional educational or policy support as depicted in Figure 4.15 below.

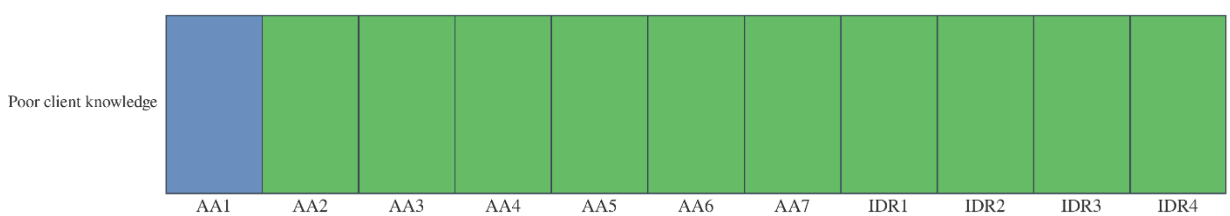


Figure 4.15: Synthesis of perspectives on poor client’s knowledge

IDRs stressed that many clients lack the technical understanding to challenge optimistic predictions or to interpret operational data. IDR1 noted that without clear education, “clients are not well equipped to know whether what they are being promised is realistic,” making them reliant on consultants whose incentives may not align with long-term performance. IDR2 linked this to the compliance driven culture of the UK, where EPC and BREEAM dominate: “clients have been told for years they are getting good buildings, but they have not been given the tools to see that this is not true." IDR3 suggested that NABERS provides a "language and a structure" that can improve understanding but warned that it takes time for clients to become confident in using it. IDR4 added that education is vital, since "if you do not understand how the target works, you cannot hold your design team to account."

AAs pointed to practical ways in which NABERS helps. AA1 argued that "the single star number is easy for clients to understand," contrasting this with abstract EPC scores. AA2 agreed, noting that NABERS "gives a simple metric that links directly to energy bills," making it relatable. AA3 observed that where NABERS is embedded, clients "become more inquisitive and start to ask better questions." AA4 emphasized that knowledge gaps remain: "clients often do not know what questions to ask," and NABERS cannot fill all gaps on its own. AA5 pointed to the role of the Rating Achievement Plan and SL style processes in educating clients about what is required after handover. AA6 described how annual re-rating creates a feedback loop, helping clients to see performance year on year. AA7 concluded that "while NABERS is not an education programme, it does make ignorance harder to sustain," since poor operational results are clearly visible in the public star rating.

In synthesis, poor client knowledge remains a structural barrier to closing the performance gap. NABERS mitigates this by providing a clear and communicable outcome metric, by embedding structured processes such as the Rating Achievement Plan, and by requiring annual re-rating that keeps performance visible. However, education and capacity building are still required if clients are to actively manage their buildings and hold consultants and contractors accountable. The scheme shines light on outcomes but cannot replace the need for clients to understand and act on what those outcomes mean.

4.3.3.3 Lack of routine post occupancy evaluation

This theme also generated 11 judgments, comprising seven (64%) green, three (27%) blue, and 1 (9%) grey as shown in Figure 4.16 below. This mixed distribution shows that while most participants see NABERS as a credible mechanism for institutionalizing POE, others identified procedural gaps or expressed uncertainty about its consistency in practice.

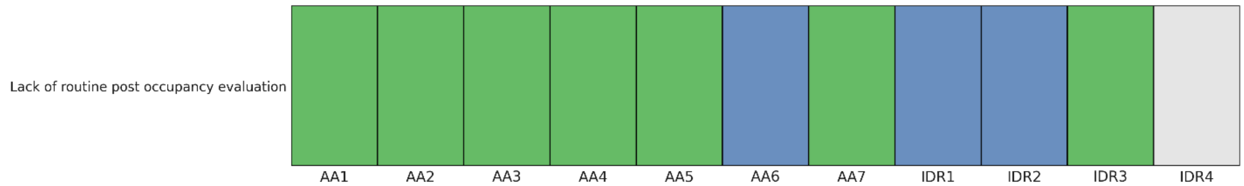


Figure 4.16: Synthesis of perspectives on lack of post occupancy evaluation

IDRs were divided with regards to NABERS' potential to address the cross-cutting factor of lack of routine post occupancy evaluation. IDR1 noted that NABERS does not explicitly require POE beyond energy monitoring, arguing that if market demand is weak, few clients will commission detailed evaluations. IDR2 agreed, explaining that in the UK "clients rarely pay for POE" and that NABERS only captures energy outcomes, not user experience. By contrast, IDR3 described POE as "a cornerstone of NABERS" in practice, citing quarterly reporting and ongoing measurement as integral parts of the scheme. IDR4 was skeptical, stating that she had "seen no evidence" that NABERS is changing evaluation culture and warning that without stronger requirements, POE remains rare.

AAs were more optimistic. AA1 emphasized that the annual re-rating creates an ongoing evaluation cycle, forcing clients to confront operational realities each year. AA2 and AA3 agreed, highlighting that NABERS "keeps the conversation alive" after handover and generates a feedback loop that is absent in traditional compliance schemes. AA4 stressed that while NABERS focuses on energy rather than wider POE dimensions, its mandatory monitoring requirements are nonetheless transformative compared to EPC or BREEAM. AA5 pointed out that POE value depends on FM teams using the data, not just collecting it, but saw NABERS as embedding this discipline. AA6 was more cautious, noting that NABERS leaves responsibility with building owners and operators, and suggesting that POE should be explicitly mandated. AA7 concluded that while NABERS is not a full POE tool, its regular ratings make ignorance of performance harder to sustain. In conclusion, participants agreed that POE remains underdeveloped in the UK but differed on how far NABERS change this. IDRs highlighted NABERS' limited scope and questioned whether it really embeds routine evaluation, while Assessors argued that annual ratings and mandatory monitoring represent a significant cultural shift.

4.3.3.4 Poor Communication and Collaboration

This theme comprising 11 judgements was the most blue-oriented. IDRs offered mixed perspectives on the extent to which NABERS addresses this weakness. With six judgments (54.5%) coded blue and five (45.5%) green as presented in Figure 4.17 below. This highlights that participants regard collaboration as requiring procedural reforms beyond what NABERS currently delivers.

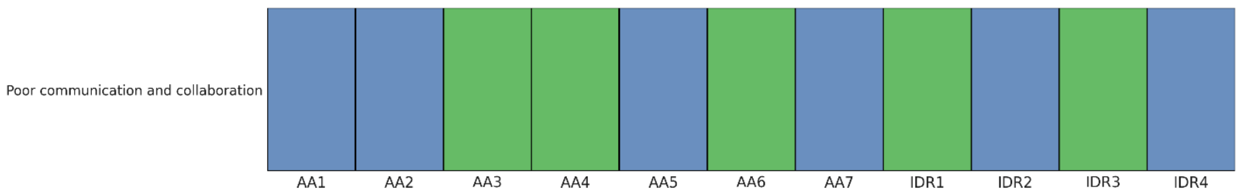


Figure 4.17: Synthesis of perspectives on poor communication and collaboration

IDR1 suggested that the scheme helps by “bringing the team together, the design team and the client, in a common goal to get the NABERS rating,” though he noted that architectural issues often surface late at stage four when they are difficult to resolve. IDR2 judged collaboration to be improving but still “disjointed,” arguing that fundamentals are not being addressed early enough and predicting that practice would improve within the next two to three years. IDR3 pointed to early appointment as critical: “if we are appointed early in the design, we can communicate more clearly,” while IDR4 was skeptical, stating “I do not think NABERS addresses this issue and more could be done.”

AAs provided a wider range of views. AA1 downplayed NABERS impact, saying “it is still just a design process,” since assessment can be done without intensive collaboration. AA2 saw some positive effects through post occupancy processes but stressed that outcomes depended on client engagement. By contrast, AA3 judged NABERS to be “the gold standard of how the design team and procurement process needs to work,” because responsibilities are clearly set out. AA4 strongly endorsed its impact, arguing that NABERS “really drives a lot of discussions and collaboration between different stakeholders” at the design stage. AA5 suggested that mandatory publication of ratings would further strengthen collaboration and transparency by raising the stakes. AA6 supported this view, observing that NABERS projects require coordination across engineers, architects, contractors and FMs, improving teamwork significantly compared to business as usual.

AA7 cautioned that such effects are mainly visible in the DfP pathway rather than in NABERS In Use.

In conclusion, participants agreed that poor communication and collaboration remains structural problems in UK, but several stressed that NABERS introduces mechanisms that partially mitigate them. IDRS and shared targets foster more integrated teamwork where applied, though skepticism remains about reach and consistency. AAs described clearer lines of responsibility and stronger dialogue in NABERS projects but also noted that impact is uneven and highly dependent on client motivation and whether ratings are public. Overall, NABERS was seen as a catalyst for better collaboration in some contexts, but not yet a comprehensive solution.

4.3.3.5 Lack of Accountability

This theme produced 11 judgments, with eight (73%) green and 3 (28%) blue as shown in Figure 4.18 below. The majority of participants therefore considered NABERS an effective accountability mechanism, though a smaller share still pointed to the need for stronger policy interventions.

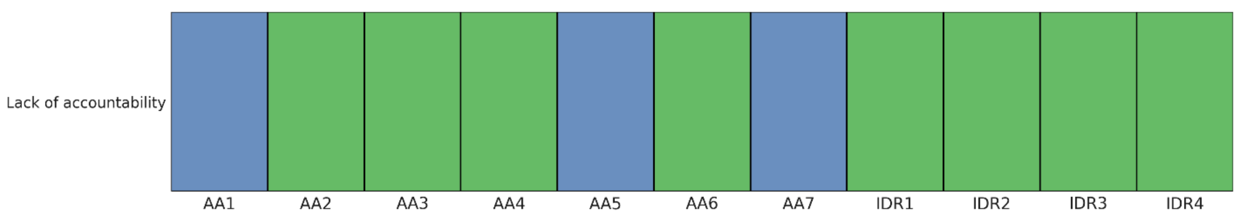


Figure 4.18: Synthesis of perspectives on lack of accountability

Interviewees repeatedly described NABERS as a framework that partially addresses lack of accountability. IDRs generally saw NABERS as strengthening accountability. IDR1 explained that the process “creates a sense of shared responsibility, we are all in a team, we are all responsible for getting the final rating.” IDR2 emphasized that NABERS requires sign off and validation: “you need an engineer to sign off, you cannot just say we have done it,” forcing contractors to take responsibility for what is delivered. IDR3 pointed to the visibility of the outcome, noting that because the rating is public, “the owner is ultimately accountable,” though support comes through post occupancy evaluation and data. IDR4 gave a clear endorsement, judging that NABERS does address accountability in ways that EPC or BREEAM do not.

AAs offered a more varied set of perspectives. AA1 argued that accountability remains fragmented between clients, contractors and facilities managers depending on the project. By contrast, AA2 stressed that linking rental value to ratings creates a financial form of accountability, describing it as "all about the money." AA3 described NABERS as providing "very clear accountability," while AA4 agreed that responsibilities under NABERS are more specific than in conventional schemes. AA5 struck a more cautious note, pointing out that while trackers exist, if the target is not achieved "there is nothing to enforce, it is just a commercial issue," since designers cannot control tenant behaviour in use. AA6 was strongly positive, arguing that NABERS removes unknowns and means "if you want to keep your ratings, you need to be accountable for everything." AA7 concluded that NABERS helps make accountability visible, but whether someone is actually held to account depends on governance and project structures.

In synthesis, accountability is widely regarded as weak in current UK practice, but NABERS provides mechanisms that improve clarity and visibility. Through IDRs, engineers sign off, public star ratings and links to rental value, the scheme creates stronger incentives for contractors, clients and owners to take responsibility. However, several experts cautioned that accountability remains conditional: outcomes depend on contracts, governance and whether clients choose to act on the information provided. NABERS therefore makes responsibility more visible and harder to avoid but does not guarantee enforcement.

4.3.3.6 Regulatory pressure (over-optimistic predictions)

This theme yielded 11 judgments, of which 10 (91%) were green and 1 (9%) blue as shown in Figure 4.19 below. Participants thus generally considered NABERS effective at disciplining over-optimistic assumptions through measured outcomes and benchmarks, with limited recognition of a continuing need for policy reinforcement.

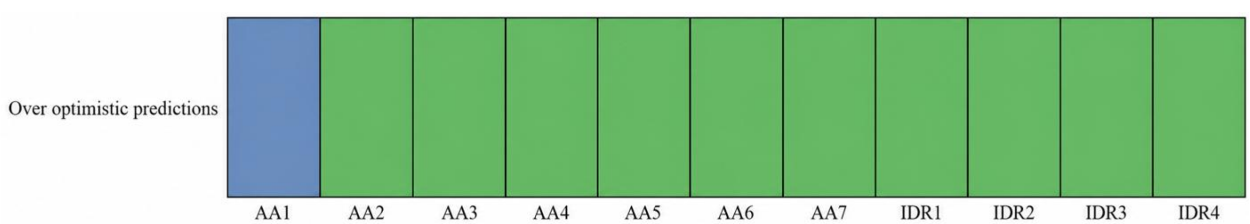


Figure 4.19: Synthesis of perspectives on over-optimistic prediction

Most IDRs and AAs agreed that NABERS represents a clear break from compliance-based modelling. IDR1 argued that while compliance modelling is a tick-box exercise often built on optimistic assumptions, NABERS reframes building energy simulation tools as risk-analysis tools that test controls and scenarios to determine the likelihood of achieving a rating. He recalled that in the early stages some teams aimed for unrealistic 5.5-star ratings, but expectations are now more grounded. IDR2 stressed that clients and design teams often push for five stars, but IDRs should insist on more realistic targets to avoid reputational risks, describing NABERS as shifting thinking away from regulation toward actual energy outcomes. IDR3 reinforced that NABERS and compliance models serve different purposes, and NABERS is not constrained by regulatory pressures. IDR4 agreed, stating that NABERS goes beyond regulations and should be treated as the toughest hurdle rather than bent to regulatory requirements.

AAs echoed these points but added perspectives on UK regulatory alignment. AA1 observed that regulatory pressure is not strong in the UK yet, though some planning authorities are beginning to reference NABERS in policies. AA2 argued that NABERS eliminates compliance gaming by focusing on metered outcomes. AA3 suggested that NABERS is now shaping regulation itself, citing City of London planning policy recommending NABERS for offices. AA4 highlighted that EPCs, and Part L models rely on generic assumptions, while NABERS is project specific and includes sensitivity analysis to produce more realistic predictions. AA5 cautioned that optimism remains a risk, since modelling margins provide confidence but can still be stretched to claim high ratings. AA6 strongly endorsed NABERS as more detailed than regulatory models, accounting for unregulated loads and complex simulations. AA7 concluded that NABERS goes beyond regulations by validating predictions against in use ratings, exposing optimism if it exists.

Participants agree that regulatory modelling encourages over-optimistic predictions and contribute little to real energy outcomes. NABERS was seen as a corrective, focusing on measured data, project specific modelling and sensitivity analysis. IDRs highlighted its role in managing client expectations and preventing reputational risk, while AAs stressed its ability to eliminate gaming and even influence regulation itself. Nevertheless, some cautioned that optimism persists and that enforcement and realistic target setting remain critical. Overall, NABERS was judged to reduce but not eliminate over optimistic predictions by reframing energy performance as an outcome to be achieved and verified rather than a regulatory box to be ticked.

4.3.3.7 Thematic Findings and Co-occurrence Analysis

At the level of individual causes, poor client knowledge and over-optimistic predictions were green leaning (91%), alongside the cultural causes of quality, integrity, and responsibility (82%). By contrast, Poor communication and collaboration attracted the largest share of Blue (54.5%), followed by post-occupancy evaluation (27%) and Lack of accountability (27%). Post-occupancy evaluation was also the only cause to receive a grey judgment, reflecting uncertainty.

Differences were also apparent across participants as shown in Table 4.20 below. AA3, AA4, and IDR3 coded all eight root causes green, demonstrating strong confidence in NABERS. AA2, AA6 and IDR1 were also highly green leaning, each with seven judgments (87.5%) green. AA5, AA7, IDR2 and IDR4 coded six judgments (75%) green, with the remainder divided between blue (for AAs) and grey (for IDR4). In contrast, AA1 gave only 1 of 8 judgments (12.5%) as green, with 4 (50%) blue and 3 (37.5%) uncategorized, while IDR2 distributed judgments across 3 (37.5%) green, 2 (25%) blue, and 3 (37.5%) uncategorized. These two participants accounted for almost all uncategorized judgments and a disproportionate share of blue judgements, reflecting more cautious or critical perspectives on NABERS' capacity to address cross-cutting factors.

Table 4.20: Participant-level Distribution of Judgments

Participant	**Addressed by NABERS (Green)**	**Policy Intervention required (Blue)**	**Uncertain position (Grey)**	**Elaboration without a category (Uncategorized)**	**Total**
AA1	1	4	–	3	8
AA2	7	1	–	–	8
AA3	8	–	–	–	8
AA4	8	–	–	–	8
AA5	6	2	–	–	8
AA6	7	1	–	–	8
AA7	6	2	–	–	8
IDR1	7	1	–	–	8
IDR2	3	2	–	3	8
IDR3	8	–	–	–	8
IDR4	6	1	1	–	8
Overall	**67**	**14**	**1**	**6**	**88**

Patterns of co-occurrence reinforce these findings as shown in *Table 4.21* below. Green responses clustered strongly between poor client knowledge and over-optimistic predictions, and across the cultural triad of quality, integrity, and responsibility. This indicates that participants perceived NABERS' coverage as systemic across these related causes. Blue responses clustered most strongly between Poor communication and collaboration and lack of accountability, with occasional co-occurrences linking them to post-occupancy evaluation. These patterns highlight that gaps are perceived as interdependent and require bundled interventions. Co-occurrence of uncategorized causes were limited to the industry culture, reflecting participants who elaborated but did not assign codes.

Table 4.21: Distribution of judgments across themes

Category	Total Judgments	Addressed by NABERS (Green)	Policy Intervention required (Blue)	Uncertain position (Grey)	Elaboration without a category (Uncategorized)
Overall industry culture (Quality, Integrity, Responsibility)	33	27	–	–	6
Poor client knowledge	11	10	1	–	–
Lack of routine post-occupancy evaluation	11	7	3	1	–
Poor communication and collaboration	11	5	6	–	–
Lack of accountability	11	8	3	–	–
Regulatory pressure	11	10	1	–	–
Overall (All Themes)	**88**	**67**	**14**	**1**	**6**

4.3.4 Combined Synthesis

Across all stages and all root causes identified and discussed, 295 individual judgments were made by participants on NABERS' potential to mitigate the EPG. When aggregated, the distribution was 153 green (52%), 45 blue (15%), five grey (2%), one red (0.3%), and 91 uncategorized (31%). This high-level pattern highlights three important dynamics. First, a majority of responses were coded green, demonstrating that participants consistently perceive NABERS as already addressing, or having strong potential to address, a wide range of root causes. Second, blue judgments account for just over 15% of the total, marking a clear minority but signalling recurrent gaps where participants believe supplementary policy or procedural interventions are required. Third, the sizeable proportion of uncategorized causes, concentrated in the commissioning and operational stages, shows that participants often elaborated without assigning a definitive category, reflecting either uncertainty in applying the framework or hesitation in attributing responsibility to NABERS.

When the data is compared across stages, distinct profiles emerge. In the design stage, green was dominant at 64%, with policy-sensitive gaps (15% blue) centred around analytical and modelling issues such as stochastic processes and software limitations. This stage exhibited very few uncategorized or uncertain judgments. By contrast, the commissioning and operational stages were characterized by hesitation: 61% of responses were uncategorized, with green judgments comprising only 22%. Where categories were applied, participants highlighted both NABERS' coverage of installation-related issues and recurring gaps in skills, accountability, and monitoring, reflected in the 15% blue responses. Finally, in the cross-cutting factors, confidence in NABERS was strongest, with 76% of judgments coded green, and the balance split between 16% blue and a small share of uncategorized and grey. Here, NABERS was seen as particularly effective in addressing issues of client knowledge, cultural drivers, and over-optimistic predictions, while communication and accountability remained policy-sensitive bottlenecks.

At the level of individual causes, the highest concentrations of green judgments were found for oversized HVAC components, undefined functions or requirements, client knowledge, over-optimistic predictions, and the cultural drivers of quality, integrity, and responsibility. In each case, participants showed widespread confidence in NABERS' ability to address these areas through measured outcomes, verification, and disclosure. Conversely, the greatest concentrations of blue judgments were found in software limitations, poor communication and collaboration, lack of accountability, and organisational capacity issues such as lack of skilled operators and poor-quality systems. These areas consistently signalled the need for further policy or procedural interventions. Grey judgments were rare, recorded primarily in relation to efficient but complex design and post-occupancy evaluation, while the only red responses arose in functional and operational changes, where two judgments classified certain factors (such as operating hours and temperature setpoints) as beyond control.

Patterns at the participant level also show variation. Some interviewees, such as IDR5, IDR3, AA3, and AA4, coded the majority or entirety of their judgments green, showing strong confidence in NABERS' coverage. Others, such as IDR2 and AA1, were far more cautious, leaving many responses uncategorised or highlighting policy gaps instead of attributing them to NABERS. These profiles illustrate that while there was broad agreement on NABERS' role, individual participants differed in their willingness to assign clear coverage versus emphasising systemic or contextual limitations.

Co-occurrence analysis across the full dataset reinforces the stage-level findings. In green responses, clusters were evident around design robustness and cultural drivers, showing that participants did not perceive NABERS' strengths as isolated but as systemic across groups of related causes. In blue responses, strong clustering occurred around analytical issues (stochastic, software limitation), organisational weaknesses (skills, accountability, poor communication), and monitoring gaps, confirming that participants saw these not as isolated gaps but as interlinked policy challenges. White co-occurrences were particularly common in commissioning and operational causes, where hesitation and elaboration without categorisation were concentrated.

The synthesis of all 295 judgments shows that participants broadly trust NABERS to address many drivers of the EPG, especially at the design stage and in relation to cross-cutting factors. However, consistent clusters of blue responses identify analytical limitations, organisational capacity, accountability, and communication as areas requiring targeted reinforcement. The large share of uncategorized responses in the commissioning and operational stages further suggests a need for greater clarity in role allocation and evaluation processes. Overall, the findings indicate that while NABERS is widely regarded as a robust framework, its effectiveness will be maximised when combined with systemic reforms that integrate policy, procedural, and accountability mechanisms across the building lifecycle.

To conclude, the root causes of the EPG were identified a priori through the SLR and through the RCA conducted where causes are traced to their roots based on a saturation approach, and the purpose of the interviews was not to generate new themes but to evaluate the extent to which NABERS addresses these predefined causes. As such, saturation was assessed in terms of interpretive and evaluative convergence, rather than the emergence of new themes.

Saturation was examined through two complementary approaches. First, a cumulative assessment of interview data was undertaken to identify whether additional interviews introduced new evaluative dimensions. The colour-coded matrices (Figures 4.5–4.18) provided structured evidence of convergence across participants. While variation started to be observed in the assignment of colours reflecting the quantitative judgements, qualitative analysis demonstrates that these differences do not necessarily reflect the emergence of new issues, but rather alternative interpretations of the same underlying causes. For example, design-stage issues related to inefficient design, especially issues related to lack of buildability, simplicity and issues in sequencing were consistently mostly coded grey, blue or left uncategorized, with variation arising from differing views on whether these should be addressed at the design or construction-stage leading into the conclusion that while such issues as they start to emerge from design stage, stakeholders involved in the design and construction stages must be aware of such issues from the point at which a designer anticipates those issues might cascade into a construction stage. This claim is justified based on quantitative judgements on poor communication and collaboration issues in the cross-cutting factors wherein six out of the 11 participants claimed that more policy intervention is needed to address such an issue as apparent in Figure 4.14.

Similarly, in commissioning and operational stages, strong repetition of key issues, particularly lack of skilled operators and lack of accountability was observed across AAs, with consistent explanatory narratives. In cross-cutting factors, apparent differences in judgment (e.g., whether NABERS has already influenced industry culture or has the potential to do so) reflect temporal positioning rather than conceptual divergence, as all participants acknowledged its role in shaping practices, albeit to varying degrees depending on its maturity in the UK context and its current scope for offices.

Importantly, the absence of new root causes should not be interpreted as a limitation of the data, but rather as a consequence of the study design, which deliberately constrained the analytical framework to established causes identified in the SLR. Within this framework, saturation and sample size are based at the point at which additional interviews no longer yield new interpretive positions or explanatory mechanisms. Evidence from both the cumulative assessment and the matrices indicates that this point was reached within the 12 interviews conducted.

Nevertheless, it is acknowledged that saturation was not uniform across all aspects. In particular, insights related to commissioning were more limited, as AAs had less direct experience in this stage. Furthermore, given the relatively recent implementation of NABERS in the UK, variability in participant perspectives reflects differing levels of exposure and experience. These aspects are therefore interpreted as providing an exploratory outline rather than a fully saturated domain. Therefore, the combination of diminishing analytical returns, repetition of explanatory patterns, and convergence in interpretation across participants provides robust evidence that the data collected were sufficient to address RO2 and part of RO5 by having enough evidence to support creation of mitigation strategies presented in under section 4.4.

4.4 Policy Interventions Towards Mitigating the EPG

This section synthesizes the insights gathered from interviews to propose a comprehensive mitigation strategy addressing the EPG causes associated with design, commissioning and operational stages and cross-cutting factors. It consolidates intervention pathways from the qualitative findings discussed in previous sections.

The mitigation framework builds on the methodological integration of RCA using the "5 Whys" technique to uncover systemic interdependencies and expert interviews, which revealed the procedural, contractual, and behavioural mechanisms sustaining the EPG. The diagrams which will be presented under this section are not static illustrations, but analytical syntheses derived from that process. Rather than focusing on quantitative coding, this section privileges qualitative reasoning. The aim is to translate the expert narratives and causal diagrams into a policy-oriented strategy for NABERS that strengthens the link between design intent and operational outcomes.

4.4.1 Design Stage Mitigation Strategies

The design phase represents the most decisive intervention point in shaping operational energy performance. It is where intentions, assumptions, and performance expectations are fixed and where the potential for later efficiency losses is embedded. Several IDRs described how design teams, operating under uncertainty about occupancy and climatic conditions, systematically oversize equipment "to be safe." One reviewer remarked that "the fear of underperformance outweighs the concern for efficiency." The problem is therefore *behavioural* rather than purely technical. The mitigation logic illustrated in Figure 4.20 reveals the compounded influence of design decisions, technological assumptions, and professional practices. The diagram emphasises that oversizing need not undermine performance if controls are sequenced and verified. As one assessor noted, "systems can be oversized but still efficient when control logic is optimised." NABERS' focus on measured outcomes provides the policy platform for this principle: control verification should be embedded as a formal requirement during the design review and sign-off stages.

The diagram further highlights "lack of appropriate details" and "unclear operational requirements." Interview evidence confirms that drawings and specifications often omit critical metering strategies, control sequences, and maintenance access provisions. NABERS' *DfP* pathway has demonstrated that such omissions diminish sharply when IDRS occur before RIBA stage three. Thus, early IDR involvement must be institutionalised, not left to client discretion.

Another recurring design-stage issue is lack of simplicity. Interviewees observed that "simplicity is not the same as efficiency." Complex systems may yield superior control but demand skilled operation. The mitigation proposed in the diagram is to produce operator-centred documentation that translates technical control logic into accessible operational language. This simple measure bridges the design–operation communication gap that otherwise leads to misuse of control interfaces post-handover. Finally, buildability concerns tight tolerances, sequencing constraints, and proper installation recur throughout the expert testimony. Passive and hybrid systems, though efficient in theory, can fail in practice when installed under humid or variable conditions. The diagram and corresponding narrative call for risk-based buildability reviews within the DfP process to ensure that energy-efficient designs remain robust under real-world conditions.

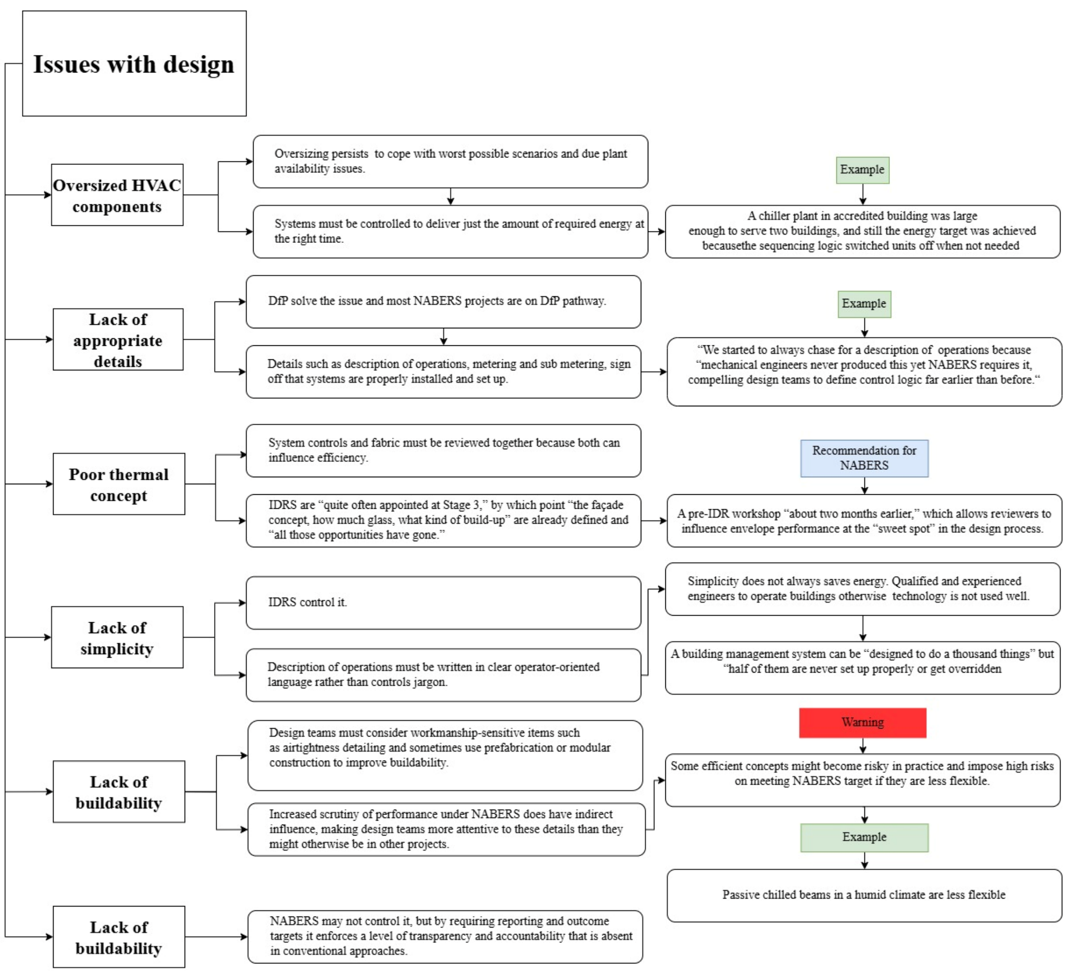


Figure 4.20: Mitigation strategies for issues with efficient and inefficient design

The second diagram in this stage (Figure 4.21) addresses uncertainties in building energy modelling. Experts consistently identified that energy simulations central to NABERS assessments suffer from uncertainty in occupancy assumptions, plug-load densities, equipment schedules, and system performance curves. These are not marginal inaccuracies; they compound into substantial over- or under-predictions of annual consumption. Previous findings show that practitioners frequently rely on default input values due to incomplete brief information or lack of coordination with tenants. IDRs emphasised that occupants have a tendency of overstating their requirement. Mitigation measures captured here include early inter-disciplinary workshops to validate occupancy and equipment assumptions, explicit documentation of data sources, and replacement of defaults with evidence-based inputs.

The diagram further builds on the concept of off-axis scenario testing a practice borrowed from risk engineering where stochastic parameters (such as variable weather years or occupancy schedules) are perturbed to test design resilience. Several participants endorsed this method as a way to move from deterministic prediction to risk-analysis modelling. Policy should codify this approach by requiring that DfP submissions demonstrate sensitivity analysis and disclose uncertainty bounds. A second root cause is software limitation. Issue of software limitation was found to be mitigated through multi-platform verification. For example, checking results across simulation engines or supplementing them with manufacturer data. The policy implication is clear: NABERS should develop modelling guidance that recognises multi-tool validation and mandates justification for simplifications.

Finally, experts drew attention to data transparency. Design teams sometimes “tune” assumptions to meet contractual targets. By integrating measured post-occupancy data into feedback loops, NABERS can progressively improve modelling practices across projects, converting individual experiences into institutional learning.

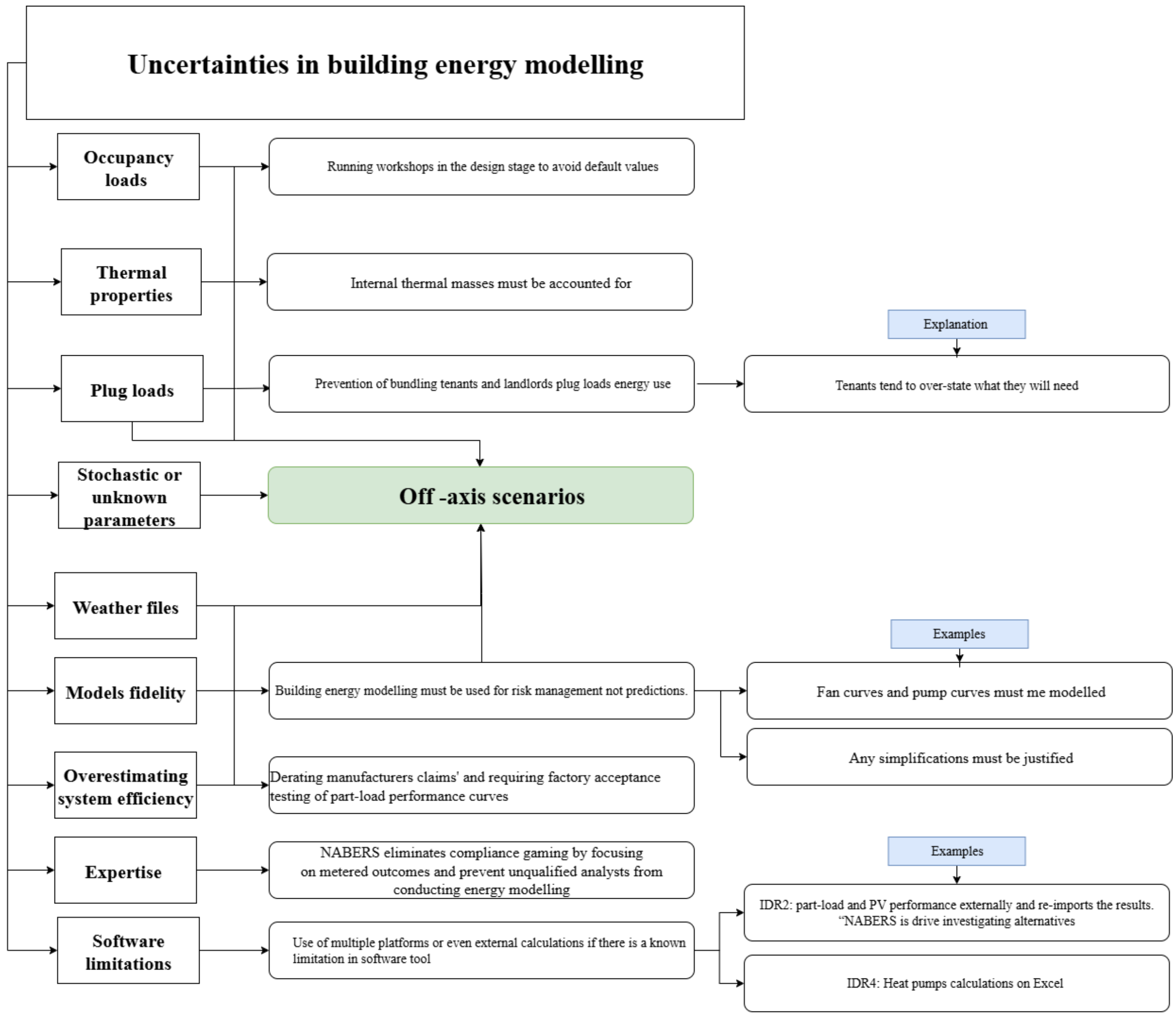


Figure 4.21: Mitigation strategies to address issues related to uncertainties in building energy modelling

The third diagram (Figure 4.22) present participant informed preventative measures for the emergence of inaccurate performance targets and undefined operational requirements, linking design ambition with operational realism. The interviews show that overly ambitious targets are often established to satisfy client expectations or marketing objectives, rather than empirical evidence. As one IDR remarked, "clients want the star before understanding what it means." This creates a misalignment between contractual commitments and deliverable performance.

The commitment agreement was repeatedly cited as a transformative instrument because it converts aspirational targets into binding deliverables. However, experts noted that operational dependencies such as future occupancy density, maintenance regimes, and tenant equipment remain outside the control of design teams. Therefore, design-stage policy must embed these dependencies contractually, ensuring that the NABERS commitment explicitly identifies operational assumptions and allocates responsibility for maintaining them post-handover. The mitigation logic in the diagram emphasises collaboration continuity: design commitments must feed forward into commissioning documentation and operational guidelines. Without this continuity, design intent becomes decoupled from management practice.

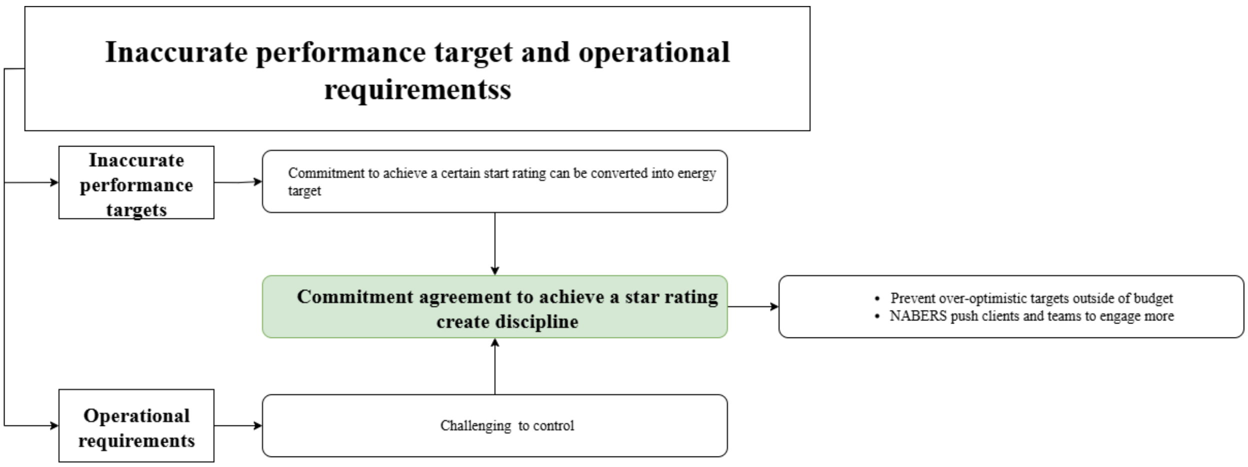


Figure 4.22: Mitigation strategies to address inaccuracy of performance target and undefined operational requirement

4.4.2 Commissioning and Operational Stages Mitigation Strategies

The diagram (Figure 4.23) presented under this section captures the transition from verified design to lived building performance. Insights derived from section 4.3.2 confirm that this stage concentrates the highest density of causes leading to the EPG, not because of technological limitations but because of organisational discontinuities. Once the design team withdraws, performance responsibility diffuses across contractors, facilities managers, and occupants.

Despite advanced controls and metering, final commissioning documentation is often incomplete or disconnected from the design logic. Experts recounted cases where set-points were changed arbitrarily to appease comfort complaints, eroding the calibrated balance achieved in simulation.

The diagram identifies poor commissioning quality, deficient documentation, and absence of functional verification as root causes. The proposed procedural remedies:

- Mandating commissioning plans developed during design,
- Requiring third-party verification of system performance, and

- Linking commissioning completion to the NABERS rating process.

The principle is simple: energy performance cannot be separated from commissioning completeness. Policy must therefore make verified commissioning a pre-condition for NABERS certification. The next branch of the diagram addresses operational degradation. Participants highlighted that FMs "are not trained to interpret energy data," and therefore cannot act effectively on feedback. Control interfaces are complex, while contractual incentives rarely reward efficiency.

The diagram identifies lack of skilled operators, weak accountability, and absence of performance feedback loops as persistent failures. Proposed mitigation strategy here suggests that NABERS framework should establish accreditation pathway for FMs and operators, trained in energy interpretation, control tuning assuring annual re-rating procedures. Furthermore, the diagrams and interviews point to sensor accuracy and data reliability as foundations of measured performance. Inadequate calibration and neglected maintenance undermine trust in metrics. Hence, NABERS should impose minimum sensor and equipment standards and require periodic verification of monitoring systems as part of the annual rating.

A crucial insight from experts is that NABERS' greatest strength lies in its cyclical measurement and disclosure. Annual re-rating transforms compliance into a learning process. However, unless re-rating is mandatory, the feedback loop collapses after the first certification. Making annual re-rating compulsory and linking responsibility of achieving targets to NABERS accredited FMs and operators creates continuity of responsibility and provides ongoing motivation for facility teams to address drift. Moreover, by publishing re-rating results, the system would normalise transparency and foster a culture of continuous improvement.

The final step in the diagram highlights control optimisation using measured data to refine sequencing and set points. This aligns with one AA viewpoint that "visibility drives accountability": once energy data are shared, behaviour changes naturally. Policy must ensure that feedback is not only presented but actively used through structured reporting, benchmarking, and professional review.

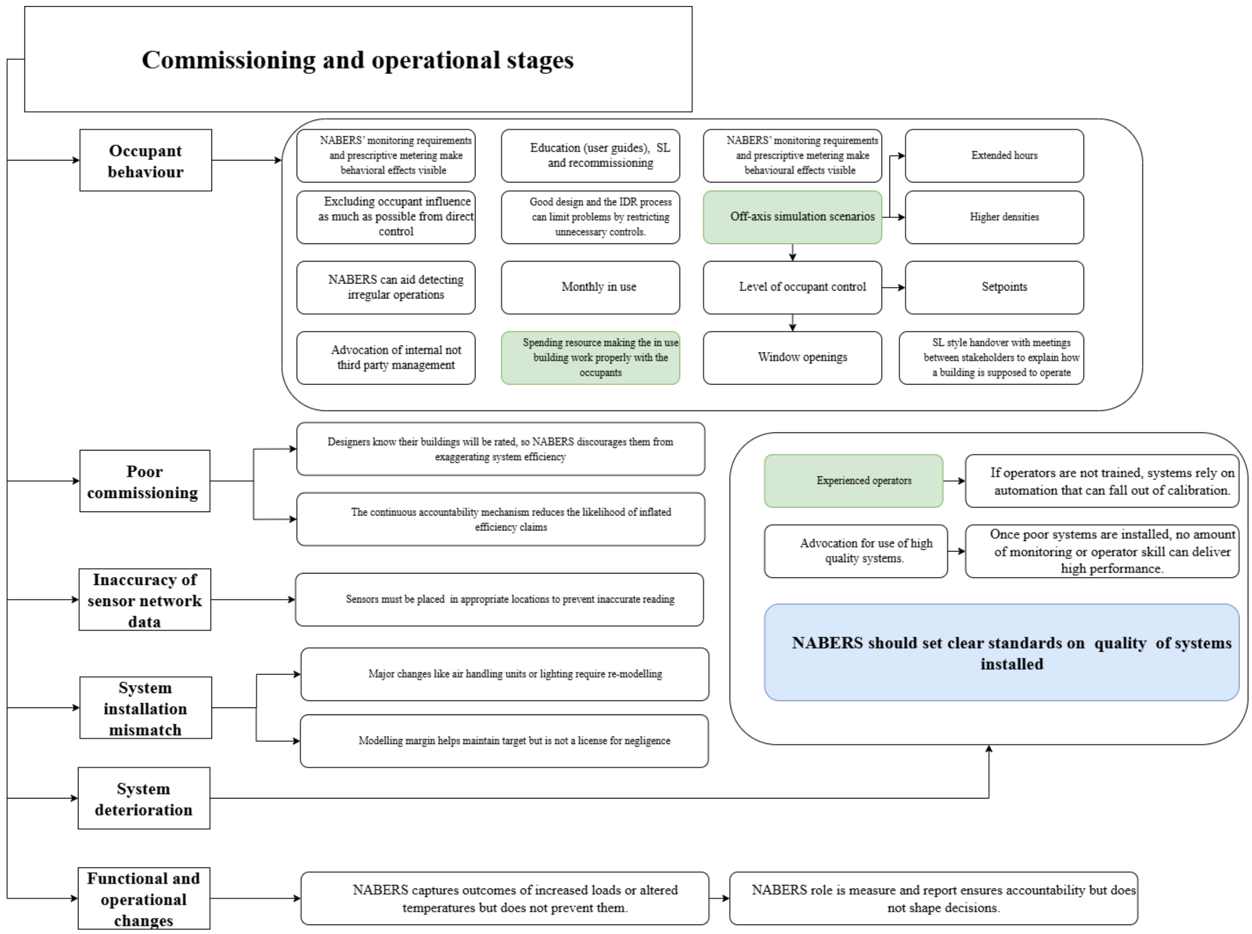


Figure 4.23: Mitigation strategies to address issues in commissioning and operational stages

4.4.3 Cross-Cutting Factors

Interviewees observed that NABERS is "reshaping industry norms." However, cultural inertia persists, especially in organizations accustomed to compliance-based environments. The diagram under Figure 4.24 underscores that ratings must be public and comparable; training must be accessible; and performance must translate into tangible benefits for clients and professionals. Many clients, particularly landlords, have limited understanding of the mechanisms that drive energy performance. The cross-cutting diagram identifies client education as a central mitigation strategy. NABERS should formalize client onboarding sessions at commitment signing, explaining performance logic, operational assumptions, and responsibilities. Embedding this understanding early helps avoid the pattern of "star-chasing" only and make clients ask better questions.

The absence of routine POE is a major institutional gap. The diagram links this deficiency to weak learning cycles across projects. NABERS' annual re-rating can act as a proxy for POE, but the process should be expanded to include qualitative occupant feedback and operational diagnostics. Making POE compulsory would institutionalize learning and drive evidence-based evolution of design standards. Perhaps the most persistent cross-cutting cause is fragmented communication. Experts emphasized that information "drops off a cliff" after handover. The diagram shows that improved communication through shared data platforms and regular coordination meetings enhances accountability and trust. NABERS could reinforce this by requiring collaboration protocols and mandatory sharing of operational data. Finally, accountability and over-optimistic predictions remain intertwined. While NABERS mitigates optimistic predictions through measured outcomes, it must also align with planning and building control regimes. Integrating NABERS verification into statutory compliance processes would close the gap between policy intent and operational reality.

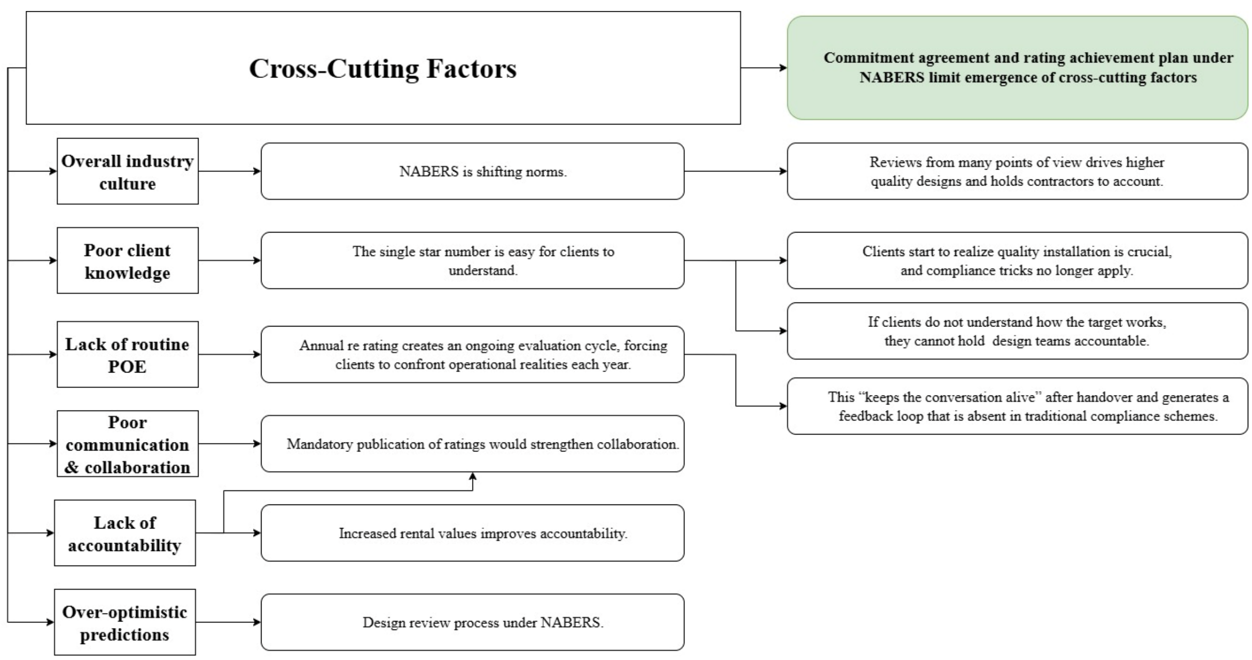


Figure 4.24: Preventative measures for emergence of cross-cutting factors

4.5 Key Takeaways and Summary of Recommendations

The analysis and causal mapping presented across sections 4.5.1 to 4.5.3 reveal that mitigating the EPG demands not only better technical design and operation but also systemic reform of the institutional environment within which buildings are conceived, delivered, and managed. The evidence confirms that the most persistent causes of underperformance stem from procedural discontinuities, unclear responsibility boundaries, and fragmented learning cycles. NABERS, by virtue of its performance-based philosophy, offers a unique opportunity to convert these weaknesses into enforceable governance mechanisms. However, its effectiveness depends on extending its remit beyond measurement, into design assurance, commissioning verification, and professional capacity-building.

The following summary outlines the key policy takeaways and stage-specific recommendations developed from the diagrams and expert insights. Collectively, they define a multi-layered mitigation framework capable of transforming the UK's approach from a *predict-and-comply* model to a *measure-and-improve* system of continuous accountability and feedback.

Firstly, design phase determines long-term operational success. Most EPG drivers originate here within flawed assumptions, incomplete detailing, or late design reviews. Early interventions and risk-based assurance are therefore the most powerful mitigation levers.

For commissioning and operational stage, even the most robust designs can underperform without effective commissioning, skilled operators, and continuous feedback. This stage must be recast as a period of *verification and learning* rather than mere handover, with clear professional accountability and annual measurement. For cross-cutting factors, performance outcomes depend on institutional context. Cultural transformation, regulatory alignment, and capacity building are essential to embed accountability and trust across the value chain.

The following policy-oriented actions are either introduced by the author or interpreted and derived from participants insights:

- **Institutionalise Pre-IDR Review Workshops:** Introduce mandatory independent design reviews before Stage 3 to ensure façade, thermal, and control strategies are scrutinised prior to major design lock-ins.

- **Mandate DfP:** Require DfP adoption across NABERS projects, embedding operational assumptions, metering strategies, and verification early in procurement.
- **Risk-Based Buildability Reviews:** Apply contextual risk assessments to innovative or passive systems to safeguard durability and adaptability under real conditions.
- **Enforce Control Logic Verification:** Make verified control sequencing part of design certification, ensuring that even oversized systems operate efficiently.
- **Adopt Multi-Platform Modelling Validation:** Mandate cross-checking of simulations across tools, with transparent documentation of deviations and assumptions.
- **Mandate Sensitivity and Uncertainty Analysis:** Reframe modelling as risk analysis by requiring off-axis scenario testing and explicit uncertainty ranges in all submissions.
- **Verify Manufacturer Performance Data:** Require third-party validation of equipment performance and part-load efficiency prior to specification approval.
- **Establish Operator-Centred Design Documentation:** Replace overly technical control descriptions with user-oriented handover documentation to bridge the knowledge gap between design and operations teams.
- **Link Commissioning Verification to NABERS Certification**: No rating should be issued without independently verified commissioning and functional performance testing.
- **Develop pathway for accredited NABERS FMs and Operators:** Establish a structured accreditation pathway to professionalise the operation of NABERS-rated buildings.
- **Mandate Annual Re-Rating:** Require compulsory yearly performance assessments.
- **Enforce Minimum Standards for Sensors and Monitoring Systems:** Standardise sensor accuracy, calibration, and maintenance protocols as part of NABERS compliance evidence.
- **Integrate POE into the Rating Cycle**: Institutionalise POE as a dual quantitative–qualitative mechanism for operational learning.
- **Create a Centralised NABERS Performance Database**: Maintain a public repository for rating and commissioning data to support benchmarking and research.
- **Require Operational Feedback Reports**: Introduce obligations to submit annual reports summarising operational challenges, corrective actions, and lessons learned.
- **Introduce Performance-Linked Incentives**: Reward sustained or improved performance through recognition, certification renewal benefits, or financial incentives.

- **Mandate Public Disclosure of Ratings**: Make NABERS results publicly accessible to strengthen transparency and market motivation.
- **Establish Formal Communication Protocols Across Lifecycle Stages**: Require structured handovers, shared digital data platforms, and post-completion coordination reviews.
- **Institutionalise Client Education and Onboarding**: Conduct standardised induction sessions for clients and asset managers at commitment signing to improve understanding and engagement.
- **Embed Accountability into Contracts and Leases**: Specify responsibility for achieving and maintaining ratings in professional and tenancy agreements.
- **Create a Continuous Professional Development scheme**: Incorporate NABERS principles into the Continuing Professional Development (CPD) requirements of professional institutions, fostering long-term competence and culture change.
- **Strengthen Knowledge Transfer Mechanisms**: Use POE and re-rating results to update NABERS guidance and industry benchmarks annually, embedding feedback into policy evolution.

The collective evidence demonstrates that the EPG is fundamentally a governance and accountability challenge, not merely a technical one. Persistent inefficiencies arise from fragmented project processes, late-stage interventions, and limited feedback mechanisms between design, operation, and policy.

The mitigation strategies developed here reframes NABERS as a multi-stage governance platform that orchestrates continuous accountability across design, commissioning, and operation. By embedding pre-emptive design reviews, verified commissioning, accredited FMs, and public transparency, NABERS can evolve from a rating tool into an institutional driver of performance integrity. Importantly, the apparent uncertainties in expert judgments are not signs of doubt but reflections of context-specific insight and cautious interpretation. Many reveal untapped opportunities of areas where better procedural clarity, training, and contractual structures could unlock further gains.

If implemented comprehensively, the policy framework proposed in this section would shift the UK building sector's culture from "predict and comply" to "measure and improve." It would mitigate the EPG not through one-off compliance exercises but through sustained institutional learning, professional accountability, and evidence-driven adaptation.

If major markets such as the United States and Europe continue to rely on design-oriented schemes such as LEED and BREEAM, the broader impact of NABERS on reducing the Energy Performance Gap will remain limited. This is because the EPG cannot be resolved through isolated national initiatives alone but requires a collective shift toward verified outcome-based policies and, in its core, outcome-based certification systems. Although NABERS has demonstrated clear success in Australia and is now influencing practice in the United Kingdom, its wider global impact remains uncertain. Meaningful progress depends on a shared international commitment to measured operational performance and on a regulatory mindset that treats energy performance as an ongoing responsibility rather than a one-off achievement. Incremental adjustments within existing design-led frameworks will therefore be insufficient without a transition toward policies that prioritise verified operational outcomes.

The analysis presented throughout this chapter demonstrates that mitigating the EPG requires a fundamental realignment of how building performance is conceived, delivered and evaluated. While the global expansion of NABERS is not guaranteed, several structural trends increase the likelihood that its principles will influence international practice. Policy makers are increasingly recognising that the Energy Performance Gap reflects governance weaknesses more than technical errors, and this is creating growing interest in regulatory frameworks that enforce accountability across the lifecycle of buildings. Even in contexts where NABERS is not formally adopted, its methods provide a transferable foundation capable of supporting the evolution of existing rating systems. The potential for NABERS to reshape global practice lies not only in its brand but in the capacity of its principles to embed measurable outcomes, transparent disclosure and long term stewardship within any regulatory environment.

The findings also illustrate several lessons that dominant schemes such as BREEAM and LEED can draw from the NABERS approach. The emphasis on measured rather than predicted performance demonstrates the limitations of compliance pathways that rely primarily on modelling assumptions. The routine practice of annual re-rating shows the importance of continuous verification in preventing performance drift. The strong integration of commissioning, operator competence and clearly defined responsibilities highlights the need to treat the operational stage as central to performance rather than peripheral. These lessons reveal that the strengths of NABERS are not tied to a specific national context but reflect a coherent governance model that could enhance the effectiveness of any performance standard.

The evidence shows that the Energy Performance Gap is fundamentally a policy challenge. Technical solutions are necessary but insufficient unless supported by institutional arrangements that create stability, transparency and accountability across design, commissioning and ongoing operation. The transition from a predict and comply culture to a measure and improve system requires policy frameworks that mandate operational verification, formalise learning cycles and allocate responsibility for maintaining performance over time. NABERS offers a clear illustration of how such a framework can function in practice. By embedding verified commissioning, annual performance assessment, skilled operation and transparent reporting into one coherent system, NABERS demonstrates how policy can transform industry culture and align market incentives with energy outcomes.

The key message emerging from this chapter is therefore that building performance cannot be separated from governance quality. Reducing the EPG requires policy structures that ensure every stage of the building lifecycle contributes to measurable and reliable energy outcomes. Whether through NABERS or through reforms to existing schemes informed by its principles, success ultimately depends on institutionalizing accountability, enabling continuous improvement and grounding performance claims in verified operational evidence.

5. Complexity Demonstration

This chapter addresses the third RO of the thesis to deepen understanding of the EPG by clarifying the ambiguity surrounding its magnitude and critique the reliability of reported EPG magnitudes. Section 5.1 delineates the detailed methodology, encompassing a comprehensive description of the case study building, the simulation inputs and assumptions, and the design of the comparative analytical framework. Section 5.2 presents empirical findings on inter-tool gaps, and the EPG between measured and simulated energy use at annual and monthly scales. Finally, section 5.3 consolidates insights derived from the analysis within the chapter and make links with insights derived from previous chapters of the thesis.

5.1 Detailed Methodology

This part of the thesis draws on collaborative work conducted with the division of energy and building design at Lund University. The case study was carried out on the Energy and Building Design (EBD) Laboratory at Lund University, Sweden, located at the Faculty of Engineering (LTH). An in-depth case study was conducted in collaboration with two master's students (Odeh, Parvaz et al. 2025, Parvaz and Jakkula 2025). It consisted of the creation of building energy simulation models of the same case study building using four building energy modelling platforms: DesignBuilder, IDA-ICE, ClimateStudio, and Integrated Environmental Solutions–Virtual Environment (IES-VE). By applying harmonized input data across the four tools and benchmarking simulation outputs against measured performance, the chapter explores both inter-tool gaps and discrepancies between actual and simulated energy performance and the influence of changing heating setpoints on simulation outcomes.

As part of this collaboration, guidance was provided to the MSc students on the development of simulation scenarios and on the methodological requirements necessary to ensure comparability across four dynamic simulation tools. This included instruction on maintaining identical geometric representations in each modelling environment and aligning all input data and underlying assumptions. It was also clarified that no operational performance data would be available; therefore, the simulation work was required to rely exclusively on the architectural drawings and building services documentation for the case study building. This framework was established to reflect early-stage design conditions and to isolate the influence of modelling decisions on predicted energy performance. The selection of the LTH building complex followed a convenience sampling strategy, enabled by the availability of detailed design documentation, direct access to the buildings, and opportunities for close collaboration with the research team. This choice was informed by a much broader international search for suitable case studies, which included several prominent digital twin demonstrators, such as candidate buildings in Sydney and the UK, that claimed advanced digital capabilities but ultimately lacked sufficient transparency or accessible data for thorough analysis. By contrast, the modest yet well documented set of four shoebox type buildings at LTH provided a rare opportunity to examine, with clarity and precision, how modelling assumptions and building characteristics contribute to the observed EPG.

5.1.1 EBD Laboratory

The laboratory is a purpose-built research and teaching facility designed as a controlled environment for experimental studies of building energy use, daylighting, and thermal comfort. The heated zones of the EBD laboratory cover approximately 230.53 m² of floor area and include specialized spaces such as the measurement room (23.25 m²), climatic chamber and measurement boxes with connecting corridor (111.5 m²), and the solar central & HVAC zone (36.5 m²). Smaller functional spaces such as the entrance, store, toilet, daylight rooms, and two calorimetric cells are also included within the heated area. The unheated zones comprise a stair room, an external store, and a crawlspace, totalling 66.96 m². Although excluded from the thermal simulations, the unheated spaces were modelled to ensure continuity of the building envelope preventing distortions in heat transfer calculations. Figure 5.1 below shows different façade views of the EPD laboratory and different interior spaces that serve various functions.

Figure 5.1: North and south facades of the EBD Laboratory (top left: north facade; top right: south facade), circulation area (bottom right), one of the lighting cells (bottom middle), and one of the thermal cells (bottom left)

5.1.2 Simulation Inputs and Assumptions

The full input schedule used to harmonise the four simulation models is provided in Appendix C, including the building fabric, glazing, doors, infiltration, thermal-bridging adjustment, HVAC assumptions, ventilation rates, heat-recovery efficiency, heating and cooling setpoints, internal loads, occupancy densities, weather data, ground temperatures and output categories. The building envelope was modelled using construction-specific data. The external walls of the main building are 297–307mm thick with U-values between 0.118 and 0.122 W/m²K. By contrast, the calorimetric chambers employ thinner walls with U-values of 0.534 and 1.455 W/m²K, while the staircase and external store are poorly insulated, with U-values as high as 3.812 W/m²K. The roof of the main building is highly insulated (U = 0.118 W/m²K), whereas the staircase roof, a thin steel plate, has a U-value of 5.882 W/m²K. Ground slabs vary between 0.172 and 0.403 W/m²K depending on thickness, and internal walls and slabs typically fall around 0.27–0.29 W/m²K, with one thinner slab (22 mm) under the measurement room, toilet, and store yielding a much higher

U-value of 2.95 W/m²K. Basement walls range between 0.134 and 0.383 W/m²K depending on orientation. Windows, including frames, were assigned a U-value of 1.6 W/m²K, and doors a U-value of 1.058 W/m²K. Infiltration was set at 1 ACH under 50 Pa, and thermal bridging was accounted for by applying a 15% adjustment.

Heating and cooling are provided by a ground source heat pump with a coefficient of performance of 4.2. The system is coupled to an air handling unit (AHU) with variable air volume (VAV) control, operating under occupancy schedules. Three simulation scenarios were conducted by altering heating setpoints (18°C, 21°C and 24°C), while cooling setpoint was fixed at 26°C. Mechanical ventilation was defined at 0.35 L/s·m² plus 7 L/s per person, with a heat recovery efficiency of 65%. Natural ventilation and humidity control were deactivated across all scenarios.

Internal loads were standardized across all models. Lighting was set to 8.5 $kWh \cdot m^{-2} \cdot a^{-1}$, equipment to 5.3 $kWh \cdot m^{-2} \cdot a^{-1}$, and hot water to 0.6 $kWh \cdot m^{-2} \cdot a^{-1}$. Occupancy densities were differentiated by functional use. The measurement room, entrance, solar central & HVAC, and climatic chamber were assigned 0.061 persons/m², while the toilet and store were assigned 0.001 persons/m². The calorimetric and daylight zones were modelled without occupancy due to their specialized usage. This zoning ensured operational diversity was reflected while maintaining consistent assumptions across all tools. Weather data was sourced from the Sveby reference file for Lund 2023. Monthly ground temperatures ranged from 2.32°C in February to 16.49°C in August, with an annual average near 10 °C. This dataset was applied uniformly across all tools to eliminate climate input variability. Outputs were collected for total annual energy use and end-use categories including heating, cooling, lighting, equipment, and hot water. By harmonizing these inputs, differences in outcomes could be attributed to the intrinsic behaviour of the tools rather than to inconsistent assumptions.

5.1.3 Comparative Analysis

The comparative analysis was structured across four stages to test reliability of simulation outcomes as indicated in Figure 5.2 below. Firstly, measured electricity consumption for the years 2023 and 2024 was compared to highlight annual and monthly variability in actual energy use. Secondly, simulation results from the four tools were compared across different simulation scenarios. This analysis was conducted both annually and by end-use category, encompassing lighting, equipment, hot water, fans, heating, and cooling. In the third stage, simulated results were benchmarked against measured consumption for both 2023 and 2024 across all scenarios, enabling calculation of the EPG between measured and simulated electricity consumption. Here, both annual totals and monthly consumption patterns were assessed to evaluate how well the tools tracked seasonal dynamics as well as yearly aggregates. Finally, in the fourth stage, a focused comparison was conducted at the 21 °C heating setpoint, chosen for its proximity to actual indoor operating conditions. At this stage, fidelity was examined not only in terms of average annual alignment but also monthly variability, thereby identifying whether apparent annual accuracy masked temporal inconsistencies.

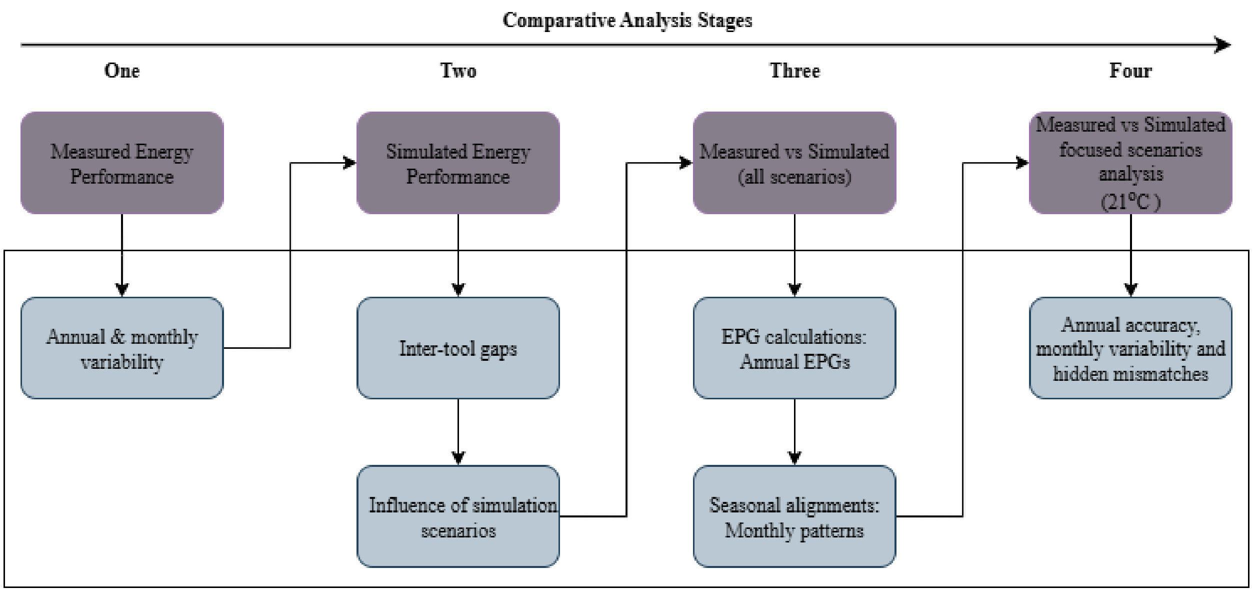


Figure 5.2: Stages of comparative analysis

To quantify monthly agreement between measured and simulated electricity use, the 21°C scenario was further assessed using mean bias error (MBE) and coefficient of variation of the root mean square error (CVRMSE). These metrics were calculated using monthly electricity consumption values, with measured 2023 electricity use treated as the reference series. MBE was used to indicate the direction and average magnitude of bias, while CVRMSE was used to assess the degree of monthly deviation after normalisation by the mean measured consumption. This additional diagnostic step was included because annual totals alone can conceal compensating monthly under- and over-predictions.

5.2 Results & Analysis

In 2023, the total recorded electricity consumption was 11,235 kWh, whereas in 2024, it increased substantially to 18,344 kWh (Figure 5.3). This represents an absolute increase of 7,109 kWh, corresponding to a 63.3% rise in total annual consumption. The lowest monthly consumption in 2023 occurred in June, with a value of 291 kWh, while the highest was observed in January at 1608 kWh. In 2024, the lowest consumption was again recorded in August at 754 kWh, whereas the peak consumption occurred in January, reaching 2,590 kWh. The marked increase in 2024 is attributable to a change of use within the laboratory during that year, which altered operational patterns and directly influenced observed energy demand. This operational shift subsequently motivated the investigation into alternative simulation scenarios and the testing of different setpoint values in order to assess how these modelling assumptions affect the magnitude of the EPG.

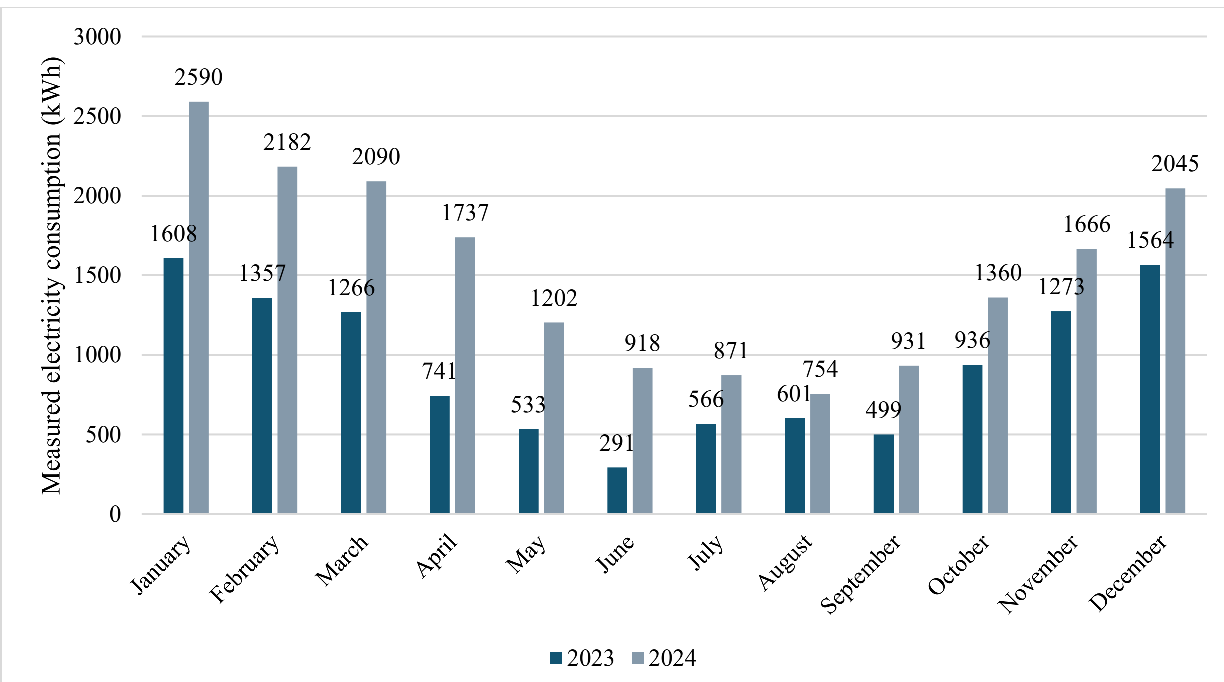


Figure 5.3: Measured electricity consumption in 2023 and 2024

5.2.1 Inter-Tool Gaps

The simulated electricity consumption represents total energy use, encompassing all end use energy categories. Energy consumption increases across all simulation tools as the indoor temperature setpoint rises, though the magnitude of this increase varies by tool. Between 18 °C and 24 °C, IDA-ICE shows an increase of 3,287 kWh (from 7,379 kWh to 10,666 kWh), DesignBuilder records a rise of 2,491 kWh (from 7,064 kWh to 9,555 kWh). ClimateStudio increases by 2,252 kWh (from 6,905 kWh to 9,157 kWh), and IES-VE shows the highest increase of 7,073 kWh (from 8,786 kWh to 15,859 kWh) as depicted in Figure 5.4 below.

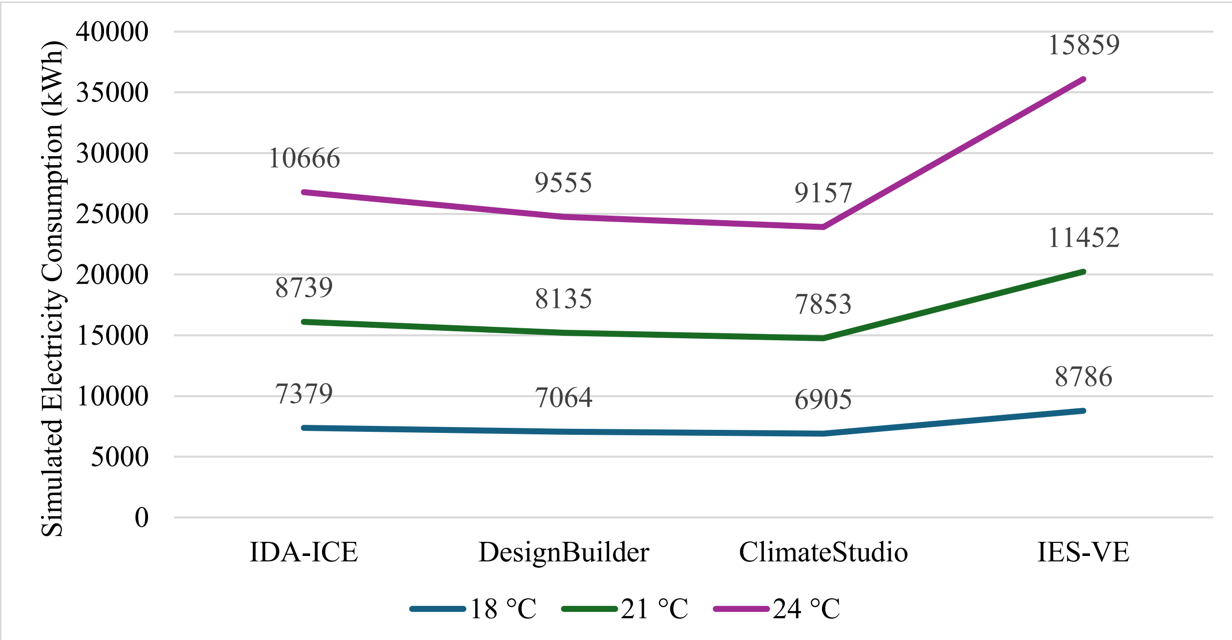
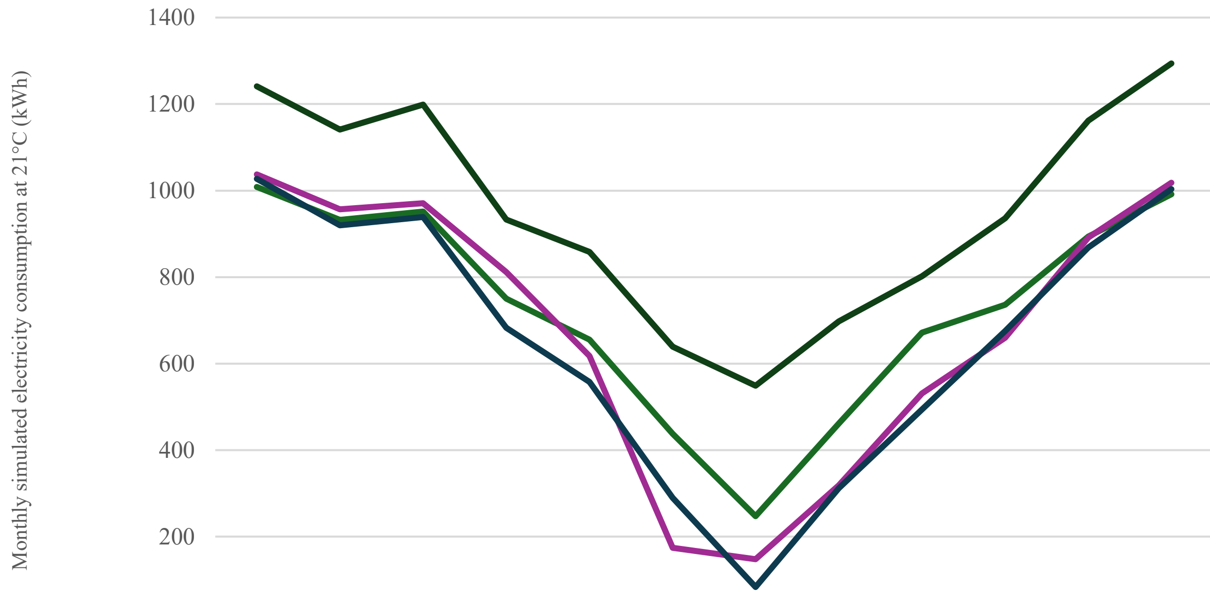


Figure 5.4: Annual simulated electricity consumption under three simulation scenarios

Figure 5.5 shows the variation in predicted electricity consumption across the different simulation tools and scenarios. At a temperature setpoint of 21°C. IES-VE consistently produces the highest estimates throughout the year, with values peaking at 1294 kWh in December and reaching a minimum of 549 kWh in July. In contrast, ClimateStudio generally reports the lowest consumption, ranging from 84 kWh in July to 1027 kWh in January. IDA-ICE and DesignBuilder fall in between, with relatively close estimates during the colder months, though DesignBuilder tends to diverge more noticeably in summer.

The differences between the tools are most pronounced during the summer months. In July, for instance, IES-VE estimates 549 kWh while ClimateStudio estimates only 84 kWh, resulting in a gap of 465 kWh, equivalent to approximately 554%. Similarly large disparities are observed in June and August, with percentage differences exceeding 260% and 120%, respectively. Even in the winter months, where estimates converge more closely, IESVE remains consistently higher. For example, in January, its estimate exceeds ClimateStudio by 214 kWh, a 20.8% difference. The trend continues throughout the year, with percentage differences ranging from around 24% to over 50% during transitional months such as April and May. These findings underscore that the magnitude of predicted electricity use can differ substantially based on simulation tool utilized even under standardized input data.



	January	February	March	April	May	June	July	August	September	October	November	December
IDA-ICE	1008	933	952	750	656	438	247	462	671	736	894	991
DesignBuilder	1037	957	971	811	618	174	148	318	531	660	891	1018
ClimateStudio	1027	920	939	683	557	290	84	312	494	676	869	1004
IESVE	1241	1141	1199	933	858	639	549	698	802	936	1162	1294

Figure 5.5: Monthly simulated electricity consumption at 21°C (kWh)

5.2.1.1 End-use Simulated Electricity Consumption

Figure 5.6 that shows the annual simulated electricity consumption for lighting, equipment, hot water, and fans, as calculated by the four tools, demonstrate a high degree of consistency among tools, with IESVE generally aligning except in the case of fan energy. For lighting, estimates range from 1919 kWh (DesignBuilder) to 2043 kWh (ClimateStudio), a difference of 124 kWh, or approximately 6.5%. Equipment consumption varies slightly more, with ClimateStudio again predicting the highest usage at 1294 kWh and IDA-ICE the lowest at 1212 kWh, a gap of 82 kWh or about 6.8%. The values for hot water are remarkably close across all tools, ranging from 135 kWh (ClimateStudio) to 141 kWh (DesignBuilder), a minimal difference of just 6 kWh or 4.4%.

The most significant disparity appears in the simulation of fan energy consumption. While IDA-ICE, DesignBuilder, and ClimateStudio produce nearly identical values of 1147 kWh, 1147 kWh, and 1150 kWh respectively, IES-VE estimates fan consumption at only 637 kWh. This results in a difference of 513 kWh, corresponding to a 44.7% lower prediction relative to the average of the other tools. This discrepancy stands in stark contrast to the relative agreement observed in other end-use categories and suggests a fundamentally different approach in IESVE's modeling of HVAC auxiliary systems. While lighting, equipment, and hot water predictions show only minor variation across platforms, the variation in fan energy highlights the critical role of tool-specific features especially in the case of IES-VE in predicting fan energy consumption.

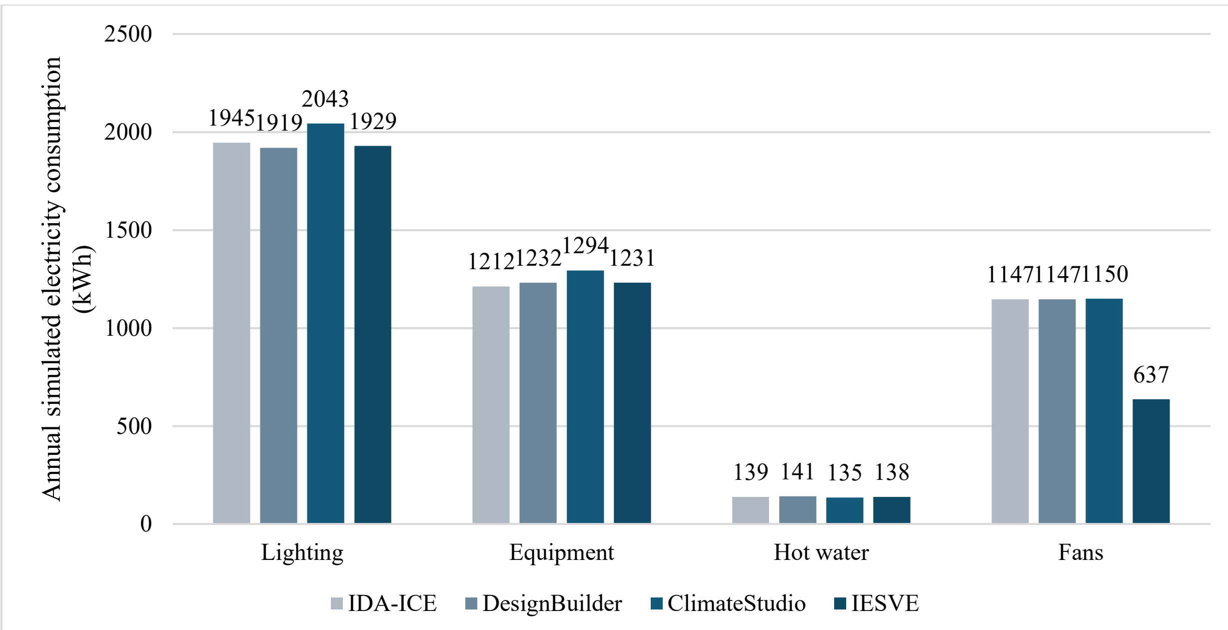


Figure 5.6: Annual simulated end use electricity consumption at 21°C (kWh)

Figure 5.7 shows the monthly simulated values for fans electricity consumption at a temperature setpoint of 21 °C, revealing a consistent trend across the year, with notable discrepancies between IES-VE and the other tools. IDA-ICE, DesignBuilder, and ClimateStudio provide nearly identical monthly estimates throughout, indicating a high level of agreement in their modeling of fan energy use. Their value remains stable, with minor seasonal fluctuations. For example, all three tools report 103 kWh in January and May, and 107 kWh in September and October. In contrast, IES-VE consistently predicts significantly lower monthly values across all months. The most pronounced differences are observed in the spring and summer months. For instance, in April, IES-VE estimates 48 kWh compared to 100 kWh by the other tools, a reduction of 52%, while in July, it predicts 43 kWh versus 63 kWh, a 31.7% difference. The smallest discrepancy occurs in

December, where IES-VE estimates 72 kWh compared to 99 kWh from the others, still reflecting a gap of 27.3%.

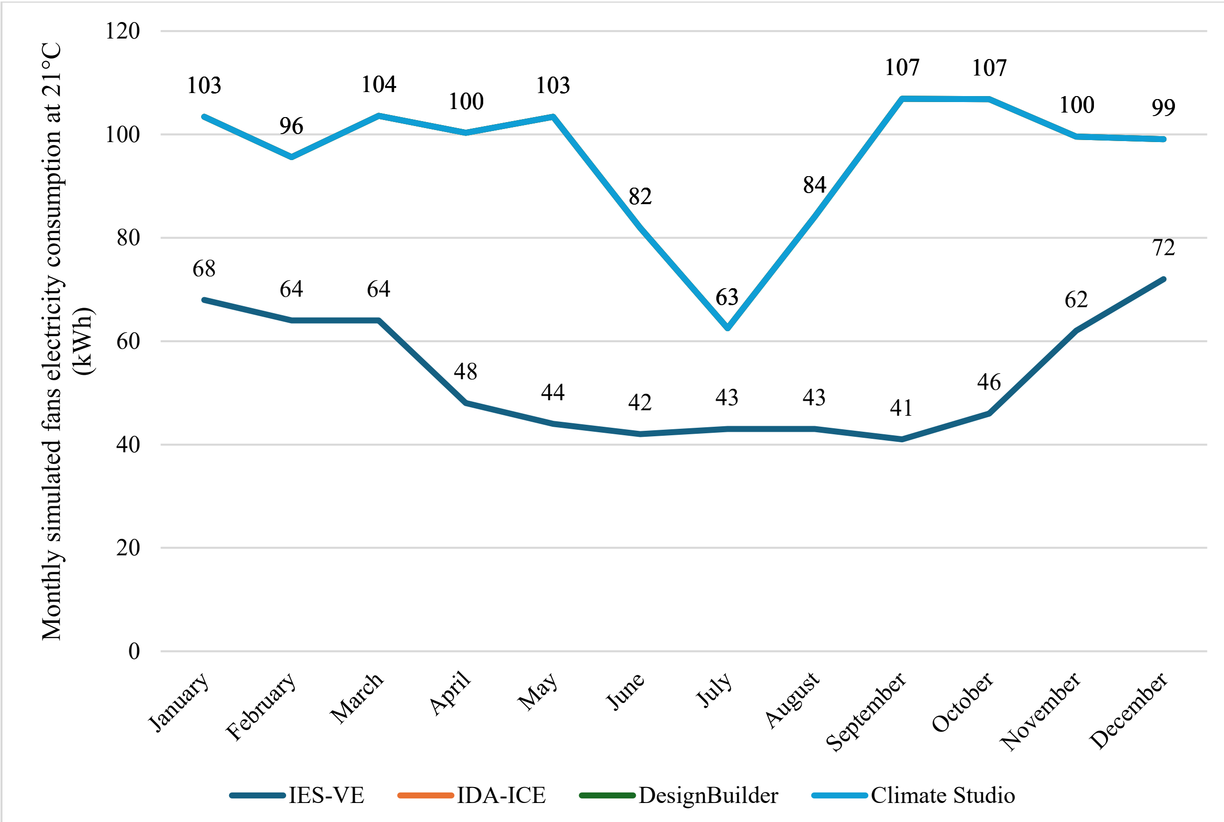


Figure 5.7: Monthly simulated fans electricity consumption at 21°C (kWh)

Figure 5.8 illustrates the variation in heating electricity consumption across the three temperature setpoint scenarios, highlighting substantial divergence between the simulation tools. For the 18 °C scenario, IES-VE predicts the highest energy use at 4676 kWh, while ClimateStudio provides the lowest estimate at 2188 kWh with a difference of 2488 kWh, or approximately 113.7%. At the 21 °C setpoint, IES-VE again produces the highest estimate at 7142 kWh, and ClimateStudio the lowest at 3123 kWh, representing a difference of 4019 kWh, or roughly 128.7%. In the 24 °C scenario, IES-VE reports the highest consumption at 10,981 kWh, while ClimateStudio remains the lowest at 4328 kWh, representing a difference of 6653 kWh, or about 153.7%. Unlike other end-use categories with minimal discrepancies, these results indicate substantial differences in heating energy predictions among the tools, particularly with IES-VE consistently producing significantly higher values. This suggests that the choice of simulation tool may considerably influence the estimation of heating energy demand, especially under different indoor temperature setpoints.

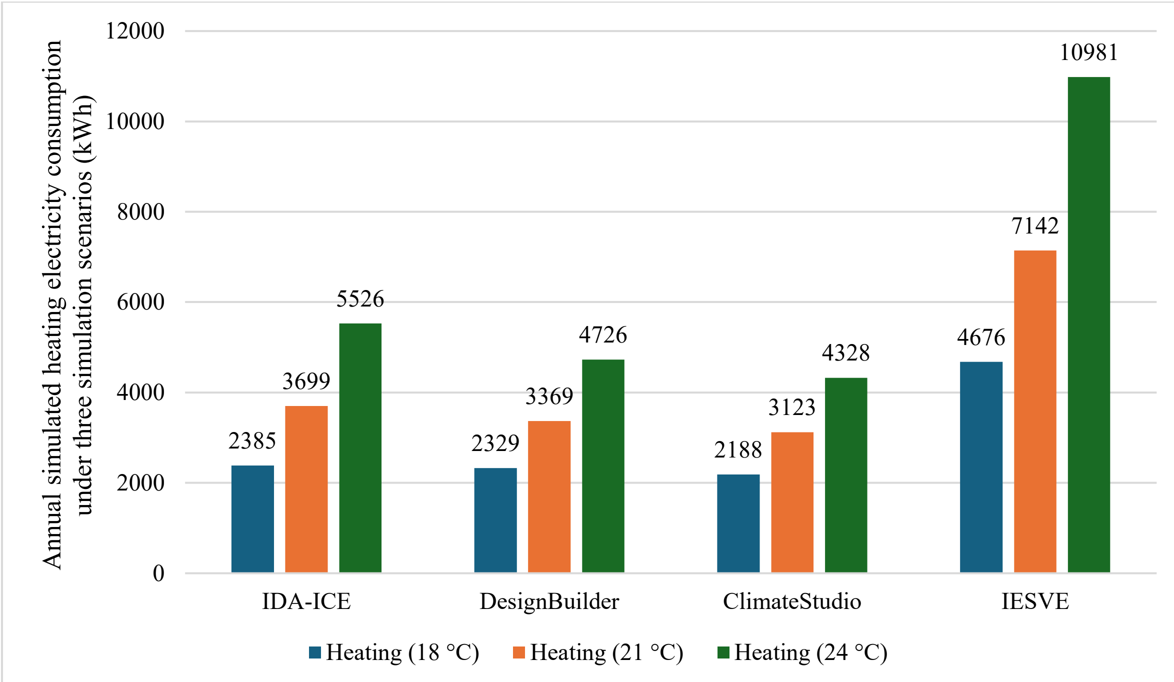
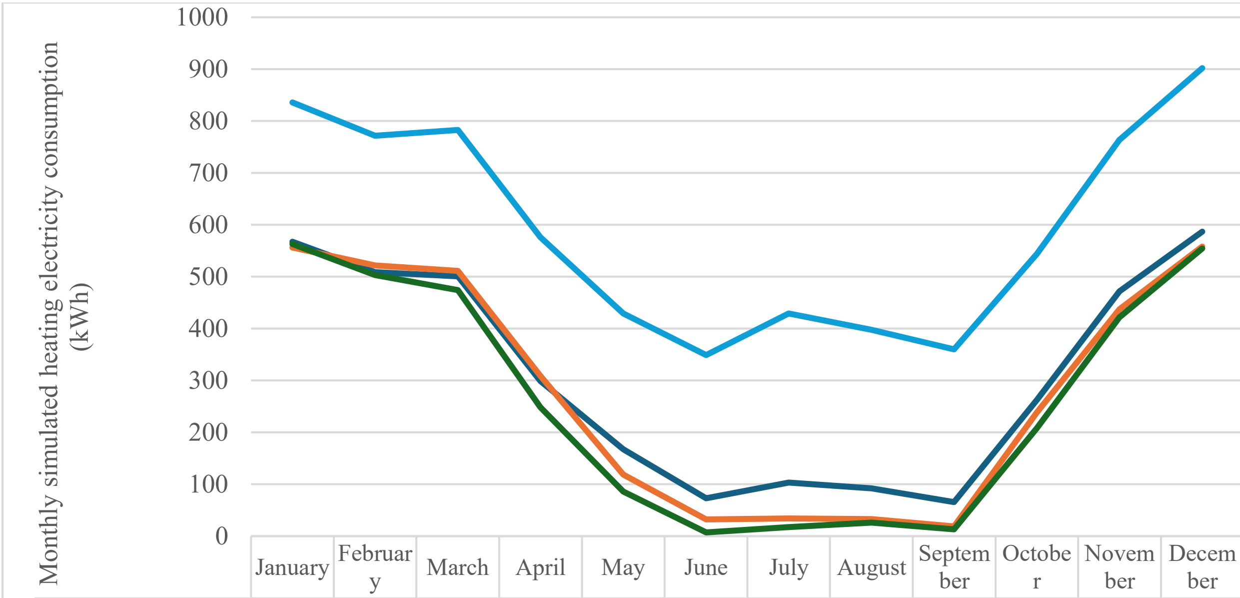


Figure 5.8: Annual simulated heating electricity consumption under three simulation scenarios (kWh)

Figure 5.9 depicts the monthly heating electricity consumption estimates for the four simulation tools, highlighting the seasonal variation and the consistently higher values predicted by IES-VE. IES-VE predicts the highest consumption across all months, ranging from 349 kWh in June to 902 kWh in December. In contrast, ClimateStudio generally produces the lowest estimates, ranging from 7 kWh in June to 563 kWh in January. For instance, in January, IES-VE's estimate is 836 kWh compared to ClimateStudio's 563 kWh, a difference of 273 kWh or approximately 48.5%. The largest discrepancy is observed in June, where IES-VE predicts 349 kWh, while ClimateStudio estimates only 7 kWh, a difference of 342 kWh or nearly 4900%. Other tools such as IDA-ICE and DesignBuilder provide intermediate values, with IDA-ICE consistently closer to ClimateStudio's estimates and DesignBuilder slightly higher but still substantially lower than IES-VE.



	January	February	March	April	May	June	July	August	September	October	November	December
IDA-ICE	567	509	501	299	168	73	103	92	66	263	472	587
DesignBuilder	556	521	511	309	119	32	34	33	19	239	437	558
Climate Studio	563	503	474	249	86	7	17	26	13	208	422	554
IESVE	836	772	783	576	429	349	429	398	360	544	764	902

Figure 5.9: Monthly simulated heating electricity consumption (kWh)

Figure 5.10 presents annual simulated cooling electricity consumption estimates from the tools across the different temperature scenarios, exhibiting significant variation. IDA-ICE consistently predicts the highest consumption, ranging from 547 kWh at 18 °C to 693 kWh at 24 °C. DesignBuilder reports intermediate values, increasing from 295 kWh at 18 °C to 389 kWh at 24 °C. ClimateStudio estimates the lowest cooling consumption, ranging from 95 kWh to 207 kWh. IES-VE values lie between DesignBuilder and IDA-ICE, with estimates increasing from 336 kWh at 18 °C to 531 kWh at 24 °C. The greatest absolute difference is observed at 24 °C, where the highest estimate (IDA-ICE, 693 kWh) exceeds the lowest (ClimateStudio, 207 kWh) by 486 kWh, corresponding to a difference of approximately 235%.

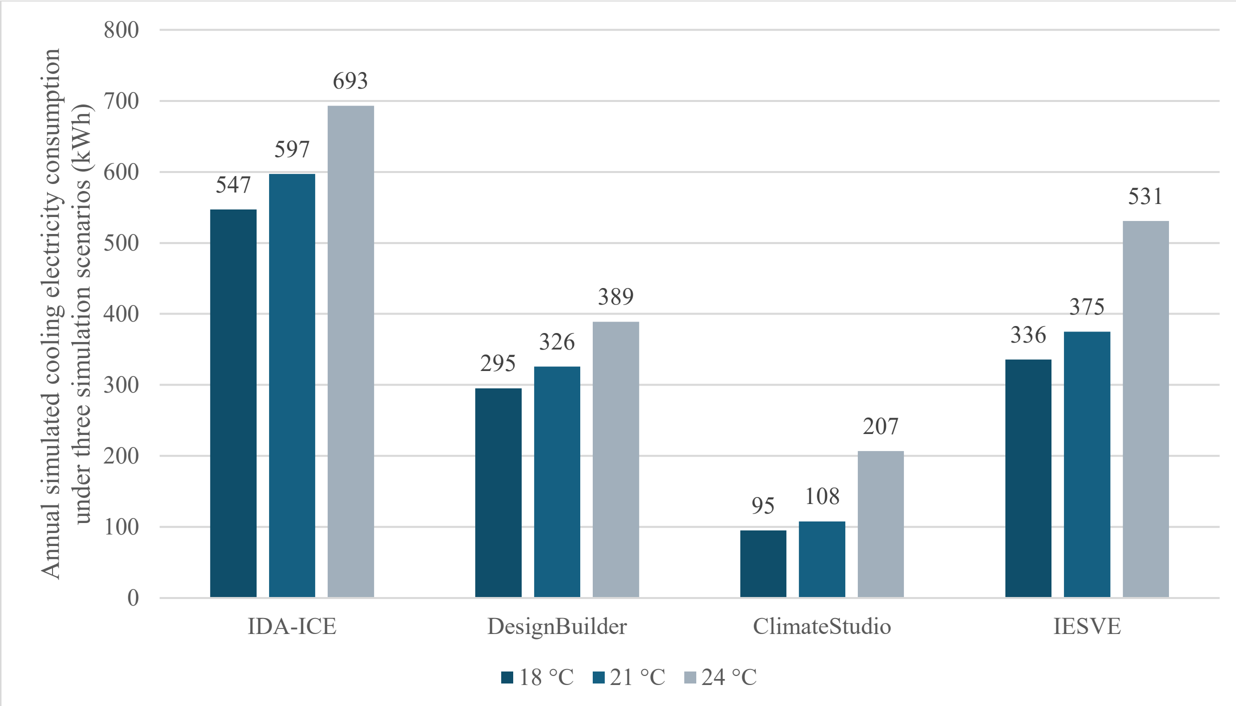
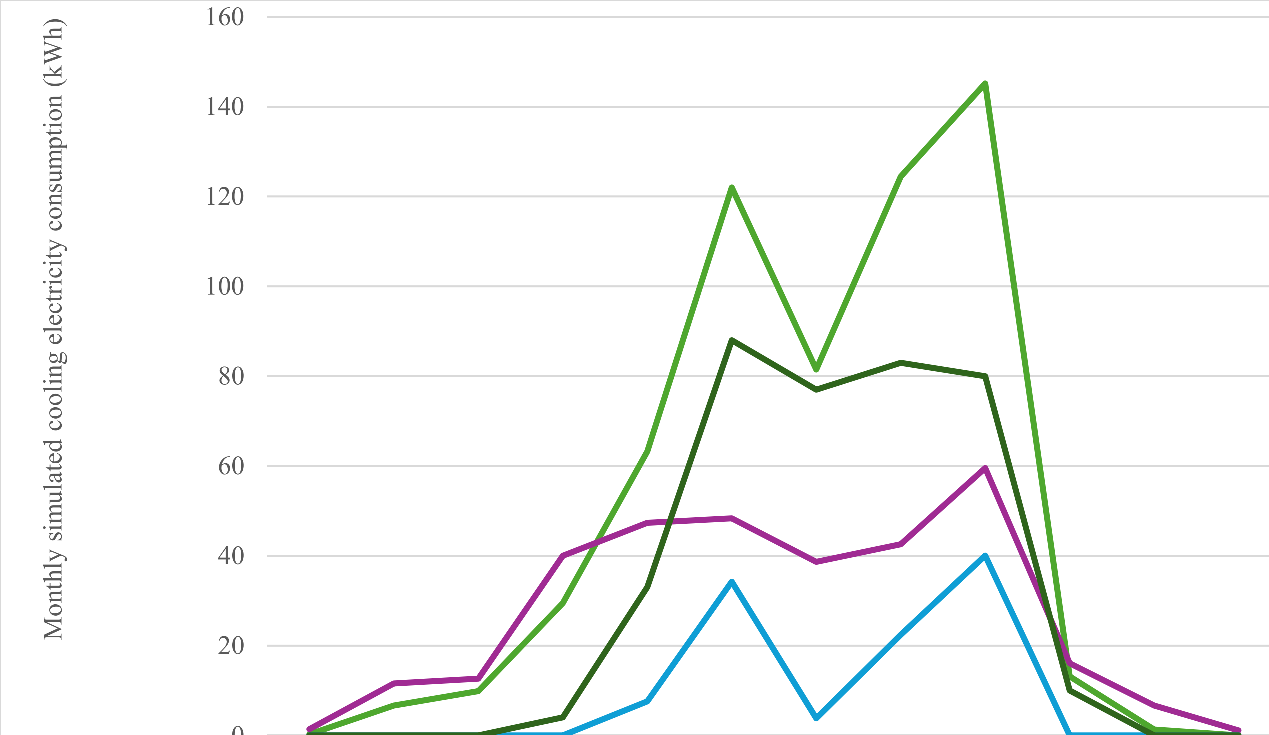


Figure 5.10: Annual simulated cooling electricity consumption under three simulation scenarios (kWh)

Figure 5.11 highlights the monthly variation in simulated cooling electricity consumption across four tools, revealing distinct seasonal peaks and differences between models. IDA-ICE generally predicts the highest cooling consumption during warmer months, peaking at 145 kWh in September, while DesignBuilder follows with moderately high values, reaching 60 kWh in the same month. ClimateStudio consistently estimates the lowest cooling demands, often near zero during cooler months and peaking at 40 kWh in September. IESVE presents intermediate values, with a maximum of 88 kWh in June and around 80 kWh in September. For example, in June, IDA-ICE predicts 122 kWh, compared to 34 kWh from ClimateStudio and 88 kWh from IESVE, representing differences of 88 kWh (approximately 259%) and 34 kWh (approximately 39%) respectively. During cooler months such as January and February, cooling consumption is negligible across all tools. These discrepancies, particularly the higher estimates by IDA-ICE relative to ClimateStudio, highlight the variability in cooling load predictions depending on the simulation software employed.



	January	February	March	April	May	June	July	August	September	October	November	December
IDA-ICE	0	7	10	29	63	122	81	124	145	13	1	0
DesignBuilder	1	12	13	40	47	48	39	43	60	16	7	1
Climate Studio	0	0	0	0	8	34	4	22	40	0	0	0
IESVE	0	0	0	4	33	88	77	83	80	10	0	0

Figure 5.11: Monthly simulated cooling electricity consumption (kWh)

5.2.2 Annual Measured Versus Simulated

Table 5.1 below compares measured electricity consumption for the 2023 and 2024 with simulated values under the use of different tool and simulation scenarios. Each cell represents the ratio of measured consumption to simulated output, providing an indicator of how closely the simulation matches real performance. We also compute the absolute and percentage change in the simulation gap from 2023 to 2024.

In 2023, the simulation scenario that exhibited one of the narrowest EPG between measured and simulated electricity consumption was IDA-ICE at 24°C setpoint, with a factor of 1.05. This indicates that the simulated energy use was only 5% lower than the actual measured consumption (11,235 kWh vs. 10,666 kWh), making it the most accurate representation among all models and scenarios for that year. A similarly close match was observed with IES-VE at a 21 °C setpoint, which produced a ratio of 0.98, suggesting a slight overestimation of energy use. While IES-VE at 24 °C recorded a lower ratio of 0.71. In 2024, IES-VE at 24 °C once again demonstrated the closest agreement with actual energy consumption, producing a ratio of 1.16, meaning it predicted 15,859 kWh compared to a measured value of 18,344 kWh. IDA-ICE at 24°C exhibited a ratio of 1.72, though the gap had widened notably from the previous year.

The largest discrepancies in 2024 occurred in the DesignBuilder and ClimateStudio simulations at the 18°C setpoint, both reporting a factor of 2.60. This indicates that measured energy use exceeded simulated values by 160%, showing substantial underestimation at lower setpoints. Across all tools and scenarios, the measured–simulated gap increased markedly from 2023 to 2024, with factor values rising consistently by 63–66%. This may reflect changes in operation or occupancy not captured in the models, limited responsiveness to real-world dynamic loads, or reduced system efficiency in 2024. Despite this wider discrepancy, IES-VE at 24°C and 21°C remained closest to measured performance. By contrast, ClimateStudio and DesignBuilder at 18°C showed significant underprediction, indicating a need to calibrate assumptions, particularly around HVAC performance under reduced heating setpoints.

Table 5.1: Annual Measured-to-Simulated Electricity Ratios

Simulation scenario	**Simulation tool**	**Measured/Simulated 2023**	**Measured/Simulated 2024**	**Absolute increase (kWh)**	**% increase**
24 °C	IDA-ICE	1.05	1.72	+5,678	+63.4%
	DesignBuilder	1.18	1.92	+5,789	+63.3%
	ClimateStudio	1.23	2.00	+6,187	+63.2%
	IES-VE	0.71	1.16	+4,725	+66.4%
21 °C	IDA-ICE	1.29	2.10	+5,605	+63.3%
	DesignBuilder	1.38	2.26	+6,209	+63.3%
	ClimateStudio	1.43	2.34	+6,490	+63.2%
	IES-VE	0.98	1.60	+5,599	+63.5%
18 °C	IDA-ICE	1.52	2.49	+6,965	+63.4%
	DesignBuilder	1.59	2.60	+7,280	+63.6%
	ClimateStudio	1.59	2.60	+7,280	+63.2%
	IES-VE	1.28	2.09	+6,101	+63.4%

5.2.3 Monthly Measured Versus Simulated (21°C)

Table 5.2 presents the monthly measured-to-simulated ratios for each software tool. Across both years and all tools, the simulated electricity use was generally lower than the measured, reflected by ratios consistently above one particularly in the summer months.

Table 5.2: Monthly Measured-to-Simulated Electricity Ratios (Setpoint 21 °C)

Month	IDA-ICE		DesignBuilder		ClimateStudio		IES-VE	
	2023	2024	2023	2024	2023	2024	2023	2024
January	1.60	2.57	1.55	2.50	1.57	2.52	1.30	2.09
February	1.45	2.34	1.42	2.28	1.48	2.37	1.19	1.91
March	1.33	2.20	1.30	2.15	1.35	2.23	1.06	1.74
April	0.99	2.32	0.91	2.14	1.09	2.54	0.79	1.86
May	0.81	1.83	0.86	1.95	0.96	2.16	0.62	1.40
June	0.67	2.10	1.67	5.28	1.003	3.17	0.46	1.47
July	2.30	3.53	3.82	5.89	6.74	10.37	1.03	1.59
August	1.30	1.63	1.89	2.37	1.93	2.42	0.86	1.08
September	0.74	1.39	0.94	1.75	1.01	1.88	0.62	1.16
October	1.27	2.06	1.42	2.07	1.39	2.01	1.00	1.45
November	1.42	1.86	1.43	1.87	1.47	1.92	1.10	1.43
December	1.58	2.06	1.54	2.01	1.59	2.04	1.21	1.58

While annual averages provide a basic overview of how well each model predicts energy consumption over the year, they fail to reflect monthly inconsistencies that are critical for model validation. Table 5.3 summarizes the range of monthly ratio values for each simulation tool and the annual average for both years. Notably, IES-VE appears to have the narrowest average gap in 2023 (0.98), but its range (0.46–1.30) reveals significant monthly variability, which is a key insight discussed further in the following section (5.3).

Table 5.3: Annual Statistics of Simulation Gap (Measured/Simulated)

Simulation Tool	Range in 2023	Range in 2024	Average in 2023	Average in 2024
IDA-ICE	0.67–2.29	1.39–3.53	1.29	2.10
DesignBuilder	0.86–3.8	1.75–5.89	1.38	2.26
ClimateStudio	0.96–6.74	1.88–10.40	1.43	2.34
IES-VE	0.46–1.30	1.08–2.09	0.98	1.60

Table 5.4 shows that IES-VE produced the closest annual total to measured 2023 electricity use, with an annual difference and MBE of +1.9%. By contrast, IDA-ICE, DesignBuilder and ClimateStudio underpredicted annual electricity use by −22.2%, −27.6% and −30.1%, respectively. However, the CVRMSE values show that monthly variation remained substantial for all tools. IES-VE had the lowest CVRMSE at 24.6%, but this still indicates that annual agreement should not be interpreted as full temporal accuracy. This supports the argument that annual EPG values can conceal compensating monthly deviations.

Table 5.4: Monthly error metrics for simulated electricity use against measured 2023 electricity consumption at 21°C

Simulation Tool	MBE	CVRMSE
IDA-ICE	−22.2%	35.7%
DesignBuilder	−27.6%	36.2%
ClimateStudio	−30.1%	37.8%
IES-VE	+1.9%	24.6%

5.3 Discussion

This chapter demonstrates that the EPG is not a fixed value but is shaped by modelling choices, tool selection, and the temporal resolution of analysis. It also shows that the "ground truth" of measurement may be subject to significant variations. Even with harmonized inputs, inter tool variability remained substantial. Apparent agreement at the annual scale, as with IES-VE in 2023, concealed significant monthly fluctuations. Under and over predictions cancelled each other out when averaged, producing the illusion of accuracy. These reversals suggest that annual convergence may be coincidental rather than indicative of model reliability.

The inter-tool comparison revealed a notable divergence between IES-VE and the other three simulation tools, particularly in relation to heating electricity consumption. This divergence should not be interpreted as a general indication that IES-VE overpredicts all end uses, nor as evidence that the tool is invalid. The pattern varies by end use. Cooling electricity does not follow the same ranking, and fan electricity also shows a different relationship between the tools. This suggests that the discrepancy is end-use-specific and is more likely to reflect differences in the way equivalent modelling assumptions are translated into each software environment, including

HVAC/control representation, plant assumptions, auxiliary-energy treatment, software defaults and calculation procedures.

IES-VE, like other established dynamic simulation tools, has published validation evidence against ASHRAE Standard 140 benchmark procedures (Oleiwi, Mohamed et al. 2019). This is important because the results reported in this chapter may appear, at first sight, to contrast with the formal validation evidence available for IES-VE. However, ASHRAE Standard 140 is a standardised method of test for evaluating building energy analysis software under defined benchmark conditions; it is not a guarantee that all applied real-building models will produce equivalent results across platforms. The present study is different in purpose and scope. It examines an applied case-study building with complex envelope characteristics, ground-source heat-pump assumptions, heat recovery, occupancy schedules and tool-specific modelling workflows. The observed discrepancy should therefore not be interpreted as evidence that IES-VE is invalid, but as evidence that, under the specific modelling assumptions and implementation pathways used in this case study, tool-specific HVAC representation and default calculation procedures can materially affect simulation outcomes.

To minimise avoidable input-related variability, the four models were developed using a common set of assumptions, documented in Appendix C. These included construction-specific envelope data, glazing and door properties, infiltration, thermal-bridging adjustment, HVAC assumptions, ventilation rates, heat recovery efficiency, heating and cooling setpoints, internal loads, occupancy densities, weather data, ground temperatures and output categories. These steps provided a consistent basis for comparison and reduced the likelihood that differences were caused by intentionally inconsistent input assumptions.

Nevertheless, the analysis cannot fully isolate the internal causes of the IES-VE discrepancy. Heat-balance and hourly diagnostic analyses could not be added because the necessary comparable time-series outputs were not available across all tools. The thesis therefore confines the analysis to the available annual and monthly measured/simulated outputs and avoids making unsupported claims about the internal causes of the discrepancy. Comparable hourly indoor air temperature profiles, surface temperature profiles, equipment-power profiles, heating and cooling power time series, HVAC-control outputs and heat-balance components were not available in a consistent form across the four simulation tools. Consequently, it was not possible to calculate MBE or CVRMSE for these unavailable hourly variables, nor to decompose the discrepancy into fabric losses, ventilation losses, solar gains, internal gains, plant effects or control effects.

The findings should therefore be interpreted cautiously. The available annual and monthly end-use outputs allow the discrepancy to be localised at a broad level, particularly in relation to heating and fan energy, but they do not allow the underlying causal mechanism to be proven. The discrepancy is therefore best understood as an example of applied modelling uncertainty: even where input assumptions are harmonised, different dynamic simulation tools may produce materially different estimates because of differences in algorithms, defaults, HVAC abstraction, control logic and implementation pathways. This reinforces the wider argument of the chapter that EPG magnitudes should not be reported as single deterministic values without transparent documentation of modelling assumptions, software settings, temporal resolution and uncertainty.

This interpretation is also consistent with the concerns raised by IDR2 and IDR4 interviewed as presented in Section 4.3.2. They noted difficulties in representing heat pumps in IES-VE and described cases where users calculate their performance externally on Excel spreadsheets or explore alternative platforms such as DesignBuilder. These views should not be treated as independent validation evidence against IES-VE, but they help contextualise why software choice, HVAC representation and modelling workflow may influence confidence in predicted outcomes.

A POE of the thermal test cells in 2022 revealed that several building elements deviated from the initial design assumptions (Strandberg-de Bruijn, Donarelli and Balksten 2019). For example, the external walls of test cells one and two were assumed to have U-values of 1.46 W/m^{2}K and 0.53 W/m^{2}K but were measured at 1.60 W/m^{2}K and 0.91 W/m^{2}K, respectively. Discrepancies were also found in the thermal conductivity, specific heat capacity, and density of external wall materials, with only the lime render density and hemp-lime in the second cell matching assumptions. The POE further highlighted differences in the operation of the HVAC systems. While the prediction stage assumed a heating setpoint of 21°C, actual operation shows that the temperature setpoint throughout the year was maintained at 20°C ± 1°C. The POE further showed that the cooling system remained inactive throughout the year, despite being modelled, and the ventilation system, assumed at 0.35 l/s/m^2 + 7 l/s/person, was also inactive. Weather data from Sveby, however, closely matched on-site measurements, with only minor differences attributed by authors to sensor inaccuracy (Strandberg-de Bruijn, Donarelli and Balksten 2019).

Measured annual heating electricity use for the first test cell in 2022 was 495 $kWh \cdot m^{-2} \cdot a^{-1}$. Initial predictions were substantially higher: IDA-ICE at 1064 $kWh \cdot m^{-2} \cdot a^{-1}$ and Climate Studio at 1012 $kWh \cdot m^{-2} \cdot a^{-1}$ for the first cell, with similar overpredictions for the second cell. Measured outcomes corresponded to roughly half of these predictions. Following POE adjustments, simulated heating electricity use dropped to 434 $kWh \cdot m^{-2} \cdot a^{-1}$ in IDA-ICE and 383 $kWh \cdot m^{-2} \cdot a^{-1}$ in Climate Studio, reducing inter-tool differences to 13%. Relative to these adjusted predictions, the measured use was 14% higher than IDA-ICE and 29% higher than Climate Studio. For the second cell, measured use was 254 $kWh \cdot m^{-2} \cdot a^{-1}$ compared to adjusted predictions of 302 $kWh \cdot m^{-2} \cdot a^{-1}$ (IDA-ICE) and 270 $kWh \cdot m^{-2} \cdot a^{-1}$ (Climate Studio), showing underpredictions of 16% and 6% respectively (Strandberg-de Bruijn, Donarelli and Balksten 2019).

5.3.1 Key Takeaways

Since first indications of the EPG, most of our accumulated knowledge regarding its magnitude has generally been based on compliance modelling outcomes and investigations of regulatory and static EPGs, as in the case of this investigation. In this chapter, only two dimensions of uncertainty origins were tested (tool selection and simulation scenarios) with identical models in terms of geometry and input data, which still demonstrated how misleading EPG values could be if a simulation model is not properly calibrated. The narrowest annual ratios revealed significant variations when investigated monthly. The monthly range revealed both under- and over-predictions. These opposing monthly under- and over-predictions cancelled each other out when averaged, creating a false sense of accuracy. Higher variations could possibly be observed if investigated at weekly or daily resolutions.

It may be reasonably argued that variations in the magnitude of the EPG are to be expected, given the differences in different typologies of the EPG, quantification methodologies and simulation tools and scenarios employed, even when measured performance remains constant across cases. Nevertheless, to effectively address numerous and well-documented challenges associated with the EPG, it is imperative to standardize the methodology under which the EPG is calculated and reported. Such standardization efforts are essential for accurately assessing the scale and implications of the EPG at building, local, municipal, national, and international levels. These implications include, but are not limited to, the management of energy supply and demand in the built environment and the widely recognized difficulty of meeting carbon reduction targets, objectives that are unlikely to be achieved without a fundamental transformation in building design, construction practices, and operational strategies.

If current methods of calculating and reporting the EPG persist without revision and the phenomenon continues to be observed as a discrepancy between measured and simulated energy performance, while neglecting the complexities inherent in predictions and measurements, the resulting body of knowledge will be fundamentally misleading in certain areas. Under such conditions, future investigations on the magnitude of the EPG could unintentionally act as sources of complication rather than illumination, limiting the advancement of knowledge of the phenomenon and preventing researchers from drawing meaningful correlations and conclusions when investigating the magnitude of the EPG across studies and on a wide scale. Furthermore, relying solely on discrepancies between measured and simulated energy performance when interpreting the EPG contradicts the fact that the phenomenon constitutes a regulatory-driven, socio-technical-economic issue in which information, knowledge, technology, behaviour, experience, economic considerations, and policy are deeply interwoven and collectively influential.

While it is essential to investigate EPGs arising from discrepancies between measured and simulated energy performance, it's more critical to adopt a more holistic and policy-oriented approach to understanding and addressing the EPG. Such an approach should account for regulatory, environmental and economic considerations and the diverse interests and priorities of different stakeholders. In contrast to compliance modelling, which is frequently regarded as a box-ticking exercise that inadequately represents the actual performance of buildings, the EPG between measured and target energy consumption offers a more meaningful metric. This is because it provides a basis for identifying performance deviations and implementing targeted mitigation strategies and energy-saving measures. Within this context, dynamic simulation tools should not be viewed solely as instruments for regulatory compliance; rather, they serve as valuable tools for risk analysis, enabling the exploration of alternative design strategies and operational scenarios to achieve predefined energy performance targets.

In such a case, the calibration process could be more valuable if it is not regarded simply as a process undertaken to close or mitigate the EPG. This is often a reductionist claim that ignores the nature of the phenomenon, the complexity of accurately accounting for all uncertainties in measurement and predictions, and the cascading gaps and conditions under which the calibration process can be terminated. Instead, calibration becomes a process in which various simulation scenarios can be tested to investigate mitigation strategies for the EPG that align with the financial capabilities of owners where a mitigation strategy that might work in one building might be difficult to implement in another.

In a policy-oriented context, where building typology and geographical location are influential factors in understanding the EPG, professional bodies such as the Chartered Institution of Building Services Engineers (CIBSE) in the UK have introduced energy benchmarks categorized as ‘typical’ and ‘good practice’ to facilitate a more informed assessment of building performance. These benchmarks serve as reference points to contextualize the EPG and highlight the extent of performance deviations. However, significant variability can arise when the EPG for a single building is assessed against both benchmark types. For instance, the measured energy use may fall below the typical benchmark, suggest a favourable outcome, while exceeding the good practice benchmark, indicating underperformance. Such inconsistencies obscure the true magnitude of the EPG and further create false sense of observing buildings either as energy efficient or inefficient.

From an investor’s standpoint, the EPG is not solely an energy-related discrepancy, but also a financial concern manifesting as a variance between projected and actual operational expenditures. From an environmental perspective, the EPG poses a substantial barrier to achieving carbon reduction targets, and there is an urgent need to investigate the impact of the EPG on climate objectives and to identify the energy efficiency measures required across different building types to meet decarbonization goals.

This chapter has demonstrated that the magnitude of the EPG is not a fixed value, but a moving one shaped by modelling choices, temporal resolution, and the benchmarks against which measured performance is compared against. Even with harmonized inputs, inter-tool variability and month-to-month reversals between underprediction and overprediction revealed how annual averages can mask poor temporal fidelity, while comparisons against policy thresholds produced favourable outcomes for the same building that appeared unfavourable when assessed against simulation baselines. The results and analysis highlight the need for a consistent procedure that specifies definitions, benchmarks, and reporting granularity which enables understanding of the EPG in a more holistic way and propose various mitigation strategies based on the typology investigated. The POE evidence shows that discrepancies in material properties and system operation are major contributors to the observed gaps. Deviations in U values, inactive cooling and ventilation systems, and differences in HVAC setpoints demonstrate how reliance on assumed conditions can distort predictions. This highlights the need to integrate empirical verification into simulation studies in order to avoid misplaced confidence in modelled outcomes.

To address these challenges, standardization in the definition, calculation, and reporting of the EPG is essential. Without consistent methodologies, cross study comparisons risk producing misleading conclusions and limiting the value of research for policy and practice. Standardization would also support more reliable decision making for policymakers, designers, and investors by clarifying the scale of the problem and aligning assessment with energy and carbon reduction objectives. Calibration plays a critical role in improving model fidelity but should not be understood merely as a mechanism to narrow the gap. Instead, calibration should be seen as an exploratory process that allows practitioners to test alternative design and operational scenarios, assess sensitivities, and evaluate mitigation strategies. Framed in this way, calibrated models serve as decision support and risk analysis process.

The implications of the EPG extend across multiple stakeholder groups. For designers, the results underscore the importance of tool selection and calibration. For policymakers, variability challenges the reliability of compliance-based assessments. For investors and building owners, the EPG translates directly into discrepancies between projected and actual operational expenditures. From an environmental perspective, unresolved gaps undermine progress toward decarbonization goals, while from a governmental perspective they complicate the management of energy supply and demand, creating uncertainty for grid capacity and infrastructure planning. Recognizing the EPG as a socio technical and economic phenomenon is therefore essential in order to design responses that extend beyond technical interventions and incorporate wider policy, behavioural, and market dimensions.

This chapter demonstrates that the EPG is not a fixed value but is shaped by modelling choices, tool selection, temporal resolution, and benchmarks. Even with harmonized inputs, inter-tool variability remained substantial. Apparent annual agreement, as with IES VE in 2023, concealed significant monthly reversals where under- and overpredictions cancelled each other out, producing the illusion of accuracy. The narrowest annual ratios still showed wide monthly variations, and finer resolutions (weekly or daily) may reveal even greater discrepancies.

Variations in the magnitude of the EPG are expected, given differences in typology, quantification methods, and scenarios. To address well-documented challenges, standardization of definitions, calculation, and reporting of the EPG is imperative. Consistent methodologies are essential for cross-study comparison, accurate assessment at multiple scales (building to international), and effective alignment with carbon reduction targets and energy management goals. If current methods persist focusing only on discrepancies between measured and simulated performance without addressing underlying complexities, future investigations risk producing misleading conclusions. The EPG must be understood as a socio-technical-economic phenomenon influenced by regulation, knowledge, technology, behaviour, and policy, not just as a modelling mismatch.

A holistic, policy-oriented approach is needed when addressing the EPG, one that considers regulatory, environmental, and economic perspectives. Unlike compliance modelling outcomes, which are often treated as a box-ticking exercise, the EPG between measured and regulatory-driven target energy consumption provides a meaningful metric for identifying deviations and informing mitigation strategies. Dynamic building energy simulation tools should therefore be viewed not only as compliance instruments but also as tools for risk analysis and the exploration of more efficient design alternatives, operational strategies and energy saving measures.

Calibration plays a critical role in improving model fidelity but should not be reduced to a gap-closing exercise. Instead, it should be an exploratory process that allows testing of design and operational scenarios, assessing sensitivities, and identifying mitigation strategies aligned with the financial capacities of building owners.

Benchmarks further complicate interpretation: measured energy use may appear favourable against a “typical” benchmark yet unfavourable against a “good practice” benchmark. Such inconsistencies obscure the true magnitude of the EPG and risk categorizing buildings incorrectly as efficient or inefficient. The implications of the EPG cut across stakeholders:

1. **Building energy analysts and FMs** must account for tool selection and proper calibration of simulation models and must collaborate to explore different strategies of mitigating the EPG at a building level.

2. **Investors and building owners** encounter financial impacts through discrepancies in projected versus actual operating costs.

3. **Environmental and governmental bodies** must contend with barriers decarbonization and uncertainty in energy supply, demand and grid planning.

In conclusion, standardization, holistic calibration, and recognition of the socio-technical-economic nature of the EPG are critical. This will support more reliable decision-making and ensure a policy is used effectively to meet energy and carbon reduction objectives.

6. Standardization Efforts

The overarching aim of this chapter is to examine how a standardized approach for quantifying and reporting the EPG can be developed and applied, drawing directly on insights provided by expert participants. In doing so, the chapter addresses RO4: Assess the potential for standardization in the quantification and reporting of the energy performance gap, identifying pathways toward greater consistency and comparability. It builds upon the findings presented in chapters Three and Five, which revealed the difficulty of establishing reliable estimates of EPG magnitude and highlighted the importance of transparent quantification and reporting in demonstrating the significance of the phenomenon. By clarifying the principles and methods through which the EPG should be quantified and reported, the framework introduced in this chapter aims to support evidence-based decision-making among designers, policymakers, and operators seeking to mitigate the EPG.

The detailed research methodology is outlined under Section 6.1, followed by the results and analysis of semi-structured interviews wherein the case for standardization is established (Section 6.2). Building upon these findings, the framework was further discussed in a panel discussion, as reported in Section 6.3. A synthesis of the findings is then presented under Section 6.4.

6.1 Methodology

A three-stage sequential exploratory methodology was implemented as presented under Figure 6.1. Each stage informed the next:

- Stage I (Section 6.2): Semi-structured interviews to establish the case for standardisation.
- Stage II (Section 6.3): Refinement of preliminary framework through expert panel discussion.
- Stage III (Section 6.4): Insights from both phases were synthesised to develop a framework for quantification and reporting of the EPG.

Prior to stage one, a preliminary conceptual framework guided questioning for the first phase of the investigation as illustrated under stage one within Figure 6.1. The EPG at building level was defined as the difference between measured and regulated target energy performance, with dynamic-simulation tools being considered as means of achieving such targets. At the level of individual energy systems, such as lighting, heating, or ventilation, the EPG was defined as the deviation between measured and simulated energy performance, with acceptable deviations influenced by the degree of occupant control and the specific characteristics of the systems involved. This stage therefore revealed the understanding needed to address the following through interviews guided by three thematic sections:

1. Theme one: Perceived importance of EPG standardisation to establish the case for investigations which will follow.
2. Theme two: Determination of suitable benchmarks for quantification of the EPG.
3. Theme three: Judgement on deviations and extent to which deviations between predicted and actual energy performance can be deemed acceptable, with the aim of establishing criteria for determining the presence and significance of the EPG.

The collective findings from the interviews and the panel discussion were synthesised, leading to the creation of the framework that operationalised the key concepts identified throughout the iterative process. This sequence of refinement reflects the principles of iterative and reflexive research design, which emphasise the cyclical development of theory through continuous engagement with empirical evidence and conceptual reflection (Srivastava and Hopwood 2009, Maxwell 2013). It is consistent with approaches used in grounded theory and design-based research, where understanding evolves through constant comparison, feedback, and theoretical integration (Charmaz 2014, Corbin and Strauss 2015).

This process ensures that both the refined framework is conceptually robust, empirically grounded, and contextually appropriate. This method is recognised as an important feature of rigorous inquiry in sustainability and building performance research, where theoretical models must be tested and adapted in relation to practical realities (Brown and Vergragt 2016, Creswell and Plano Clark 2018). The iterative process therefore strengthened the validity of the investigation and enhanced the practical relevance of its outcomes, linking conceptual development with real-world applications.

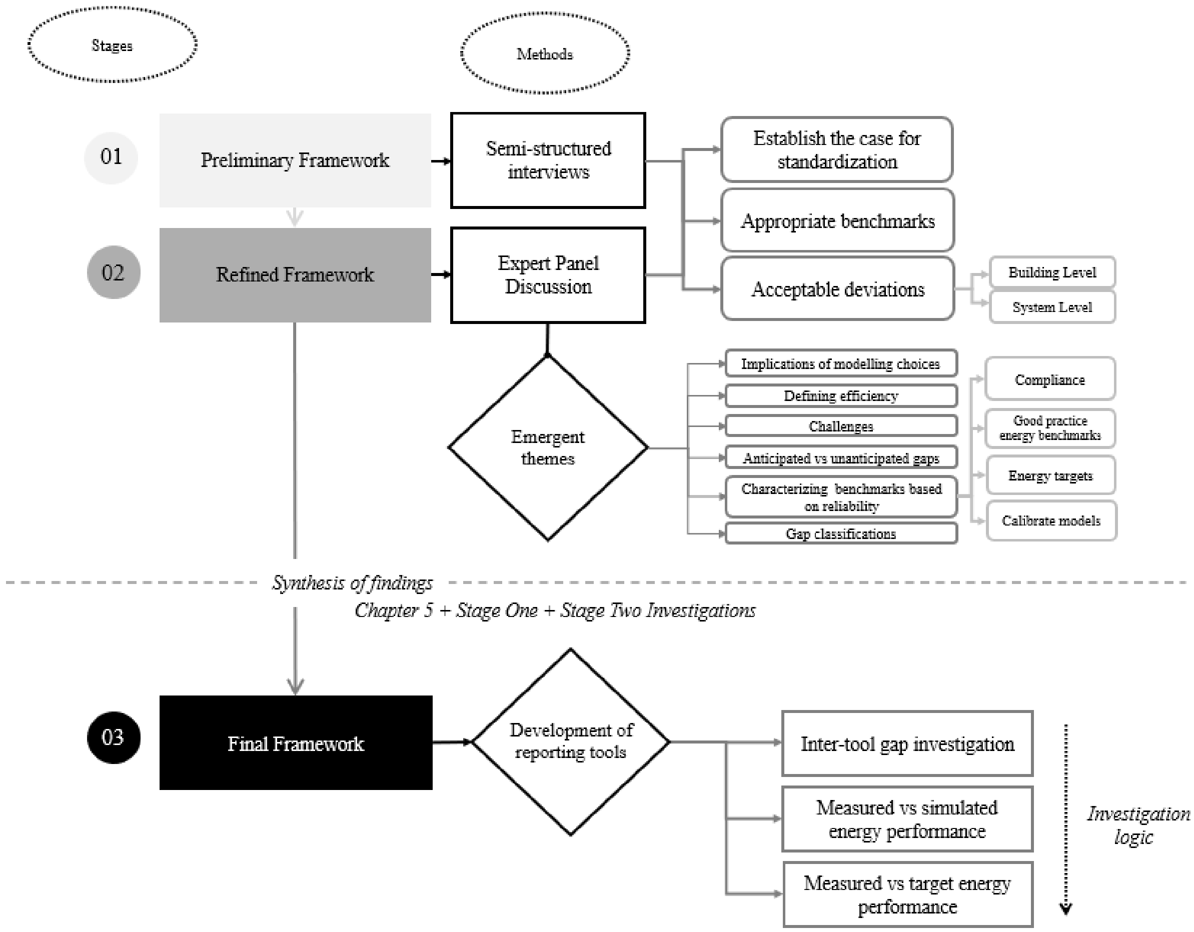


Figure 6.1: Methodologies adopted to achieve the aim of the chapter divided into three stages

6.1.1 Participants Selection

A purposive sampling strategy identified five experts with complementary academic, industry, and policy experience relevant to building energy performance and the EPG:

- Stage I: Three participants (SS1–SS3) provided in-depth insights from engineering practice, sustainability strategy, and public-sector implementation. Their collective expertise ensured the investigation captured practical challenges in applying EPG standards in real-world projects.
- Stage II: Two internationally renowned experts (EP1–EP2) were consulted. Both have been instrumental in developing major building-performance simulation tools, including ESP-r and EnergyPlus, and bring extensive experience in energy-systems modelling, professional training, and applied research. Their contributions provided authoritative technical insight and

helped validate the methodology against current industry practice, ensuring integration of advanced simulation knowledge with real-world applicability.

Through the second stage of the research, discussions with EP1 and EP2 revealed several critical insights that extended far beyond the scope of the intended discussion plan. Their deep technical expertise in building-performance simulation, EPG and current industry practices uncovered important dimensions of the EPG that had not been explicitly addressed earlier. Their contributions proved pivotal in validating certain elements of the thesis, even though this had not been the initial intention. The conversations brought to light a range of emergent themes, particularly around the differentiation between anticipated and unanticipated EPGs, as well as the complexities of judgement involved in model calibration and interpretation. Notably, both participants reframed the concept of efficiency in the context of the EPG, highlighting that the presence of a performance gap can distort perceptions of a building's true efficiency or inefficiency. Their perspectives enriched the research by bridging theoretical considerations with the practical and methodological challenges of accurately quantifying and interpreting building performance.

6.2 Establishing the Case for Standardization

In the first stage of investigation, all three participants emphasized the urgent need for a standardized and transparent approach to quantifying and reporting the EPG. Although their perspectives differed in emphasis, ranging from refining existing benchmarks to developing new ones, they proposed complementary approaches to establishing consistent energy performance benchmarks for EPG calculation. Section 6.2.1 summarizes the three participants' perspectives on the appropriate energy benchmarks to use for the quantification of the EPG, while Section 6.2.2 outlines their views on what constitutes acceptable and unacceptable deviations that determine the presence of an EPG and its significance.

6.2.1 Benchmarks

SSI3 challenged the prevailing use of the term *"performance gap"*, contending that in many cases it is more accurately described as a compliance gap. He explained that regulatory compliance models such as Simplified Building Energy Models (SBEM) are designed to demonstrate adherence to minimum building standards rather than to reflect actual operational performance. Consequently, comparing in-use energy data to such models is, in his words, quals "*comparing* *apples with pears."* He argued that robust performance evaluation requires standardized frameworks such as NABERS, which enable consistent normalization for variables including occupancy, weather and usage patterns. This, as he suggested, is what makes NABERS an effective mechanism for calculating the EPG within the office sector. However, he noted the limitation that NABERS currently applies only to specific typologies, calling for the development of similar sector-specific methodologies, particularly for residential, healthcare and higher education buildings, to facilitate accurate benchmarking and decarbonization tracking across the built environment.

SSI1 emphasized the urgent need to establish reliable, quantitative in-use data as the foundation for developing credible energy benchmarks. He argued that current industry practices, particularly those relying on standardized model-based tools such as Energy Performance Certificates (EPCs), fail to reflect actual operational realities. Instead, he advocated for more granular and meaningful benchmarking approaches that capture real energy use across different building types.

In his view, the development of such benchmarks depends fundamentally on improved data sharing, supported by both cultural change and legislative enforcement. He proposed the creation of a centralized database of operational performance data to facilitate transparency, feedback, and accountability. However, he acknowledged that significant barriers persist, primarily arising from commercial sensitivities and a deeply ingrained reluctance within the industry to disclose performance information. Consultants and contractors, he noted, often fear exposure or professional liability when measured outcomes diverge from predicted values.

SSI1 argued that genuine progress in performance measurement and benchmarking would likely require a policy shift toward a performance-based framework similar to NABERS, which mandates annual performance reporting. Such an approach, he suggested, would promote continuous improvement and ensure accountability over time, but would also represent a *"huge change"* in industry culture and practice.

He further highlighted that benchmarking challenges are compounded by the diversity of building operations, even within a single typology. For example, speculative offices, bespoke corporate headquarters, and call centres differ significantly in function, occupancy, and load profiles, while some high-end developments incorporate atypical energy demands such as data centres or large catering facilities. Comparing such varied buildings against generic benchmarks risks producing misleading conclusions.

According to SSI1, progress toward standardized, performance-based benchmarking will therefore require both incremental steps and structural reform. Sectors like offices have advanced due to better data access and sharing, whereas others, such as healthcare, continue to lag behind. He concluded that the absence of shared, high-quality in-use data not only impedes benchmarking but also restricts designers' feedback loops and learning opportunities. Legislative enforcement, in his view, remains essential to overcoming these barriers and achieving credible, transparent, and effective benchmarking across the industry.

SSI2 echoed these concerns, underscoring that *"there isn't a standard method"* for calculating and reporting the EPG. However, she highlighted that the UK Net Zero Carbon Building Standard (NZCBS) represents a significant step toward resolving this gap. As an initiative jointly governed by professional institutions such as CIBSE and RIBA, the NZCBS consolidates benchmark data across sectors and integrates best-practice insights from in-use exemplars. SSI2 cautioned against *"reinventing the wheel,"* advocating instead for alignment with such emerging standards that already benefit from broad institutional backing. Crucially, she emphasized that performance targets must be embedded from the project inception: *"you need to set that target right at the start and then design the building to meet that target."*

These perspectives, as summarized in Table 6.1, reveal a critical tension between the call for entirely new, bespoke benchmarks tailored to specific building types (SSI1), the pragmatic adoption of a national, consensus-driven standard that ensures comparability and avoids fragmentation (SSI2), and the proposal to extend NABERS beyond its current office-building scope (SSI3).

Table 6.1: Synthesis of perspectives on appropriate benchmarks for calculating the EPG

Participant	Proposed benchmark	Elaborations
SSI1	Creation of new, data-driven benchmarks developed through the systematic collection and analysis of in-use performance data.	Legislative enforcement and cultural change to ensure participation and data sharing.
SSI2	Adoption of a nationally coordinated benchmark consolidating data from professional bodies.	The UK Net Zero Carbon Building Standard (NZCBS)
SSI3	NABERS	Extension of NABERS beyond offices

6.2.2 Setting Acceptable Deviations

SSI1 argued that defining acceptable deviations from predicted or target energy use is inherently complex and highly dependent on building type, operational intensity, and client use. He noted that an office designed for a client with energy-intensive operations could exhibit a substantially higher Energy Use Intensity (EUI) than a typical office of similar size and form, without necessarily indicating poor performance. In such cases, he suggested that "a percentage deviation may be more practical than an absolute difference in EUIs," although he cautioned that thresholds must vary by context. He illustrated this variability by noting that "if we are somehow able to quantify that a 10% variance is acceptable, it might be 15% to 20% on a different type of office block," depending on how the building is used and designed.

SSI1 linked this variability to the design and procurement context, explaining that speculative developments may warrant greater tolerance than bespoke client-led projects, where end-user engagement allows more accurate prediction of operational patterns. Ultimately, he maintained that acceptable thresholds can only be defined credibly through the accumulation and analysis of large-scale benchmarking datasets that compare predicted and actual energy performance. As he explained, "once you have a big enough dataset, then you can start to see what the current percentage or EUI variance is, and then over time adjust those variances."

SSI2 was less definitive, initially acknowledging that she could not specify what deviation would be acceptable. However, she proposed learning from empirical evidence in exemplary projects, particularly PH. Rather than prescribing a fixed threshold, she recommended uncovering what is acceptable through what is achievable. In this sense, SSI2 framed acceptable deviation not as a static figure but as an evidence-based measure derived from best-practice projects, implying that what is considered acceptable should evolve with industry capability and available performance data. Their perspectives highlight that defining acceptable deviation is both a contextual and evidence-driven challenge: SSI1 stresses variability across building types and the need for aggregated datasets to establish realistic thresholds, while SSI2 points to high-performing exemplars as a guide to determine what is achievable. A detailed comparison of their perspectives is presented in Table 6.2.

At the energy system level, both SSI1 and SSI2 emphasized that setting acceptable deviations is considerably more complex than at the building level, since performance is shaped not only by technical efficiency but also by system integration, occupant behaviour, and operational governance. SSI1 supported categorizing systems according to levels of occupant control, described as high, medium, or low, and observed that "systems are made up of component parts from different manufacturers," which makes it challenging to link measured outcomes neatly to a single specification. He advocated benchmarking subsystems such as HVAC, lighting, and plug loads, supported by granular metering and smart building data. For example, he noted that people counters and booking-linked controls can be used to normalize consumption and identify anomalies, particularly when different floors serve different functions.

SSI2, by contrast, viewed system-level EPG less as a question of numeric allowance and more as an issue of procedural governance and user management. She emphasized adopting structured frameworks and developing risk registers to anticipate and mitigate underperformance. Rather than permitting wider deviations where occupant control is high, she proposed constraining user behaviour through setpoint caps, user training, and facilities management oversight. She also recommended regular monitoring, suggesting that each month an "energy champion" review performance to identify anomalies such as heating left on while doors are open. In her view, the EPG is as much about people and communication as technology, underscoring that social and procedural interventions are critical to narrowing system-level EPGs. Table 6.2 below provides a summary of the two perspectives.

Table 6.2: Comparison of SSI1 and SSI2 perspectives on acceptable deviations in energy performance

Theme	SSI1 Perspective	SSI2 Perspective
Approach to defining acceptable deviations	• Context-specific and data-driven, with thresholds varying by building type and use.	• Evidence-based and evolving, derived from empirical data in high-performing exemplar projects rather than fixed thresholds. • Acceptability should be guided by what is demonstrably achievable.
Basis for determining thresholds	• Large-scale benchmarking datasets. • Over time, aggregated data would allow thresholds to be refined and normalized across typologies.	• Recommends empirical investigation using best-practice examples such as PH projects to establish what is realistically achievable, rather than prescribing numerical limits upfront.
Contextual factors	• Emphasizes that building type, client operations, and design intent strongly influence acceptable deviation.	• Frames acceptable deviation as dynamic, evolving alongside industry learning and improvements in performance data. • Highlights the importance of context and project ambition.
System-Level Deviation	• Deviations at the system level are complex due to varied system architectures, component integration, and user control.	• Focuses on procedural governance and user management rather than numeric tolerances. • Advocates frameworks such as SL and risk registers, emphasizing behavioural controls, training, and monthly performance reviews by an 'energy champion'.

6.3 Expert Panel

In this stage the framework was refined and discussed with two internationally well-known experts in the building energy simulation community. The discussion started by presenting issues related to the ambiguity associated with the actual magnitude of the EPG leading into the emergence of various themes such as the influence of modelling choices on the size of the EPG as outline under section 6.3,1, defining efficiency in the context of the EPG (6.3.2) and anticipated versus unanticipated gaps (6.3.5) and classifying EPGs (6.3.6) which have emerged due to challenges of applying proposed roadmap for standardizing the EPG.

6.3.1 Modelling Choices and Ambiguity of the Size of the EPG

Both participants agreed that modelling choices exert a significant influence on the outcomes of simulation models and, by extension, on the apparent magnitude of the EPG. Their reflections highlight both the case-specific nature of modelling decisions and the broader challenge of reducing variability and uncertainty.

EP1 emphasized that sensitivity to modelling choices is highly contextual. In some buildings, small changes to assumptions such as temperature setpoints can produce significant shifts in predicted performance, whereas in others the effect is marginal. He observed that *“five different people approaching the same building will have different approaches to conducting the simulation,”* with each analyst prioritizing different parameters and kinds of feedback. This leads to divergent results even for ostensibly the same problem. At the same time, EP1 stressed that modelling inherently involves simplification. He noted that *“large sections of heat-flow paths”* are frequently omitted from building representations, meaning that numerical engines cannot deal with things that are excluded. Analysts must therefore decide which elements to include, which to simplify, and which to ignore. These modelling choices are not trivial: they shape what the model can and cannot represent, and they can produce substantial differences in predicted performance. Nonetheless, EP1 also underlined that the iterative use of simulation to test variations such as changing window types to investigate whether performance improves is a normal and valuable part of the design process. In this sense, modelling choices not only introduce uncertainty but also provide a structured way of exploring design alternatives.

EP2 reinforced the point about variability of modelling outcomes. He referred to studies in which different groups were asked to simulate the same building, yet the results diverged by as much as ±20%. For him, this highlights the difficulty of separating building energy analysts from modelling outcomes. He argued that it would be *"incredibly useful"* for the community to develop ways of reducing the uncertainty and errors introduced by these subjective modelling decisions, thereby improving consistency and reliability in simulation outputs.

6.3.2 Defining Efficiency in the Context of the EPG

The participants also reflected on how energy efficiency should be judged in practice, particularly given that both favourable and unfavourable EPGs can present misleading impressions of energy efficiency. EP1 suggested that efficiency is best understood relative to the expectations of a building owner or facilities manager. If a building consistently performs at or below the level of energy expenditure that has been budgeted to keep it operational, then it can reasonably be regarded as energy efficient in that context. Conversely, if actual costs are substantially higher than expected or if the building repeatedly fails to meet agreed performance targets, then it should be considered inefficient. For EP1, therefore, efficiency is a pragmatic and context-specific judgement tied to whether a building aligns with the financial and operational thresholds set by those responsible for its management.

EP2 proposed a complementary approach. He recommended beginning with EUI as a baseline indicator. A high EUI does not necessarily mean a building is inefficient, but it should prompt further analysis of operational drivers. For instance, a 24-hour hospital or a data centre will naturally consume more energy than a typical office building due to their intensive internal loads. To distinguish true inefficiency, EP2 emphasized the need to compare such buildings with others of similar type, using measured data wherever possible to establish benchmarks. In this way, efficiency is judged not against abstract thresholds but against realistic sector-specific norms.

Both perspectives underline that efficiency cannot be inferred from the magnitude of the EPG alone. Instead, it requires contextual interpretation: for EP1, alignment with budgeted expectations and operational targets, and for EP2, benchmarking against comparable buildings with similar functions and loads.

6.3.3 Challenges

EP1 identified a series of interrelated challenges that hinder efforts to standardize the EPG. He emphasized that the viability of any standardization initiative ultimately depends on political will and institutional commitment. In his view, progress *"depends on how much policymakers actually want to carry on paying attention"* and on the quality of feedback received from the industry and how they act upon this feedback. Without consistent engagement from policymakers and regulatory bodies, existing benchmarks risk remaining underutilized, and poor-performing buildings may continue to be tolerated.

A fundamental difficulty, according to EP1, lies in determining who should hold the controlling authority for EPG standardization and enforcement. He questioned *"who is the responsible agent that will be paying attention to the matter?"* and *"what form of sanctions can they enforce if buildings do not perform as intended?"* He warned against placing excessive trust in bureaucratic processes, noting that responsibility often defaults to building operators or owners rather than regulators, resulting in fragmented accountability and limited enforcement capacity.

EP1 also highlighted the challenge of achieving consensus across a diverse and fragmented sector. The simulation and engineering communities consist of many actors with differing priorities, making agreement on a single methodology unlikely. He articulated several unresolved questions that illustrate the uncertainty surrounding what should be standardized. Should the process of creating simulation models be standardized, and should there be agreed rules for the level of detail required? Should measurement methods be standardized, and what parameters should be measured? What defines the baseline or reality case against which models are compared? Should buildings be tested under a range of weather conditions to account for climatic variability?

These questions, in EP1's view, reveal that the concept of standardization itself remains ambiguous. Without clarity on its scope and intent, defining acceptable deviations risks inconsistency or superficiality. He proposed that meaningful progress would require long-term observation of building performance data, potentially over a decade, to capture natural fluctuations and establish realistic performance ranges. Deviations that fall outside these empirically derived ranges could then be more reliably classified as performance gaps rather than modelling anomalies.

EP1 cautioned, however, that achieving such standardization depends not only on technical consensus but also on cultural and institutional alignment. Standardization requires broad agreement among stakeholders to adopt a common methodology, yet many practitioners may be reluctant to alter established practices or accept stricter modelling and reporting requirements. This inertia, combined with competing professional interests, limits the sector's capacity to implement consistent approaches to measuring and reporting the EPG.

Finally, EP1 warned that focusing narrowly on energy metrics risks overlooking other critical performance dimensions. He argued that there are multiple interrelated gaps, including comfort gap and that to *"focus solely on energy performance is to examine only one dimension of a broader and more complex performance matrix."* He gave examples of buildings that meet energy targets but fail to deliver adequate thermal or visual comfort, often due to undersized systems or misaligned environmental controls. Such cases demonstrate, in his view, that performance gaps extend beyond energy consumption and that any credible standardization framework must acknowledge these interdependencies.

On the other hand, EP2 also underscored the fragmented nature of the field, describing the current state of practice as *"all over the place."* He noted that in the United States, for example, very few officials have the technical expertise required to properly review compliance regulations, which undermines the effectiveness of regulatory processes. He argued that the importance of the EPG depends entirely on objectives. If the goal is absolute energy reduction, then the EPG is a critical indicator. However, if the goal is to comply with a specific voluntary framework such as PH, the meaning and relevance of the EPG will differ. For him, the first step in any standardization effort must therefore be to clarify what stakeholders are trying to achieve and how the gap will be defined in relation to that goal.

6.3.4 Different Benchmarks Under which the EPG is Quantified

The framework presented to the participants illustrated how the EPG is typically quantified in literature, with measured energy outcomes compared against compliance modelling outcomes, good practice energy benchmarks, sustainability targets, performance targets, or calibrated energy simulation models. The participants reflected on the strengths and limitations of each approach, offering insights into which benchmarks might provide the most credible basis for standardization.

Compliance modelling: Both participants were strongly critical of compliance modelling as a benchmark. EP1 described it as a *"box ticking exercise,"* driven by political and regulatory expediency. He argued that its outputs are based on oversimplified assumptions and bear no correspondence to any form of reality. He provided the example of compliance consultants where an entire hospital west facing façade is abstracted despite the fact rooms serve different purposes, a practice described to have *"no basis in physics"* but necessary to satisfy regulatory requirements within tight time and resource constraints. In his view, compliance tools are misused when treated as predictors of actual performance. At best they are regulatory artefacts designed to demonstrate minimum compliance with regulatory requirements. He also criticized the widespread production of low-cost energy performance certificates, questioning the credibility of results produced in minutes. EP2 supported this critique, emphasizing that compliance models do not reflect building performance but only demonstrate that a design meets or exceeds code requirements. He argued that comparing compliance outcomes with measured performance, in his words is like comparing *"apples and oranges."* For him, this misinterpretation has distorted discussions of the EPG, particularly where compliance outcomes are used to secure green building points rather than to represent operational reality.

Good practice energy benchmarks: EP1 noted that values of good practice benchmarks are drawn from accumulated experience and knowledge of how different building types typically perform in use. He considered them useful as a rough reality check when validating a simulation model, helping to identify whether results are wildly different from expected norms and thus flagging potential modelling errors. However, he cautioned against over reliance, stressing that good practice values are often less ambitious than best practice performance and what engineers can achieve and are shaped by conservative assumptions designed to avoid liability rather than to represent optimal outcomes. As such, he regarded them as rough quality assurance benchmarks than a credible reference for quantifying EPGs.

Energy and Sustainability Certification Scheme Targets: EP1 distinguished between highly efficient standards such as PH and more conventional building types. In his experience delivering PH workshops, he observed that the EPG is relatively small because the standard is so stringent about construction quality, meaning even minor faults such as a single thermal bridge have a large and visible impact on performance. By contrast, in older or more leaky buildings, small defects are obscured within overall inefficiency, and the EPG appears less sensitive to construction quality. This comparison illustrated his view that the magnitude of the EPG is closely tied to the rigor of the target itself and the degree to which it drives quality assurance during delivery. EP1 identified the gap between measured and performance targets as the most important type of gap from his perspective. He argued that when clients and design teams establish explicit goals, whether aligned with NABERS ratings or bespoke performance objectives, the existence of a shared commitment significantly increases the likelihood of narrowing the EPG. By contrast, in cases where compliance is the only goal, there is little incentive to minimize discrepancies.

He suggested that performance targets could be operationalized not only at an annual level but also on a shorter timescale such as daily or weekly, thereby allowing building managers to monitor deviations in near real time and respond to anomalies such as unexpected occupancy changes or unusual weather events. For him, the central issue is whether the design team have provided clients with accurate expectations and whether the building delivers on those commitments. In his view, this client-oriented accountability makes performance targets the most meaningful benchmark of all. Where explicit targets are linked to post occupancy verification, as in NABERS where certification is contingent on measured energy data, he argued that accountability is stronger, and the EPG is narrower.

Calibrated models: The panel acknowledged the potential of calibrated simulation models to reliability, though with important caveats. EP1 offered an extended reflection on what proper calibration involves. He described it as a process that starts with a model built from plans and sections then adjusted until the simulation aligns with measured outcomes. Then, measurements of the actual building's response to deliberate stimuli such as heat injections or window openings. For him this transforms simulation from a static design exercise into a dynamic process. By inserting this feedback loop engineers can detect specific faults, such as missing insulation in one façade zone, that explain discrepancies between predictions and reality. After several cycles the result is a well calibrated model that tracks actual building performance.

EP1 stressed the operational value of such models. Once calibrated, they can be used by FMs as reliable decision support tools, for example when converting rooms into conference space with higher occupancy. Because the rest of the building has already been validated against measured performance, managers can have confidence that the model will produce accurate forecasts for new scenarios. He suggested that this approach directly mitigates the EPG and, when repeated across multiple projects of similar type, generates patterns of learning that reduce uncertainty over time. In his view, calibration adds an essential extra step that transforms a model into a living asset rather than a disposable deliverable.

EP2 agreed that calibrated models can be powerful but emphasized that their accuracy is often undermined by uncertainties in construction quality. He identified airtightness as one of the most critical unknowns, noting that calibration frequently relies on adjusting infiltration rates because direct testing is rarely affordable. For him, the complexity of full building simulations means uncertainties can never be fully eliminated, though simplified calculations may sometimes be sufficient.

Expanding this reflection, EP1 also considered the potential role of digital twins. Drawing on the example of the aircraft industry, he explained that a genuine digital twin is one that operates in real time, such that the digital representation of an aircraft at any given moment mirrors the temperatures, pressures, and forces measured in the physical system. By contrast, most building energy models are highly abstract, shaped by assumptions and simplifications, and thus cannot be considered true digital twins. He illustrated his understanding of a more faithful twin through an experiment in which he monitored thermal and electrical behaviour in real time, then iteratively adjusted an ESP-r model until it tracked the measured values. This, in his view, came much closer to the true definition of a digital twin, as the physical and digital artefacts aligned closely at each moment in time. EP1 questioned whether the construction industry genuinely seeks to achieve this level of alignment, noting that many models branded as *"digital twins"* do not in fact mirror real-time conditions. For him, this definitional ambiguity risks diluting their value. Nevertheless, he argued that if real-time simulation were more rigorously implemented in buildings, it could directly support attaining performance targets by creating a continuous feedback loop in which predicted and measured outcomes are constantly compared. This would allow deviations to be identified and corrected as they occur, reducing unanticipated EPGs and enhancing long-term accountability. *Figure 6.2* below provides a comparison of perspectives on different energy benchmarks under which the EPG is calculated.

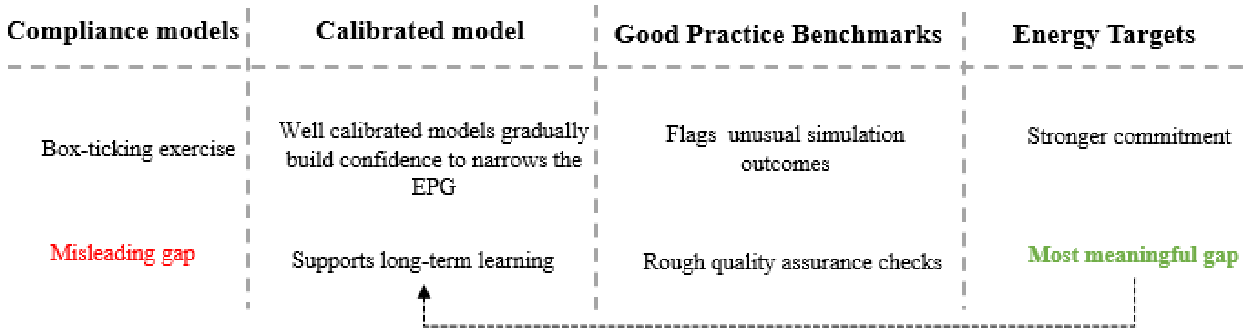


Figure 6.2: Distinction between different energy benchmarks under which the EPG is calculated and their representativeness of actual EPGs

6.3.5 Anticipated and Unanticipated Performance Gaps

In addition to commenting on benchmark selection, the participants introduced a distinction between *anticipated* and *unanticipated* performance gaps, which adds an important layer of nuance to how the EPG is conceptualized. EP1 explained that an unanticipated performance gap arises in cases where expectations are high and outcomes diverge unexpectedly from those expectations. For example, in a PH building the standard sets very stringent requirements, so even if performance falls slightly short of the target, the building is still likely to outperform typical stock and deliver a comfortable environment. He described it as a case when design teams overlook critical factors where unanticipated failures can emerge. He added, situations where highly insulated buildings risk overheating because of solar gains, producing outcomes very different from what occupants and clients expect. These unanticipated gaps are particularly significant because they directly affect bills, comfort, and user satisfaction, and they erode confidence in design claims.

By contrast, an *anticipated performance gap* occurs where underperformance is already understood and expected. EP1 illustrated this with reference to a 1912 building that struggled to heat adequately, rising only from 10 to 14 degrees after prolonged heating. In such cases the poor performance is already recognized and known, and in such case, there is no EPG, and the gap is the known fact of the undersized radiators and insufficient insulation. The gap therefore does not represent a misalignment between expectation and outcome, but rather a known and accepted deficiency. This distinction as illustrated in *Figure 6.3* below highlights that not all performance gaps carry the same weight in practice. Anticipated gaps represent acknowledged shortfalls in known contexts, whereas unanticipated gaps represent failures that matter more because they disrupt expectations and impose unforeseen costs or discomfort.

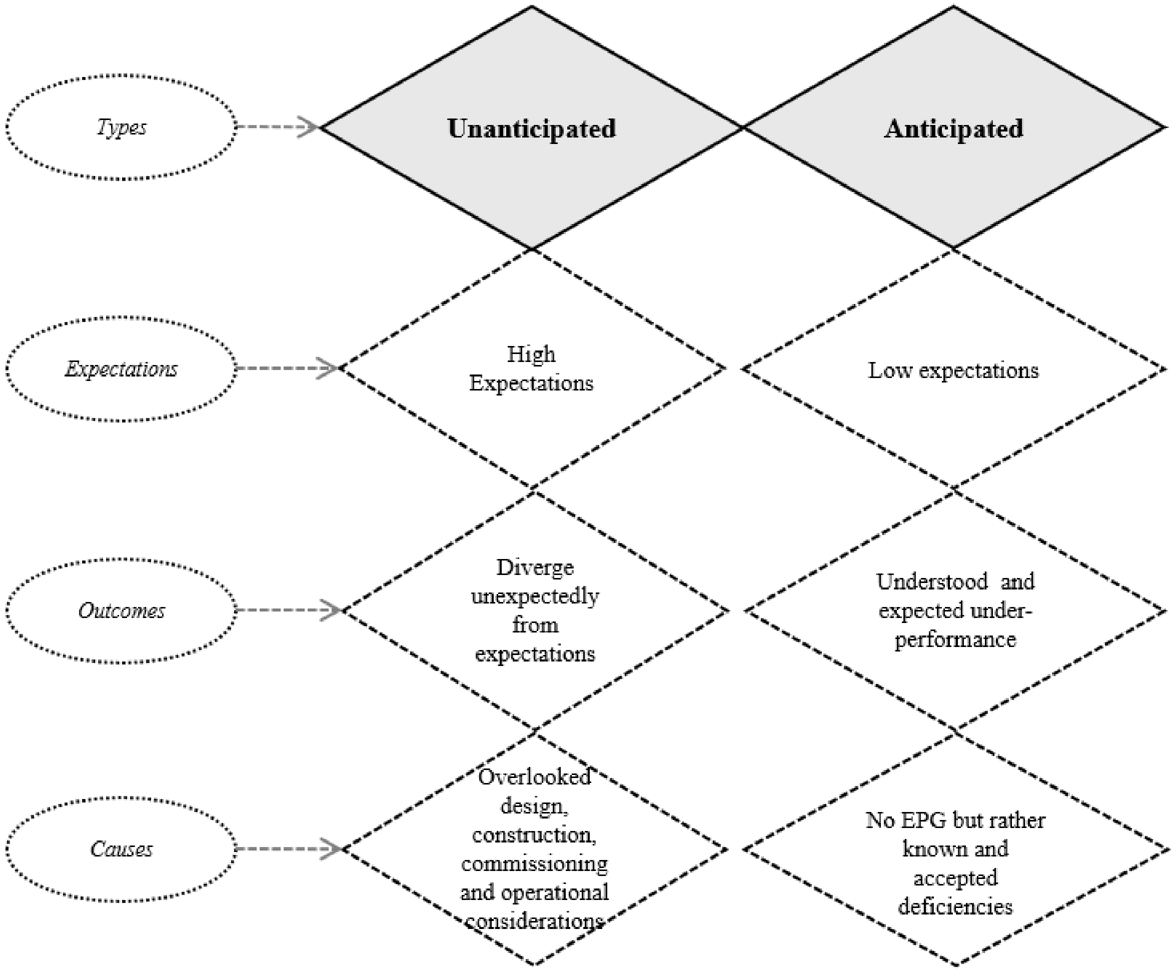


Figure 6.3: Anticipated versus unanticipated gaps

6.3.6 Classifications of the Energy Performance Gap

The participants reflected on how the EPG might be classified. Both proposed typologies that distinguish between different contexts, actors, and levels of analysis, underscoring the difficulty of giving all EPGs equally.

Four distinct classes were introduced, each associated with different stakeholders and forms of feedback. The first is a societal or regional gap, which is of interest to authorities responsible for planning infrastructure such as electricity grids or gas distribution networks. Understanding the discrepancy between predicted and actual energy use at this level supports effective long-term planning. The second class is a professional gap, relevant within professional organizations. Here the standard or benchmark provides feedback to practitioners, encouraging higher quality work and strengthening accountability in design processes. The third class is at the interface between the design team and the building owner or facilities manager. This gap highlights whether a building, once delivered, meets the expectations and commitments made during design, thereby shaping trust and responsibility between professional teams and clients.

The fourth class is the operational gap, observable in real time through detailed monitoring. This involves examining how a building responds dynamically to factors such as occupancy changes, window openings, or weather fluctuations. EP1 distinguished between real-time measures, which capture deviations during specific events such as cold snaps or mornings with high heating demand, and static measures, which reflect performance aggregated over days, weeks, or seasons. He suggested that this type of classification aligns closely with the principles of digital twins, enabling continuous error detection and adaptive management by revealing how performance patterns drift relative to expectations.

EP1 concluded that recognizing multiple classes of performance gap is crucial. Attempting to define a single EPG across all contexts is highly complex. Instead, a typology that specifies whether the gap is defined at regional, professional, contractual, or operational levels provides a more credible and actionable framework. EP2 also highlighted the importance of contextualizing the gap, offering the example of utility-driven incentives. In this case, the EPG is evaluated in terms of how much energy a building saves relative to a benchmark, demonstrating that yet another classification of the gap emerges when viewed from an incentive or policy perspective.

The different perspectives and classifications summarized in *Figure 6.4* below suggest that the EPG is not a single, monolithic concept but rather a family of related classifications that vary according to scale, purpose, and stakeholder perspective. Recognizing this diversity may allow for more tailored and effective strategies for quantification, reporting and mitigation.

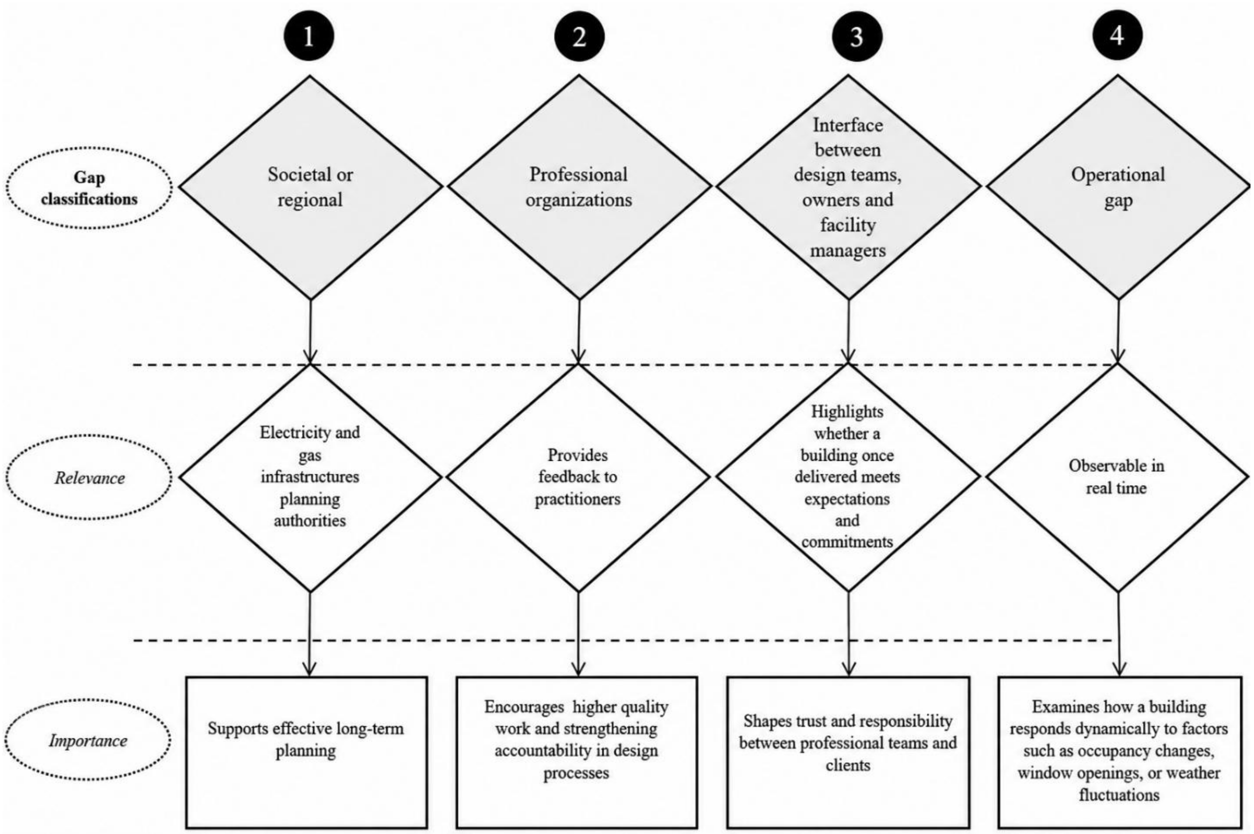


Figure 6.4: Classifications of the EPG, their relevance and importance

6.4 Summary

While the concept of a standardized methodology is attractive in principle, in practice, it is difficult to enforce due to diverse driving forces, varying building contexts, and the limitations of current practices. The challenges are not solely technical but also cultural and institutional. The reluctance to share building performance data driven by legal liability concerns and commercial sensitivities presents a major barrier. Even when data is available, it is often inconsistent across sectors and countries. The widespread use of compliance modelling, which tends to prioritize regulatory approval, further exacerbates the challenge of accurately quantifying the EPG. These models often rely on oversimplified assumptions, omitting crucial factors such as occupant behaviour and unregulated loads, which significantly affect actual performance. Beyond energy, there also exist comfort gaps, an issue that must be acknowledged in future standardization efforts. The EPG cannot be fully understood in isolation; it should be positioned within a broader matrix of building performance indicators.

Ultimately, standardization efforts must address several key areas: the modelling process, measurement methods, acceptable deviations, and data collection protocols. Even if these foundational issues are resolved, two major challenges remain rooteed in the following questions:

1. Who enforces standardization?
2. What form of sanctions could be introduced if buildings do not perform as intended?

The interviews and panel discussions reveal a strong and consistent consensus on the need for standardization in calculating and reporting the EPG. All participants agreed that the current state of practice is fragmented, inconsistent, and in many cases misleading. The rationale for standardization stems from the recognition that existing benchmarks, whether compliance models, energy performance certificates, or good practice benchmarks fail to provide reliable comparability across projects. Compliance modelling in particular was criticized as a "box-ticking exercise" that produces outputs unrelated to actual building performance. Without standardized procedures, interviewees stressed, the sector is left "comparing apples with pears," undermining both credibility and opportunities for learning. A standardized methodology must therefore not only establish common reporting rules but also embed safeguards for comparability, normalization, and uncertainty disclosure.

Although the case for standardization was uncontested, the experts diverged on what form it should take. One perspective (SSI1) called for the development of entirely new benchmarks derived from granular in-use datasets, supported by legislative enforcement to overcome the cultural and commercial barriers to data sharing. Another perspective (SSI2) cautioned against *"reinventing the wheel"* and argued instead for building on consensus-driven initiatives already underway, notably the UK Net Zero Carbon Building Standard. This approach emphasizes broad institutional collaboration and integration of benchmarks from project inception. A third voice (SSI3) highlighted the strengths of existing NABERS-style, but noted its limited sectoral scope, calling for extension beyond office buildings. This tension between creating bespoke sector-specific datasets and adopting unified national standards underscores the central design challenge for the proposed framework: it must be rigorous enough to capture sectoral variation yet simple and consensual enough to support comparability and wide adoption. The panel discussions reinforced that adoption ultimately depends on political will, regulatory enforcement, and credible sanctions for non-compliance. Standardization, in this view, is not only a technical exercise but also a governance and policy challenge, one that requires both institutional authority and cultural change in professional practice.

A major point of debate concerned how to define *“acceptable deviations”*. SSI1 argued that deviation thresholds should be expressed in percentage terms, given the variability across building types, client operations, and occupancy patterns. He stressed that only large-scale benchmarking datasets can credibly establish realistic thresholds over time. SSI2, by contrast, suggested that acceptable deviation should be derived from best-practice exemplars, such as PH projects, rather than fixed thresholds. Both views highlight that acceptable deviation is a contextual and evolving threshold, shaped by building typology, operational intensity, and the maturity of available data. A standardized framework must report both absolute and percentage deviations while embedding mechanisms for evidence-based adjustment as datasets grow. Participants pointed out that acceptable performance deviations vary by building type and context. In tailored buildings (e.g., offices with specific end users), performance expectations can be tightly defined, and deviations more critically assessed. In contrast, speculative buildings or those with unknown usage patterns warrant a broader tolerance. This complexity calls for a tiered benchmarking framework, one that differentiates between system-level and building-level performance and accounts for degrees of occupant control.

At the system level, covering major building systems such as lighting, fans, boilers, pumps and façade-related components, two complementary pathways emerged. SSI1 emphasized a data-driven approach, using subsystem metering and smart building technologies to quantify deviations, normalise consumption and identify anomalies across different levels of occupant control. SSI2 advanced a governance-led approach, focusing on procedural controls, commissioning, SL and facilities management. For her, system-level gaps are addressed less through numeric allowances than through operational safeguards such as setpoint caps, risk registers and regular monitoring by “energy champions.”.

The expert panel added further depth by interrogating the role of modelling choices and uncertainty. Both experts highlighted that outcomes could vary significantly depending on assumptions, analyst choices, and simplifications. Small changes to setpoints or envelope assumptions can lead to divergent predictions, with studies showing up to ±20% variation across analysts simulating the same building. These reflections underscore the need for the framework to explicitly capture modelling assumptions, high-impact variables, and uncertainty ranges. Efficiency, moreover, cannot be inferred from the magnitude of the EPG alone. EP1 argued that efficiency should be judged relative to operational budgets and client expectations, while EP2 insisted on benchmarking against comparable buildings within the same typology.

The panel also examined the credibility of different benchmarks. Compliance models were rejected as unsuitable for EPG quantification, while good practice benchmarks were regarded as useful but conservative quality assurance checks. Performance targets, whether based on NABERS ratings, PH standard, or bespoke client goals were judged the most meaningful benchmarks because they align accountability with explicit commitments. Calibrated models were seen as highly valuable when developed iteratively and linked to empirical data, transforming static simulations into dynamic tools for diagnosis and decision-making. However, their reliability depends on context, construction quality, and the resolution of critical unknowns such as airtightness. Digital twins were considered a promising but still ambiguous extension of calibration, with their value contingent on whether they truly replicate real-time performance rather than abstract representations.

Finally, the experts introduced new conceptual distinctions that add nuance to the framework. EP1 differentiated between anticipated and unanticipated gaps: the former represents known shortfalls in acknowledged contexts, while the latter represent expectation failures that impose unforeseen costs and erode trust. Both experts also emphasized the importance of multi-scalar classification, distinguishing between societal or regional gaps, professional gaps, contractual gaps between designers and clients, and operational gaps observable in real time. These classifications underline that the EPG is not a single, monolithic concept but a family of related constructs that vary by scale, audience, and purpose. Any standardized methodology must therefore be flexible enough to accommodate these classifications while providing clarity about which type of gap is being reported.

6.4.1 Framework

The findings presented in this chapter indicate that standardizing the EPG in terms of quantification and reporting remains both conceptually and practically challenging. As discussed in the preceding sections, multiple methodological inconsistencies, data quality limitations, and contextual variations in modelling practice have hindered efforts toward achieving a unified and replicable approach. These challenges underline the complexity of standardising EPG assessment across diverse building types, climates, and regulatory environments.

In response to these challenges, the framework presented in Figure 6.5 has been developed as a synthesis of data and insights derived from the preceding investigations. The framework provides a structured pathway for systematically quantifying, analysing, and mitigating the EPG through a holistic structure composed of interrelated analytical components that collectively support the quantification, interpretation, and mitigation of the EPG. These components can be applied flexibly and iteratively depending on data availability, modelling approach, and assessment objectives. It integrates lessons from the second segment of the SLR (Section 3.3), Chapter Five investigation of the ambiguity surrounding the actual magnitude of the EPG and the present chapter's exploration of a coherent methodology for EPG quantification and reporting. The framework comprises three key analytical components:

- **Component One: Inter-tool gap investigation**

The process commences with an inter-tool gap investigation, positioned at the far left of the framework. This stage seeks to identify the simulation tools and scenarios that constantly produce the most accurate and representative models of actual building conditions. By comparing outcomes of tools different simulation tools, this investigation evaluates compliance, performance, and calibrated model's outputs under identical input data. The objective is to minimise tool-related discrepancies and to establish a simulation platform that reliably reflects empirical building behaviour. This step serves as the foundation upon which subsequent analyses are constructed, ensuring that the ensuing evaluation of energy performance is grounded in the most technically robust model available.

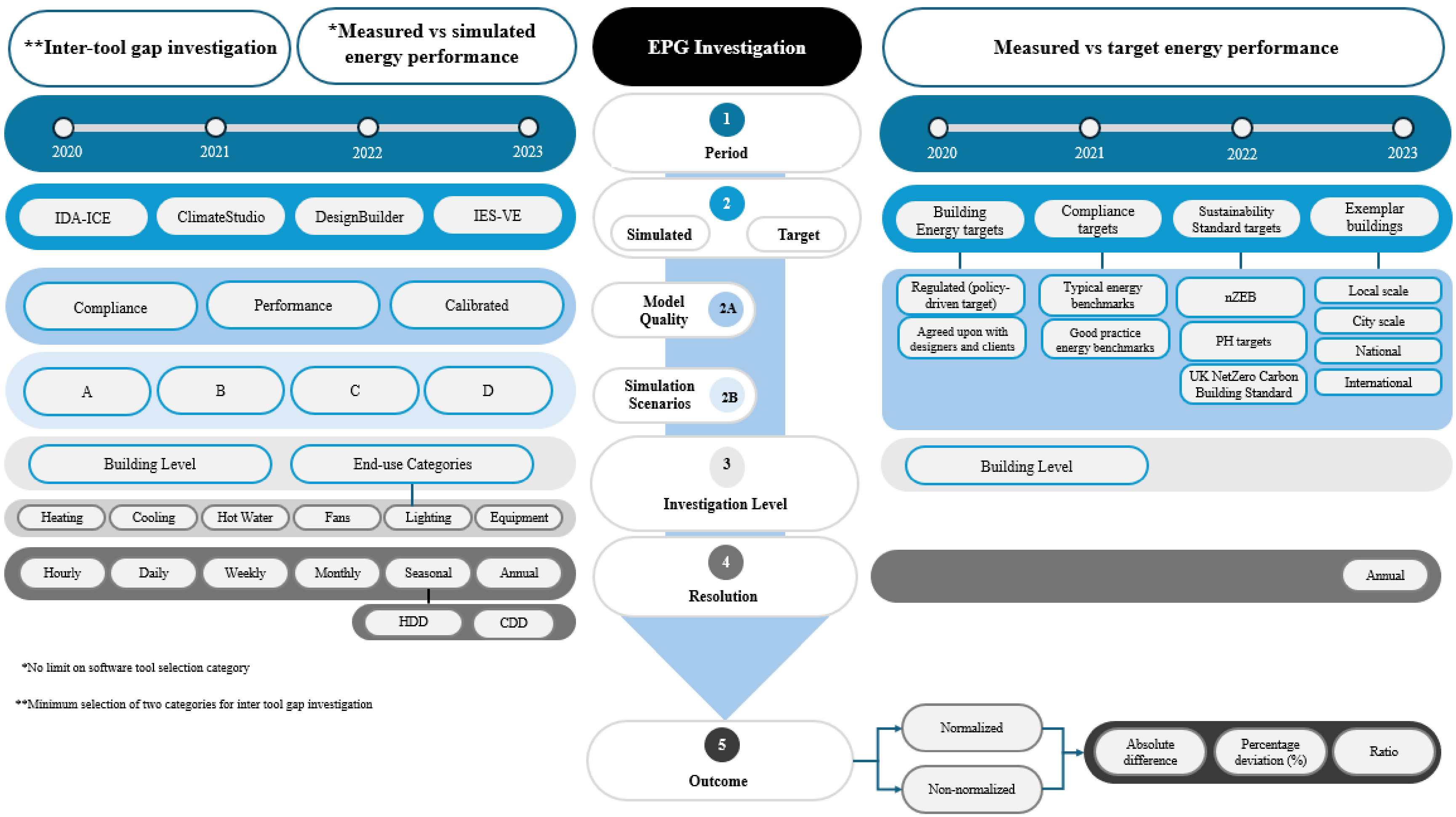


Figure 6.5: Investigation, quantification and reporting framework of the EPG

- **Component Two: Measured versus simulated energy performance**

This component involves a detailed comparison between measured and simulated energy performance after identification of most accurate and representative simulation model to operational conditions. This stage tests the validity of the selected model under real-world operating conditions. By contrasting monitored energy data with predicted outcomes, this analysis quantifies the divergence between modelled expectations and actual building performance. It thereby allows for the calibration and refinement of simulation assumptions, identifying the extent to which the EPG arises from modelling limitations versus operational or behavioural factors.

- **Component Three: Measured versus target energy performance**

The investigation under this component evaluates measured versus target energy performance. This step aligns empirical building data with relevant energy performance targets, which may include regulatory compliance standards, sustainability certification benchmarks (e.g., *nZEB*, *PH*, or *UK Net Zero Carbon Building Standard*), and more importantly bespoke client-defined objectives. By integrating both top-down policy requirements and bottom-up project constraints, this phase ensures that the EPG analysis responds appropriately to building, local, national, and international contexts. This enables the framework to support comparative assessments across multiple scales ranging from single-building evaluations to sector-wide policy benchmarking.

The magnitude of the EPG within the proposed framework is evaluated using three principal indicators with each indicator presenting two distinct values, one calculated before normalisation based on total energy use and another after normalisation based on energy intensity per unit of floor area as follows:

1. Energy Difference (kWh and kWh·m^{-2}·a^{-1})

This indicator shows the discrepancy between measured and simulated or target energy performance as absolute difference. It presents one value representing the total energy difference in kilowatt hours and another value representing the normalised difference in kilowatt hours per square metre. This dual representation enables comparison between buildings of different sizes and supports consistent evaluation across projects.

2. Percentage Deviation (%)

This metric expresses the relative magnitude of the EPG as a proportion of the reference energy value. It shows two values, one derived from total energy use and another derived from normalised energy intensity, allowing the analysis to reflect how the deviation changes before and after normalisation.

3. Performance Ratio

The EPG ratio indicates the proportional relationship between measured and simulated or target energy performance. It also shows two values, one based on total energy use and another based on normalised energy, thereby reflecting both the whole building and area adjusted perspectives of performance accuracy.

The six indicators provide a comprehensive understanding of the EPG by illustrating how the EPG manifests at both absolute and normalised scales. This dual representation ensures that the framework captures the full range of variability in energy performance across different building sizes, typologies, and assessment contexts. Furthermore, these quantitative indicators provide the analytical basis for interpreting the scale of performance discrepancies and for prioritising potential interventions. Critically, this stage introduces a feedback mechanism, in which insights from measured-versus-target comparisons are reintegrated into the simulation environment. The recalibrated model is then used to test alternative energy-saving measures or operational adjustments under various simulation scenarios. This iterative process allows for the identification of cost-effective and technically feasible strategies that not only reduce the measured-versus-simulated gap but also ensure continued compliance with regulatory or client-specific requirements.

By applying the same analytical process across multiple temporal datasets, it becomes possible to monitor whether the EPG is widening or narrowing over time. This longitudinal monitoring function provides essential insight into the effectiveness of implemented measures, enabling researchers and practitioners to distinguish between short-term improvements and sustained performance gains. Moreover, it allows identification of the simulation scenarios and operational strategies that consistently demonstrate the most positive contribution to mitigating the EPG.

In essence, the framework functions not merely as an analytical structure but as a dynamic decision-support mechanism. It operationalises the iterative relationship between simulation outcomes, measured performance, and a variety of performance targets, offering a replicable yet adaptable approach to energy performance evaluation. By bridging the gap between simulation outcomes, monitored data, and performance targets, the cyclical feedback structure promotes continuous improvement, transforming energy performance evaluation from a static assessment into an ongoing process of optimisation and refinement.

7. Synthesis of Knowledge

This chapter integrates the insights derived from the preceding investigations, providing a synthesis of knowledge that unites the multi-strand contributions of this thesis. The research has critically examined the EPG as a complex, multi-dimensional, and socio-technical phenomenon, exploring its underlying causes, magnitude, mitigation strategies, and proposing a pathway for standardisation in its quantification and reporting. The aim of this synthesis is to consolidate the evidence gathered across the thesis to generate an integrated understanding of the EPG that advances both theoretical and practical knowledge.

7.1 Integration of Key Findings

The SLR established a comprehensive empirical foundation by mapping both the root causes and the magnitude of the EPG across multiple contexts. It revealed that the EPG cannot be attributed to isolated technical deficiencies but emerges instead from interacting causal networks that span the design, construction, commissioning, and operational stages of the building lifecycle. The SLR identified patterns of failure such as miscommunication of performance targets, inadequate detailing, and poor integration of design intent that recur systematically in practice. It also demonstrated that the EPG is reinforced by *cross-cutting institutional factors* including weak accountability, poor collaboration, and insufficient client knowledge.

Equally significant, the SLR exposed *wide* inconsistency in the way the magnitude of the EPG is calculated and reported. Discrepancies arise from divergent baseline assumptions, differences in modelling practice, and variations in measurement methodologies, leading to global figures that range from marginal to extreme deviations between predicted and actual performance and inconsistences hindering the ability to draw meaningful correlations and conclusions on the seriousness of the phenomenon on a wide scale. This heterogeneity underscores the difficulty of comparing studies and the absence of a unified reporting standard. The SLR therefore established two fundamental premises that underpin the remainder of the thesis: (1) the EPG is inherently complex and multi-causal; and (2) progress in its mitigation requires methodological standardisation and policy realignment toward measured outcomes.

Building on this foundation, Chapter Four explored the potential for mitigating the EPG through approaches that emphasize measured performance outcomes rather than predicted values, corresponding to the second RO of the thesis. Specifically, the chapter focused on NABERS and its role as an outcome-based framework. The chapter combined RCA with semi-structured interviews with NABERS IDRs and AAs to assess how NABERS addresses the underlying causes of the EPG across the building lifecycle.

Across all stages and identified root causes, participants made 295 individual assessments of NABERS' potential to address identified root causes. When combined, the findings demonstrated that NABERS has significant potential to mitigate key contributors to the EPG, particularly design stage root causes and cross-cutting factors. Quantitative coding of 295 individual assessments from participants revealed that 52% of judgments were green, indicating that NABERS strongly addresses specific root causes; 15% were blue, highlighting areas where additional policy intervention is required; and 31% were uncategorized, often reflecting uncertainty in assigning responsibility, particularly in commissioning and operational stages.

This overall pattern reveals three main findings. First, the large share of green responses shows that participants generally view NABERS as already addressing or having strong potential to address many of the identified root causes. Second, the blue responses, which account for just over 15%, indicate a smaller but recurring set of areas where participants believe that additional policy or procedural action is needed. Third, the relatively high proportion of uncategorized causes, which were mostly found in the commissioning and operational stages, suggests that participants often elaborated without assigning a clear category. This may reflect hesitation to attribute responsibility directly to NABERS. The analysis highlighted stage-specific strengths and limitations:

- **Design stage:** Green responses dominated (64%), showing NABERS' strong potential in addressing design stage causes, with blue responses (15%) related to analytical and modelling challenges such as stochastic processes and software limitations. Very few responses in this stage were grey or unclassified.
- **Commissioning and operational stages:** This stage showed more hesitation with a higher proportion of uncategorized responses accounting for 61%. However, participants acknowledged that NABERS covers installation-related issues but still identified gaps in skills,

accountability, and monitoring suggesting uncertainty or reliance on complementary mechanisms to achieve outcomes and only 22% of responded coded green. On the other hand, 15% of blue responses indicated gaps in skills, accountability, and monitoring.

- **Cross-cutting factors:** The cross-cutting factors showed the strongest confidence in NABERS, with 76% of responses coded green, 16% blue, and only a small proportion uncategorized or grey. In this category, participants viewed NABERS as particularly effective in improving client knowledge, addressing cultural influences, and managing overly optimistic performance predictions, although issues of communication and accountability remained areas requiring further policy attention.

Furthermore, the findings revealed that the strength of NABERS lies in its emphasis on measured energy performance and transparent reporting. However, the research also identified critical limitations. While NABERS introduces mechanisms such as commitment agreements and independent design reviews, its effectiveness is contingent on stakeholder understanding, data integrity, and integration with procurement processes. The analysis suggested that outcome-based policies can narrow the EPG only when embedded within a broader culture of performance accountability and reinforced by regulatory requirements for transparency in operational data.

Chapter Four provided empirical evidence that policy interventions can reshape industry practices from prediction-based compliance toward performance-based, thereby aligning regulatory intent with real-world outcomes. It also demonstrated that policies alone are insufficient unless supported by institutional learning and feedback mechanisms. Whilst Chapter Four did not directly evaluate alternative frameworks such as PH or SL, although both have been referenced for their potential to mitigate the EPG. PH is particularly relevant to addressing construction-stage root causes, while SL targets operational-stage causes. These frameworks can complement NABERS, particularly in stages where NABERS alone is less influential. The SLR (Section 3.3) also identified studies quantifying EPG outcomes in buildings following PH and SL approaches

Chapter Five examined uncertainty in the prediction of building energy performance through a comparative modelling exercise involving four simulation tools: IES VE, DesignBuilder, ClimateStudio, and IDA ICE, applied to the Energy and Building Design Laboratory at Lund University and compared against available measured electricity data. The results showed substantial variability between the tools, with differences emerging even when the main input assumptions were harmonized. When compared with measured data, all models showed some level of inconsistency, reinforcing the argument that the prediction of absolute building energy performance involves an inherent degree of uncertainty.

However, this finding should not be interpreted as diminishing the value of white-box, physics-based building energy modelling. A traditional and important use of such models is not necessarily to predict absolute operational energy consumption with complete accuracy, but to compare the relative performance of alternative design or operational scenarios under controlled assumptions. For example, a model may be used to estimate whether additional insulation, an improved glazing specification, or an alternative control strategy is likely to reduce energy use relative to a baseline scenario, rather than to claim that the exact future metered consumption of the building can be predicted. In this sense, the value of simulation lies partly in its capacity to support comparative reasoning, sensitivity analysis and decision-making, even where absolute predictions remain uncertain.

The findings of Chapter Five therefore support arguments in the literature that complete closure of the EPG may be conceptually difficult or unattainable (Patel and Green 2020, Mahdavi and Berger 2024). Variability arises not only from modelling assumptions and tool-specific calculation procedures but also from stochastic human behaviour, contextual factors such as microclimate, and changing operational conditions. The results suggest that the EPG should not be understood as an error to be eliminated through calibration. Rather, it should be examined systematically to identify the factors that influence building energy use and to support mitigation strategies that are appropriate to the type of gap being addressed, as well as to the practical and financial capacity to intervene.

This component of the research therefore contributes directly to RO3 by demonstrating that the reported magnitude of the EPG is sensitive to modelling choices, tool selection and the basis of comparison. It also clarifies that uncertainty in absolute prediction does not remove the usefulness of simulation for comparative assessment. Instead, it highlights the need to report clearly whether simulation results are being used as absolute predictions, relative comparisons, calibration exercises, or exploratory scenario tests. This distinction provides empirical justification for the standardised reporting framework proposed in Chapter Six, where transparency regarding modelling purpose, assumptions, uncertainty and comparison basis becomes central to interpreting the reported magnitude of the EPG.

If simulation results are interpreted without clarifying their purpose, assumptions, uncertainty and comparison basis, the reported magnitude of the EPG may be misleading. Greater transparency in these areas is therefore necessary to improve the interpretation, comparability and usefulness of EPG reporting. This is particularly important if the field is to develop a more reliable understanding of the true trend of the EPG over time. Despite extensive research efforts aimed at mitigating the influence of the EPG, inconsistent definitions, modelling purposes, assumptions and reporting practices can make it difficult to determine whether reported changes in EPG magnitude reflect genuine performance improvement or are instead a consequence of methodological variation. Clearer reporting is therefore essential not only for interpreting individual case studies, but also for assessing whether mitigation efforts are producing meaningful reductions in the EPG at a broader scale.

Chapter Six responded directly to the inconsistencies revealed in Chapters Three and Five by developing and validating a framework for standardised quantification and reporting of the EPG. Drawing on insights from literature, simulation evidence, and expert consultation, the framework established a structured methodology for defining baselines, reporting parameters, and distinguishing between the EPG as a discrepancy between measured and simulated energy performance and a more meaningful discrepancy between measured and target energy performance.

Through expert interviews and panel validation, the framework was refined to ensure both academic robustness and practical applicability. Experts emphasised that consistency in reporting must coexist with flexibility to account for contextual variation. This outcome marked a transition from diagnostic to constructive contribution. The research moved beyond identifying problems to proposing a transferable mechanism that could improve comparability across studies, enhance policy evaluation, and support accountability in practice. The framework stands as a key contribution to knowledge, bridging the gap between empirical evidence and institutional implementation.

7.2 Cross-Cutting Themes

A unifying theme across all investigations is the recognition of the EPG as a phenomenon rather than a purely technical error. The convergence of design, construction, operation, and policy findings illustrates that EPGs emerge from the interaction of human behaviour, institutional processes, and technological limitations. Design miscommunication, contractor substitution, operational mismanagement, and modelling assumptions all represent nodes within a broader socio-technical system. This framing challenges reductionist narratives that attempt to locate the EPG within single disciplines. Instead, it positions the phenomenon within the domain of complex adaptive systems, where interdependencies and feedback loops produce non-linear outcomes. Recognising the EPG as socio-technical implies that effective mitigation requires interventions that integrate technical precision, human competence, and institutional alignment.

Another recurrent theme is the persistence of uncertainty. Despite advances in modelling tools, measurement technologies, and policy mechanisms, uncertainty pervades every stage of the building lifecycle. The comparative simulation analysis revealed that even under identical assumptions, predicted outcomes vary substantially. Similarly, the SLR showed that empirical studies report widely divergent magnitudes of the EPG under comparable conditions. This persistence underscores that uncertainty is not an anomaly but an intrinsic feature of the system. Consequently, the pursuit of zero performance gap is both unrealistic and epistemologically unsound. Instead, uncertainty must be explicitly acknowledged, quantified, and managed. The proposed standardisation framework contributes to this by offering a means of transparent uncertainty communication, allowing stakeholders to interpret EPG values within credible confidence bounds.

The research consistently found that *lack of accountability* and *absence of feedback mechanisms* amplify the EPG. Whether in design misalignment, construction quality, or post-occupancy operation, failures were traced to weak responsibility structures and fragmented communication. Investigation of NABERS' potential to mitigate the EPG demonstrated that formalising accountability through measured performance and public disclosure can drive behavioural change and institutional learning. Feedback loops are equally essential. Without systematic post-occupancy evaluation, the sector remains trapped in cycles of error replication. Integrating continuous feedback into design and operation through mechanisms such as the SL and outcome-based certification can progressively narrow the EPG by transforming isolated experiences into shared knowledge.

The absence of standardised definitions and transparent reporting emerged as a barrier to progress. The framework developed in Chapter 6 directly addresses this deficiency, proposing uniform metrics and explicit documentation of boundary conditions. Standardisation thus becomes not merely a methodological improvement but a precondition for collective learning within the field.

7.3 Theoretical and Methodological Implications

The synthesis reveals that the EPG should be reconceptualized not as a discrete measurement discrepancy but as a manifestation of systemic complexity. This reconceptualization aligns with contemporary complexity theory, which views socio-technical systems as dynamic, adaptive, and path-dependent. The thesis extends this theoretical stance to the domain of building performance, demonstrating that deviations between predicted and measured outcomes are emergent properties of interrelated processes rather than isolated errors. This theoretical shift has significant implications. It challenges the binary framing of success or failure embedded in compliance-based approaches and promotes an epistemology of probabilistic understanding and continuous calibration. Recognising the EPG as emergent rather than exceptional reframes mitigation from error correction to system learning.

Methodologically, the thesis highlights the importance of integrative mixed methods research in exploring complex building performance issues. The combination of SLR, qualitative interviews, simulation analysis, and expert validation offered both comprehensive scope and detailed understanding. Each method shed light on distinct aspects of the EPG, including causal complexity, stakeholder perceptions, empirical variability, and the potential for standardisation, ultimately converging into a coherent analytical framework. The empirical evidence from simulation modelling and the theoretical framing of complexity converge toward an epistemology of managed uncertainty. Rather than pursuing deterministic prediction, this approach advocates transparent acknowledgement of uncertainty ranges and their implications for design and policy.

Standardisation, in this context, becomes a mechanism not to eliminate uncertainty but to structure its representation in a way that supports informed decision-making.

7.4 Policy and Practice Implications

A central implication of the research is the need for regulatory paradigms to evolve from prediction-based compliance to measured-performance governance. As shown in Chapter 4, outcome-based schemes such as NABERS introduce accountability through operational measurement yet remain supplementary rather than integral to national building codes. Embedding such frameworks within regulation could bridge the gap between design intent and operational reality. This transition would require harmonisation between compliance modelling and in-use verification, potentially through dual-stage certification processes that assess both design quality and operational outcome. Transparent disclosure of measured data, supported by legal or contractual obligations, would enhance trust, stimulate competition based on performance, and reinforce a culture of continuous improvement.

Institutional learning emerges as a prerequisite for sustainable performance improvement. The research emphasises that data collection, monitoring, and feedback must extend beyond individual projects to inform sector-wide learning. Mandating POE and integrating its outcomes into future design briefs can transform isolated experiences into structured knowledge. Furthermore, the adoption of SL when combined with performance contracts and outcome-based ratings can institutionalise feedback processes, ensuring that lessons from operation directly inform design and policy. This would gradually reduce the recurrence of known performance shortfalls.

Across the building lifecycle, professional competence and collaboration were found to be decisive factors. The research revealed that even robust frameworks cannot compensate for inadequate skills or poor communication among stakeholders. Therefore, policy initiatives must be complemented by capacity-building programmes, interdisciplinary education, and collaborative procurement models that align incentives with performance outcomes. Encouraging integrated design and delivery teams, supported by digital tools for information sharing, can mitigate the fragmentation that currently amplifies the EPG. In parallel, clear delineation of roles and responsibilities throughout the process enhances accountability and reduces the diffusion of responsibility observed in conventional practice.

The proposed standardisation framework provides a foundation for transparent and comparable reporting of EPG data. Its adoption could enable benchmarking across regions and building types, facilitating empirical evaluation of policy impacts and technological interventions. Integration of this framework into regulatory and voluntary schemes would establish a consistent evidence base for decision-making. Moreover, the requirement for public disclosure of performance data already embedded in schemes such as NABERS should be extended to broader contexts. Data transparency not only incentivises improvement but also fosters public trust in sustainability claims, addressing the "credibility gap" identified early in the literature.

Through its multi-method, multi-scale investigation, the thesis advances understanding of the EPG as a manifestation of systemic complexity and provides both theoretical and practical pathways for its mitigation. The synthesis underscores that mitigating the EPG is not a matter of calibration alone but of cultural, institutional, and epistemological transformation.

A considerable portion of the research was devoted to identifying a genuine Digital Twin case study, involving prolonged engagement with practitioners in Scandinavia, leading simulation experts in the UK, and groups in Australia who claim to have developed operational Digital Twins. Despite this extensive effort, it became evident that what is labelled a Digital Twin in current practice is, in most cases, little more than an uncalibrated or loosely calibrated simulation model. This reflects a widespread misinterpretation of the core principles that define a true Digital Twin, including continuous data integration, dynamic responsiveness, and rigorous and ongoing validation. Consequently, the idea that Digital Twins presently offer a practical route for addressing the EPG resembles more of a mirage than a viable industry pathway. This has important implications for policy and practice: without clearer definitions, stronger standards, and a more disciplined application of DT methodologies, the anticipated benefits of digitalisation in building performance assessment will remain largely unattainable.

8. Conclusion

This chapter describes the key conclusions, lessons and limitations of the thesis with a link to context, aim and objectives. The aim of the thesis was to critically investigate the EPG in buildings, conceptualizing it as a complex and multifaceted phenomenon. It seeks to enhance understanding of its magnitude, assess possibilities for its mitigation, and to advance the development of standardized approaches for its quantification and reporting. Rather than viewing the EPG solely as a discrepancy between simulated and measured energy performance, the thesis framed it as a phenomenon shaped by technical, behavioural, regulatory, and contextual factors. It further considered manifestations of EPGs that are of relevance to different stakeholder groups, with the objective of deepening understanding of the phenomenon, its severity, and context-specific mitigation strategies aligned with the type of EPG which must be mitigated. To achieve this aim, the thesis pursued the following ROs:

- RO1: Demonstrate the complexity of the energy performance gap by exploring its root causes and magnitude.
- RO2: Explore the potential for mitigating the energy performance gap through approaches that emphasize measured performance outcomes rather than predicted values.
- RO3: Deepen the understanding of the energy performance gap by clarifying the ambiguity surrounding its magnitude.
- RO4: Assess the potential for standardization in the quantification and reporting of the energy performance gap, identifying pathways toward greater consistency and comparability.
- RO5: Develop recommendations for more effective mitigation of the energy performance gap and provide guidance for advancing standardization efforts.

Chapter Eight: Conclusion

8.1 Contribution to Knowledge

The thesis extends the body of knowledge by critically interrogating how the EPG is understood, quantified, and reported, while also assessing the potential of outcome-based frameworks. Furthermore, instead of focusing on a limited subset of root causes, the thesis provides the first comprehensive exploration of potential solutions, and preventative actions addressing all root causes that contribute to the EPG across the design, commissioning, operational stages including cross-cutting factors. By adopting a holistic, stage-based perspective, this research identifies interdependencies and feedback loops that previous research have largely overlooked. The key contributions of the thesis are:

1. **Largest and most comprehensive cross-context synthesis of the magnitude of the EPG:**

The thesis provides the largest systematic syntheses of reported EPG magnitudes across multiple geographical locations and quantification scenarios under Section 3.3. By bringing together evidence from a wide range of contexts, the thesis demonstrates how the EPG widens, narrows, or varies under different conditions. This approach provides new insights into why the EPG cannot be understood in a uniform way, emphasizing the need for a standardized approach for its quantification and reporting. Furthermore, while the focus is on quantifying the magnitude of the EPG, Section 3.3 indirectly identifies pathways through which the EPG could be mitigated by exploring different EPG magnitudes with different procurement methods, ventilation strategies applied or SL as in the case of Thompson, Burman et al. (2022) under Section 3.3.1.

While there is published research that focus on the magnitude of the EPG with measured and predicted energy data for buildings that exceed the sample size reviewed under Section 3.3, such studies differ in their approach and coverage of geographical locations. Such studies as in the case of Firth, Allinson and Watson (2024), for example, whilst being informative sources of demonstrating the size of the EPG at a city or local cale, they are different in their scope and approach providing evidence grounded in regulatory and static perspectives.

2. First systematic assessment of NABERS in the UK context

The thesis offers the first independent academic assessment of NABERS since its introduction in the UK by firstly identifying all potential causes that contribute to the emergence of the EPG under Section 3.2 of the thesis then mapping all identified root causes based on the stage in which they emerge including cross-cutting factors under Section 4.2. Under section 4.3, all identified root causes in design, commissioning, operational and cross-cutting factors where thoroughly discussed with participants who were directly involved in delivering NABERS projects providing evidence-based policy interventions and mitigation strategies that extended beyond the scope of NABERS as summarized under Section 4.4 of the thesis. While NABERS has been recognized as a successful outcome-based framework in Australia, its applicability and effectiveness in mitigating the EPG in the UK remained underexplored. By evaluating NABERS within the UK context, the thesis provided applied insights into the transferability of the scheme, informing both academic debate and policy discourse on outcome-based frameworks and their potential to narrow the EPG.

3. Advancement of standardization in EPG quantification and reporting

A further contribution lies in pioneering the discussion of standardized approaches to quantifying and reporting the EPG. Chapter five and Section 3.3 of the thesis demonstrate how inconsistencies in definitions, different prediction methods and simulation scenarios undermine comparability across studies and limit the reliability of aggregated findings. It proposes pathways toward greater consistency in reporting the EPG, laying the groundwork for more transparent, comparable, and actionable knowledge. This has implications for improving both academic research quality and industry practice, particularly in the design of policies and performance-based frameworks. The value of such an effort lies in addressing knowledge gaps concerning the notion of the EPG wherein its discourse was acknowledged to be meaningful if the prediction of buildings' energy performance is explicitly assumed to constitute the prediction of future energy use not for compliance purposes. In such a case, EPG figures observed through the second segment of the SLR (Section 3.3) and Chapter five established and justified the case for standardization in Chapter 6 of the thesis.

8.2 Key Lessons

By thoroughly investigating the causes, magnitude, mitigation strategies through policy disclosure and the potential for standardizing the EPG, different key lessons can be drawn as listed below which also serve as recommendations for future investigations of the phenomenon:

1. Frameworks such as NABERS provide valuable insights into the potential of outcome-based certifications, as demonstrated through semi-structured interviews with IDRs and AAs in Section 4.3. However, the findings also indicate that such frameworks are inherently limited int their ability to capture all determinants of building performance, particularly those related to occupant behaviour and real-world variability. Consequently, energy performance cannot be modelled or predicted with absolute precision, and outcome-based approaches should be understood as mechanisms for improving alignments with expected operational performance rather than eliminating discrepancies entirely.
2. Research should clearly differentiate and categorize what is currently perceived and known to be the causes that contribute to the EPG (based on findings from the SLR in Sections 3.2 and 3.3, and further supported by emergent themes from expert panel discussions in Section 6.3, including the distinction between anticipated and unanticipated EPGs) into the following:

- Factors that contribute to discrepancies between measured and simulated energy performance: Such factors are the inherent and unavoidable outcomes of simplifications and assumptions in building energy simulation models.
- Factors the influence energy use in buildings: Such factors once identified and characterized based on the stage in which they emerge, could be considered as the actual causes that contribute to the emergence of the EPG in buildings. Once such factors are identified and solved, any residual EPGs would normally be the result of inaccurate simulation models or changes in operational requirements. Once this distinction is established, research efforts could be more specific in their focus whether the aim is to narrow the EPG resulting from discrepancies between measured and simulated energy performance to improve modelling practices or identifying deficiencies that influence energy use in buildings which hinder achievement of various energy efficiency or decarbonization targets.

3. Evidence synthesised in the SLR on EPG magnitude (Section 3.3) shows that reported values are highly dependent on methodological choices, including baseline definitions and calculation

approaches. As a result, previously reported EPG magnitudes cannot be interpreted as universally representative indicators, but rather as context-specific observations influenced by modelling assumptions and reporting practices. However, such investigations, if detailed enough, could aid in the following:

- Informing the potential of a mitigation strategy in case of availability of measured energy performance before and after a strategy is applied and validated across different building types.
- Exploring the influence of using different types of energy systems and adoption of various energy saving measures on narrowing the EPG down.

4. While NABERS demonstrates potential to mitigate aspects of the EPG, findings from Sections 4.3.1–4.3.3 indicate that its effectiveness is contingent upon broader industry conditions, including stakeholder competence, operational practices, and long-term engagement with performance evaluation. This suggests that outcome-based frameworks alone are insufficient without corresponding improvements in delivery processes and accountability structures.
5. Insights derived from interviews with AAs (Section 4.3.2) indicate that no single framework can comprehensively address the EPG due to the inherent complexity of the construction and operation processes. The multiplicity of actors, procurement routes, and contextual variables limits the extent to which any individual framework can fully resolve performance discrepancies. As evidenced in Section 4.3.2, the EPG cannot be eliminated in real-world contexts. Occupant behaviour introduces variability that cannot be fully predicted or controlled within modelling processes, reinforcing the view that some degree of performance discrepancy should be considered inherent to building operation.
6. Findings from Section 4.3.2 further demonstrate that it is not feasible to capture all energy performance-related variables within a single modelling or evaluation framework. This limitation contributes to persistent discrepancies between predicted and measured performance and highlights the need for cautious interpretation of modelling outputs.
7. The investigation of cross-cutting factors in Section 4.3.3, particularly those related to over-optimistic predictions, indicates that modelling accuracy achieved in controlled academic environments does not directly translate to industry practice. Constraints related to time, cost, and data availability necessitate simplifications that introduce additional uncertainty in real-world applications.

8. Ratings should not be heavily influenced by external variables that are beyond the control of designers or operators. This concern emerged from insights provided by AAs, as discussed in Section 4.3.2, and is grounded in the qualitative evidence collected through the semi-structured interview methodology outlined in Section 4.2. However, as reflected in both practitioner perspectives and expert panel discussions in Section 6.3, these mechanisms are not always transparent to users and introduce additional layers of complexity. The findings suggest that while such adjustments are necessary to account for external influences, their implementation raises questions regarding interpretability and practical usability. In this context, the pursuit of absolute precision may not always be proportionate to its benefits, as increasing methodological complexity can lead to greater time and cost demands without necessarily improving decision-making outcomes. Instead, a balance is required between accuracy and usability, ensuring that rating systems remain both technically robust and practically applicable in real-world contexts.
9. By integrating insights from Sections 4.3.1–4.3.3 with expert panel findings (Section 6.3), the EPG cannot be more accurately understood as a failure of a specific stage in a project's lifecycle, such as attributing it solely to a failure of design-stage predictions prediction. This reinforces the need for lifecycle-based accountability rather than attribution of discrepancies to isolated stages.
10. Insights from IDRs (Section 4.3.1) indicate that improvements in modelling accuracy are closely linked to the availability of high-quality operational data. However, even with improved data availability, external variables and system uncertainties ensure that predictive accuracy will remain inherently limited. As more operational data becomes available, the accuracy of modelling is expected to improve. Current uncertainties remain in estimating the energy use of systems such as servers, lifts, and escalators, which have historically been poorly represented in simulation tools. Once sufficient data is collected, it can inform better modelling assumptions, leading to a gradual reduction of the EPG. Systems such as heating and cooling are modelled with greater confidence due to long-term experience and established methodologies, which differs from year to year and contributes to modelling uncertainty.

8.3 Research Limitations

The research recognises a number of methodological and scope-related limitations that have shaped its findings. These limitations suggest several opportunities for future research to build upon and refine the work presented in this thesis:

- While the causes of the EPG span all stages of the construction process, interviews were not conducted with stakeholders directly involved in the construction stage of NABERS-certified projects. This limits the extent to which first-hand implementation experiences were captured. This limits the findings by restricting insight into how construction-stage practices differ between NABERS and conventional projects, particularly in assessing which root causes are effectively addressed during implementation. Also, how the experience of construction managers or quality assurance engineers varied between conventional and NABERS projects.
- Commissioning and operational stage–root causes were examined collectively rather than separately; a more refined approach would have involved engaging commissioning and value engineers for the commissioning stage, and facilities managers alongside NABERS-accredited assessors for the operational stage. This aggregation was necessary due to the availability of participants but limits the precision of the findings, as accredited assessors are not directly involved in commissioning processes. As a result, conclusions regarding the effectiveness of NABERS in addressing commissioning-related causes may lack the depth and accuracy that could have been achieved through direct engagement with commissioning specialists.
- Because NABERS has only recently been introduced in the UK and few certified case studies currently exist, the research would have benefited from an in-depth panel discussion with stakeholders who have participated in delivering NABERS-certified projects to explore how their experiences differ from conventional projects. Such a forum could have facilitated consensus-building on cross-cutting factors such as lack of accountability and cultural change within the industry. Its absence limits the ability to validate findings collectively and to strengthen recommendations, particularly those presented in Section 4.4.
- Direct observation through site visits or engagement with occupants in certified offices would also have provided valuable practical insights into the lived experience and post-occupancy performance of NABERS-rated buildings. This limitation restricts the ability to capture user-centred perspectives, particularly in relation to occupant behaviour and comfort. Inclusion of

occupant feedback could have enabled exploration of subjective or perception-based performance gaps, which fall outside the scope of this research but represent an important complementary dimension presented under Figure 1.1 and acknowledged for its importance in Chapter Six by EP1.

- The simulation-based investigation was constrained by the absence of sub-metering data for the case study building. Consequently, the simulation models could not be calibrated, and the analysis was limited to compliance-level models. Furthermore, the assessment of the EPG focused on annual and monthly inter-tool and measured-versus-simulated gaps, which may not fully represent the true magnitude of the EPG. This limitation may also have influenced the variation in results obtained across the four simulation tools, which, if explored further, could have strengthened the case for standardization discussed in Chapter Six.
- Due to time constraints, several recommendations arising from the expert consultations were not explored in depth. Specifically, the potential to expand NABERS beyond the office sector and to undertake a detailed evaluation of the UK NZCBS could not be pursued further. Greater emphasis could also have been placed on refining the standardisation framework developed in Chapter Six and expanding the number of participants to achieve stronger consensus and validation. This reflects a scope- and time-related limitation, as these investigations were undertaken in the final stages of the research, restricting the extent to which they could be explored in depth. Further examination would likely have relied primarily on secondary data, which may have limited the richness and contextual depth of the findings, although it could have strengthened the applicability of the results across a wider range of building types and policy contexts.
- Although interviews and expert discussions highlighted the strong potential of the PH standard to address construction stage root causes, the thesis did not examine PH as a complementary standard to NABERS. This was a deliberate scope decision to maintain focus on NABERS; however, it limits the ability to assess how multiple standards could interact to mitigate the EPG.

8.4 Future Research

Building on the foundations established in this thesis, several opportunities exist for future work to advance understanding of the EPG and the role of performance-based rating systems:

- Engage construction-phase practitioners: Interviewing contractors, site managers, and construction engineers on NABERS-certified projects would provide critical insights into construction-stage root causes and practical implementation challenges.
- Deepen commissioning versus operational analyses: Research that separately investigates commissioning-stage specialists, such as commissioning engineers and value engineers, and operational stakeholders, such as facilities managers and NABERS assessors, could yield more nuanced root-cause mapping.
- Conduct stakeholder panels on emerging UK NABERS practice: As more NABERS accredited case studies are completed, future studies could convene multi-stakeholder panels to examine how experiences differ from conventional projects and what lessons are emerging.
- Include site visits and occupant studies: Observational research and post-occupancy engagement would offer richer insights into the lived experience and performance outcomes of NABERS-rated offices.
- Develop and test calibrated simulation models: Access to detailed sub-metering data would enable calibrated modelling approaches that more accurately quantify the EPG and test the robustness of compliance-based models.
- Refine and validate standardisation frameworks: Future work could focus on strengthening the frameworks proposed in chapter six through iterative development and wider expert consultation.
- Explore broader applications of NABERS and related frameworks: Investigating the potential extension of NABERS into sectors beyond offices, as well as conducting thorough assessments of the UK NZCBS, would help evaluate their alignment and practical compatibility.
- Examine multi-standard synergies, such as between NABERS and PH: Given the potential of the PH standard to address construction-stage issues, future research could explore how PH and NABERS interact or complement one another, including through stakeholder engagement across both domains.

Chapter Eight: Conclusion

At its core, the EPG phenomenon arises from a comparison between two imperfect but essential aspects of building science: the measurement of buildings and the modelling of their expected performance. This comparison is extremely challenging. The measurement of buildings is inherently difficult because it must be carried out in situ and involves large, complex, and continuously changing environments. Modelling and simulation, on the other hand, are based on abstractions that simplify reality in order to make it analysable and predictable. Because of these differences in nature and purpose, some level of divergence between measured and modelled outcomes is not only unsurprising but practically unavoidable. For this reason, the presence of an EPG should not be interpreted as a failure of either measurements or modelling, but rather as an indication of the limitations and assumptions that shape both sides of the comparison. One of the real concerns surrounding the EPG should therefore focus on the quality of building science efforts. This includes the fidelity, consistency, and robustness of measurement practices, as well as the transparency, standardisation, and methodological rigour of modelling and simulation approaches. By strengthening both measurement and prediction, uncertainties on each side are gradually reduced, allowing for a clearer understanding of the true magnitude of the EPG. As these uncertainties diminish, time and resources can be directed more effectively toward practical energy saving measures that contribute to mitigating various EPGs other than the discrepancy between measured and simulated energy performance.

References

A’zami, A. (2005). Badgir in traditional Iranian architecture. International Conference on Passive and Low Energy Cooling for the Built Environment, Santorini, Greece.

Ahmed, K., T. Hasu and J. Kurnitski (2022). "Actual energy performance and indoor climate in Finnish NZEB daycare and school buildings." JOURNAL OF BUILDING ENGINEERING **56**.

Ahn, K. U., D. W. Kim, Y. J. Kim, S. H. Yoon and C. S. Park (2016). "Issues to Be Solved for Energy Simulation of An Existing Office Building." SUSTAINABILITY **8**(4).

Al-Rifai, M. (2025). "Unlocking insights: evaluating 5Whys methodology for root cause analysis and employee problem-solving skills in 91 case studies." International Journal of Lean Six Sigma **16**(5): 1276-1304.

Alam, M. and M. R. Devjani (2021). "Analyzing energy consumption patterns of an educational building through data mining." Journal of Building Engineering **44**.

Allen, J., R. Cohen and G. Hammond (2020). "From Energy World to today: the legacy of Milton Keynes and the Energy Cost Index." Global Transitions **2**: 140-151.

Amasyali, K. and N. M. El-Gohary (2018). "A Review of Data-driven Building Energy Prediction Methods." Renewable and Sustainable Energy Reviews **81**: 1192-1205.

Anelkovi, A. S., M. Kljaji, D. Macura, V. Munan, I. Mujan, M. Tomi, e. Vlaovi and B. Stepanov (2021). "Building energy performance certificate—a relevant indicator of actual energy consumption and savings?" Energies **14**(12).

Arbulu, M., S. Perez-Bezos, A. Figueroa-Lopez and X. Oregi (2024). "Opportunities and Barriers of Calibrating Residential Building Performance Simulation Models Using Monitored and Survey-Based Occupant Behavioural Data: A Case Study in Northern Spain." Buildings **14**(7): 1911 (1919 pp.).

Arjunan, P., K. Poolla and C. Miller (2020). "EnergyStar++: Towards more accurate and explanatory building energy benchmarking." Applied Energy **276**: 115413.

ASHRAE (2020). ANSI/ASHRAE Addendum a to ANSI/ASHRAE Standard 140-2017: Standard Method of Test for the Evaluation of Building Energy Analysis Computer Programs. Atlanta, GA, American Society of Heating, Refrigerating and Air-Conditioning Engineers.

ASHRAE (2021). ANSI/ASHRAE Standard 90.1: Energy Standard for Buildings Except Low-Rise Residential Buildings. Atlanta, American Society of Heating, Refrigerating and Air-Conditioning Engineers.

Aste, N., R. S. Adhikari, M. Buzzetti, C. Del Pero, H. E. Huerto-Cardenas, F. Leonforte and A. Miglioli (2022). "nZEB: bridging the gap between design forecast and actual performance data." Energy and Built Environment **3**(1): 16-29.

Bahadori-Jahromi, A., R. Salem, A. Mylona, A. U. Hasan and H. X. Zhang (2022). "The Effect of Occupants' Behaviour on the Building Performance Gap: UK Residential Case Studies." SUSTAINABILITY **14**(3).

Bai, Y., W. Pan and C. Yu (2024). Impact of passive design measures on energy performance gap: A case study of high-rise residential buildings in Hong Kong. Energy Proceedings, Scanditale AB.

Bai, Y., C. Yu and W. Pan (2023). Quantifying Building Energy Performance Gap Using a Multi-indicator Framework. Environmental Science and Engineering.

Bai, Y., C. Yu and W. Pan (2024). "Systematic examination of energy performance gap in low-energy buildings." Renewable and Sustainable Energy Reviews **202**.

Baldwin, R., M. Cave and M. Lodge (2012). Understanding Regulation: Theory, Strategy, and Practice. Oxford, Oxford University Press.

Bannister, P. and C. Flores (2024). NABERS Nearly 25 Years on – Program Overview.

Bao, X. and H. Meng (2023). A Design Method to Diagnose Predicted and Measured Energy Performance Gap of Central Cooling Plant Based on Operation Mode. 2023 9th International Conference on Energy Materials and Environment Engineering (ICEMEE 2023), 8-10 June 2023, France, EDP Sciences.

Battista, A. and D. Torre (2023). "Mixed methods research designs." Medical Teacher **45**(6): 585-587.

Belpoliti, V. (2021). "Comparative analysis of simulated and monitored energy performance of a prototype house. Lesson learnt from the project KNOW HOWse for Solar Decathlon Middle East 2018 in Dubai." Journal of Building Engineering **41**.

Benedict, S. T., A. Regnier, C. Koenig and Z. Dehouche (2024). Whole-Building Calibrated Energy Performance Simulation towards Energy Sufficiency, Climate Resilience and Net Zero Carbon: Case Study of a Medium-Sized Hotel. ASHRAE Transactions, American Society of Heating Refrigerating and Air-Conditioning Engineers.

Bennett, G. A. and et al. (1977). CAL-ERDA Program Manual. Los Alamos, NM, Los Alamos Scientific Laboratory.

Beu, L. (2014). "So, who is accountable?" ASHRAE Journal **56**(1): 68-72.

Bicchieri, C. (2006). The Grammar of Society: The Nature and Dynamics of Social Norms. Cambridge, Cambridge University Press.

Bordass, B., R. Cohen and J. Field (2004). Energy performance of non-domestic buildings: closing the credibility gap. Building performance congress.

Bordass, B., A. Leaman and J. Eley (2002). "A guide to feedback and post-occupancy evaluation." Building Research & Information **30**(2): 145-157.

Boubekri, M. (2008). Daylighting, Architecture and Health. Burlington, Architectural Press.

Brady, L. and M. Abdellatif (2017). "Assessment of energy consumption in existing buildings." Energy and Buildings **149**: 142-150.

Brady, L. and R. Hanmer-Dwight (2016). The management and control of energy at the design stage of buildings. Proceedings of the 32nd Annual ARCOM Conference, ARCOM 2016.

Bramer, W. M., M. L. Rethlefsen, J. Kleijnen and O. H. Franco (2017). "Optimal database combinations for literature searches in systematic reviews." Systematic Reviews **6**: 245.

Bros-Williamson, J., J. Stinson and J. Currie (2015). "Energy performance evaluation of a passive house built to Scottish Building standards." International Journal for Housing Science and Its Applications **39**(4): 225-236.

Brown, H. S. and P. J. Vergragt (2016). "From consumerism to wellbeing: Toward a cultural transition?" Journal of Cleaner Production **132**: 308-317.

Brunsson, N. and B. Jacobsson (2000). A World of Standards. Oxford, Oxford University Press.

Buhl, F. W., R. B. Curtis, S. D. Gates, J. J. Hirsch, M. Lokmanhekim, S. P. Jaeger, A. H. Rosenfeld, F. C. Winkelmann, B. D. Hunn, M. A. Roschke, H. D. Ross and G. S. Leighton (1979). DOE-2: A New State-of-the-Art Computer Program for Energy Utilization Analysis of Buildings. Proceedings of the Second International CIB Symposium on Energy Conservation in the Built Environment. Copenhagen, Denmark.

Burman, E., J. Kimpian and D. Mumovic (2018). "Building Schools for the Future: Lessons Learned From Performance Evaluations of Five Secondary Schools and Academies in England." FRONTIERS IN BUILT ENVIRONMENT **4**.

Burman, E., D. Mumovic and J. Kimpian (2014). "Towards measurement and verification of energy performance under the framework of the European directive for energy performance of buildings." Energy **77**: 153-163.

Cali, D., T. Osterhage, R. Streblow and D. Muller (2016). "Energy performance gap in refurbished German dwellings: Lesson learned from a field test." ENERGY AND BUILDINGS **127**: 1146-1158.

Carpino, C., E. Loukou, P. Heiselberg and N. Arcuri (2022). "Energy performance gap of a nearly Zero Energy Building (nZEB) in Denmark: The influence of occupancy modelling." JOURNAL OF PLANNING LITERATURE **37**(1): 200-201.

Carrier (2014). "Weather Makers to the World."

Casian, C., C. Berger, A. Mahdavi, U. Pont and M. Schuss (2023). Projected Versus Actual Energy Performance Improvement Due to Thermal Retrofit of Buildings: A Case Study. AIP Conference Proceedings.

Cavka, B. T., H. B. Cavka and M. M. Salehi (2022). "An investigation of the design process's effect on a high-performance building's actual energy system performance." JOURNAL OF INTEGRATED DESIGN & PROCESS SCIENCE **26**(1): 85-100.

Charlier, D. (2021). "Explaining the energy performance gap in buildings with a latent profile analysis." Energy Policy **156**.

Charmaz, K. (2014). Constructing Grounded Theory. London, Sage Publications.

Chen, C. (2011). Residential Passive House Development in China.

Chen, X., T. Guo, M. Kriegel and P. Geyer (2022). "A hybrid-model forecasting framework for reducing the building energy performance gap." Advanced Engineering Informatics **52**: 101627.

Cholewa, T., C. A. Balaras, S. Nižetić and A. Siuta-Olcha (2020). "On calculated and actual energy savings from thermal building renovations – Long term field evaluation of multifamily buildings." Energy and Buildings **223**.

CIBSE (2013). TM54: Evaluating Operational Energy Performance of Buildings at the Design Stage. London, Chartered Institution of Building Services Engineers.

Clarke, J. A. and D. McLean (1988). ESP: A Building and Plant Energy Simulation System. Strathclyde, Energy Simulation Research Unit, University of Strathclyde.

Cohen, R. and B. and Bordass (2015). "Mandating transparency about building energy performance in use." Building Research & Information **43**(4): 534-552.

Colclough, S., R. O. Hegarty, M. Murray, D. Lennon, E. Rieux, M. Colclough and O. Kinnane (2022). "Post occupancy evaluation of 12 retrofit nZEB dwellings: The impact of occupants and high in-use interior temperatures on the predictive accuracy of the nZEB energy standard." Energy and Buildings **254**.

Coleman, S. and J. B. Robinson (2018). "Introducing the qualitative performance gap: stories about a sustainable building." Building Research & Information **46**(5): 485-500.

Corbin, J. and A. Strauss (2015). Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. Thousand Oaks, CA, Sage Publications.

Coyne, B. and E. Denny (2021). "Mind the Energy Performance Gap: testing the accuracy of building Energy Performance Certificates in Ireland." Energy Efficiency **14**(6): 57 (28 pp.).

Cozza, S., J. Chambers, A. Brambilla and M. K. Patel (2021). "In search of optimal consumption: A review of causes and solutions to the Energy Performance Gap in residential buildings." Energy and Buildings **249**.

Cozza, S., J. Chambers, C. Deb, J. L. Scartezzini, A. Schluter and M. K. Patel (2020). "Do energy performance certificates allow reliable predictions of actual energy consumption and savings? Learning from the Swiss national database." Energy and Buildings **224**: 110235 (110214 pp.).

Cozza, S., J. Chambers, C. Gambato, G. Branca, A. Geissler and M. K. Patel (2019). Energy consumption of high-performance buildings: Design vs. Reality. CLIMATE RESILIENT CITIES - ENERGY EFFICIENCY & RENEWABLES IN THE DIGITAL ERA (CISBAT 2019). **1343**.

Cozza, S., J. Chambers and M. K. Patel (2020). "Measuring the thermal energy performance gap of labelled residential buildings in Switzerland." Energy Policy **137**.

Crawley, D. B., L. K. Lawrie, F. C. Winkelmann, W. F. Buhl, Y. J. Huang, C. O. Pedersen, R. K. Strand, R. J. Liesen, D. E. Fisher, M. J. Witte and J. Glazer (2001). "EnergyPlus: Creating a New-Generation Building Energy Simulation Program." Energy and Buildings **33**(4): 319-331.

Creswell, J. W. and V. L. Plano Clark (2018). Designing and Conducting Mixed Methods Research. Thousand Oaks, CA, Sage Publications.

de Wilde, P. (2014). "The gap between predicted and measured energy performance of buildings: a framework for investigation." Automation in Construction **41**: 40-49.

de Wilde, P. (2014). "The gap between predicted and measured energy performance of buildings: A framework for investigation." Automation in Construction **41**: 40-49.

Deb, C., F. Zhang, J. Yang, S. E. Lee and K. W. Shah (2017). "A review on time series forecasting techniques for building energy consumption." Renewable and Sustainable Energy Reviews **74**: 902-924.

Department for Business, E. and S. Industrial (2017). The Clean Growth Strategy: Leading the way to a low carbon future. London, HM Government.

Doan, D. T., A. Ghaffarianhoseini, N. Naismith, T. Zhang, A. Ghaffarianhoseini and J. Tookey (2017). "A critical comparison of green building rating systems." Building and Environment **123**: 243-260.

Doggett, A. M. (2005). "Root cause analysis: A framework for tool selection." Quality Management Journal **12**(4): 34-45.

Drgoňa, J., J. Arroyo, I. Figueroa and M. Wetter (2020). "Reinforcement Learning and Deep Learning for Building Control: A Review." Energy and Buildings **224**: 110047.

Duncan, T., C. Hayles and J. Littlewood (2024). A Novel Case Study Methodology for Affordable Housing In-Depth Post-occupancy Evaluation in Wales, UK. Sustainability in Energy and Buildings 2023. Smart Innovation, Systems and Technologies (378). Berlin, Germany, Springer Nature: 745-757.

Eisman, A. B. and et al. (2022). "Practical Guide to Achieve Rigor and Data Integration in Mixed Methods Research." Frontiers in Psychology **13**.

El-Shorbagy, A. M. (2010). "Design with Nature: Windcatcher as a Paradigm of Natural Ventilation Device in Buildings." International Journal of Civil and Environmental Engineering **10**(3): 21-26.

Eon, C., J. K. Breadsell, J. Byrne and G. M. Morrison (2020). "The discrepancy between as-built and as-designed in energy efficient buildings: A rapid review." Sustainability (Switzerland) **12**(16).

eQUEST (2005). eQUEST Version 3.55.

European, C. (2018). Directive (EU) 2018/844 of the European Parliament and of the Council amending Directive 2010/31/EU on the energy performance of buildings and Directive 2012/27/EU on energy efficiency. Brussels, European Commission.

Everett, R., A. Horton and J. Doggart (1985). Linford Low Energy Houses [Project Summary]. Open University Energy Research Group Project Summary. Milton Keynes, Open University Energy Research Group.

Fedoruk, L. E., R. J. Cole, J. B. Robinson and A. Cayuela (2015). "Learning from failure: understanding the anticipated–achieved building energy performance gap." Building Research & Information **43**(6): 750-763.

Few, J., D. Manouseli, E. McKenna, M. Pullinger, E. Zapata-Webborn, S. Elam, D. Shipworth and T. Oreszczyn (2023). "The over-prediction of energy use by EPCs in Great Britain: A comparison of EPC-modelled and metered primary energy use intensity." ENERGY AND BUILDINGS **288**.

Filippi, M. and E. Sirombo (2019). "Energy and Water Monitoring for a Large Social Housing Intervention in Northern Italy." Frontiers in Energy Research **7**.

Firth, S. K., D. Allinson and S. Watson (2024). "Quantifying the spatial variation of the energy performance gap for the existing housing stock in England and Wales." JOURNAL OF BUILDING PERFORMANCE SIMULATION.

Flourentzos, F., Y. Ivanov Antonov and S. Pantet (2019). Understand, simulate, anticipate and correct performance gap in NZEB refurbishment of residential buildings. International Conference on Climate Resilient Cities - Energy Efficiency and Renewables in the Digital Era 2019, CISBAT 2019, September 4, 2019 - September 6, 2019, Lausanne, Switzerland, IOP Publishing Ltd.

Foo, E. and P. Bannister (2022). Design for Performance: Lessons from the NABERS UK Independent Design Review Process. London, UK, Better Buildings Partnership.

Fritzson, P. (2003). Principles of Object-Oriented Modeling and Simulation with Modelica 2.1, Wiley.

Fumo, N. (2014). "A review on the basics of building energy estimation." Renewable and Sustainable Energy Reviews **31**: 53-60.

Galvin, R. (2014). "Making the 'rebound effect' more useful for performance evaluation of thermal retrofits of existing homes: Defining the 'energy savings deficit' and the 'energy performance gap'." Energy and Buildings **69**: 515-524.

Gao, Y., P. Xu, H. Yu and X. Xu (2024). "A novel blockchain-based system for improving information integrity in building projects from the perspective of building energy performance." Environmental Impact Assessment Review **109**.

Geng, Y., W. J. Ji, B. R. Lin, J. J. Hong and Y. X. Zhu (2018). "Building energy performance diagnosis using energy bills and weather data." ENERGY AND BUILDINGS **172**: 181-191.

Ghaemmaghami, P. S. and M. Mahmoudi (2005). "Wind Tower: A Natural Cooling System in Iranian Traditional Architecture." Proceedings of Passive and Low Energy Cooling.

Gill, P., K. Stewart, E. Treasure and B. Chadwick (2008). "Methods of data collection in qualitative research: Interviews and focus groups." British Dental Journal **204**(6): 291-295.

Gram-Hanssen, K. and S. Georg (2018). "Energy performance gaps: promises, people, practices." BUILDING RESEARCH AND INFORMATION **46**(1): 1-9.

Grossmann, D., R. Galvin, J. Weiss, R. Madlener and B. Hirschl (2016). "A methodology for estimating rebound effects in non-residential public service buildings: Case study of four buildings in Germany." Energy and Buildings **111**: 455-467.

Gui, X. and Z. Gou (2020). "Association between green building certification level and post-occupancy performance: Database analysis of the National Australian Built Environment Rating System." Building and Environment **179**: 106971.

Gupta, R., A. Howard and A. Kotopouleas (2019). Meta-study of the energy performance gap in UK low energy housing. 2019 ECEEE Summer Study on Energy Efficiency: Is Efficient Sufficient?, June 3, 2019 - June 8, 2019, Presqu'ile de Giens, France, European Council for an Energy Efficient Economy.

Gupta, R. and M. Kapsali (2016). Evaluating the 'as-built' performance of an eco-housing development in the UK, SAGE Publications Ltd.

Gupta, R., M. Kapsali and M. Gregg (2017). "Comparative building performance evaluation of a 'sustainable' community centre and a public library building." BUILDING SERVICES ENGINEERING RESEARCH & TECHNOLOGY **38**(6): 691-710.

Gusenbauer, M. and N. R. Haddaway (2020). "Which academic search systems are suitable for systematic reviews or meta-analyses?" Research Synthesis Methods **11**(2): 181-217.

Guyot, D., F. Giraud, F. Simon, D. Corgier, C. Marvillet and B. Tremeac (2020). "Building energy model calibration: A detailed case study using sub-hourly measured data." ENERGY AND BUILDINGS **223**.

Hamburg, A., K. Kuusk, A. Mikola and T. Kalamees (2020). "Realisation of energy performance targets of an old apartment building renovated to nZEB." ENERGY **194**.

Hamdaoui, M. A., M.-H. Benzaama, Y. El Mendili and D. Chateigner (2021). "A review on physical and data-driven modeling of buildings hygrothermal behavior: Models, approaches and simulation tools." Energy and Buildings **251**: 111343.

Hammond, G. P. and J. Li (2022). "The development of UK building energy performance metrics: from the Milton Keynes Energy Cost Index to modern benchmarking schemes." Energies **15**(3): 895.

Hansen, A. R. and K. Gram-Hanssen (2023). Over- and underconsumption of residential heating: Analyzing occupant impacts on performance gaps between calculated and actual heating demand. Journal of Physics: Conference Series.

Hattell, M. J. and W. B. Morrison (1995). The Commonwealth Building: Groundbreaking History with a Groundwater Heat Pump, ASHRAE.

Hennink, M. and B. N. Kaiser (2022). "Sample sizes for saturation in qualitative research: A systematic review of empirical tests." Soc Sci Med **292**: 114523.

Hennink, M. M., B. N. Kaiser and V. C. Marconi (2017). "Code Saturation Versus Meaning Saturation: How Many Interviews Are Enough?" Qualitative Health Research **27**(4): 591-608.

Hensen, J. L. M. (2000). Past, Present and Future Plans of the International Building Performance Simulation Association – IBPSA. 27 October, Eindhoven, International Building Performance Simulation Association - Benelux.

Hernandez-Cruz, P., C. Giraldo-Soto, C. Escudero-Revilla, J. M. Hidalgo-Betanzos and I. Flores-Abascal (2023). "Energy efficiency and energy performance gap in centralized social housing buildings of the Basque Country." Energy and Buildings **298**.

Hinge, A. (2019). Building Energy Performance Gap Issues: An International Review. C. International Partnership for Energy Efficiency. Paris, OECD/IPEEC.

Hongling, C., L. Yishu, W. Yang, T. Maoyi, Y. Zhenhua, Z. Lin, L. Nan, W. Yangfeng, Z. Yiming, C. Chunhua, C. Yahong and Z. Siyan (2025). "Mixed methods research in general practice: Key points of design and implementation and case analysis." Chinese General Practice Journal **2**(3).

Horrigan, M., E. Murphy and J. Odonnell (2017). Bridging the environmental and energy performance gap in buildings through simulation, measurement and data analysis. 15th International Conference of the International Building Performance Simulation Association, Building Simulation 2017, August 7, 2017 - August 9, 2017, San Francisco, CA, United states, International Building Performance Simulation Association.

Howlett, M. (2019). Designing Public Policies: Principles and Instruments. London, Routledge.

Hussein, F. and K. Alhebsi (2025) "Adaptive Re-Use of Cultural Heritage Sites: A Strategy for Circular Economy." Sustainability **17** DOI: 10.3390/su17146403.

Ing Liang, W., A. C. Menoncin Loper, E. Kruger and F. K. Mori (2020). "Energy performance evaluation and comparison of sampled Brazilian bank buildings with the existing and proposed energy rating systems." Energy and Buildings **225**: 110304 (110312 pp.).

Jain, N., E. Burman, D. Mumovic and M. Davies (2021). "Managing energy performance in buildings from design to operation using modelling and calibration." Building Services Engineering Research and Technology **42**(5): 517-531.

Jain, N., E. Burman, S. Stamp, D. Mumovic and M. Davies (2020). "Cross-sectoral assessment of the performance gap using calibrated building energy performance simulation." Energy and Buildings **224**.

Janser, M., M. Hubbuch and L. Windlinger (2020). Call for a Definition and Paradigm Shift in Energy Performance Gap Research. WSBE 20 - World Sustainable Built Environment - Beyond2020, 2-4 Nov. 2020, UK, IOP Publishing.

Ji, C., T. Hong, H. Kim and S. Yeom (2022). "Effect of building energy efficiency certificate on reducing energy consumption of non-residential buildings in South Korea." Energy and Buildings **255**.

Jiang, L., L. Zhang, W. Lu, J. Xu and D. Luo (2025) "Study on Performance Index of Near-Zero-Energy Consumption Residence in Towns of Southern Jiangsu Province." Buildings **15** DOI: 10.3390/buildings15111922.

Jones, G. and L. Bouamane (2012). Power from Sunshine. USA, Harvard Business School.

Jones, P., L. Xiaojun, E. C. Bassas, E. Perisoglou and J. Patterson (2020). "Energy-positive house: performance assessment through simulation and measurement." Energies **13**(18): 4705 (4722 pp.).

Jones, R. V., A. Fuertes and P. De Wilde (2015). The gap between simulated and measured energy performance: A case study across six identical New-Build flats in the UK. 14th Conference of International Building Performance Simulation Association, BS 2015, December 7, 2015 - December 9, 2015, Hyderabad, India, International Building Performance Simulation Association.

Jradi, M., K. Arendt, F. C. Sangogboye, C. G. Mattera, E. Markoska, M. B. Kjaergaard, C. T. Veje and B. N. Jorgensen (2018). "ObepME: An online building energy performance monitoring and evaluation tool to reduce energy performance gaps." Energy and Buildings **166**: 196-209.

Jradi, M., N. Liu, A. Johansen, K. Arendt, C. G. Mattera, M. B. Kjaergaard, C. Veje and B. N. Jorgensen (2020). Dynamic Energy Model-Based Automatic Building Performance Testing for

Continuous Commissioning. PROCEEDINGS OF BUILDING SIMULATION 2019: 16TH CONFERENCE OF IBPSA**:** 822-829.

Juntunen, M. and M. Lehenkari (2019). "A narrative literature review process for an academic business research thesis." Studies in Higher Education **46**(2): 330-342.

Kallio, H., A. M. Pietilä, M. Johnson and M. Kangasniemi (2016). "Systematic methodological review: developing a framework for a qualitative semi-structured interview guide." Journal of Advanced Nursing **72**(12): 2954-2965.

Kampelis, N., K. Gobakis, V. Vagias, D. Kolokotsa, L. Standardi, D. Isidori, C. Cristalli, F. M. Montagnino, F. Paredes, P. Muratore, L. Venezia, M. K. Dracou, A. Montenon, A. Pyrgou, T. Karlessi and M. Santamouris (2017). "Evaluation of the performance gap in industrial, residential & tertiary near-Zero energy buildings." Energy and Buildings **148**: 58-73.

Kampelis, N., G. I. Papayiannis, D. Kolokotsa, G. N. Galanis, D. Isidori, C. Cristalli and A. N. Yannacopoulos (2020). "An integrated energy simulation model for buildings." Energies **13**(5).

Kane, T., S. K. Firth, V. Dimitriou, M. Coleman and T. Hassan (2015). Exploring the performance gap in UK homes: New evidence from smart home and smart meter data. 14th Conference of International Building Performance Simulation Association, BS 2015, December 7, 2015 - December 9, 2015, Hyderabad, India, International Building Performance Simulation Association.

Kang, Y., V. W. C. Chang, D. Chen, V. Graham and J. Zhou (2021). "Performance gap in a multi-storey student accommodation complex built to Passivhaus standard." Building and Environment **194**.

Kang, Y., N. Ma, V. Bunster, V. W-C Chang and J. Zhou (2022). "Optimizing the Passive House Planning Package simulation tool: A bottom-up dynamic approach to reduce building performance gap." Energy and Buildings **276**.

Katic, D., H. Krstic and S. Marenjak (2021). "Energy Performance of School Buildings by Construction Periods in Federation of Bosnia and Herzegovina." Buildings **11**(2): 42 (21 pp.).

Kelly, G. E., C. Park, D. R. Clark and W. B. May, Jr. (1984). HVACSIM+: A Dynamic Building/HVAC Control System Simulation Program. Proceedings of the Workshop on HVAC Controls Modeling and Simulation. Georgia Institute of Technology.

Khatri, P., V. Dabas, A. Thomas, G. Russo and A. Kallmuenzer (2025). "Unearthing the Literature Review Dilemma: A Comparative of Typologies of Literature Reviews." Business Ethics, the Environment & Responsibility **35**(1): 391-414.

Kim, Y. K., L. Bande, K. A. T. Aoul and H. Altan (2019). Using Energy Audit with POE Study to Reduce Energy Performance Gap in an Office Building in UAE. ZEMCH International Conference.

Kotarela, F., A. Kyritsis, R. Agathokleous and N. Papanikolaou (2023). "On the exploitation of dynamic simulations for the design of buildings energy systems." Energy **271**.

Kraus, S., M. Breier and S. Dasí-Rodríguez (2020). "The art of crafting a systematic literature review in entrepreneurship research." International Entrepreneurship and Management Journal **16**(3): 1023-1042.

Kühnl-Kinel, J. (2000). The History of Ventilation and Air Conditioning. CERN ST Workshop, Chamonix, France.

Kuivjogi, H., A. Uutar, K. Kuusk, M. Thalfeldt and J. Kurnitski (2021). "Market based renovation solutions in non-residential buildings – Why commercial buildings are not renovated to NZEB." Energy and Buildings **248**.

Kusuda, T. (1976). NBSLD: The Computer Program for Heating and Cooling Loads in Buildings. Washington, DC, National Bureau of Standards.

Leaman, A. and B. Bordass (2001). "Assessing building performance in use 4: The Probe occupant surveys and their implications." Building Research & Information **29**(2): 129-143.

Lehmann, U., J. Khoury and M. K. Patel (2017). Actual energy performance of student housing: case study, benchmarking and performance gap analysis. CISBAT 2017 INTERNATIONAL CONFERENCE FUTURE BUILDINGS & DISTRICTS - ENERGY EFFICIENCY FROM NANO TO URBAN SCALE. **122:** 163-168.

Li, J., L. D. Iulo and U. Poerschke (2023). A Review of the Energy Performance Gap between Predicted and Actual Use in Buildings. Building Simulation Conference Proceedings.

Liang, J., Y. Qiu and M. Hu (2019). "Mind the energy performance gap: Evidence from green commercial buildings." Resources, Conservation and Recycling **141**: 364-377.

Lin, B. and Z. Chen (2022). "Net zero energy building evaluation, validation and reflection – A successful project application." Energy and Buildings **261**.

Lizana, J., S. Wheeler, E. Azizi, C. Halloran, J. Wheeler, D. C. H. Wallom and M. McCulloch (2024). "Integrated post-occupancy evaluation and intervention that achieve real-world zero-carbon buildings." Energy and Buildings: 113766.

Lomas, K. J., P. Steadman and M. Brown (2019). "Energy performance of UK housing: tracing the lineage of building energy rating schemes from Milton Keynes to EPCs." Energy and Buildings **202**: 109365.

Loukou, E., P. K. Heiselberg, R. L. Jensen and H. Johra (2019). Energy performance evaluation of a nearly zero energy building and the reasons for the performance gap between expected and actual building operation. 1st Nordic conference on Zero Emission and Plus Energy Buildings, 6-7 Nov. 2019, UK, IOP Publishing.

Mahdavi, A. and C. Berger (2024). "Ten questions regarding buildings, occupants and the energy performance gap." Journal of Building Performance Simulation.

Mahood, Q., D. Van Eerd and E. Irvin (2014). "Searching for grey literature for systematic reviews: challenges and benefits." Research Synthesis Methods **5**(3): 221-234.

Manfren, M., P. A. B. James and L. Tronchin (2022). "Data-driven building energy modelling: An analysis of the potential for generalisation through interpretable machine learning." Renewable and Sustainable Energy Reviews **167**: 112686.

Maracchini, G., A. Latini, E. Di Giuseppe, A. Gianangeli and M. D'Orazio (2024). What Matters the Most? The Role of Actual Occupancy Patterns and Automatic Model Calibration in Reducing the Building Energy Performance Gap in an Italian Case Study. Sustainability in Energy and Buildings 2023. Smart Innovation, Systems and Technologies (378). Berlin, Germany, Springer Nature**:** 237-247.

Marchi, L., E. Antonini and S. Politi (2021). "Green Building Rating Systems (GBRSs)." Encyclopedia **1**(4): 998-1009.

Marshall, A., R. Fitton, W. Swan, D. Farmer, D. Johnston, M. Benjaber and Y. Ji (2017). "Domestic building fabric performance: Closing the gap between the in situ measured and modelled performance." Energy and Buildings **150**: 307-317.

Martín-Martín, A., E. Orduna-Malea, M. Thelwall and E. Delgado López-Cózar (2018). "Google Scholar, Web of Science and Scopus: a systematic comparison of citations in 252 subject categories." Journal of Informetrics **12**(4): 1160-1177.

Mattimoe, R., M. T. Hayden, B. Murphy and J. Ballantine (2021). "Approaches to Analysis of Qualitative Research Data: A Reflection on the Manual and Technological Approaches." Accounting, Finance & Governance Review **27**.

Maxwell, J. A. (2013). Qualitative Research Design: An Interactive Approach. Thousand Oaks, CA, Sage Publications.

Mazzeo, D., N. Matera, C. Cornaro, G. Oliveti, P. Romagnoni and L. De Santoli (2020). "EnergyPlus, IDA ICE and TRNSYS predictive simulation accuracy for building thermal behaviour evaluation by using an experimental campaign in solar test boxes with and without a PCM module." Energy and Buildings **212**: 109812.

McElroy, D. J. and J. Rosenow (2019). "Policy implications for the performance gap of low-carbon building technologies." Building Research & Information **47**(5): 611-623.

Menezes, A. C., A. Cripps, D. Bouchlaghem and R. Buswell (2012). "Predicted vs. actual energy performance of non-domestic buildings: Using post-occupancy evaluation data to reduce the performance gap." Applied Energy **97**: 355-364.

Mikhail, M., D. Mather, P. Parker and K. Kapsis (2023). "Net-positive office commissioning and performance gap assessment: Empirical insights." Energy and Buildings **279**.

Milek, K. (2012). "Floor formation processes and the interpretation of turf buildings." Journal of Anthropological Archaeology **31**: 119-137.

Miles, M. B., A. M. Huberman and J. Saldaña (2014). Qualitative Data Analysis: A Methods Sourcebook. Thousand Oaks, CA, SAGE Publications.

Miller, W., B. Sodagar, D. Whaley, K. Bamdad and S. Zedan (2021). "Mind the gap: A comparison of socio-technical limitations of national house rating systems in the UK and Australia." Journal of Building Engineering **43**.

Mirzaie, S. (2024). "From PROBE to net zero: A study of soft landings' contribution." BUILDING SERVICES ENGINEERING RESEARCH & TECHNOLOGY.

Mirzaie, S. and G. Menzies (2017). How can the combination of breeam and soft landings successfully deliver a low energy, comfortable building? 33rd International on Passive and Low Energy Architecture Conference: Design to Thrive, PLEA 2017, July 2, 2017 - July 5, 2017, Edinburgh, United kingdom, NCEUB 2017 - Network for Comfort and Energy Use in Buildings.

Mitchell, R. and S. Natarajan (2020). "UK Passivhaus and the energy performance gap." Energy and Buildings **224**: 1042-1055.

Moeller, S. (2024). "Is it a match? Smart home energy management technologies and user comfort practices in German multi-apartment buildings." Energy Research and Social Science, **118**, Article 103794.

Moeller, S., I. Weber, F. Schroder, A. Bauer and H. Harter (2020). "Apartment related energy performance gap - How to address internal heat transfers in multi-apartment buildings." Energy and Buildings, **215**, Article 109887.

Mohareb, E., A. Hashemi, M. Shahrestani and M. Sunikka-Blank (2017). "Retrofit planning for the performance gap: results of a workshop on addressing energy, health and comfort needs in a protected building." Energies **10**.

Mojic, I., M. Lehmann, S. Van Velsen and M. Haller (2019). ImmoGap - Analysis of the performance gap of apartment buildings. 13th REHVA World Congress, CLIMA 2019, May 26, 2019 - May 29, 2019, Bucharest, Romania, EDP Sciences.

Mongeon, P. and A. Paul-Hus (2016). "The journal coverage of Web of Science and Scopus, a comparative analysis." Scientometrics **106**(1): 213-228.

Moreno-Rangel, A. (2021) "Passivhaus." Encyclopedia **1**, 20-29 DOI: 10.3390/encyclopedia1010005.

Motuzien, V., V. Lapinskien, G. Rynkun and J. Bielskus (2021). "Energy Performance Gap Analysis in Energy Efficient Residential Buildings in Lithuania." Environmental and Climate Technologies **25**(1): 610-620.

Motuziene, V., V. Lapinskiene and G. Rynkun (2024). "Optimizing Ventilation Systems for Sustainable Office Buildings: Long-Term Monitoring and Environmental Impact Analysis." Sustainability **16**(3): 984 (916 pp.).

Munn, Z., M. D. J. Peters, C. Stern, C. Tufanaru, A. McArthur and E. Aromataris (2018). "Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach." BMC Medical Research Methodology **18**(1): 143.

Munsami, K., D. Prasad and D. Lan (2017). "The Role of Post Occupation Evaluation in Achieving High Performance Buildings through Diagnostics." Procedia Engineering **180**: 356-364.

Murugaiah, U., S. J. Benjamin, M. S. Marathamuthu and S. Muthaiyah (2010). "Scrap loss reduction using the 5-whys analysis." International Journal of Quality & Reliability Management **27**(5): 527-540.

Nabers (2023). Become a NABERS Accredited Assessor, NABERS.

Nigumann, E., T. Kalamees, K. Kuusk and P. Pihelo (2024). "Circular Renovation of an Apartment Building with Prefabricated Additional Insulation Elements to Nearly Zero Energy Building." Journal of Sustainable Architecture and Civil Engineering **34**(1): 22-34.

Norford, L. K., R. H. Socolow, E. S. Hsieh and G. V. Spadaro (1994). "Two-to-one discrepancy between measured and predicted performance of a 'low-energy' office building: insights from a reconciliation based on the DOE-2 model." Energy and Buildings **21**(2): 121-131.

Odeh, K., M. Parvaz, J. Jakkula, P. de Wilde, L. B. McElroy, A. Moreno-Rangel and B. Kumar (2025). Exploring the ambiguity associated with the actual magnitude of the energy performance gap in buildings. EG-ICE 2025.

Office, B. S. (1992). BLAST 3.0 Users Manual. Urbana-Champaign, IL, Department of Mechanical and Industrial Engineering, University of Illinois.

Ohno, T. (1988). Toyota Production System: Beyond Large-Scale Production. Portland, OR, Productivity Press.

Oleiwi, M. Q., M. F. Mohamed, M. K. A. M. Sulaiman, A. I. Che-Ani and S. N. Raman (2019). "Thermal environment accuracy investigation of Integrated Environmental Solutions-Virtual Environment (IES-VE) software for double-story house simulation in Malaysia." ARPN Journal of Engineering and Applied Sciences **14**(11): 3659-3665.

Oxlade, C. (2012). Solar Power. London, Capstone.

Ozarisoy, B. and H. Altan (2022). "Bridging the energy performance gap of social housing stock in south-eastern Mediterranean Europe: Climate change and mitigation." Energy and Buildings **258**.

Padey, P., K. Goulouti, G. Wagner, B. Perisset and S. Lasvaux (2021). "Understanding the reasons behind the energy performance gap of an energy-efficient building, through a probabilistic approach and on-site measurements." Energies **14**(19).

Paez, A. (2017). "Gray literature: An important resource in systematic reviews." Journal of Evidence-Based Medicine **10**(3): 233-240.

Page, M. J., J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M. M. Lalu, T. Li, E. W. Loder, E. Mayo-Wilson, S. McDonald, L. A. McGuinness, L. A. Stewart, J. Thomas, A. C. Tricco, V. A. Welch, P. Whiting and D. Moher (2021). "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews." BMJ **372**: n71.

Palladino, D. (2023). "Energy performance gap of the Italian residential building stock: Parametric energy simulations for theoretical deviation assessment from standard conditions." Applied Energy **345**: 121365.

Palma, P., J. P. Gouveia and S. G. Simoes (2019). "Mapping the energy performance gap of dwelling stock at high-resolution scale: Implications for thermal comfort in Portuguese households." Energy and Buildings **190**: 246-261.

Palmiter, L., T. Wheeling and R. Judkoff (1982). SERIRES: Residential Energy Simulator, Solar Energy Research Institute (now NREL).

Pan, W. and H. Garmston (2012). "Building regulations in energy efficiency: Compliance in England and Wales." Energy Policy **45**: 594-605.

Parvaz, M. O. F. M. and J. P. Jakkula (2025). Building Energy Performance Gap: Different Approaches for Mitigation Master's thesis.

Patel, H. and S. D. Green (2020). "Beyond the performance gap: reclaiming building appraisal through archival research." Building Research & Information **48**(5): 469-484.

Pătrașcu, G. (1984). Arhitectura și tehnica populară. Bucharest, Editura Tehnică.

Petri, I., S. Kubicki, Y. Rezgui, A. Guerriero and H. Li (2017). "Optimizing energy efficiency in operating built environment assets through building information modeling: A case study." Energies **10**(8).

Pitts, A. (2017) "Passive House and Low Energy Buildings: Barriers and Opportunities for Future Development within UK Practice." Sustainability **9** DOI: 10.3390/su9020272.

PRISMA (2024). "PRISMA 2020 checklist."

Pritchard, R. and S. Kelly (2017). "Realising operational energy performance in non-domestic buildings: Lessons learnt from initiatives applied in Cambridge." Sustainability (Switzerland) **9**(8).

Rahimi, S. and M. Khatooni (2024). "Saturation in qualitative research: An evolutionary concept analysis." Int J Nurs Stud Adv **6**: 100174.

Reguis, A., M. Tunzi, B. Vand, P. Tuohy and J. Currie (2023). "Energy performance of Scottish public buildings and its impact on the ability to use low-temperature heat." Energy and Buildings **290**.

Rethlefsen, M. L., S. Kirtley, S. Waffenschmidt, A. P. Ayala, D. Moher, M. J. Page and J. B. Koffel (2021). "PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews." Systematic Reviews **10**(1): 39.

Rosenow, J., R. Cowart and S. Thomas (2017). "Market-based instruments for energy efficiency: A global review." Energy Policy **98**: 1-19.

Rotimi, A., A. Bahadori-Jahromi, A. Mylona, P. Godfrey and D. Cook (2017). "Estimation and validation of energy consumption in UK existing hotel building using dynamic simulation software." Sustainability (Switzerland) **9**(8).

Rouleau, J., L. Gosselin and P. Blanchet (2018). "Understanding energy consumption in high-performance social housing buildings: A case study from Canada." Energy **145**: 677-690.

Ruslin, S. Mashuri, M. S. A. Rasak, F. Alhabsyi and H. Syam (2022). "Semi-structured interview: A methodological reflection on the development of a qualitative research instrument in educational studies." IOSR Journal of Research & Method in Education **12**(1): 22-29.

Salehi, M. M., B. Terim Cavka, A. Frisque, D. Whitehead and W. K. Bushe (2015). "A case study: The energy performance gap of the Center for Interactive Research on Sustainability at the University of British Columbia." Journal of Building Engineering **4**: 127-139.

Salvia, G., E. Morello, F. Rotondo, A. Sangalli, F. Causone, S. Erba and L. Pagliano (2020). "Performance gap and occupant behavior in building retrofit: Focus on dynamics of change and continuity in the practice of indoor heating." Sustainability (Switzerland) **12**(14).

Saunders, B., J. Sim, T. Kingstone, S. Baker, J. Waterfield, B. Bartlam, H. Burroughs and C. Jinks (2018). "Saturation in qualitative research: Exploring its conceptualization and operationalization." Quality & Quantity **52**: 1893-1907.

Sayigh, A. and A. Marafia (1998). "Thermal comfort and the development of bioclimatic concept in building design." Renewable and Sustainable Energy Reviews **2**: 3-24.

Schoonenboom, J. and R. B. Johnson (2017). "How to Construct a Mixed Methods Research Design." Kölner Zeitschrift für Soziologie und Sozialpsychologie **69**(Suppl 2): 107-131.

Schultz, A. L. and J. McAllister (2024). A comparison of coverage and subject areas in Compendex, Inspec and the Web of Science Core Collection for use as an engineering database. ASEE Midwest Section Conference.

Sear, F. (1982). Roman Architecture. New York, Cornell University Press.

Seguel-Vargas, S., C. Rubio-Bellido, L. Pereira-Ruchansky and A. Perez-Fargallo (2024). "Heating Energy Performance Gap in Vulnerable Households: Identification and Impact of Associated Variables." Energies **17**(19): 4995 (4921 pp.).

Serrat, O. (2017). The Five Whys Technique. Knowledge Solutions: Tools, Methods, and Approaches to Drive Organizational Performance. O. Serrat. Singapore, Springer Singapore**:** 307-310.

Sharma, D. L. R., S. Bidari, D. Bidari, S. Neupane and R. Sapkota (2023). "Exploring the Mixed Methods Research Design: Types, Purposes, Strengths, Challenges, and Criticisms." Global Academic Journal of Linguistics and Literature **5**(1): 3-12.

Shi, X., B. Si, J. Zhao, Z. Tian, C. Wang, X. Jin and X. Zhou (2019). "Magnitude, causes, and solutions of the performance gap of buildings: A review." Sustainability (Switzerland) **11**(3).

Shrubsole, C., I. G. Hamilton, N. Zimmermann, G. Papachristos, T. Broyd, E. Burman, D. Mumovic, Y. Zhu, B. Lin and M. Davies (2019). "Bridging the gap: The need for a systems thinking approach in understanding and addressing energy and environmental performance in buildings." INDOOR AND BUILT ENVIRONMENT **28**(1): 100-117.

Skjott Linneberg, M. and S. Korsgaard (2019). "Coding qualitative data: a synthesis guiding the novice." Qualitative Research Journal **19**(3): 259-270.

Solar Energy, L. (2000). TRNSYS Version 15 User Manual and Documentation. Madison, WI, Solar Energy Laboratory, Mechanical Engineering Department, University of Wisconsin.

Srivastava, P. and N. Hopwood (2009). "A practical iterative framework for qualitative data analysis." International Journal of Qualitative Methods **8**(1): 76-84.

Strandberg-de Bruijn, P., A. Donarelli and K. Balksten (2019) "Full-scale Studies of Improving Energy Performance by Renovating Historic Swedish Timber Buildings with Hemp-lime." Applied Sciences **9** DOI: 10.3390/app9122484.

Sun, K. and T. Hong (2017). "A framework for quantifying the impact of uncertainty in building energy models." Applied Energy **203**: 868-880.

Szulgowska-Zgrzywa, M., E. Stefanowicz, A. Chmielewska and K. Piechurski (2023). "Detailed Analysis of the Causes of the Energy Performance Gap using the Example of Apartments in Historical Buildings in Wroclaw (Poland)." Energies: 1814 (1819 pp.).

Szulgowska-Zgrzywa, M., E. Stefanowicz, K. Piechurski, A. Chmielewska and M. Kowalczyk (2020). "Impact of Users' Behavior and Real Weather Conditions on the Energy Consumption of Tenement Houses in Wroclaw, Poland: Energy Performance Gap Simulation Based on a Model Calibrated by Field Measurements." ENERGIES **13**(24).

Tagliabue, L. C., S. Maltese, F. Re Cecconi, A. L. C. Ciribini and E. De Angelis (2018). BIM-based interoperable workflow for energy improvement of school buildings over the life cycle. 35th International Symposium on Automation and Robotics in Construction and International AEC/FM

Hackathon: The Future of Building Things, ISARC 2018, July 20, 2018 - July 25, 2018, Berlin, Germany, International Association for Automation and Robotics in Construction I.A.A.R.C).

Templier, M. and G. Paré (2015). "A framework for guiding and evaluating literature reviews." Communications of the Association for Information Systems **37**(1): 112-137.

Thompson, D., E. Burman, D. Mumovic and M. Davies (2022). "Managing the risk of the energy performance gap in non-domestic buildings." Building Services Engineering Research and Technology **43**(1): 57-88.

Tierney, G. and S. Tennant (2015). House building in Scotland: The sustainability performance gap. 31st Annual Association of Researchers in Construction Management Conference, ARCOM 2015, September 7, 2015 - September 9, 2015, Lincoln, United kingdom, Association of Researchers in Construction Management.

Tuohy, P. G. and G. B. Murphy (2015). "Closing the gap in building performance: Learning from BIM benchmark industries." Architectural Science Review **58**(1): 47-56.

van den Brom, P., A. Meijer and H. Visscher (2020). "Actual energy saving effects of thermal renovations in dwellings: Longitudinal data analysis including building and occupant characteristics." A+BE | Architecture and the Built Environment **10**(03): 89-128.

van Dronkelaar, C., M. Dowson, C. Spataru and D. Mumovic (2016). "A Review of the Regulatory Energy Performance Gap and Its Underlying Causes in Non-domestic Buildings." Frontiers in Mechanical Engineering **1**.

van Someren, K. L., C. P. Beaman and L. Shao (2017). "Determining the Difference between Predicted vs. Actual Lighting Use in Higher Education Corridors." Frontiers in Mechanical Engineering **Volume 3 - 2017**.

Vivian, J., L. Carnieletto, M. Cover and M. De Carli (2023). "At the roots of the energy performance gap: Analysis of monitored indoor air before and after building retrofits." Building and Environment **246**.

Wang, D., X. F. Pang, W. Wang, Z. W. Qi, Y. Ji and R. X. Yin (2020). "Evaluation of the dynamic energy performance gap of green buildings: Case studies in China." BUILDING SIMULATION **13**(6): 1191-1204.

Wei, T., W. Li and Y. Guo (2019). "Transfer Learning for Building Energy Prediction." Energy and Buildings **199**: 527-544.

West, R. E. and F. Martin (2023). "What type of paper are you writing? A taxonomy of review and theory scholarship distinguished by their summary and advocacy arguments." Educational technology research and development **72**(5): 2443-2476.
Wilde, P. and R. Jones (2014). The building energy performance gap: up close and personal.
Xu, X. and P. X. W. Zou (2020). "Analysis of factors and their hierarchical relationships influencing building energy performance using interpretive structural modelling (ISM) approach." Journal of Cleaner Production **272**.
Yoo, Y. S., H. Shin, D. W. Kim and C. S. Park (2023). Performance Gap Analysis in The Context of Building Energy Efficiency Certificate in South Korea. Building Simulation Conference Proceedings.
Yoo, Y. S., H. Shin, D. W. Kim and C. S. Park (2024). "Performance gap analysis for Korean building energy efficiency certification." ENERGY AND BUILDINGS **314**.
Yu, J., W.-S. Chang and Y. Dong (2022) "Building Energy Prediction Models and Related Uncertainties: A Review." Buildings **12**, 1284 DOI: 10.3390/buildings12081284.
Zaid, S. M. and A. Kiani (2016). Energy Prediction versus Energy Performance of Green Buildings in Malaysia. Comparison of Predicted and Operational Measurement of GBI Certified Green Office in Kuala Lumpur. 4th International Building Control Conference 2016 (IBCC 2016), 7-8 March 2016, France, EDP Sciences.
Zaid, S. M., A. K. Rad and N. Zainon (2017). "Are green offices better than conventional? Measuring operational energy consumption and carbon impact of green office in Malaysia." Facilities **35**(11-12): 622-637.
Zare, N., S. M. E. Saryazdi, A. M. Bahman, A. Shafaat and M. Sartipipour (2023). "Investigation of heating energy performance gap (EPG) in design and operation stages of residential buildings." Energy and Buildings **301**.
Zhan, S., M. Zhu, S. Cheng and A. Chong (2024). "Bridging performance gap for existing buildings: The role of calibration and the cascading effect." Building Simulation.
Zhan, S., M. Zhu, S. Cheng and A. Chong (2025). "Bridging performance gap for existing buildings: The role of calibration and the cascading effect." Building Simulation **18**(1): 123-140.
Zheng, Z., J. Zhou, Z. Jiaqin, Y. Yang, F. Xu and H. Liu (2024). "Review of the building energy performance gap from simulation and building lifecycle perspectives: Magnitude, causes and solutions." Developments in the Built Environment **17**.

Zou, P. X. W. and M. Alam (2020). "Closing the building energy performance gap through component level analysis and stakeholder collaborations." Energy and Buildings **224**.

Zou, P. X. W., M. Alam and L. Hebert (2019). Closing the Gap Between Design and Reality of Building Energy Performance. Australia, Sustainable Built Environment National Research Centre (SBEnrc).

Zou, P. X. W., D. Wagle and M. Alam (2019). "Strategies for minimizing building energy performance gaps between the design intend and the reality." Energy and Buildings **191**: 31-41.

Zou, P. X. W., X. Xiaoxiao, J. Sanjayan and W. Jiayuan (2018). "Review of 10 years research on building energy performance gap: Life-cycle and stakeholder perspectives." Energy and Buildings **178**: 165-181.

Appendices

Appendix A: PRISMA 2020 Checklist Mapping for the SLR

Table A.1: Mapping of PRISMA 2020 checklist items to thesis sections

This appendix maps the PRISMA 2020 checklist items to the sections of the thesis in which they are addressed. The checklist was used as a reporting aid rather than as a rigid methodological protocol.

PRISMA 2020 item	Topic	How addressed in the thesis	Thesis location	Applicability
1	Title	The review is identified as a systematic literature review within the thesis methodology and Chapter 3.	Chapter 2, Section 2.1; Chapter 3, Section 3.1	Applicable
2	Abstract	The thesis abstract summarises the research aim, mixed-methods approach, SLR, root-cause analysis and key findings. A separate PRISMA abstract checklist was not used because the review forms part of a doctoral thesis rather than a standalone systematic review article.	Abstract	Partly applicable
3	Rationale	The rationale for conducting the SLR is explained in relation to the need to establish a robust evidence base on the causes and magnitude of the EPG.	Chapter 3, Section 3.1	Applicable
4	Objectives	The objectives of the SLR are stated as investigating the causes contributing to the emergence of the EPG and its magnitude across global contexts.	Chapter 3, Section 3.1	Applicable
5	Eligibility criteria	Eligibility decisions are described through the date range, language restrictions, full-text availability, relevance to EPG causes and/or magnitude, and inclusion in the relevant review strand.	Chapter 3, Section 3.1.1; Figure 3.1	Applicable
6	Information sources	The review used Web of Science, Scopus and Engineering Village. Targeted grey literature was also considered where directly relevant and methodologically transparent.	Chapter 3, Section 3.1.1	Applicable
7	Search strategy	The search string combined the terms building, energy and performance gap using Boolean logic, with truncation applied to capture lexical variants.	Chapter 3, Section 3.1.1	Applicable
8	Selection process	Records were imported into EndNote and screened through identification, screening, eligibility and inclusion stages. The process is summarised in the PRISMA flow diagram.	Chapter 3, Section 3.1.1; Figure 3.1	Applicable
9	Data collection process	The review used two strands: saturation-based synthesis for root causes and exhaustive extraction for EPG magnitude.	Chapter 3, Section 3.1.1	Applicable
10a	Data items: outcomes	The main data items were causes contributing to the EPG and EPG magnitudes.	Chapter 3, Sections 3.1.1, 3.2 and 3.3	Applicable
10b	Data items: other variables	Contextual variables, including geographical context, building type and regulatory or	Chapter 3, Sections 3.1.2 and 3.3	Applicable

PRISMA 2020 item	Topic	How addressed in the thesis	Thesis location	Applicability
		methodological context, were considered where relevant to reported EPG magnitudes.		
11	Study risk of bias assessment	A formal intervention-specific risk-of-bias assessment was not conducted because the review did not evaluate intervention effects. Instead, rigour was pursued through transparent search reporting, duplicate removal, explicit inclusion criteria, documented screening stages and consistent recalculation of EPG ratios.	Chapter 3, Section 3.1.1	Not directly applicable
12	Effect measures	Effect measures were not applied because the review did not conduct a statistical meta-analysis. The quantitative synthesis focused on reported EPG magnitudes and recalculated EPG ratios.	Chapter 3, Section 3.1.2	Not directly applicable
13a	Synthesis methods: eligibility for synthesis	Eligibility for synthesis was determined separately for the two review strands: saturation-based synthesis for root causes and exhaustive extraction for reported magnitudes.	Chapter 3, Section 3.1.1	Applicable
13b	Synthesis methods: preparation of data	The approach to extracting, recalculating and normalising EPG values is explained.	Chapter 3, Section 3.1.2	Applicable
13c	Synthesis methods: tabulation and visual display	The identification, screening and inclusion process is shown in Figure 3.1. Included evidence is then tabulated and discussed in the causes and magnitude sections.	Figure 3.1; Chapter 3, Sections 3.2 and 3.3	Applicable
13d	Synthesis methods: synthesis rationale	The use of saturation for root causes and exhaustive extraction for magnitude is justified by the dual purpose of the SLR.	Chapter 3, Section 3.1.1	Applicable
13e	Synthesis methods: heterogeneity	Statistical heterogeneity was not assessed because no meta-analysis was undertaken. Variation in reported EPG magnitudes was instead addressed through contextual comparison and normalisation of EPG ratios.	Chapter 3, Sections 3.1.2 and 3.3	Not directly applicable
13f	Synthesis methods: sensitivity analyses	Sensitivity analyses were not conducted because the review did not use statistical meta-analytic models.	Chapter 3, Section 3.1.1	Not directly applicable
14	Reporting bias assessment	Formal reporting-bias tests were not conducted. The review reduced potential publication and coverage bias by using multiple databases and targeted grey-literature supplementation.	Chapter 3, Section 3.1.1	Partly applicable
15	Certainty assessment	A formal certainty assessment was not conducted because the review did not assess intervention effects. The strength of the evidence base was instead supported through transparent source selection and consistent extraction procedures.	Chapter 3, Section 3.1.1	Not directly applicable
16a	Study selection: search and selection results	The number of records identified, screened, excluded and included is reported in the PRISMA flow diagram.	Chapter 3, Section 3.1.1; Figure 3.1	Applicable
16b	Study selection:	Exclusion decisions are summarised numerically in the PRISMA flow diagram; individual excluded	Figure 3.1	Applicable

PRISMA 2020 item	Topic	How addressed in the thesis	Thesis location	Applicability
	excluded studies	studies are not listed because the review is embedded within a doctoral thesis rather than reported as a standalone SLR.		
17	Study characteristics	Characteristics of included studies are discussed in relation to EPG causes, reported magnitudes, building types and geographical contexts.	Chapter 3, Sections 3.2 and 3.3	Applicable
18	Risk of bias in studies	Formal risk-of-bias assessment was not applied because the review was not an intervention-effect review. Limitations relating to source type and evidence boundaries are acknowledged in the methodology.	Chapter 3, Section 3.1.1	Not directly applicable
19	Results of individual studies	Individual study evidence is synthesised in the root-cause taxonomy and in the magnitude analysis tables.	Chapter 3, Sections 3.2 and 3.3	Applicable
20a	Results of syntheses: summary characteristics	The root-cause strand reports categories of causes, while the magnitude strand reports EPG values across geographical and regulatory contexts.	Chapter 3, Sections 3.2 and 3.3	Applicable
20b	Results of syntheses: synthesis results	The results are presented through thematic synthesis for causes and quantitative comparison of reported EPG magnitudes.	Chapter 3, Sections 3.2 and 3.3	Applicable
20c	Results of syntheses: heterogeneity	Statistical heterogeneity was not assessed. Contextual variation is discussed through comparison across locations, building types and calculation approaches.	Chapter 3, Section 3.3	Not directly applicable
20d	Results of syntheses: sensitivity analyses	Sensitivity analyses were not conducted because no statistical meta-analysis was undertaken.	Chapter 3, Section 3.1.1	Not directly applicable
21	Reporting biases	Formal reporting-bias assessment was not undertaken. Potential source and publication biases are acknowledged through the discussion of academic and grey-literature boundaries.	Chapter 3, Section 3.1.1	Partly applicable
22	Certainty of evidence	A formal certainty of evidence framework was not applied because the review did not evaluate intervention effects.	Chapter 3, Section 3.1.1	Not directly applicable
23a	Discussion: interpretation	The findings are interpreted in relation to the wider EPG literature and the thesis research objectives.	Chapter 3, Sections 3.2, 3.3 and 3.4	Applicable
23b	Discussion: limitations of evidence	Limitations relating to source type, academic focus, grey literature and adjacent approaches such as NABERS, Passivhaus, Soft Landings and POE are acknowledged.	Chapter 3, Section 3.1.1; Chapter 8, Section 8.3	Applicable
23c	Discussion: limitations of review process	The methodological boundary of not treating NABERS, Passivhaus, Soft Landings and POE as separate review domains is acknowledged at the point of method selection.	Chapter 3, Section 3.1.1	Applicable
23d	Discussion: implications	Implications of the SLR for the later empirical and standardisation stages of the thesis are discussed.	Chapter 3, Section 3.4; Chapters 4–6; Chapter 7	Applicable

PRISMA 2020 item	Topic	How addressed in the thesis	Thesis location	Applicability
24a	Registration and protocol	The review was not registered as a standalone systematic review protocol because it was conducted as part of a doctoral thesis methodology rather than as a separately registered systematic review.	Appendix A	Not applicable
24b	Protocol access	No separate public review protocol is available.	Appendix A	Not applicable
24c	Protocol amendments	No registered protocol amendments apply.	Appendix A	Not applicable
25	Support	Institutional and supervisory support are acknowledged in the thesis acknowledgements. No separate review-specific funding is reported.	Acknowledgements	Applicable
26	Competing interests	The thesis declaration confirms the originality of the work and appropriate acknowledgement of sources. No competing interests are reported for the SLR.	Declaration of Authorship	Applicable
27	Availability of data, code and materials	Extracted study evidence is presented in the Chapter 3 tables and synthesis sections. Where additional materials are not publicly deposited, they may be made available according to university requirements.	Chapter 3, Sections 3.2 and 3.3; Appendix A	Applicable

Appendix B: Consent Form Template

University of Strathclyde Glasgow

Consent Form

Name of department: Architecture
Title of the study: Adoption of Outcome-Based Polices and Digital Twins to Mitigate the Building Energy Performance Gap

- I confirm that I have read and understood the Participant Information Sheet for the above project and the researcher has answered any queries to my satisfaction.
- I confirm that I have read and understood the Privacy Notice for Participants in Research Projects and understand how my personal information will be used and what will happen to it (i.e. how it will be stored and for how long).
- I understand that my participation is voluntary and that I am free to withdraw from the project at any time, up to the point of completion, without having to give a reason and without any consequences.
- I understand that I can request the withdrawal of some information and that whenever possible researchers will comply with my request. This includes the following personal data:
    - audio recordings of interviews that identify me;
    - my personal information from transcripts.
- I understand that anonymised data (i.e. data that do not identify me personally) cannot be withdrawn once they have been included in the study.
- I understand that any information recorded in the research will remain confidential and no information that identifies me will be made publicly available.
- I consent to being a participant in the project.
- I consent to being audio recorded as part of the project
- I consent to remain anonymous in the final research papers

(PRINT NAME)	
Signature of Participant:	Date:

Appendix C: Harmonised Input Assumptions and Output Categories for the Simulation Models

Table C.1: Building fabric and envelope assumptions

Building element	Location / item	Thickness	U-value	Unit
External wall	Main building, south	297	0.1223	mm; W/m²K
External wall	Main building, north	297	0.1223	mm; W/m²K
External wall	Main building, east	307	0.1182	mm; W/m²K
External wall	Main building, west	307	0.1182	mm; W/m²K
External wall	Calorimetric cell 1	151	1.455	mm; W/m²K
External wall	Calorimetric cell 2	161	0.534	mm; W/m²K
External wall	Staircase and external store, east-west	120	2.653	mm; W/m²K
External wall	Staircase and external store, north	12	3.812	mm; W/m²K
Roof	Entire building	300	0.1176	mm; W/m²K
Roof	Staircase steel plate	0.5	5.882	mm; W/m²K
Ground slab	Main building	305	0.1721	mm; W/m²K
Ground slab	Staircase and external store	185	0.4034	mm; W/m²K
Internal wall	Test cell	120	0.285	mm; W/m²K
Internal wall	Climatic chamber and other spaces	144	0.271	mm; W/m²K
Internal slab	Daylight zones	220	0.279	mm; W/m²K
Internal slab	Calorimetric zones	120	0.285	mm; W/m²K
Internal slab	Measurement room, toilet and store	22	2.95	mm; W/m²K
Internal ceiling	Test cells	120	0.285	mm; W/m²K
Basement / ground wall	South	370	0.219	mm; W/m²K
Basement / ground wall	North	450	0.383	mm; W/m²K
Basement / ground wall	East/west	470	0.134	mm; W/m²K
Window	Windows including frame	—	1.6	W/m²K
Door	External/internal doors	—	1.058	W/m²K
Infiltration	Pressure test assumption	—	1	ACH at 50 Pa
Thermal bridging	Global adjustment	—	15	% adjustment

Table C.2: Zone areas included in the simulation models

Zone category	Zone	Area	Unit
Heated zone	101 – Entrance	7.74	m²
Heated zone	102 – Measurement Room	23.25	m²
Heated zone	103 – Toilet	4.84	m²
Heated zone	104 – Store	3.50	m²
Heated zone	106 – Daylight 1	10.80	m²
Heated zone	107 – Daylight 2	10.80	m²
Heated zone	108 – Calorimetric 1	10.80	m²
Heated zone	109 – Calorimetric 2	10.80	m²
Heated zone	116 – Solar Central & HVAC	36.50	m²
Heated zone	110 – Climatic Chamber, Measurement Boxes and Corridor	111.50	m²
Total heated floor area	**All heated zones**	**230.53**	**m²**
Unheated zone	201 – Stair Room	8.83	m²
Unheated zone	115 – External Store	17.36	m²
Unheated zone	116 – Crawlspace	40.77	m²
Total unheated area	**All unheated zones**	**66.96**	**m²**

Table C.3: HVAC, ventilation and control assumptions

Category	Parameter	Value	Unit / note
Heating/cooling system	Ground-source heat pump	Applied across tools	Common system assumption
Heat-pump performance	Coefficient of performance	4.2	COP
Air-side system	AHU with VAV control	Applied across tools	Occupancy controlled
Heating setpoint scenario 1	Heating setpoint	18	°C
Heating setpoint scenario 2	Heating setpoint	21	°C
Heating setpoint scenario 3	Heating setpoint	24	°C
Cooling setpoint	Fixed cooling setpoint	26	°C
Mechanical ventilation	Area-based ventilation rate	0.35	L/s·m²
Mechanical ventilation	Occupant-based ventilation rate	7	L/s/person
Heat recovery	Heat recovery efficiency	65	%
Natural ventilation	Operation	Always off	Applied across scenarios
Humidity control	Operation	Always off	Applied across scenarios

Table C.4: Internal loads and occupancy assumptions

Category	Zone / parameter	Value	Unit / note
Lighting	Area-normalised annual electricity	8.5	$kWh \cdot m^{-2} \cdot a^{-1}$
Equipment	Area-normalised annual electricity	5.3	$kWh \cdot m^{-2} \cdot a^{-1}$
Hot water	Area-normalised annual electricity	0.6	$kWh \cdot m^{-2} \cdot a^{-1}$
Occupancy	Entrance	0.061	persons/m²
Occupancy	Measurement Room	0.061	persons/m²
Occupancy	Toilet	0.001	persons/m²
Occupancy	Store	0.001	persons/m²
Occupancy	Daylight 1	N/A	No occupancy modelled
Occupancy	Daylight 2	N/A	No occupancy modelled
Occupancy	Calorimetric 1	N/A	No occupancy modelled
Occupancy	Calorimetric 2	N/A	No occupancy modelled
Occupancy	Solar Central & HVAC	0.061	persons/m²
Occupancy	Climatic Chamber, Measurement Boxes and Corridor	0.061	persons/m²

Table C.5: Weather and ground-temperature assumptions

Category	Parameter / month	Value	Unit / note
Weather file	Lund 2023	Sveby reference file	Applied across all tools
Ground temperature	January	5.40	°C
Ground temperature	February	2.98	°C
Ground temperature	March	2.32	°C
Ground temperature	April	2.89	°C
Ground temperature	May	6.16	°C
Ground temperature	June	9.83	°C
Ground temperature	July	13.27	°C
Ground temperature	August	15.77	°C
Ground temperature	September	16.49	°C
Ground temperature	October	15.32	°C
Ground temperature	November	12.50	°C
Ground temperature	December	8.94	°C

Table C.6: Output categories used for comparison

Category	Output variable	Temporal resolution	Note
Whole-building electricity	Total simulated electricity consumption	Annual and monthly	Compared across tools and against measured electricity use
Heating	Heating electricity consumption	Annual and monthly	Used to investigate IES-VE divergence
Cooling	Cooling electricity consumption	Annual and monthly	Used to compare seasonal cooling disagreement
Fan energy	Fan electricity consumption	Annual and monthly	Used to investigate HVAC auxiliary-energy differences
Lighting	Lighting electricity consumption	Annual	Used as a consistency check across tools
Equipment	Equipment electricity consumption	Annual	Used as a consistency check across tools
Hot water	Domestic hot-water electricity consumption	Annual	Used as a consistency check across tools
Measured electricity	Actual electricity consumption	Annual and monthly	Used as the reference series for measured-versus-simulated comparison

The inputs reported in Appendix C represent the harmonised assumptions used to develop the four simulation models. Some implementation differences remained unavoidable because each tool uses different modelling interfaces, default calculation procedures, HVAC abstraction methods and output definitions. Therefore, the appendix supports transparency and reproducibility of user-defined assumptions, but it does not imply that the internal algorithms of the four tools are identical.