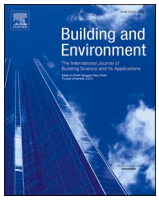
Building and Environment 264 (2024) 111903

Contents lists available at ScienceDirect

Building and Environment

journal homepage: www.elsevier.com/locate/buildenv



Ten questions concerning statistical data analysis in human-centric buildings research: A focus on thermal comfort investigations

Matteo Favero [a,*], Salvatore Carlucci [b], Giorgia Chinazzo [c], Jan Kloppenborg Møller [d], Marcel Schweiker [e], Marika Vellei [f,g], Andrew Sonta [a]

[a] ETHOS Lab, School of Architecture, Civil and Environmental Engineering (ENAC), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[b] Department of Theoretical and Applied Sciences, University of Insubria, Varese, Italy
[c] Department of Civil and Environmental Engineering, Northwestern University, Evanston, USA
[d] Department of Applied Mathematics and Computer Science, Technical University of Denmark, DTU, Lyngby, Denmark
[e] Healthy Living Spaces lab, Institute for Occupational, Social and Environmental Medicine, Medical Faculty, RWTH Aachen University, Pauwelsstr. 30, 52074, Aachen, Germany
[f] Univ. Bordeaux, CNRS, Bordeaux INP, I2M, UMR 5295, F-33400, Talence, France
[g] Arts et Metiers Institute of Technology, CNRS, Bordeaux INP, Hesam Universite, I2M, UMR 5295, F-33400, Talence, France

ARTICLE INFO

Keywords:
Thermal comfort
Human-centric research
Statistical data analysis
Simulations
Causal thinking
Statistical thinking

ABSTRACT

Given the large amount of time we spend indoors, designing and operating buildings that are safe, comfortable, and conducive to productivity and well-being is essential. To achieve this goal, in the past decades, research has been conducted to investigate the influence of the indoor environment on occupants. Thermal comfort has been the subject of most investigations in this field. However, despite being a consolidated research topic since the 1920s, statistical practices for analysing thermal comfort data often rely on simplified premises, which may be due to several possible factors (e.g., limited computational capabilities and lack of training). Consequently, important aspects of data analysis are often absent or overlooked. Recent statistics and statistical software advances have provided more options for effectively modelling complex issues. However, properly using these tools requires a solid understanding of statistical analysis, increasing the risk of misuse in practice. This paper presents ten questions highlighting the most critical issues regarding statistical analysis for thermal comfort research and practice. The first four questions provide general perspectives concerning statistical data analysis, while the remaining ones address specific problems related to thermal comfort research, but that can extend to all human-centric research in the built environment. Additionally, the last five questions demonstrate the practical significance of analysis pitfalls (i.e., sampling variability, selection bias, variable selection, clustered/nested observations, and measurement error) through examples with synthetic data. This study provides insights into the current statistical 'habits' in thermal comfort research and, more importantly, help researchers better define and conduct their statistical analyses.

1. Introduction

People in developed countries spend a significant portion of their lives inside buildings. Therefore, it is necessary to design and operate buildings that are safe, comfortable, conducive to productivity, and aspire to improve occupants' health and well-being. Several parameters influence how the built environment affects its occupants. Among them, thermal comfort has been the subject of numerous research investigations in the past century, leading to the compilation of international standards (e.g., ASHRAE 55:2020 [1], ISO 7730:2005 [2] and EN 16798–1:2019 [3]) to guide professionals in designing and maintaining thermally comfortable indoor climates. Evaluating thermal comfort involves first defining the indoor thermal parameters and, subsequently, quantifying their influence on the occupants, often through statistical analyses.

In thermal comfort investigations, an often-used technique is regression analysis. This kind of analysis can be tied back to the approach introduced by Bedford's research in the 1930s ([4] cited in Ref. [5]). Bedford was the first to apply multiple regression to heating and ventilating research to improve the prediction of subjective human

* Corresponding author. Passage du Cardinal 13b, CH-1700, Fribourg, Switzerland.
E-mail address: matteo.favero@epfl.ch (M. Favero).

https://doi.org/10.1016/j.buildenv.2024.111903
Received 4 March 2024; Received in revised form 26 July 2024; Accepted 28 July 2024
Available online 5 August 2024

thermal responses from environmental data. Several years later, probit analysis, an approach for analysing binary data, was also applied to thermal comfort data derived from laboratory experiments (e.g., in 1953 by Chrenko [6] cited in Ref. [5]) and field studies (e.g., in 1959 by Webb [7] cited in Ref. [5]). This method can be easily applied to the ASHRAE 7-point thermal sensation scale, And, indeed, Fanger used a similar approach to derive the predicted percentage of dissatisfied (PPD) [8]. However, before the advent of computers, probit analysis was a time-consuming process, since each calculation had to be iterated until the result converged on the final value. Additionally, based on McIntyre's work in 1978 [9], ordinal data measured on the ASHRAE scale has been treated as a continuous variable, which is also legitimised in the ISO 10551:2019 [10]. Even though the arguments used to justify this approach are disputable (as discussed in Ref. [11]), linear regression is still nowadays applied to subjective thermal comfort data (e.g., thermal sensation vote) measured on an ordinal scale. This is an example of the lack of updated practices, demonstrating that important aspects of data analysis are often missing or overlooked. Recent developments in statistics and the increasing availability of statistical software have given researchers more options to model complex issues. Examples of methods are beta regression, which models continuous but bounded data, and multilevel models, which deal with data with a hierarchical or clustered structure (and therefore, dependent observations). However, to leverage these tools effectively, it is crucial to have a robust understanding of statistical analysis. If not utilised skilfully, these tools can pose a higher potential for misuse. Stark and Saltelli [12], referring to the poor practice of statistics as '*cargo-cult statistics*', emphasise that plug-and-play data-analysis software encourages the ritualistic miming of statistics rather than conscientious practice. They also state that many statistics courses, especially those for non-specialists, teach mechanical calculations without due consideration for scientific context, experimental design, assumptions, limitations, or interpretation of results. Statistics is more than just tools and mathematical formulae; it is an evidence-based *way of thinking*. While this discipline comes with a toolbox containing many different tools, knowing the contents of a toolbox requires what has been called *statistical thinking*, that is, '*the art of choosing a proper tool for a given problem*' [13].

Poor statistical practice can dramatically increase the probability of incorrectly published findings [14]. Munafò et al. [15] highlighted various threats to the efficiency of knowledge accumulation and science's ability to self-correct, such as lack of replication [16], hypothesising after the results are known (HARKing) [17], poor study design, low statistical power [18], analytical flexibility [19], *p*-hacking [20], publication bias [21] and lack of data sharing [22]. The book by Humphreys, Nicol and Roaf [5] offers valuable advice for statistical data analysis related to thermal comfort, mainly focusing on the adaptive approach. They discuss, among other topics, linear and probit regression and the presence of measurement error in regression analysis. The authors justify adding four chapters on statistical methods in their book by stating that although many excellent books in the statistical literature exist, in their experience, '*highly intelligent people without a mathematical education find them impenetrable*', highlighting a lack of statistical training in the field. In our experience, this aspect is also reflected in the thermal comfort literature —where statistical concerns are often relegated to the limitations section of an article or overlooked altogether— with only a few papers addressing statistical issues. For instance, Humphreys and Nicol [23] discuss the impact of measurement and formulation errors on thermal comfort indices, Sun et al. [24] demonstrate how causal reasoning can reveal hidden assumptions and interpretations of statistical analysis, while Pan et al. [25] explore common methodological pitfalls for causal inference in the field of cross-modal research (e.g., causal and predictive research, generalisability, measurement error, and violation of statistical assumptions). Moreover, the growing emphasis on personalised comfort models [26] has driven an increase in the utilisation of machine learning (e.g., Ref. [27]) and data-driven algorithms (e.g., Ref. [28]) to predict individual comfort responses. Therefore, as reliance on data-driven models for guiding decisions related to environmental control systems and user comfort grows, enhancing statistical learning in thermal comfort becomes imperative. However, while previous studies shed light on individual statistical concerns in thermal comfort data analysis, comprehensive guidance is lacking.

This paper aims to share critical insights from experts in human-centric building research regarding improving statistical analysis in the field. We focus our efforts on the statistical analysis of existing data, that is, already available data. As such, while important, specific discussions about experimental design (e.g., sensor installation, HVAC control, survey design and administration) and research methods at large are outside this article's scope. The ten questions answered here are chosen to lead the reader through a holistic and critical reflection on the current typical application of statistical analysis to thermal comfort data, not prescribing statistical methods. The first three questions in the manuscript provide an overview of the fundamental concepts of defining and understanding the question being asked, choosing a framework to answer the question, and (the ever-present but rarely considered) causal thinking. The following two questions describe regression analysis and how it can be extended to model different aspects of a data set. These questions emphasise this powerful tool's full utilisation and proper application, especially in thermal comfort studies. Subsequently, the last five questions demonstrate, through examples with synthetic data, the tangible effects on data analysis of more specific issues, such as sampling variability, selection bias, variable selection, clustered/nested observations, and measurement error. Although real-world data are available (e.g., ASHRAE Global Thermal Comfort Database II), we decided to rely on synthetic data as they allowed us to highlight the implication of selected issues (and only one at a time) in the data analysis, which is impossible to obtain with real data. Specifically, these examples use a 'true' model to generate a synthetic data set, which is then analysed using a correctly and incorrectly specified statistical model to highlight their implications for results and conclusions. This could not have been accomplished with real-world data, which is prone to being affected by multiple issues simultaneously. To foster an understanding of the issues discussed in the paper and the related simulation examples, we created a supplementary online document.[1] This document provides a step-by-step guide to the simulated examples and offers more details on the problems addressed in the paper.

2. Question 1: What is the data analytic question type?

Any data analysis starts with a question, which should be specific and focused. In formulating the question, it is important to understand the goal of answering it, as this plays a significant role in interpreting the results. Although this statement seems trivial and obvious (especially in a research context), improperly specifying the question is a common error. According to Leek and Peng [29], the most common mistake in data analysis is misidentifying the question being addressed. This confusion is central to the replication crisis, distorted press releases describing scientific results, and controversial claims of false research findings. Leek and Peng broadly classified the different types of data analysis into six types: (i) descriptive, (ii) exploratory, (iii) inferential, (iv) predictive, (v) causal and (vi) mechanistic. Each of these six basic types of data analysis has different goals, which are described along with thermal comfort examples in Fig. 1. Some of the mistakes are so common that Leek and Peng [29] coded them in standard phrases. For example, '*correlation does not imply causation*' describes an inferential question mistaken for a causal question or 'data dredging' to describe an

[1] https://mfavero.quarto.pub/10q-simulation-examples/.

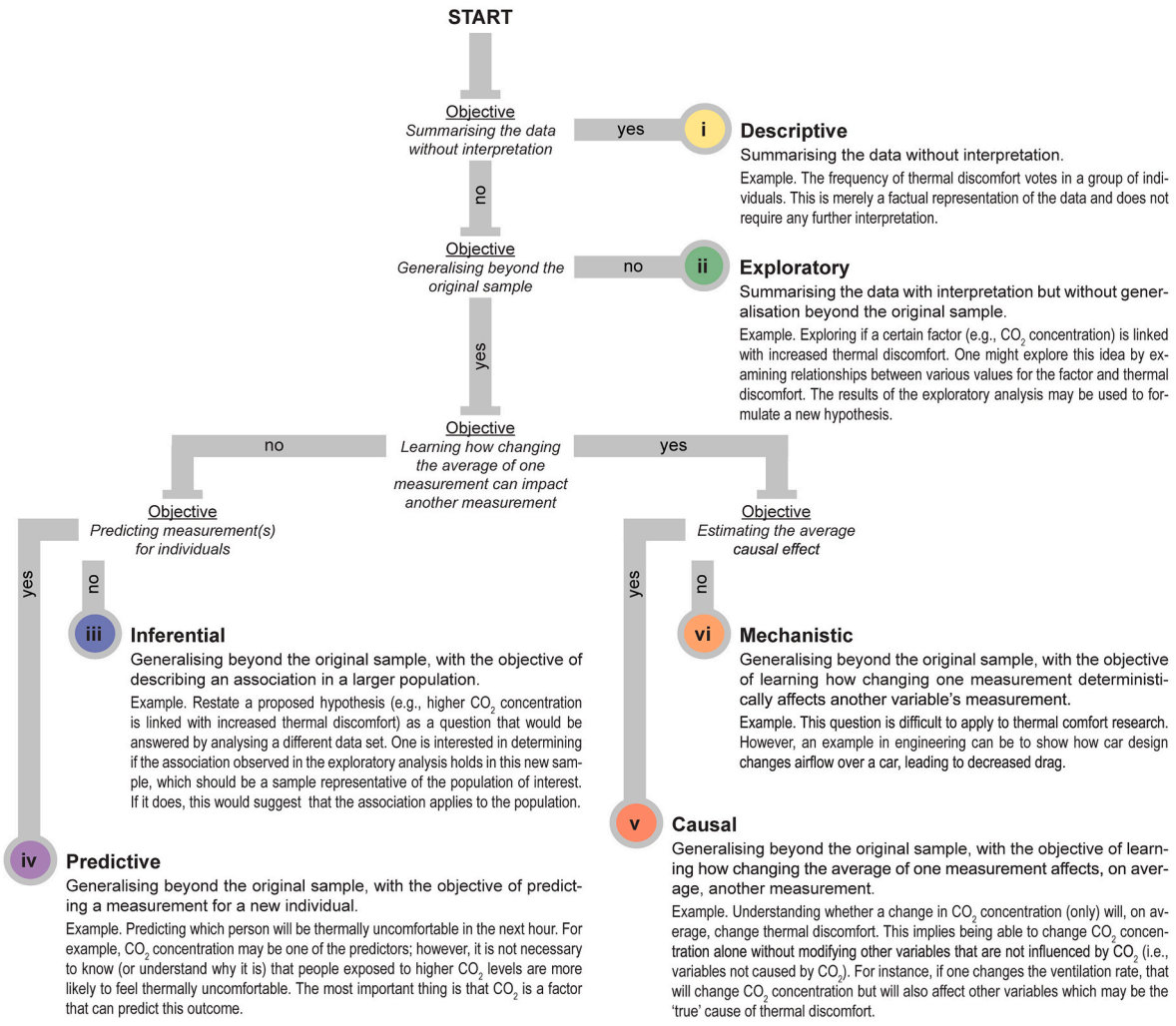


Fig. 1. – Broad classification of objectives for different analysis types (adapted from Ref. [29]). (Note. A 'mechanistic' model is a 'causal' model. Here, the two terms are used to distinguish the goal of determining a deterministic (i.e., 'mechanistic') effect *versus* an average (i.e., 'causal') effect).

exploratory question perceived as inferential.

To a similar view is Shmueli [30], who distinguishes three types of modelling[2] to reflect the distinct scientific goals they aim at. Specifically: (a) descriptive modelling, (b) explanatory modelling, and (c) predictive modelling. The author highlights that although explanatory and predictive modelling are often confused, their objectives differ and significantly impact each step of the statistical modelling process and its consequences.

In thermal comfort research, it is not a trivial endeavour to determine how many times the misunderstanding of the data analytic question type occurs, since it is uncommon for a thermal comfort study to indicate precisely which type of analysis it performs. While identifying the question being asked can sometimes be inferred by the analysis performed, it can become a complex task, especially when numerous analyses are conducted within a single study or on the same data set. However, in the current literature, it is possible to identify incorrect interpretations of the data analytic question type (i.e., using language and analysis tools for a different purpose than the intended one). For example, it is common to see association-based statistical models (e.g., regression) applied to thermal comfort data (collected from either field or lab studies), which are then interpreted using causal terminology (e.g., *x* affects *y*). If the objective is not causal inference, it is inappropriate to label regression coefficients as 'effects' (however, they can always be interpreted as average comparisons (page 85 of Ref. [31]). This confusion may arise because causal and associational concepts are sometimes conflated. A helpful distinction between causal and associational concepts can be found in Pearl [32].

Different data analysis questions have different objectives, which affect statistical modelling and the conclusion that can be drawn. Therefore, in future thermal comfort studies, it would be crucial to accurately label each step according to its original purpose to depict the data analysis precisely. In addition, unless one intends to collect the data needed for the study (e.g., through a new experiment), the available data determines the type of questions that can be posed. Consequently, **having data is not enough; having the right data for the question at hand is imperative**. This dependency may be especially relevant for routinely collected data gathered without specific *a priori* research questions or when analysing data from repositories and databases.

3. Question 2: Why does the interpretation of probability matter?

An editorial in *Science* defined statistics as '*the science of learning from data, and of measuring, controlling, and communicating uncertainty*' [33], a definition also used by the American Statistical Association (ASA), the

[2] Shmueli [30] intentionally chose the term 'modelling' over 'models' to highlight the entire process involved, from goal definition, study design, and data collection to scientific use.

world's largest community of statisticians [34]. To understand and quantify uncertainty, statisticians (and mathematicians) use probability. Although they (largely) agree on what the laws of probability are, there is far less agreement on what the term 'probability' actually means. In his book 'The Foundations of Statistics', the mathematician and statistician Leonard J. Savage (page 2 of Ref. [35]) stated that there is general agreement on the purely mathematical properties of probability, but controversy arises when interpreting the axiomatic concept of probability, that is, determining its extra-mathematical properties. Savage referred to the well-known probability axioms introduced by the mathematician Andrey Kolmogorov in 1933 [36], which defined the properties of probability, but it did not address the questions of how probabilities ought to be understood in the real world or where they came from. According to the mathematician Aubrey Clayton (page 47 of Ref. [37]), any method of assigning numbers to subsets of a sample space that satisfies Kolmogorov's axioms can be called a probability, and all the mathematical properties follow automatically. Therefore, the disagreement arises from the 'freedom' in interpreting probability.

In statistical inference, there are three main and distinct 'branches': (i) likelihoodist, (ii) Bayesian and (iii) frequentist. These three approaches use the likelihood function,[3] which is central to estimating unknown parameters, but they use it differently and for different purposes. The likelihoodist approach involves presenting data as evidence, for pairs of simple statistical hypotheses. It does so by using the Law of Likelihood: the evidence x (the data or observable variable) supports one hypothesis θ_1 (the parameter or unobservable variable) against another θ_2, if and only if the likelihood ratio (i.e., the ratio of their likelihoods) is greater than 1 (and *vice versa* if less than 1). If this ratio is 1, the evidence is indifferent. On the other hand, the Bayesian approach updates probabilities using the likelihood function rather than treating it as a standalone object of interest. In light of the evidence x (the data or observable variable), following Bayes's theorem, the degree of *a priori* belief in any proposition θ (the prior probability of the parameters or unobservable variables) is updated. This results in a posterior probability: the probability of a hypothesis (the parameter or unobservable variable) given the evidence (the data or observable variable). The frequentist approach is quite different from the two previous approaches. The main objective is to create procedures with long-run frequency guarantees. As such, they do not evaluate the epistemic states of individual hypotheses but offer guidance regarding long-run error rates on decisions about how to behave regarding hypotheses, no matter what the truth may be. As Neyman and Pearson (page 74 of Ref. [38]) stated, '*without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong*'.

Royall (page 122 of Ref. [39]) makes a valuable distinction between the three branches by comparing them in terms of answering three separate questions:

1. 'How should I interpret this body of observations as evidence?' (likelihoodist)
2. 'What should I believe?' (Bayesian)
3. 'What should I do?' (frequentist)

The important aspect is that one approach cannot answer all these questions. Several tools are available in the statistical toolbox, and no tool is superior to the others.

3.1. A frequentist overview

The frequentist framework is commonly used in thermal comfort research. As such, it will be further described here. Within the frequentist branch, there are two frameworks: Fisher's null hypothesis testing (based on p-value) and the Neyman–Pearson decision theory (based on statistical power, type I error (α), type II error (β)). These two frameworks are different and entail distinct ideas —as clearly expressed by the different views of their creators [40]. However, they have been incongruently combined in what is nowadays known as 'null-hypothesis significance testing' (NHST). For more details, the reader is referred to Refs. [13,41].

Perhaps unsurprisingly, many misconceptions regarding p-values, confidence intervals, and power can be found in the scientific literature. Greenland et al. [42] summarise and clearly explain many false beliefs concerning p-values, confidence intervals, and power in a guide to avoid and spot misinterpretations. Concerning p-values, ASA released a statement [43] including principles underlying its proper use and interpretation. However, it is important to emphasise that criticisms of the p-values (e.g., their dependency on the investigator's sampling intentions) are actual properties of the framework from which they originated. As such (valid) p-values behave exactly as they should [44]. In addition, mistakes in what can be concluded from p-values (e.g., by setting the p-value threshold at .05 as sufficient evidence to reject hypotheses), are not properties of p-values but a researcher's choice. As Fisher (page 42 of [45]) stated, '*no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas*'. No tool is immune to misinterpretation and misuse, and regardless of the specific tools chosen, it is its inappropriate or erroneous application that harms science.

In thermal comfort research, as for the data analytic question type, understanding whether the framework used (e.g., frequentist) is a conscious choice or 'cargo-cult statistics' is hard to define. However, **it is the researcher's responsibility to select and justify the framework used to answer the specific question(s) of interest**. Given all the issues above, a general suggestion in thermal comfort research may be to shift the focus from tests to estimates. This would involve using point estimates, such as effect size, and interval estimates, such as confidence intervals, along with precise p-values (not whether p-values are above or below some thresholds). Importantly, the selection of effect size should not be based on statistical significance. Especially in low-power studies (i.e., low signal and high noise), statistically significant results are subject to type M ('magnitude') and type S ('sign') errors, where type M error is the factor by which the magnitude of an effect might be overestimated, and type S error is the probability of an estimate being in the wrong direction [46]. In addition, statistics terminology can be highly confusing, often using common words in technical ways that bear little resemblance to their conventional meanings. Words like likelihood, significant, and confidence have vastly different connotations in statistical contexts. To avoid inappropriate use of terminology, it is advisable to refer to a (valid) glossary of statistical terms (e.g., Refs. [47,48]).

[3] The likelihood function $\mathscr{L}(\theta|y)$ (often denoted as $\mathscr{L}(\theta)$ or $\mathscr{L}(y|\theta)$) is a function of θ (the parameter or unobservable variable) with y (the data or observable variable) fixed. This function, evaluated only for the specific y observed, is examined to see how it varies as θ is changed. It is essential to mention that the likelihood is not a probability —Fisher called it 'likelihood' to emphasise this critical distinction— since it does not obey Kolmogorov's axioms (e.g., it does not need to integrate/sum to 1).

4. Question 3: What is 'causal thinking', and why should it not be overlooked?

The phrase '*correlation does not imply causation*'[4] is well-known in research. An inflated mention of this sentence has led to a virtual ban on causal talk over the years. Fortunately, the 'causal revolution' instigated (in large part) by Judea Pearl completely reshaped the causality debate, and now the use of causal inference can also be found in the building design process (e.g. Ref. [50]). Recently, Sun et al. [24] have shown how causal thinking can be used to uncover hidden assumptions and interpretations of statistical analysis in building science, which echoes the importance of causality discussed in the related field of cross-modal research by Pan et al. [25]. Indeed, causal thinking —the process of identifying the relationship between a cause and its effect— is not limited to causal inference but underlies all types of data analysis. Even seemingly straightforward descriptive analysis (which aims to summarise the data without interpretation, see Fig. 1) requires it. Generally, in a descriptive question, the sample is not the target; the population is. Understanding how the sample differs from the population might allow for extrapolation from one to the other. However, to do so, one needs to understand certain aspects of the sample, such as what produced the data and whether the process was biased, which was the missing data process (if any), etc. Identifying biases and the causes of missingness requires causal thinking, even though the goal is not causal inference. In addition, it is common to hear researchers refer to '*modelling the data*' instead of '*modelling the data-generating process*'. Ideally, the ultimate aim of science is to model the phenomenon under scrutiny and not fitting curves to data. **The data are a means of obtaining information about that phenomenon**. Keeping causality in mind can be one way to remember this.

An example of the utility of causal thinking in thermal comfort is in the prediction versus prevention of thermal discomfort. Identifying whether a person is at risk of discomfort (i.e., prediction) is different from identifying the (best) strategy to prevent a person from being uncomfortable (i.e., prevention). The two are both forms of prediction, but while the latter requires causality (to prevent thermal discomfort, it is necessary to intervene on its causes), the former does not (to predict thermal discomfort, it is sufficient to use variables associated with it). **In a building control logic that aims to maintain thermal comfort conditions, it is paramount to have a model designed for prevention and not just prediction. In other words, a causal model is needed for building control logic.** This is based on the fact that causes are not in the data (i.e., data cannot speak for themselves) but derive from external knowledge, and any causal conclusion (e.g., an answer to a 'what if?' question) must be based on some causal assumption —'*no causes in, no causes out*' [51]. For example, this is how any building performance simulation (BPS) tool works. BPS tools are computer-based mathematical models in which fundamental physical principles encode causality. As such, they can answer 'what if?' questions (e.g., if the thermal transmittance of a building's roof is changed, what will the building's energy consumption be?).

One effective way to imbue causal thinking into a problem is to use graphical models [52]. Directed acyclic graphs (DAGs) are examples of graphical models in which letter-and-arrow pictures summarise our existing scientific knowledge. The letters represent the variables (the quantities of interest), while the arrows indicate the established or suspected causal linkages between these variables —namely, which variable 'listens' to which others [53]. Fig. 2 shows, in a DAG format, three elemental relations that form the building blocks of a DAG [54].

The uses of DAGs are manifold. To begin with, they are an excellent tool for communication. DAGs help to organise the expert knowledge visually, clarifying conceptual problems, thus enhancing communication among researchers. For example, an omitted arrow between two variables means no direct influence of one variable on another (e.g., from X to Y in Fig. 2), while a present arrow remains totally agnostic about the magnitude of the effect and its functional form. Additionally, by exposing the chosen scientific model to criticism, a DAG compels researchers to justify the scientific model's choice, be specific in the question they pose, and be open about the assumptions they are willing to make. Clarifying these matters upfront can enable a productive debate among anyone with domain knowledge, even without the relevant specific statistical knowledge. Translating the scientific model into a statistical model comes later and requires additional/different expertise. Finally, researchers can use DAGs as an inferential tool, allowing them to (i) estimate effect sizes even when working with non-experimental data [55–59], (ii) derive testable implications from the assumptions they make [55,60,61], and (iii) test for external validity (i.e., generalisability) [62]. In addition, DAGs have been useful in describing typical biases (e.g., Refs. [63,64]), finding adjustment variables (e.g., Ref. [65]), and elucidating apparent paradoxes (e.g., Simpson's paradox in Refs. [66,67]).

Causal thinking is ever present in scientific and thermal comfort research, even if the goal of the work is not causal inference. Being clear about one's scientific model, in a way that clearly represents causal influence, helps to avoid statistical pitfalls in subsequent analysis.

5. Question 4: What is regression analysis, and how to properly apply it in thermal comfort studies?

Regression analysis is '*the blanket name for a family of data analysis techniques that examine relationships between variables*' [68], which can be called, in general and 'neutral' terms,[5] regressand(s) —$Y(s)$— and regressor(s) — $X(s)$. Here, regressand can be plural because it is possible to model more than one Y (e.g., thermal preference and thermal acceptability) simultaneously in the same regression model. Models with multiple regressands are known as multivariate, and models with multiple regressors are known as multivariable. Regression analysis is a widely adopted technique in thermal comfort studies. It has been used, for example, to establish a relationship between the thermal environment and human response (in both laboratory and field studies) and to derive thermal comfort models (e.g., the adaptive thermal comfort models [69]). In addition, it is a powerful, flexible, transparent and multipurpose method which, in the context of data analysis (see Fig. 1), can be used to answer diverse data analytic question types (e.g., descriptive, exploratory, prediction, and causal). Certainly, various methods can be used to answer a question of interest, not just regression analysis. Different approaches in the thermal comfort literature, some of which belong to machine learning (ML), can be found. However, given the popularity of regression analysis in thermal comfort research and the fact that it can also be used in some ML approaches (e.g., support vector regression and decision tree regression), improving its understanding

[4] In general, correlation implies association, but not causation; conversely, causation implies association, but not correlation [49]. Although correlation and association are often used interchangeably in everyday speech, they have different meanings in statistical terms. Association is a very general relationship where one variable provides information about another. Instead, correlation implies a specific type of association, that is, an increasing or decreasing trend. For continuous data, Pearson's correlation (which measures linear trends) is an example, while for ordinal data, Spearman's (rank) correlation (which measures monotonic trends) is an example.

[5] In the literature, there are many terms to refer to the Y and X, such as, 'explained' and 'explanatory' variables, 'outcome' and 'covariate', 'dependent' and 'independent' variables, 'output' and 'input' variable, 'predicted' and 'predictor' variable, etc., and some of them can be confusing/poor terminology. For example, 'explanatory' may imply inappropriate causation; 'covariates' may mean that the variables co-vary, which may or may not be the case; 'independent' variables are not assumed to be independent of anything, so the term can be misleading, etc.

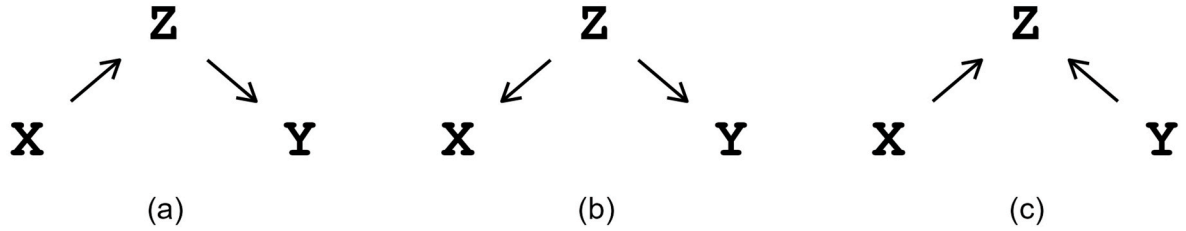


Fig. 2. – DAGs for the three elementary relations: (a) patterns of the form X⟶Z⟶Y are called 'chains' (or mediators); (b) patterns of the form X⟵Z⟶Y are called 'forks' (or common causes); (c) patterns of the form X⟶Z⟵Y are called 'colliders' (or common effects).

and correct application can undoubtedly benefit the field.

In thermal comfort, there is generally a narrow view of regression analysis, which can hinder the full utilisation and proper application of its potential. For example, regression analysis is commonly thought of as 'just' linear regression or its many special cases (e.g., ANOVA, ANCOVA, *t*-test, etc.). Linear regression is the simplest form of regression and belongs to the general linear models (see Fig. 3), where the regressand Y (a continuous and unbounded variable) is modelled given some regressors X, generally assuming a conditional normal distribution of the regressand:

$$\begin{aligned} Y_i &\sim \text{Normal}(\mu_i, \sigma^2) \\ \mu_i &= \eta_i \\ \eta_i &= \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} \end{aligned} \qquad \text{Eq. (1)}$$

where μ_i is the mean, σ is the standard deviation, and η_i is the predictor term function of some regressors $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$. The subscript i is to stress the dependency on the i^{th} observation.

Linear regression analysis relies on various assumptions, which are listed in decreasing order of importance by Gelman et al. (pages 153–155 of Ref. [31]): (i) validity, (ii) representativeness, (iii) additivity and linearity, (iv) independence of errors, (v) equal variance of errors, and (vi) normality of errors. When these assumptions break down, what steps should be taken? While there is no single answer to this question, perhaps the most direct approach is to extend the model. An example would be adding interactions to expand the model to be non-additive or adding splines to capture non-linearities. The extension may sometimes involve a change of model, as shown schematically in Fig. 3. For example, if the independence of errors is not verified in linear regression (e.g., for longitudinal data), the latter can be extended to a linear mixed model. Or, if the regressand Y is measured as an ordinal variable (violating linear regression requirement to have continuous and unbounded data), for example, an ordered multinomial model should be used. Failing to consider dependent errors and using linear regression to analyse either ordinal data or continuous but bounded data are typical issues in thermal comfort, which will be discussed in Q.9 and Q.5, respectively. In general, the assumptions underlying any method are never confirmed, making it vital to identify significant violations. However, while understanding the ideas behind a model is important, it is also essential to recognise that some assumptions may rely on the researcher's knowledge of the subject area and cannot be verified solely through data analysis [31]. As such, engaging with data and understanding its intended use is irreplaceable.

In a regression analysis, the regression coefficients can be conceptualised as effect sizes. Generally, (any) effect size measure should have an intuitive interpretation that provides a 'sense' of the degree of an effect. To enhance the interpretability of regression coefficients, it may be helpful to centre regressors that lack a meaningful zero. This can be achieved by subtracting the mean of the data or selecting a reference point (e.g., 22 °C for air temperature). Centring makes it easier to interpret the term intercept in a context where it makes no sense to consider the regressors set to zero. Standardisation using z-scores is another approach, which involves standardising the regressors by subtracting the mean and dividing by the standard deviation (sometimes it is preferable to divide by two times the standard deviation; see page 187 of Ref. [31]). However, standardisation may not always be advisable [70]. It is fundamental to remember that a direct interpretation of the regression coefficient on the scale of the data is not always possible. In linear regression, a coefficient β is the expected difference in the regressand Y, by comparing two observational units (e.g., people) that differ by one unit in the regressor x with all other regressors held constant. This interpretation is not possible when the functional form (i.e., the relationship between regressand and regressors) is non-linear. This is the case for generalised linear models (GLMs) because, on the regressand scale Y, all regressors interact with each other (also with themselves) even if the interaction term is absent. An example is logistic regression, which is linear in the parameters (i.e., the logit scale) but non-linear in the functional form (i.e., the probability scale). As such, a specified difference in one of the regressors x does not correspond to a constant difference in $\Pr(y = 1)$. A different approach to interpreting logistic regression coefficients is by using odds ratios. However, the notion of odds can be tricky to comprehend, and odds ratios can be even more confusing. Graphing the fitted model (by making separate plots as a function of the regressors of interest and holding the other regressors constant at different representative values) is a good strategy to overcome potential interpretation issues.

Sometimes, the same (or very similar) regression models can be referred to by very different names (e.g., ordered multinomial model with the logit link is known as multinomial logistic regression, polytomous logistic regression, multinomial logit, softmax regression, etc.), hindering their use and comprehension. To deal with this ambiguity, including a mathematical (or algorithmic) description of the model (e.g., Eq. (1) for the linear regression) may be helpful. **Being transparent and specific about the model formulation and modelling steps is essential**. In the context of thermal comfort, it is common only to present the final model, but doing so may result in overlooking statistical and modelling issues. A possible solution is to include all the modelling steps as an appendix or supplementary material.

For applied and conceptual issues in regression, the reader is referred to Harrell's book [71]. For a non 'conventional' book on regression (i.e., a non-mix of cookbook instruction and mathematical derivation), the reader is referred to Gelman et al. [31], which also provide, in its Appendix B, '*Ten quick tips to improve your regression modelling*'. For an application of probit regression to analyse thermal comfort data, the reader is referred to Ref. [5].

6. Question 5: Should distinct types of response variables be handled differently?

One of the goals of thermal comfort research is to increase knowledge of the relationship between the thermal environment and human response. The human response is typically assessed by a subjective evaluation of the thermal environment using rating scales (e.g., thermal sensation votes, TSVs). The ordinal scale is commonly employed, which will result in ordinal data. Ordinal data is categorical data that have naturally ordered categories (e.g., 'cold' < 'cool' < … < 'warm' < 'hot'),

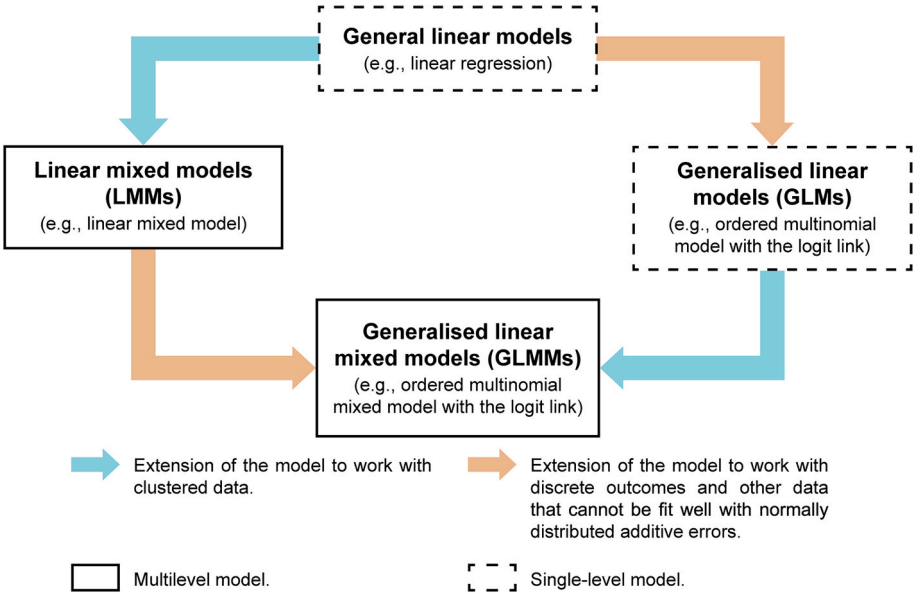


Fig. 3. – Schematic overview of different model categories for regression analysis.

but the distances between these categories are unknown (examples in thermal comfort can be found in Fuchs et al. [72] and Schweiker et al. [73]). Despite this, it is common in the literature to consider ordinal data as continuous (i.e., ordinal-as-metric) and analyse it with methods that require continuous data (e.g., linear regression). However, this practice may lead to severe errors in inference, such as inflated error rates, distorted effect-size estimates and other issues [74]. As mentioned in Q.4, a suitable approach to model ordered categorical data is to use, for example, an ordered multinomial logit or probit model. These are GLMs and, specifically, they are extensions of logistic and probit regressions for (ordered) categorical data with more than two options. These models generally have a 'latent' interpretation. The idea is that the regressand is the categorisation of a latent (not observable) continuous variable, which is assumed to follow a specific distribution. For example, if a logistic distribution is taken, this implies an ordered multinomial model with a logit link. Mathematically, this can be expressed as follows:

$$\begin{aligned} \mathbf{Y}_i &\sim \text{Multinomial}(n, \pi_i) \\ \text{logit}(\gamma_{k,i}) &= \tau_k - \eta_i \\ \eta_i &= \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} \end{aligned} \qquad \text{Eq. (2)}$$

where, for the i^{th} observation, π_i is the probability vector $\{\pi_{1,i}, \ldots, \pi_{k,i}\}$, $\gamma_{k,i}$ is the cumulative probability for the k category $(\gamma_{k,i} = \pi_{1,i} + \ldots + \pi_{k,i})$, $\{\tau_k\}$ are the strictly ordered threshold parameters, and η_i is the predictor term (without an intercept), a function of some regressors $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$. If, instead, a normal distribution is assumed for the latent variable, this implies an ordered multinomial model with a probit link. In Eq. (2), the probit function will replace the logit function. For a general discussion of the issue of analysing ordinal data as metric, the reader is referred to Liddell and Kruschke [74]; for a specific treatment of this issue in the context of thermal comfort research, the reader is referred to Favero et al. [11]. For further insight on categorical data analysis, the reader is referred to Agresti's book [75].

Subjective thermal comfort data are also measured using rating scales that are continuous but bounded. This type of data can be affected by ceiling or floor effects. Ceiling and floor effects occur when an observation (or measurement) reaches its scale's highest or lowest point. This means that the observation is censored because the exact value is unknown. The only information available is that the true value is at or above the upper threshold, or at or below the lower threshold. Linear regression is not appropriate in this case because it will overlook the fact that there is an upper/lower limit and treat all observations as actual values. As a result, it will produce biased parameter estimates (see Fig. 4 for an example).

In addition, independently of the presence of ceiling/floor effects, linear regression can produce impossible predictions (i.e., predicted values outside the observable limits) when used to predict subjective thermal comfort data. Nevertheless, this statement does not mean linear

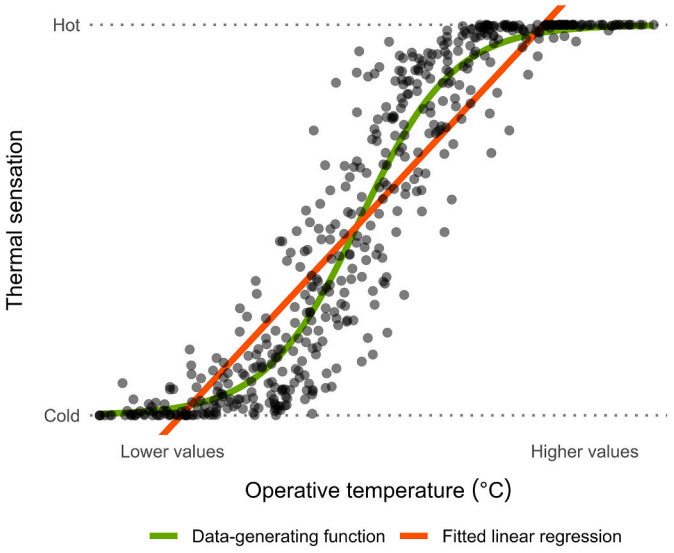


Fig. 4. – Schematic view of ceiling and floor effects. The green and orange lines represent the data-generating function and a fitted linear regression, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

regression cannot generally be applied to bounded data.[6] A suitable approach to model continuous but bounded data is beta regression, a GLM where the conditional distribution of the regressand (i.e., $Y|\mu$) is assumed to follow a beta distribution. The choice of this model is motivated primarily by the flexibility that the assumed beta distribution provides. Depending on the values of the parameters, the beta density can take on a variety of shapes, such as left-skewed, right-skewed, or the flat shape of a uniform density. However, since the probability density function of a beta distribution is defined only on the interval $(0,1)$, the regressand needs to be rescaled. For an application of beta regression to analyse subjective thermal comfort data, the reader is referred to Favero et al. [76]. Another appropriate approach to model bounded data is censored regression. The censored regression model is a generalisation of the Tobit model, proposed by Tobin in 1958 [77], and very popular in econometrics (e.g., Ref. [78]).

Different responses of interest can be measured in several ways (e.g., continuous or categorical), each of which may have a specific set of characteristics (e.g., a lower and/or upper measurement bound). For instance, one could model clothing insulation by assuming a log-normal (or gamma) distribution, which is appropriate since clothing insulation cannot be negative. **Considering the characteristics of the regressand (s) during data analysis is essential to have reasonable estimates of parameters and/or predictions**.

7. Question 6: How does sample size influence thermal comfort studies?

Statistical inference is used to gain insights from incomplete or imperfect data. Typically, a data set is just a sample of the process or population of interest, and the observed data set would differ if the data collection process were redone. The set of possible data sets that could have been observed, along with the probabilities of these possible values, is called a sampling distribution.[7] The sample is assumed to be obtained by randomly sampling data from the process or population, usually as a simple random sample, but other random sampling techniques are possible (e.g., stratified random sample and cluster random sample). For example, multistage cluster sampling (an extension of cluster sampling) could be used for a field study. Assuming that the target population consists of office workers, this sampling strategy will first involve randomly selecting clusters (i.e., office buildings). Secondly, the sample units (i.e., office workers) within the selected clusters will be randomly selected. In practice, buildings and people are seldom selected randomly due to practical constraints, such as geographical proximity, availability at a given time, or willingness to participate in the research. Random samples are important because they possess desirable statistical properties that non-random samples do not have. Examples of a non-random sample are convenience and voluntary response samples (see Q.7).

As mentioned above, an observed data set would differ if the data collection process were redone. As a result, (random) samples are inherently variable. Concerning sampling variability,[8] Cumming (page 143 of Ref. [79]) states that there are two common misconceptions when analysing data or reading research results, specifically:

1. The lack of understanding of the irregularity of randomness in the short term and its high predictability in the long term;
2. The underestimation of the extent of sampling variability.

As a result, it is common to overinterpret aspects of sample data due to sampling variability and underestimate the extent to which results might differ if an experiment is repeated (or if the results of several similar experiments are compared). Tversky and Kahneman ([80] cited in Ref. [79]) conducted a statistical cognition experiment and discovered that researchers tend to underestimate sampling variability and overestimate the chance that repeating an experiment will yield a similar result. They called these misconceptions of randomness and variability the 'law of small numbers' to highlight the wrong idea that small samples behave like very large samples.[9] One example in thermal comfort is the tendency to interpret studies with *p*-values on opposite sides of .05 (i.e., 'statistically significant' vs 'nonsignificant') as conflicting. It is never appropriate to compare *p*-values. To assess differences between studies' results, it is necessary to conduct a formal evaluation, such as estimating and testing those differences (e.g., test of heterogeneity) [81]. **It is important to recognise that sampling variability can affect all studies, even 'well-designed' laboratory experiments**. In general, the advantage of lab over field studies is that the former can greatly reduce different sources of external variability, for example, by randomisation[10] and holding factors constant (e.g., two groups should be made as similar as possible except for the tested condition). Nevertheless, sampling variability will always be present and can be reduced by increasing the sample size. Larger samples will contain less sampling variability, offering a more precise point estimate, and are more likely to be closer to the true population value (assuming no bias). However, it is essential to remember that sample size greatly influences statistical significance, as an inverse relationship exists between sample size and the standard error (i.e., the standard deviation of the sampling distribution). When the sample size is large, even tiny differences in the compared parameters (e.g., the mean of two groups) will be statistically significant. An example can be found in Altomonte et al. [82], where a negligible effect size of air temperature on satisfaction was found to be statistically significant. **Therefore, assessing practical significance (e.g., effect sizes) alongside statistical significance is crucial, as the former is not affected by sample size**. This good practice is still underused in thermal comfort research.

There is often confusion between inferential uncertainty (e.g., uncertainty in the estimate of a population average) and outcome variability (e.g., variation across the population). A high level of confidence that the average outcome for one group is greater than the average for the other should not be considered a statement about the entire distribution of outcomes. Zhang et al. [83] explored this confusion and concluded that researchers often confound the two concepts, which have

[6] An example would be the height of adult humans. Human height has a lower bound at 0, but using linear regression to model adult height can be appropriate because all the observations are 'far' from the lower bound. However, this will not be the case if one is interested in modelling human height at large, which includes infancy, childhood and puberty, not just adulthood.

[7] The sampling distribution is an abstract concept and should not be confused with the sample distribution, that is, the distribution of the observations in the data set.

[8] Sampling variability should not be confused with random sampling. Random sampling refers to selecting a sample from a process or population. In contrast, sampling variability refers to the fact that the statistical information from a sample (i.e., a statistic, such as the mean) will vary as the random sampling is repeated. As the sample size increases, sampling variability decreases.

[9] The law of large numbers assures that the distribution of sufficiently large random sample match the distribution of the underlying population closely.

[10] Randomisation (i.e., random allocation) should not be confused with random sampling. Random sampling is a way of selecting population members for a study's sample. In contrast, random assignment is a method used to place participants into groups, such as control and treatment groups, in an experimental study. The former affects external validity (or generalisation), whereas the latter affects internal validity.

previously been observed among laypeople [84]. In thermal comfort research, for example, a highly precise estimate of the average TSV for two groups (e.g., males and females) can be obtained when the sample size is large enough. However, there will generally be significant outcome variability as individual TSVs within each group will exhibit substantial variability around their respective averages. Outcome variability does not systematically decrease when sample size increases.

Discussions about sample size calculation are outside the scope of this article since its focus is on data analysis and not experimental design. For design and sample size decisions, the reader is referred to chapter 16 of Ref. [31] and, for examples within thermal comfort, Refs [85,86].

7.1. Simulation example: sampling variability

In this simulation example, the highlighted data analysis pitfall is sampling variability. Specifically, it shows the influence of random sampling variability on the results of data analysis. The data is generated based on the process outlined in Fig. 5. This data-generating mechanism simulates the population of interest (constituted by one million observations).

From this population, we obtained three thousand data sets with three different sample sizes using simple random sampling —specifically, one thousand data sets each with sample sizes of 30, 300 and 900. We assumed that the objective of the data analysis is to answer the following inferential question: '*Is there a direct association between biological sex and TSV?*'. To address this data analytic question, the synthetic data sets are analysed using ordinal regression within a frequentist framework. For more details, the reader is referred to the online document; here, only a summary is provided (see Table 1).

Fig. 6 shows the first 100 estimates (point estimate and confidence interval) of the coefficient for sex for the three different sample sizes. The inferential uncertainty shrinks (i.e., narrow confidence intervals) when the sample size increases from 30 to 900 observations. However, it can be seen that, independently of the sample size, some confidence intervals (the ones in red) do not overlap the data-generating parameter (the dashed blue line). If all the thousand simulations are considered, the frequency of the coverage of the calculated confidence intervals (i.e., how many times the confidence intervals overlap the data-generating parameter) is 92.4 %, 95.7 % and 95.8 % for the 30, 300 and 900 observations, respectively. This result is expected because it is a property[11]

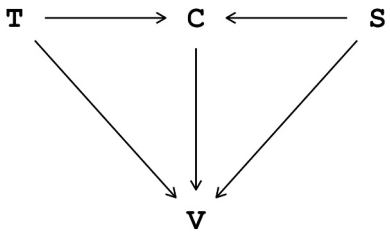


Fig. 5. – Graphical representation via DAG of the data-generating process. Indoor air temperature (T) and biological sex (S) influence TSV (V) both directly and indirectly, passing through clothing insulation (C).

[11] To calculate the 95 % confidence interval, we used the 95 % Wald confidence interval, which is a common way to do this. However, the Wald confidence interval is only guaranteed to be valid in large samples. When the sample size is small, the interval calculated this way may not be valid. In our example, with a sample size of 30, the coverage does not reach its nominal level (i.e., 95 % for a 95 % confidence interval) but is lower (i.e., 92.4 %). There are many alternatives to the Wald interval (e.g., Ref. [87]); however, an in-depth treatment of this issue is beyond the scope of this article. Nevertheless, for a data analysis, this should be considered.

Table 1
Summary description of the simulation example.

Pitfall	Forgetting the influence of sampling variability on the results and interpretation of a data analysis.
Type of analysis	Inferential.
Framework	Frequentist.
Assumptions	Random sample (simple random sampling): everyone in the population has an equal chance of being selected into the sample. Independence of observations: each observation represents independent bits of information. No confounding: the DAG includes all shared causes among the variables. No model error: perfect functional form specification. No measurement error: all variables are measured perfectly.
Variables	Indoor air temperature (T): continuous variable [unit: °C] Thermal resistance of clothing (C): continuous variable [unit: clo] Sex (S): categorical variable ['1' male; '2' female] Thermal sensation vote (V): ordinal variable ['1' cold; '2' cool; '3' slightly cool; '4' neutral; '5' slightly warm; '6' warm; '7' hot]

that a 95 % confidence interval has. However, once a specific confidence interval has been calculated (in our case, the result of one simulation), it is impossible to make any conclusions about the probability of it containing the data-generating parameter. As discussed in Q.2, probability refers to frequency in the frequentist approach. The data-generating parameter is a fixed constant, and a calculated confidence interval is also fixed. Therefore, a given confidence interval either includes the parameter or does not, with no frequency involved. Confidence intervals only quantify the uncertainty due to random error (i.e., sample variability), not systematic error (i.e., bias). According to Gelman et al. (pages 56 of Ref. [31]), the ways to account for sources of errors that are not in the statistical model are: (i) improving data collection, (ii) expanding the model, and (iii) increasing the stated uncertainty. Concerning the last point, some authors (see Ref. [88] and chapter 19 of Ref. [89]) recommend the application of quantitative bias analysis to produce intervals around the effect estimate, taking into account random and systematic sources of uncertainty.

When conducting a study, it is important to recognise that (generally) only uncertain conclusions can be drawn from data, even if statistical significance is achieved. In thermal comfort, claims on data tend to be stated deterministically by placing the results into the 'significant' or 'nonsignificant' bins. However, this practice, called *dichotomania* [90], is harmful because it overlooks the inherent variation in the human response (but more generally in the topic under study) as well as the uncertainty involved in statistical inference. Instead, focusing on uncertainty quantification is likely to result in a reduction of overly confident assertions that, upon further examination, may lack support from the data. Practical examples of this suggestion can be found in Vasishth and Gelman [91].

8. Question 7: Why does participant selection matter in thermal comfort studies?

In laboratory and field studies, defining the population of interest is necessary. A population (or population of inference) refers to a group of individuals who share at least one common characteristic, like geographic location. The subset of such a population with more specific attributes that researchers want to draw conclusions about is the target population. The study population is the portion of the target population who will actively participate in the research [92,93]. In thermal comfort research, for example:

- Population: individuals living in Copenhagen.
- Target population: office workers (20–67 years of age) with no health issues.

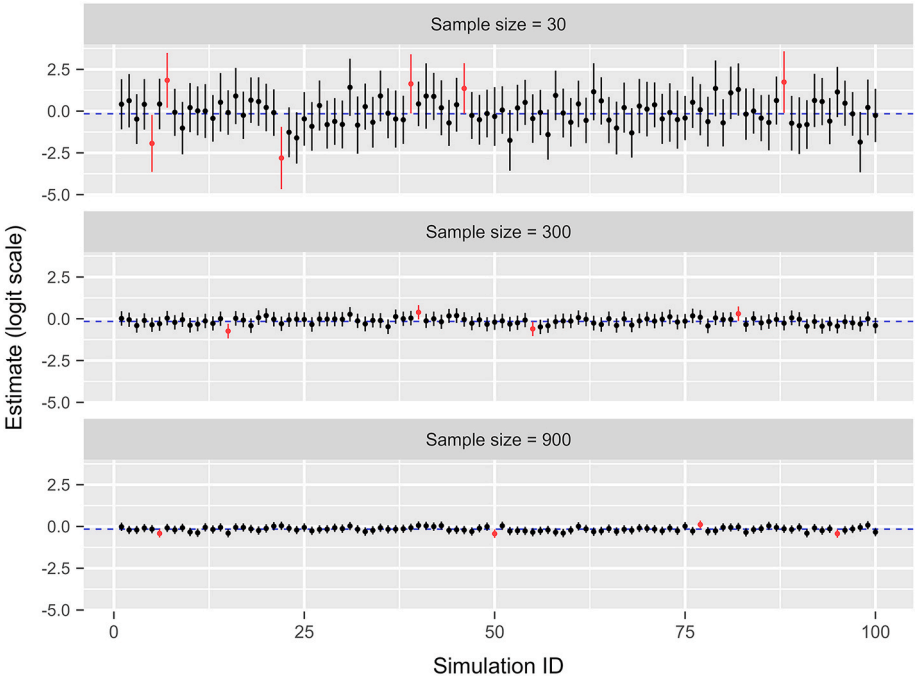


Fig. 6. – Estimates of the coefficient of sex (only the first hundreds of the thousand simulations are shown). The black dots represent the point estimate, and the black lines represent the 95 % (Wald) confidence intervals. In red are highlighted the confidence intervals that do not overlap the parameter used to generate the data (dashed blue line). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

– Study population: participants with chosen characteristics.

Generally, once the population has been selected, two approaches can be followed: (i) include the entire target population in the study (census) or (ii) select a sample of it [94]. Census-based research should be preferred whenever possible in scientific studies [95]; however, due to its cost or other practical constraints, sampling is the common (if not unique) procedure in thermal comfort research. A selection process to obtain the sample (i.e., the study population) is ubiquitous and can lead to accuracy and/or precision biases. The accuracy bias occurs when the selected participants do not represent the target population [94]. Hence, the results derived from the sample are not consistent with those that would have been obtained from the entire population [93]. However, caution should be exerted since statistical inference, linked to representativeness, does not imply scientific inference[12] [95]. Precision (and/or statistical power) bias can occur when the number of participants is less than the minimum necessary [92]. Increasing the number of participants without ensuring the representativeness of the target population will not improve the accuracy of the results; it will only increase the precision (e.g., getting a more precise estimate with a narrower confidence interval) of an inaccurate response [94] or, in other words, a more precise incorrect answer. As discussed in Q.6, increasing the sample size reduces only sampling variability, not bias.

A lack of representativeness can be the result of selection bias. There can be multiple causes for selection bias, as illustrated by Hernán et al. [64], who differentiate it from confounding (i.e., when the factor investigated and the outcome share a common cause). Examples of selection bias include sampling bias, volunteer bias, inappropriate selection of controls in a case-control study, frame coverage bias, size bias and non-response bias [64,93,94,96]. In thermal comfort research, selection bias can be quite common, considering that participant selection is mainly based on voluntary response and convenience sampling. Voluntary response and convenience sampling, together with purposive, quota and 'snowball' sampling, are considered nonprobabilistic (or non-probability) sampling. In this kind of sampling method, in contrast with the probabilistic one in which all participants are equally likely to be selected (i.e., random sampling), the study population does not represent the target population, and the results cannot be generalised [94]. Many thermal comfort studies involve the participation of students or researchers' colleagues due to their convenient accessibility. Voluntary bias could also arise in this context, as people might decide to sign up for a study for specific reasons but might not be representative of the target population [97]. For example, office workers might join a survey of the indoor environment to complain about the temperature in the building. When a single-blind experimental design is sought, it is imperative to ensure that response bias is minimised (e.g., students and colleagues must not be fully aware of the end goal of the research to avoid giving responses that seem correct to them to please the researchers). **Even when response bias is avoided, the selection bias due to convenience sampling can lead to inaccurate responses**. For example, when the target population includes healthy office workers, but only young individuals participate in the study. In this case, it is crucial to define how the study population differs systematically from the target population (e.g., in terms of age, metabolic activity, and clothing preferences) and how these systemic differences can be measured or at least considered. A discussion on inverse probability weighting, a method suggested to correct selection bias in longitudinal studies, can be found in Hernán et al. [64]. Multilevel regression and poststratification, a technique used to adjust for non-representativeness by correcting the results from non-representative samples to the target population for known systematic differences, is discussed by Kennedy and Gelman [98] and applied by Wang et al. [99].

[12] In their work, Rothman et al. [95] distinguish between the scientific objective of comprehending a phenomenon (scientific inference) and the practical aim of utilising that knowledge for specific populations (statistical inference). The former purpose is not improved by representativeness but rather relies on rigorously regulated comparisons made across various pertinent scenarios. It is the latter purpose, the practical application of science, that may require representative sampling.

8.1. Simulation example: selection bias

In this simulation example, the highlighted data analysis pitfall is selection bias. Specifically, it shows the implication of selection bias on the result of the data analysis. The data is generated based on the process outlined in Fig. 7. This data-generating mechanism simulates the population of interest (constituted by one million observations).

From this population, we obtained two data sets of ten thousand observations each: one using simple random sampling and the other using non-random sampling. In the non-random sample, participation (P) in the survey is affected by the TSV (V) by assuming that people with lower and higher TSVs are more likely to answer the survey. As a result, the sample obtained is biased. We assumed that the objective of the data analysis is to answer the following inferential question: '*What is the (population's) expected average TSV for males exposed to air temperatures between 16° C and 30° C with a clothing insulation value of 0.5 clo?*'. To address this data analytic question, the two synthetic data sets are analysed using ordinal regression within a Bayesian framework. For more details, the reader is referred to the online document; here, only a summary is provided (see Table 2).

An ordinal model has as its outcome a vector of probabilities, one for each category (i.e., seven in this case). However, we are interested in the average outcome. The mean of the probabilities can be calculated as:

$$\text{Mean Pr} = \sum_{1}^{K} \pi_k k \qquad \text{Eq. (3)}$$

where π_k is the probability of a specific category k, $k \in \{1, \ldots, K\}$.

Fig. 8 displays the average of posterior probabilities (and 95 % credible intervals) across varying air temperatures for males with a clothing insulation value of 0.5 clo. The orange line indicates the calculated mean probability from a biased sample (left side), while the green line represents the mean calculated using a random sample (right side). The black line displays the mean of the probabilities of the population, which is determined using the data-generating mechanism. This figure shows that the mean probability from the biased sample does not match the mean probability of the population. This discrepancy arises because the biased sample contains a disproportionate number of people who voted with lower and higher thermal sensations. However, the mean probability calculated from the random sample is almost a 'perfect' match with the population. The random sample can accurately recover the population mean. For both the biased and random samples, narrow 95 % credible intervals are observed due to the large sample size (i.e., ten thousand observations). Increasing the sample size will only improve precision and not accuracy. This is evident in the biased sample, where a larger sample size provides a more precise but incorrect answer.

9. Question 8: What is variable selection, and when is it useful?

Generally, variable (or feature) selection primarily focuses on removing non-informative or redundant variables from a larger set of potential regressors. Heinze et al. [100] provided an overview of various

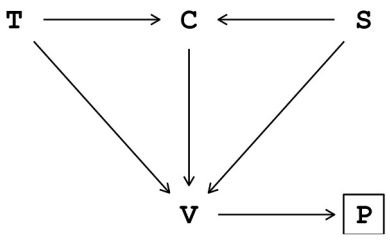


Fig. 7. – Graphical representation via DAG of the data-generating process. Indoor air temperature (T) and biological sex (S) influence TSV (V) both directly and indirectly, passing through clothing insulation (C). Participation (P) in the survey response is affected by TSV (V).

Table 2
Summary description of the simulation example.

Pitfall	Ignore the implications of selection bias (e.g., case-control bias) on the results and interpretation of a data analysis.
Type of analysis	Inferential.
Framework	Bayesian.
Assumptions	Limited random variability: large sample size. Independence of observations: each observation represents independent bits of information. No confounding: the DAG includes all shared causes among the variables. No model error: perfect functional form specification. No measurement error: all variables are measured perfectly.
Variables	Air temperature (T): continuous variable [unit: °C] Thermal resistance of clothing (C): continuous variable [unit: clo] Sex (S): categorical variable ['1' male; '2' female] Thermal sensation vote (V): ordinal variable ['1' cold; '2' cool; '3' slightly cool; '4' neutral; '5' slightly warm; '6' warm; '7' hot]

available variable selection methods that are based on significance criterion, change-in-estimate criterion, information criteria, penalised likelihood and background knowledge. These criteria can be implemented in variable selection algorithms such as backward elimination, forward selection, and stepwise selection. However, it is essential to highlight that **including/excluding variables in/from a model (i.e., performing variable selection) and the specific approach used is closely tied to the data analytic question type**. Hence, this topic has different implications depending on causal or predictive purposes.

For causal (and inferential) aims, a fundamental aspect is statistical adjustment (i.e., adjusting for a variable by including it in the model). Assuming that no selection bias (see Q.7) and measurement error (see Q.10) are present, bias can arise by including/excluding variables in/from a (regression) model. Common issues are confounding, and adjustment for mediators and colliders (see Fig. 2; for more details, the reader is referred to Cinelli et al. [54]). In thermal comfort research, for example, body mass index (BMI) is often included as a regressor. However, BMI is only an arithmetic derivation (BMI = weight/height^2); as such, it may be associated with an outcome of interest (e.g., TSV) but cannot influence the outcome because only weight and/or height can be causal [101]. Furthermore, depending on the causal structure of the data, conditioning on a variable can increase existing bias (i.e., bias amplification) [102]. Consequently, statistical adjustment requires causal thinking [65]. Since data are ignorant to causes, performing (automated) variable selection based on any criterion that is not causal thinking is very likely to produce misleading results. Although convincing arguments exist for not using stepwise variable selection (e.g., Refs. [103,104]), its use is widespread. For example, its use leads to parameter estimates that are biased away from zero, producing too low standard errors and p-values and narrower confidence intervals (page 68 of Ref. [71]).

For predictive purposes, there is no need to adjust for confounding. If the aim is to develop a model for predicting Y, all regressors (Xs) that are strong predictors of Y should be included. Since no causal interpretation is assigned to parameter estimates, the concept of adjusting for confounding does not apply. This aspect ties directly into the issue of collinearity. Collinearity is not a problem for predictive aims because it does not affect predictive power. It only affects the association between regressors and regressand, which is generally not of concern for prediction, with one exception. Collinearity becomes problematic for prediction if the level of collinearity in the sample does not reflect that of the population. Variable selection can help reduce the number of variables to include in a prediction model (especially when having all of them can be impossible for applicability in real-world scenarios [105] or result in unstable predictions), but it is important to remember that overfitting can be an issue. To solve this problem, cross-validation is an

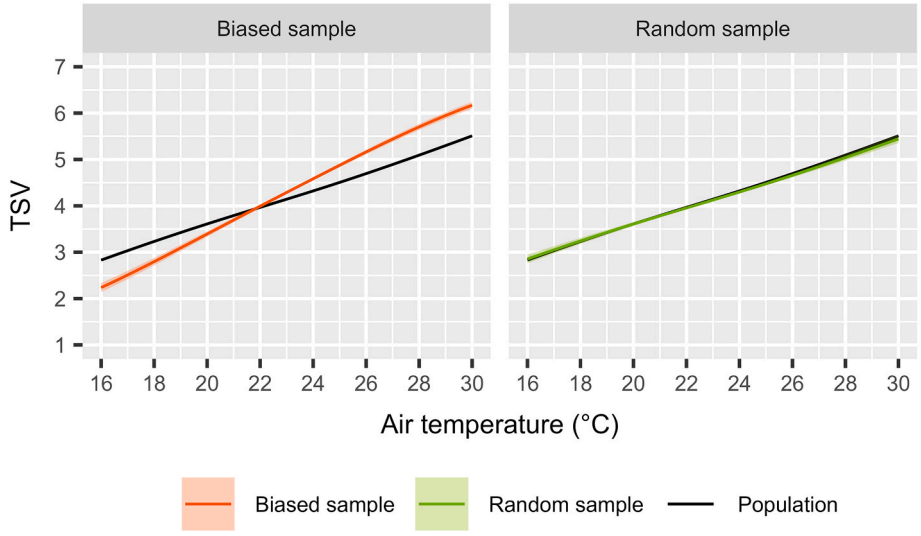


Fig. 8. – Average posterior probability (and 95 % credible intervals) across varying air temperatures for males with a clothing insulation value of 0.5 clo. The orange line indicates the calculated mean probability from a biased sample (left side), while the green line represents the mean calculated using a random sample (right side). The black lines display the mean of the probabilities of the population, which is determined using the data-generating mechanism. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

option, which can take different forms, such as k-fold and leave-one-out cross-validation. For exploratory and predictive purposes, data reduction methods can be an alternative approach to consider. Data reduction methods reduce the original data size to represent it in a much smaller space. There are many approaches to data reduction, with principal component analysis (PCA) being one example. A detailed description of variable selection and data reduction methods can be found in the books by Harrell [71], James et al. [106] and Hastie et al. [107].

9.1. Simulation example: variable selection

In this simulation example, the highlighted data analysis pitfall is variable selection. Specifically, it shows the implication of variable selection on the result of the data analysis. The data is generated based on the process outlined in Fig. 9. This data-generating mechanism simulates the population of interest (constituted by one million observations).

From this population, we obtained a data set of ten thousand observations using simple random sampling. We assumed that the objective of the data analysis is to answer the following causal question: '*What is the total causal effect of decreasing the indoor air temperature from 22°C to 20°C on the TSV?*'. To address this data analytic question, the synthetic data set is analysed using beta regression within a frequentist framework. For more details, the reader is referred to the online document; here, only a summary is provided (see Table 3).

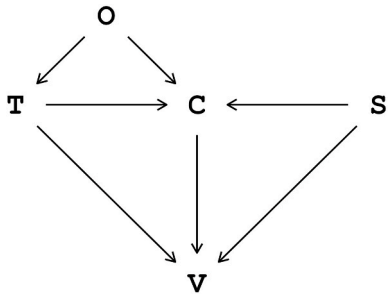


Fig. 9. – Graphical representation via DAG of the data-generating process. Indoor air temperature (T) and biological sex (S) influence TSV (V) both directly and indirectly, passing through clothing insulation (C). In addition, TSV (V) is influenced indirectly by outdoor air temperature (O) through indoor air temperature (T) and clothing insulation (C).

Table 3
Summary description of the simulation example.

Pitfall	Ignore the implications of variable selection on the results and interpretation of a data analysis.
Type of analysis	Causal.
Framework	Frequentist.
Assumptions	Random sample (simple random sampling): everyone in the population has an equal chance of being selected forinto the sample. Limited random variability: large sample size. Independence of observations: each observation represents independent bits of information. No confounding: the DAG includes all shared causes among the variables. No model error: perfect functional form specification. No measurement error: all variables are measured perfectly.
Variables	Outdoor air temperature (O): continuous variable [unit: °C] Indoor air temperature (T): continuous variable [unit: °C] Thermal resistance of clothing (C): continuous variable [unit: clo] Sex (S): categorical variable ['1' male; '2' female] Thermal sensation vote (V): continuous but bounded variable with interval (0, 1) ['0' cold; '1' hot]

To estimate the effect of interest, we need to select the variables to adjust for: the adjustment set. This selection is done in two ways: using causal thinking and backward selection with the AIC criterion. Using causal thinking, there are two possible adjustment sets to estimate the total causal effect of indoor air temperature: add outdoor temperature (Model 1) or add outdoor temperature and sex (Model 2). Adjusting for (i.e., conditioning on) sex is optional. However, conditioning on sex can help detect the effect of indoor air temperature by explaining some residual variations in TSV, thus improving precision. For this reason, we selected the adjustment set of 'Model 2'. Instead, if backward selection with the AIC criterion is used, the selected variables are indoor air temperature, clothing insulation and sex (Model 3). However, this is the

wrong adjustment set,[13] and it will not allow us to estimate the total causal effect of indoor air temperature. The AIC criterion is a predictive criterion and, as such, is a tool for pure predictive tasks (i.e., prediction in the absence of intervention). Here, the question of interest is causal, that is, prediction in the presence of intervention (i.e., an answer to a 'what if?' question). Table 4 compares the three models in terms of their AIC: the model selected by backward selection has the best AIC (i.e., the lowest value) and, as such, it is the best model for a pure predictive task. However, it will be the wrong model to answer the question of interest.

Using 'Model 2', we can calculate the total causal effect of indoor air temperature. To do so, we will use the potential outcomes framework [108] using the *g-computation algorithm formula* (*g-computation*) first described in 1986 [109] (see the online document for more details). Applying this method, the estimate of the total causal effect of decreasing the indoor air temperature from 22 °C to 20 °C is a reduction of −0.177 (95 % CI [−0.181, −0.173]) of the mean TSV.

The critical aspect to remember is that all the variables added as the adjustment set are solely there to enable the estimation of the estimate of interest. Here, outdoor air temperature and sex are selected to yield the total causal effect of indoor air temperature, and nothing else. As such, these variables should not be interpreted. The interpretation of such variables is known as 'Table 2 Fallacy' [110].

10. Question 9: Why should the data structure be considered during data analysing?

Many commonly used methods (both machine learning- and statistics-based) assume that the errors are identically and independently distributed (i.i.d.). This assumption is violated when observational units are clustered and/or nested within groups. The definition of what constitutes a group depends on the context. For instance, when a survey is conducted on occupants of several buildings, the observational unit is the occupant, while the building is the group. On the other hand, in a survey in which multiple measurements are carried out on each occupant of a single building, the observational unit is the measurement, and the occupant is the group.

As mentioned in Q.4, independence of error is one of the assumptions of linear regression and GLMs. If there is clustering in the data and it is not considered during the analysis, the standard errors of the regression coefficients will generally be underestimated. Consequently, confidence intervals will be too narrow, and *p*-values will be too small, leading to an overstatement of statistical significance. For example, consider 1000 individuals clustered in 10 groups where, for simplicity, each cluster contains 100 individuals. The standard errors of a regression model that assume independence of error are calculated on the assumption that each individual in the sample provides independent pieces of information (i.e., 1000 independent observations). However, there will be fewer than 1000 independent observations when the data are clustered. The number of independent information is called the effective sample size and depends on the degree of clustering. In the extreme case that all individuals in a group have the same value for the regressand (i.e., all errors are perfectly correlated), each group provides only one independent observation. In this case, the effective sample size will be equal to 10 (i.e., the number of clusters) rather than 1000. As mentioned in Q.6, there is an inverse relationship between sample size and standard error, where a smaller sample size will lead to a larger standard error. For a regression, coefficient estimates are labelled statistically significant (at .05 level) if they are at least two standard errors away from zero. Consequently, underestimated standard errors will lead to an overstatement of statistical significance. However, for some models, the underestimation of standard errors can generally be observed for variables measured at the group level (between-group variables). The opposite can be observed for within-group variables: standard errors are overestimated, leading to overly conservative *p*-values.

Table 4
– Models comparison.

Variable selection strategy	Selected model	df[a]	AIC
Causal thinking	Model 1: V ~ T + O	4	−16706.34
	Model 2: V ~ T + S + O	5	−16919.99
Backward selection	Model 3: V ~ T + S + C	5	−17352.15

[a] Df stands for degrees of freedom.

Multilevel models (also known as hierarchical or mixed models) are an appropriate approach for analysing clustered/nested data and provide accurate standard errors. Marginal models[14] (e.g., generalised estimating equations, GEE) are another approach to explicitly model dependency between observations in the same group, but the correlation structure is then essentially regarded as a nuisance. Although the marginal model method yields correct standard errors, multilevel modelling allows researchers to investigate the nature of between-group variability and the effects of group-level characteristics on individual outcomes.

Multilevel models can represent complex data structures using three primary forms: hierarchical (nested), cross-classified, and multiple membership. Fig. 10 shows examples of each using classification diagrams (left side) and unit diagrams (right side). These three elemental types of structures can also be combined to describe more complicated real-world situations and research designs. It is important to note that the multilevel structure is not an inherent feature of the model but rather a characteristic of the experimental/study design, which is then reflected in the data and subsequently captured by the model. For example, a situation in which multilevel modelling arises naturally is in the analysis of data obtained by stratified or cluster sampling. Therefore, **it is impossible to determine the data's structure simply by examining it, as this requires knowledge of the data and the experimental design**.

Other forms of non-independence arise from spatial and temporal patterns. In thermal comfort research, spatial patterns emerge when measurements are taken in non-homogeneous spaces (e.g., observations in different points of an open-plan office), and temporal patterns arise in longitudinal data with repeated measurements. In these situations, data exhibit autocorrelation, meaning that observations that are closer in space or time are more highly correlated. Examples of methods to model these dependency structures are spatial correlation and time series. In a multilevel model, it is possible to account for autocorrelation explicitly by defining a variance-covariance matrix that incorporates a correlation structure that approximates the patterns of dependency. An example of modelling (simple) temporal autocorrelation could be using a first-order autoregressive correlation structure, AR(1). However, if complex/general temporal correlation structures are expected, they should be handled by time series modelling tools.

Considering the many situations in which errors are dependent, multilevel modelling could be considered the 'default' starting point. For an application of multilevel regression to analyse subjective thermal

[13] Given the DAG, assuming air temperature and sex are conditioned on (i.e., included in the model), conditioning on clothing will statistically remove any association between indoor air temperature and TSV influenced by clothing insulation. Stated analogously, it will block the indirect effect of indoor air temperature (i.e., the mediating path).

[14] For regressions with correlated outcomes, there are two main approaches: marginal models (e.g., GEEs) and conditional models (e.g., LMMs and GLMMs). One of their main differences is in the interpretation of the resulting coefficient estimates. In a marginal model, the coefficients have a population-averaged interpretation, whereas in a mixed model, they have a cluster-specific interpretation. For LMMs, the two interpretations coincide, but they do not for GLMMs.

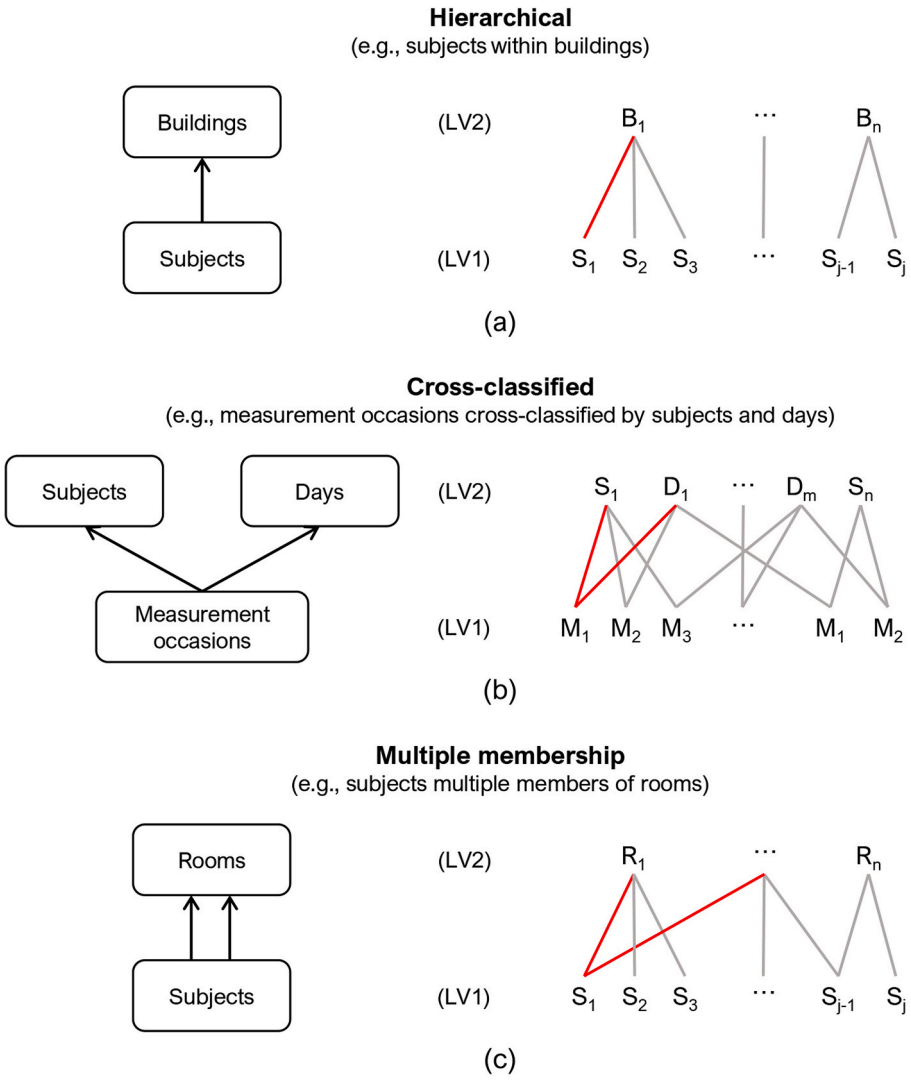


Fig. 10. – Examples of multilevel structures: hierarchical (a), cross-classified (b) and multiple membership (c) using classification diagrams (left-side) and unit diagrams (right-side). (Note. LV = level).

comfort data, the reader is referred to Favero et al. [76]. For more detail on multilevel modelling, the reader is referred to Refs. [31,111].

10.1. Simulation example: clustered/nested observations

In this simulation example, the highlighted data analysis pitfall is clustered/nested observations. Specifically, it shows the implication of having clustered/nested observations on the result of the data analysis. The data is generated based on the process outlined in Fig. 11. This data-generating mechanism simulates the population of interest (constituted by one million observations from twenty thousand buildings).

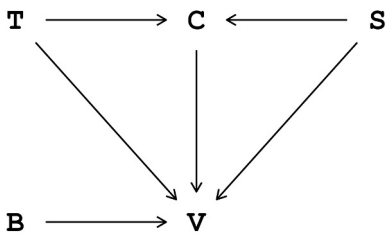


Fig. 11. – Graphical representation via DAG of the data-generating process. Indoor air temperature (T) and biological sex (S) influence thermal sensation vote (V) both directly and indirectly, passing through clothing insulation (C). In addition, thermal sensation vote (V) is influenced directly by some unmeasured characteristics of the specific building (B).

From this population, we obtain a data set of ten thousand observations using multistage sampling. We first randomly sampled 500 buildings; among the sampled buildings, we randomly sampled ten thousand observations. We assumed that the objective of the data analysis is to answer the following predictive question: '*What is the probability that the TSV for females at 25° C with a clothing insulation value of 0.6 clo in a typical (i.e., average) building is lower than 0.5 (i.e., 'neutral')?*'. To address this data analytic question, the synthetic data set is analysed using beta regression within a frequentist framework. For more details, the reader is referred to the online document; here, only a summary is provided (see Table 5).

To answer the question of interest, two beta regression models are fitted: one that ignores the cluster present in the data (Model 1) and one that considers it (Model 2).

Fig. 12 shows the predicted posterior distribution of the TSVs for 'Model 1' (orange) and 'Model 2' (green), respectively. The black line and dot at the bottom of each distribution represent the highest predictive density (HPD) and the mean, respectively. These intervals (i.e., the black lines) are defined here to span over 95 % of the distribution, representing the 95 % HPDs. Using 'Model 1', it is not possible to calculate the predicted TSVs for an average building because the model

Table 5
Summary description of the simulation example.

Pitfall	Ignore the implications of clustered/nested observations (i.e., dependent observations) on the results and interpretation of a data analysis.
Type of analysis	Predictive.
Framework	Bayesian.
Assumptions	Random sample (multistage sampling): everyone in the population has an equal chance of being selected into the sample. Limited random variability: large sample size. No confounding: the DAG includes all shared causes among the variables. No model error: perfect functional form specification. No measurement error: all variables are measured perfectly.
Variables	Indoor air temperature (T): continuous variable [unit: °C] Thermal resistance of clothing (C): continuous variable [unit: clo] Sex (S): categorical variable [‘1’ male; ‘2’ female] Building ID (B): index variable Thermal sensation vote (V): continuous but bounded variable with interval (0, 1) [‘0’ cold; ‘1’ hot]

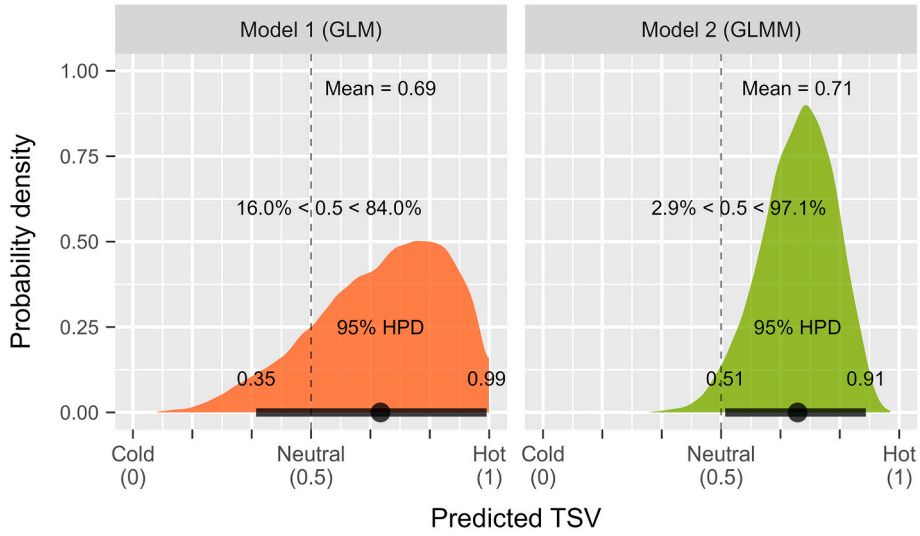


Fig. 12. – Predicted posterior distribution of the TSVs ‘Model 1’ (orange) and ‘Model 2’ (green). The black line and dot at the bottom of each distribution represent the highest predictive density (HPD) and the mode, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

does not model the variability among buildings. Consequently, it is not possible to answer the question of interest. The predicted probability of 16.0 % in Fig. 12 (left) could be thought of as that of females at 25 °C with clothing insulation of 0.6 clo across buildings on average. ‘Model 2’, on the other hand, can calculate the predicted TSVs for an average building. To do so, we need to hold the group-level residual (i.e., the group random effect) at its mean of zero, calculate the predicted TSV for the specific regressors-values (i.e., females at 25 °C with clothing insulation of 0.6 clo) and then calculate the percentage of it that is lower than 0.5 (‘neutral’). This results in a probability of 2.9 % (Fig. 12 (right)).

11. Question 10: What is measurement error, and how can it affect the data analysis?

Measurement error is an essential concept in scientific research and data analysis (it is an inherent part of any experimental or observational process) and represents the discrepancy between the result of a measurement and the true, unknown value (i.e., the measurand). Measurement error should not be confused with ‘uncertainty’, which quantifies doubt about the measurement result, or ‘accuracy’, which is a qualitative description of the closeness of a measurement.[15] Measurement error can arise from multiple factors (e.g., the measuring instrument, the measurement process, etc.) and take different forms (e.g., it can affect continuous and categorical measurements). It is usually classified into two types: (i) systematic errors and (ii) random errors. Systematic errors are consistent and repeatable discrepancies between measurements and the true value. They often result from flaws in measurement instruments, calibration, or experimental design. Random errors are unpredictable variations in measurements that occur due to inherent variability in the system or the measurement process itself. They result from factors such as electrical noise in measuring instruments, short-term fluctuations in the local environment and variability in the performance of the person carrying out the measurement. There is another category which is often referred to as error, namely gross error. Common examples of gross errors are incorrectly applied corrections and transcription errors. However, gross errors should be regarded as mistakes and not as measurement errors. The reason is that mistakes cannot be easily accounted for when evaluating uncertainty, and they should be avoided by working carefully. The international standard ISO/IEC Guide 98–3:2008 [112] provides useful indications for expressing measurement uncertainty.

In the answers to Q.7 and Q.8, we discussed two general issues that can induce biased estimates: selection bias and confounding. Measurement error is another issue that can introduce bias (specifically measurement bias). However, there is a general distinction between these biases. While there is a standard causal structure that can be used to summarise selection bias (i.e., conditioning on common effects of regressor and regressand, or of their causes) and confounding (i.e., the presence of common causes of regressor and regressand), this is not the case for measurement error. Hernán and Cole [114] provide a functional distinction of the measurement error structure of a regressor (e.g., air

[15] ISO/IEC Guide 98–3:2008 [112], which is based on JCGM 100:2008 [113], defines (i) accuracy (of measurement) as ‘closeness of the agreement between the result of a measurement and a true value of the measurand’ and (ii) uncertainty (of measurement) as ‘parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the measurand’.

temperature) and a regressand (e.g., TSV). While this classification is undoubtedly useful (see Ref. [114] for more details), it is very general in that the structure of the measurement error would at least depend on which variable is affected (in some settings, the magnitude of the measurement error may be judged, or known, to be too small to matter), and the relationship between the variables.

Humphreys and Nicol [23], using the data from the ASHRAE database I [115], explored the nature of errors for thermal comfort indices, distinguishing them in measurement error and formulaic error (i.e., error arising from an incorrect or incomplete formulation of an index). Concerning the effects of measurement error in regression, they generally state that:

1. The effect of measurement error in a regressor variable is to weaken the correlation and regression coefficients, causing them to fall below their actual values ('regression attenuation').
2. Random errors in the regressand (e.g., TSV) reduce the correlation coefficient, but the regression coefficient is not affected in a systematic manner.

However, there are some limitations to these statements. Considering the first one, Carroll et al. (page 41 of Ref. [116]) state that this conclusion, generally derived focusing on simple linear regression (as for Ref. [23]), must be qualified. Even for linear regression, the effects of measurement error vary, and its effect can range from a simple attenuation described earlier, (i) to cases in which real effects are concealed, (ii) to instances where observed data displays relationships that are absent in error-free data, and even (iii) to situations where the estimated coefficients' signs (±) are flipped compared to the scenario with no measurement error. In non-linear models, the effects of measurement error on the regressors are even more complicated. The second statement is only valid for linear regression where there is 'classical' measurement error[16]; however, different error structures are certainly possible (see Ref. [114]). In addition, as discussed in Q.5, linear regression is not appropriate for analysing subjective thermal comfort data measured on a continuous but bounded scale, and other regressions should be used. In logistic regression, for example, the error in the regressand is called misclassification. The bias introduced by misclassification is profound and can lead to a severe bias in parameter estimates (pages 345–347 of Ref. [116]).

Measurement error should not be overlooked in statistical analysis as it propagates uncertainty and can lead to significant consequences. For example, Carroll et al. (chapters 3 and 10 of Ref. [116]) demonstrate how measurement error can introduce bias in the estimation of regression coefficients and affect hypothesis testing. In addition, it will also affect sample size determination [117,118] and automated variable selection. However, Carroll et al. [116] state that when the purpose of the study is predictive, it is generally unnecessary to model measurement error, with one exception. When the aim is to predict a different population than the one used to develop the prediction model, it is crucial to account for the measurement error that can occur.

11.1. Simulation example: measurement error

In this simulation example, the highlighted data analysis pitfall is measurement error. Specifically, it shows the influence of measurement error on the results of a data analysis. The data is generated based on the process outlined in Fig. 13. This data-generating mechanism simulates the population of interest (constituted by one million observations).

From this population, we obtain a data set of ten thousand observations using simple random sampling. We defined the unobserved error e_T as normally distributed with zero mean and five different values for the standard deviation (0.1 K, 0.2 K, 0.3 K, 0.4 K, and 0.5 K, respectively). These different standard deviations reflect different random fluctuations (i.e., random error). We then added the unobserved error to the unobserved indoor air temperature (T) to obtain the observed indoor air temperature (T*). We assumed that a 4-wire Pt100 resistance temperature sensing (RTD) temperature probe (tolerance class DIN B) and a transmitter that operates with a current output signal of 4–20 mA and a temperature range of 0–200 °C were used. The combined standard uncertainty $u_c(T)$ (hypothetical but realistic) of the programmable transmitter with a Pt100 RTD is 0.23. This is used as the standard deviation for the measurement error in air temperature (i.e., $u_c(T) = 0.23 = \sigma_{T^*}$). We assumed that the objective of the data analysis is to answer the following inferential question: '*What is the association between TSV and sex, indoor air temperature and clothing insulation?*'. To address this data analytic question, the synthetic data set is analysed using beta regression within a Bayesian framework. For more details, the reader is referred to the online document; here, only a summary is provided (see Table 6).

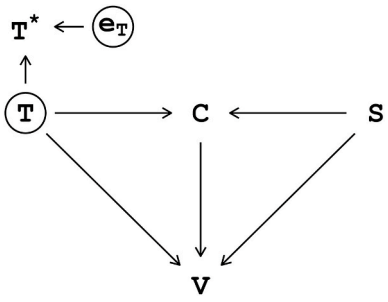


Fig. 13. – Graphical representation via DAG of the data-generating process. Indoor air temperature (T) and biological sex (S) influence thermal sensation vote (V) both directly and indirectly, passing through clothing insulation (C). The true indoor air temperature T cannot be observed, so it is circled as an unobserved node. However, we do get to observe T*, which is a function of both the true air temperature T and some unobserved error e_T. The error itself is not directly observable but could be roughly approximated by considering the possible uncertainty in the measurement process.

Three models were considered:

– 'Model 1' is the model that uses the air temperature measured without error, T. It can be considered the reference model.
– 'Model 2' is the model that uses air temperature measured with error, T*.

Table 6
Summary description of the simulation example.

Pitfall	Ignore the implications of measurement error on the results and interpretation of a data analysis.
Type of analysis	Inferential.
Framework	Bayesian.
Assumptions	Random sample (simple random sampling): everyone in the population has an equal chance of being selected into the sample. Limited random variability: large sample size. Independence of observations: each observation represents independent bits of information. No confounding: the DAG includes all shared causes among the variables. No model error: perfect functional form specification.
Variables	Unobserved indoor air temperature (T): continuous variable [unit: °C] Observed indoor air temperature (T*): continuous variable [unit: °C] Thermal resistance of clothing (C): continuous variable [unit: clo] Sex (S): categorical variable ['1' male; '2' female] Thermal sensation vote (V): continuous but bounded variable with interval (0, 1) ['0' cold; '1' hot]

[16] Classical measurement error refers to additive error uncorrelated with the regressors.

– 'Model 3' is the model that uses air temperature measured with error, T* but considering measurement error (using $\sigma_{T^*} = 0.23$).

Fig. 14 shows the estimates of the regression coefficients for three models and their 95 % credible intervals. Here, the estimated posterior distributions of the regression coefficients for 'Model 1', 'Model 2' and 'Model 3' are shown by the black, orange and green lines, respectively. With regard to the increment in measurement error (i.e., from 'No error' to sd = 0.5), the following comparison can be made:

– The bias of the sex coefficient is positive and very similar in magnitude for both 'Model 2' and 'Model 3';
– The bias of the temperature coefficient is negative (i.e., regression dilution) for 'Model 2' and 'Model 3' but overall higher in magnitude for 'Model 2'. However, in 'Model 3' considering measurement error when it does not exist (i.e., 'no error') or is lower in magnitude than the assumed one (i.e., $\sigma_{T^*} = 0.23 >$ sd $= 0.1$ and 0.2) undermines the inference by overestimating the coefficient;
– The bias of the clothing coefficient is negative and very similar in magnitude for both 'Model 2' and 'Model 3'.

Considering measurement error in statistical modelling is essential, but it is important to proceed cautiously as no 'one-size-fits-all' solutions are applicable. The presence of measurement error can lead to unpredictable effects; it can over- or underestimate the coefficients of interest depending on which variables are affected, how the measurement error is structured and its magnitude. Although directly modelling measurement error can be complicated, evaluating its implications for the results and conclusions is always possible by performing a sensitivity analysis.

12. Conclusions

In this paper, we have explored ten questions about statistical data analysis in human-centric research focusing on thermal comfort. Nonetheless, the issues highlighted in the paper are also relevant to other domains of the indoor environment and beyond. For example, in the recent review by Chinazzo et al. [119], problems of poor practice in statistical analysis, among other issues, have also been identified in multi-domain studies. This paper first covers an overview of the fundamental concepts of statistical data analysis and then, through synthetic data, highlights the tangible effects of various statistical pitfalls applied to examples from thermal comfort research. We have shown that statistical thinking is critical to avoid misleading results. Furthermore, while the diverse pitfalls are presented individually, they are combined in 'real-world' data analysis, stressing the added importance of statistical thinking.

Statistics are often perceived, selected, and reported based on preferred causal stories; stories researchers, reviewers and editors want to believe causally affect analysis choices and output interpretation. However, in this 'garden of forking paths' [120], where this path appears predetermined, it is actually due to implicit choices made along

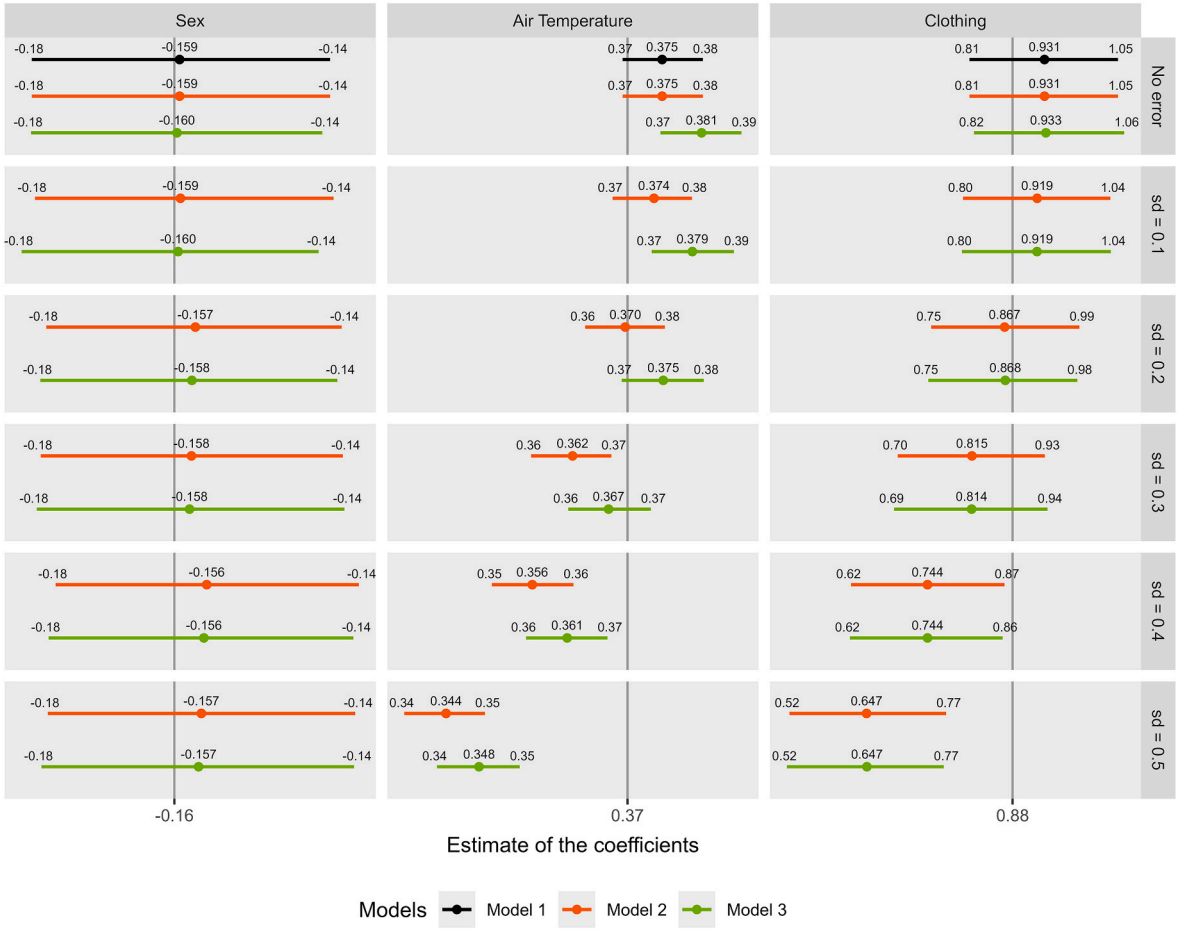


Fig. 14. – Posterior distributions for the regression coefficients for 'Model 1' (black), 'Model 2' (orange) and 'Model 3' (green). The lines and dots represent the 95 % credible interval and the mean, respectively. The solid grey lines represent the values of the coefficients used to generate the data set. (Note. On the right-side label, 'sd' stands for standard deviation and represents the different random fluctuations (i.e., random error) of the unobserved error e_T). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

the way. Therefore, we advocate for greater attention by researchers to justify and document all choices made in the statistical analysis of human-centric building analysis data. Although it is certainly true that much of the scientific work can be done with some details hidden in a 'black-box', it is important to be transparent and specific about all modelling steps. Acknowledging the word limit in journals, we recommend providing such details in the supplementary materials. To this end, we have created an online document that offers a clear and practical explanation of the modelling process for the simulation examples discussed in the paper. With human-centred approaches gaining more importance, this paper and the online document will be a valuable resource for experienced and inexperienced researchers and practitioners who need to analyse the collected data from buildings and their occupants with sound statistical methods.

CRediT authorship contribution statement

Matteo Favero: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Salvatore Carlucci:** Writing – review & editing, Writing – original draft, Validation, Formal analysis, Data curation. **Giorgia Chinazzo:** Writing – review & editing, Writing – original draft, Validation, Formal analysis, Data curation. **Jan Kloppenborg Møller:** Writing – review & editing, Writing – original draft, Validation, Formal analysis, Data curation. **Marcel Schweiker:** Writing – review & editing, Writing – original draft, Validation, Formal analysis, Data curation. **Marika Vellei:** Writing – review & editing, Writing – original draft, Validation, Formal analysis, Data curation. **Andrew Sonta:** Writing – review & editing, Writing – original draft, Validation, Supervision, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Within the manuscript there is a link to an online document that describe the simulation examples including code and data.

Acknowledgements

This work was supported by funding from EPFL to the ETHOS Lab. M. S. was supported by a research grant (21055) by VILLUM FONDEN.

References

[1] American National Standard & American Society of Heating, Refrigerating and air-conditioning engineers, thermal environmental conditions for human occupancy (ANSI/ASHRAE standard 55:2023). https://www.ansi.org/, 2023.
[2] International Organization for Standardization, Ergonomics of the Thermal Environment — Analytical Determination and Interpretation of Thermal Comfort Using Calculation of the PMV and PPD Indices and Local Thermal Comfort Criteria, 2005. ISO 7730:2005, https://www.iso.org/.
[3] European Committee for Standardization, Energy performance of buildings - ventilation for buildings - Part 1: indoor environmental input parameters for design and assessment of energy performance of buildings addressing indoor air quality, thermal environment, lighting and acoustics - module Ml-6 (EN 16798–1: 2019). https://www.cencenelec.eu/, 2019.
[4] T. Bedford, The warmth factor in comfort at work: a physiological study of heating and ventilation, H.M. Stationery Office (1936).
[5] F. Nicol, M. Humphreys, S. Roaf, Adaptive Thermal Comfort: Foundations and Analysis, Routledge, 2015, https://doi.org/10.4324/9781315765815.
[6] F.A. Chrenko, Probit analysis of subjective reactions to thermal stimuli — a study of radiant panel heating in buildings, Br. J. Psychol. 44 (1953) 248–256, https://doi.org/10.1111/j.2044-8295.1953.tb01204.x. General Section.
[7] C.G. Webb, An analysis of some observations of thermal comfort in an equatorial climate, Br. J. Ind. Med. 16 (1959) 297–310, https://doi.org/10.1136/oem.16.4.297.
[8] P.O. Fanger, Thermal Comfort: Analysis and Applications in Environmental Engineering, Danish Technical Press, Copenhagen, Denmark, 1970.
[9] D.A. McIntyre, Seven point scales of warmth, Build. Serv. Eng. 45 (1978) 215–226.
[10] International Organization for Standardization, Ergonomics of the physical environment — subjective judgement scales for assessing physical environments (ISO 10551:2019). https://www.iso.org/, 2019.
[11] M. Favero, A. Luparelli, S. Carlucci, Analysis of subjective thermal comfort data: a statistical point of view, Energy Build. 281 (2023) 112755, https://doi.org/10.1016/j.enbuild.2022.112755.
[12] P.B. Stark, A. Saltelli, Cargo-cult statistics and scientific crisis, Significance 15 (2018) 40–43, https://doi.org/10.1111/j.1740-9713.2018.01174.x.
[13] G. Gigerenzer, Mindless statistics, J. Soc. Econ. 33 (2004) 587–606, https://doi.org/10.1016/j.socec.2004.09.033.
[14] J.P.A. Ioannidis, Why most published research findings are false, PLoS Med. 2 (2005) e124, https://doi.org/10.1371/journal.pmed.0020124.
[15] M.R. Munafò, B.A. Nosek, D.V.M. Bishop, K.S. Button, C.D. Chambers, N. Percie du Sert, U. Simonsohn, E.-J. Wagenmakers, J.J. Ware, J.P.A. Ioannidis, A manifesto for reproducible science, Nat. Human Behav. 1 (2017) 21, https://doi.org/10.1038/s41562-016-0021.
[16] M.C. Makel, J.A. Plucker, B. Hegarty, Replications in psychology research: how often do they really occur? Perspect. Psychol. Sci. 7 (2012) 537–542, https://doi.org/10.1177/1745691612460688.
[17] N.L. Kerr, HARKing: hypothesizing after the results are known, Pers. Soc. Psychol. Rev. 2 (1998) 196–217, https://doi.org/10.1207/s15327957pspr0203_4.
[18] K.S. Button, J.P.A. Ioannidis, C. Mokrysz, B.A. Nosek, J. Flint, E.S.J. Robinson, M. R. Munafò, Power failure: why small sample size undermines the reliability of neuroscience, Nat. Rev. Neurosci. 14 (2013) 365–376, https://doi.org/10.1038/nrn3475.
[19] J.P. Simmons, L.D. Nelson, U. Simonsohn, False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant, Psychol. Sci. 22 (2011) 1359–1366, https://doi.org/10.1177/0956797611417632.
[20] L.K. John, G. Loewenstein, D. Prelec, Measuring the prevalence of questionable research practices with incentives for truth telling, Psychol. Sci. 23 (2012) 524–532, https://doi.org/10.1177/0956797611430953.
[21] D. Fanelli, "Positive" results increase down the hierarchy of the sciences, PLoS One 5 (2010) e10068, https://doi.org/10.1371/journal.pone.0010068.
[22] J.M. Wicherts, D. Borsboom, J. Kats, D. Molenaar, The poor availability of psychological research data for reanalysis, Am. Psychol. 61 (2006) 726–728, https://doi.org/10.1037/0003-066X.61.7.726.
[23] M.A. Humphreys, J. Fergus Nicol, Effects of measurement and formulation error on thermal comfort indices in the ASHRAE database of field studies, Build. Eng. 106 (2000) 493.
[24] R. Sun, S. Schiavon, G. Brager, E. Arens, H. Zhang, T. Parkinson, C. Zhang, Causal thinking: uncovering hidden assumptions and interpretations of statistical analysis in building science, Build. Environ. (2024) 111530, https://doi.org/10.1016/j.buildenv.2024.111530.
[25] J. Pan, A. Mahdavi, I. Mino-Rodriguez, I. Martínez-Muñoz, C. Berger, M. Schweiker, The untapped potential of causal inference in cross-modal research, Build. Environ. 248 (2024) 111074, https://doi.org/10.1016/j.buildenv.2023.111074.
[26] J. Kim, S. Schiavon, G. Brager, Personal comfort models – a new paradigm in thermal comfort for occupant-centric environmental control, Build. Environ. 132 (2018) 114–124, https://doi.org/10.1016/j.buildenv.2018.01.023.
[27] Z. Qavidel Fard, Z.S. Zomorodian, S.S. Korsavi, Application of machine learning in thermal comfort studies: a review of methods, performance and challenges, Energy Build. 256 (2022) 111771, https://doi.org/10.1016/j.enbuild.2021.111771.
[28] Y. Feng, S. Liu, J. Wang, J. Yang, Y.-L. Jao, N. Wang, Data-driven personal thermal comfort prediction: a literature review, Renew. Sustain. Energy Rev. 161 (2022) 112357, https://doi.org/10.1016/j.rser.2022.112357.
[29] J.T. Leek, R.D. Peng, What is the question? Science 347 (2015) 1314–1315, https://doi.org/10.1126/science.aaa6146.
[30] Galit Shmueli, To explain or to predict? Stat. Sci. 25 (2010) 289–310, https://doi.org/10.1214/10-STS330.
[31] A. Gelman, J. Hill, A. Vehtari, Regression and Other Stories, Cambridge University Press, 2020.
[32] J. Pearl, An Introduction to Causal Inference, vol. 6, 2010, https://doi.org/10.2202/1557-4679.1203.
[33] M. Davidian, T.A. Louis, Why statistics? Science 336 (2012) https://doi.org/10.1126/science.1218685, 12–12.
[34] Asa newsroom, American Statistical Association (n.d.). https://www.amstat.org/asa-newsroom (accessed December 18, 2022).
[35] L.J. Savage, The Foundations of Statistics, second ed., Dover Publications, New York, 1972.
[36] A. Kolmogoroff, Grundbegriffe der wahrscheinlichkeitsrechnung, 1933.
[37] A. Clayton, Bernoulli's Fallacy, Columbia University Press, 2021, https://doi.org/10.7312/clay19994.
[38] J. Neyman, E.S. Pearson, On the problem of the most efficient tests of statistical hypotheses, in: S. Kotz, N.L. Johnson (Eds.), Breakthroughs in Statistics: Foundations and Basic Theory, Springer New York, New York, NY, 1992, pp. 73–108, https://doi.org/10.1007/978-1-4612-0919-5_6.

[39] R. Royall, The likelihood paradigm for statistical evidence, in: M.L. Taper, S. R. Lele (Eds.), The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations, University of Chicago Press, 2004, https://doi.org/10.7208/chicago/9780226789583.003.0005.
[40] R. Nuzzo, Scientific method: statistical errors, Nature 506 (2014) 150–152, https://doi.org/10.1038/506150a.
[41] J.D. Perezgonzalez, Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing, Front. Psychol. 6 (2015), https://doi.org/10.3389/fpsyg.2015.00223.
[42] S. Greenland, S.J. Senn, K.J. Rothman, J.B. Carlin, C. Poole, S.N. Goodman, D. G. Altman, Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations, Eur. J. Epidemiol. 31 (2016) 337–350, https://doi.org/10.1007/s10654-016-0149-3.
[43] R.L. Wasserstein, N.A. Lazar, The ASA statement on p-values: context, process, and purpose, Am. Statistician 70 (2016) 129–133, https://doi.org/10.1080/00031305.2016.1154108.
[44] S. Greenland, Valid P-values behave exactly as they should: some misleading criticisms of P-values and their resolution with S-values, Am. Statistician 73 (2019) 106–114, https://doi.org/10.1080/00031305.2018.1529625.
[45] R.A. Fisher, Statistical Methods and Scientific Inference, Hafner Publishing Co., Oxford, England, 1956.
[46] A. Gelman, J. Carlin, Beyond power calculations: assessing type S (sign) and type M (magnitude) errors, Perspect. Psychol. Sci. 9 (2014) 641–651, https://doi.org/10.1177/1745691614551642.
[47] International Organization for Standardization, Statistics — vocabulary and symbols — Part 1: general statistical terms and terms used in probability. https://www.iso.org/, 2006.
[48] P.B. Stark, Glossary of Statistical Terms, SticiGui: Statistics Tools for Internet and Classroom Instruction with a Graphical User Interface Java Tools (n.d.). https://www.stat.berkeley.edu/~stark/SticiGui/Text/gloss.htm (accessed December 18, 2022).
[49] N. Altman, M. Krzywinski, Association, correlation and causation, Nat. Methods 12 (2015) 899–900, https://doi.org/10.1038/nmeth.3587.
[50] X. Chen, J. Abualdenien, M.M. Singh, A. Borrmann, P. Geyer, Introducing causal inference in the energy-efficient building design process, Energy Build. 277 (2022) 112583, https://doi.org/10.1016/j.enbuild.2022.112583.
[51] N. Cartwright, No causes in, No causes out, in: N. Cartwright (Ed.), Nature's Capacities and Their Measurement, Oxford University Press, 1994, https://doi.org/10.1093/0198235070.003.0003.
[52] J. Pearl, Causal diagrams for empirical research, Biometrika 82 (1995) 669–688, https://doi.org/10.1093/biomet/82.4.669.
[53] J. Pearl, D. Mackenzie, The Book of Why: the New Science of Cause and Effect, Basic books, 2018.
[54] C. Cinelli, A. Forney, J. Pearl, A crash course in good and bad controls, Socio. Methods Res. (2022), https://doi.org/10.1177/00491241221099552.
[55] J. Pearl, Causality, second ed., Cambridge University Press, Cambridge, 2009. https://www.cambridge.org/core/product/B0046844FAE10CBF274D4ACBDAEB5F5B.
[56] J. Tian, J. Pearl, A new characterization of the experimental implications of causal Bayesian networks, in: Eighteenth National Conference on Artificial Intelligence, American Association for Artificial Intelligence, USA, 2002, pp. 574–579.
[57] I. Shpitser, J. Pearl, Complete identification methods for the causal hierarchy, J. Mach. Learn. Res. 9 (2008) 1941–1979.
[58] E. Bareinboim, J. Pearl, Causal inference by surrogate experiments: z-identifiability, in: Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, AUAI Press, Arlington, Virginia, USA, 2012, pp. 113–120.
[59] J.M. Rohrer, Thinking clearly about correlations and causation: graphical causal models for observational data, Adv. Methods Practices Psychological Sci. 1 (2018) 27–42, https://doi.org/10.1177/2515245917745629.
[60] J. Tian, J. Pearl, On the testable implications of causal models with hidden variables, in: Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002, pp. 519–527.
[61] B. Chen, J. Tian, J. Pearl, Testable implications of linear structural equation models, AAAI 28 (2014), https://doi.org/10.1609/aaai.v28i1.9065.
[62] Judea Pearl, Elias Bareinboim, External validity: from do-calculus to transportability across populations, Stat. Sci. 29 (2014) 579–595, https://doi.org/10.1214/14-STS486.
[63] E. Bareinboim, J. Tian, J. Pearl, Recovering from selection bias in causal and statistical inference, AAAI 28 (2014), https://doi.org/10.1609/aaai.v28i1.9074.
[64] M.A. Hernán, S. Hernández-Díaz, J.M. Robins, A structural approach to selection bias, Epidemiology 15 (2004), https://doi.org/10.1097/01.ede.0000135174.63482.43.
[65] A.C. Wysocki, K.M. Lawson, M. Rhemtulla, Statistical control requires causal justification, Adv. Methods Practices Psychological Sci. 5 (2022) 25152459221095823, https://doi.org/10.1177/25152459221095823.
[66] M.A. Hernán, D. Clayton, N. Keiding, The Simpson's paradox unraveled, Int. J. Epidemiol. 40 (2011) 780–785, https://doi.org/10.1093/ije/dyr041.
[67] J. Pearl, Comment: understanding Simpson's paradox, Am. Statistician 68 (2014) 8–13, https://doi.org/10.1080/00031305.2014.876829.
[68] P. Lavrakas, Encyclopedia of Survey Research Methods, SAGE Publications, Inc., Thousand Oaks, 2024, https://doi.org/10.4135/9781412963947.
[69] R. Yao, S. Zhang, C. Du, M. Schweiker, S. Hodder, B.W. Olesen, J. Toftum, F. Romana d'Ambrosio, H. Gebhardt, S. Zhou, F. Yuan, B. Li, Evolution and performance analysis of adaptive thermal comfort models – a comprehensive literature review, Build. Environ. 217 (2022) 109020, https://doi.org/10.1016/j.buildenv.2022.109020.
[70] T. Baguley, Standardized or simple effect size: what should be reported? Br. J. Psychol. 100 (2009) 603–617, https://doi.org/10.1348/000712608X377117.
[71] Jr. Harrell Frank E, Regression modeling strategies: with applications to linear models, logistic and ordinal regression. And Survival Analysis, second ed., Springer International Publishing, Cham, 2015.
[72] X. Fuchs, S. Becker, K. Schakib-Ekbatan, M. Schweiker, Subgroups holding different conceptions of scales rate room temperatures differently, Build. Environ. 128 (2018) 236–247, https://doi.org/10.1016/j.buildenv.2017.11.034.
[73] M. Schweiker, X. Fuchs, S. Becker, M. Shukuya, M. Dovjak, M. Hawighorst, J. Kolarik, Challenging the assumptions for thermal sensation scales, Build. Res. Inf. 45 (2017) 572–589, https://doi.org/10.1080/09613218.2016.1183185.
[74] T.M. Liddell, J.K. Kruschke, Analyzing ordinal data with metric models: what could possibly go wrong? J. Exp. Soc. Psychol. 79 (2018) 328–348, https://doi.org/10.1016/j.jesp.2018.08.009.
[75] A. Agresti, Categorical Data Analysis, third ed., John Wiley & Sons, 2013.
[76] M. Favero, J. Kloppenborg Møller, D. Calì, S. Carlucci, Human-in-the-loop methods for occupant-centric building design and operation, Appl. Energy 325 (2022) 119803, https://doi.org/10.1016/j.apenergy.2022.119803.
[77] J. Tobin, Estimation of relationships for limited dependent variables, Econometrica 26 (1958) 24–36, https://doi.org/10.2307/1907382.
[78] J.M. Wooldridge, Econometric Analysis of Cross Section and Panel Data, second ed., The MIT Press, Cambridge, Mass, 2010.
[79] G. Cumming, Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis, Routledge, 2011.
[80] A. Tversky, D. Kahneman, Belief in the law of small numbers, Psychol. Bull. 76 (1971) 105–110, https://doi.org/10.1037/h0031322.
[81] S. Goodman, A dirty dozen: twelve P-value misconceptions, Semin. Hematol. 45 (2008) 135–140, https://doi.org/10.1053/j.seminhematol.2008.04.003.
[82] S. Altomonte, S. Schiavon, M.G. Kent, G. Brager, Indoor environmental quality and occupant satisfaction in green-certified buildings, Build. Res. Inf. 47 (2019) 255–274, https://doi.org/10.1080/09613218.2018.1383715.
[83] S. Zhang, P.R. Heck, M.N. Meyer, C.F. Chabris, D.G. Goldstein, J.M. Hofman, An illusion of predictability in scientific results: even experts confuse inferential uncertainty and outcome variability, Proc. Natl. Acad. Sci. USA 120 (2023) e2302491120, https://doi.org/10.1073/pnas.2302491120.
[84] J.M. Hofman, D.G. Goldstein, J. Hullman, How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1–12, https://doi.org/10.1145/3313831.3376454.
[85] L. Lan, Z. Lian, Application of statistical power analysis – how to determine the right sample size in human health, comfort and productivity research, Build. Environ. 45 (2010) 1202–1213, https://doi.org/10.1016/j.buildenv.2009.11.002.
[86] H. Du, Z. Lian, L. Lan, D. Lai, Application of statistical analysis of sample size: how many occupant responses are required for an indoor environmental quality (IEQ) field study, Build. Simulat. 16 (2023) 577–588, https://doi.org/10.1007/s12273-022-0970-4.
[87] G. Casella, R.L. Berger, Statistical Inference, Cengage Learning, 2021.
[88] T.L. Lash, M.P. Fox, A.K. Fink, Applying Quantitative Bias Analysis to Epidemiologic Data, first ed., Springer Publishing Company, Incorporated, 2009.
[89] K.J. Rothman, S. Greenland, T.L. Lash, Modern Epidemiology, third ed., Wolters Kluwer Health/Lippincott Williams & Wilkins, Philadelphia, Philadelphia, 2008.
[90] S. Greenland, Invited commentary: the need for cognitive science in methodology, Am. J. Epidemiol. 186 (2017) 639–645, https://doi.org/10.1093/aje/kwx259.
[91] S. Vasishth, A. Gelman, in: How to Embrace Variation and Accept Uncertainty in Linguistic and Psycholinguistic Data Analysis, vol. 59, 2021, pp. 1311–1342, https://doi.org/10.1515/ling-2019-0051.
[92] J. Martínez-Mesa, D.A. González-Chica, J.L. Bastos, R.R. Bonamigo, R.P. Duquia, Sample size: how many participants do I need in my research? An. Bras. Dermatol. 89 (2014) 609–615, https://doi.org/10.1590/abd1806-4841.20143705.
[93] N.G. Fielding, R.M. Lee, G. Blank, The SAGE Handbook of Online Research Methods, 55 City Road, 2024. London, https://sk.sagepub.com/reference/the-sage-handbook-of-online-research-methods-2e.
[94] J. Martínez-Mesa, D.A. González-Chica, R.P. Duquia, R.R. Bonamigo, J.L. Bastos, Sampling: how to select participants in my research study? An. Bras. Dermatol. 91 (2016) 326–330, https://doi.org/10.1590/abd1806-4841.20165254.
[95] K.J. Rothman, J.E. Gallacher, E.E. Hatch, Why representativeness should be avoided, Int. J. Epidemiol. 42 (2013) 1012–1014, https://doi.org/10.1093/ije/dys223.
[96] X. Wang, Z. Cheng, Cross-sectional studies: strengths, weaknesses, and recommendations, Chest 158 (2020) S65–S71, https://doi.org/10.1016/j.chest.2020.03.012.
[97] R. Rosenthal, The volunteer subject, Hum. Relat. 18 (1965) 389–406, https://doi.org/10.1177/001872676501800407.
[98] L. Kennedy, A. Gelman, Know your population and know your model: using model-based regression and poststratification to generalize findings beyond the observed sample, Psychol. Methods 26 (2021) 547–558, https://doi.org/10.1037/met0000362.
[99] W. Wang, D. Rothschild, S. Goel, A. Gelman, Forecasting elections with non-representative polls, Int. J. Forecast. 31 (2015) 980–991, https://doi.org/10.1016/j.ijforecast.2014.06.001.

[100] G. Heinze, C. Wallisch, D. Dunkler, Variable selection – a review and recommendations for the practicing statistician, Biom. J. 60 (2018) 431–449, https://doi.org/10.1002/bimj.201700067.
[101] E. Shahar, The association of body mass index with health outcomes: causal, inconsistent, or confounded? Am. J. Epidemiol. 170 (2009) 957–958, https://doi.org/10.1093/aje/kwp292.
[102] J. Pearl, Invited commentary: understanding bias amplification, Am. J. Epidemiol. 174 (2011) 1223–1227, https://doi.org/10.1093/aje/kwr352.
[103] K.L. Sainani, Multivariate regression: the pitfalls of automated variable selection, PM&R 5 (2013) 791–794, https://doi.org/10.1016/j.pmrj.2013.07.007.
[104] G. Smith, Step away from stepwise, J. Big Data 5 (2018) 32, https://doi.org/10.1186/s40537-018-0143-6.
[105] J. Guenther, O. Sawodny, Feature selection and Gaussian Process regression for personalized thermal comfort prediction, Build. Environ. 148 (2019) 448–458, https://doi.org/10.1016/j.buildenv.2018.11.019.
[106] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning: with Applications in R, second ed., Springer US, New York, NY, 2021.
[107] T. Hastie, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed., Springer Nature, Netherlands, 2009.
[108] G.W. Imbens, D.B. Rubin, Causal Inference for Statistics, Social, and Biomedical Sciences: an Introduction, Cambridge University Press, Cambridge, 2015, https://doi.org/10.1017/CBO9781139025751.
[109] J. Robins, A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect, Math. Model. 7 (1986) 1393–1512, https://doi.org/10.1016/0270-0255(86)90088-6.
[110] D. Westreich, S. Greenland, The table 2 fallacy: presenting and interpreting confounder and modifier coefficients, Am. J. Epidemiol. 177 (2013) 292–298, https://doi.org/10.1093/aje/kws412.
[111] A. Gelman, J. Hill, Data Analysis Using Regression and Multilevel/hierarchical Models, Cambridge University Press, New York, 2007.
[112] International Organization for Standardization, Uncertainty of measurement — Part 3: guide to the expression of uncertainty in measurement (ISO/IEC Guide 98-3:2008). https://www.iso.org/, 2008.
[113] Joint committee for guides in metrology, evaluation of measurement data — guide to the expression of uncertainty in measurement, JCGM 100 (2008), 2008, https://www.iso.org/sites/JCGM/GUM/JCGM100/C045315e-html/C045315e.html?csnumber=50461.
[114] M.A. Hernán, S.R. Cole, Invited commentary: causal diagrams and measurement bias, Am. J. Epidemiol. 170 (2009) 959–962, https://doi.org/10.1093/aje/kwp293.
[115] R.J. de Dear, Global Database of Thermal Comfort Field Experiments, 1998, pp. 1141–1152. https://www.scopus.com/inward/record.uri?eid=2-s2.0-0031632764&partnerID=40&md5=1d6861aaf3dc04f743cd91e239bdf006.
[116] R.J. Carroll, Measurement Error in Nonlinear Models: a Modern Perspective, second ed., Chapman & Hall/CRC, Boca Raton, 2006.
[117] O.J. Devine, J.M. Smith, Estimating sample size for epidemiologic studies: the impact of ignoring exposure measurement uncertainty, Stat. Med. 17 (1998) 1375–1389, https://doi.org/10.1002/(SICI)1097-0258(19980630)17:12<1375::AID-SIM857>3.0.CO;2-D.
[118] G.E. McKeown-Eyssen, R. Tibshirani, Implications of measurement error in exposure for the sample sizes of case-control studies, Am. J. Epidemiol. 139 (1994) 415–421, https://doi.org/10.1093/oxfordjournals.aje.a117014.
[119] G. Chinazzo, R.K. Andersen, E. Azar, V.M. Barthelmes, C. Becchio, L. Belussi, C. Berger, S. Carlucci, S.P. Corgnati, S. Crosby, L. Danza, L. de Castro, M. Favero, S. Gauthier, R.T. Hellwig, Q. Jin, J. Kim, M. Sarey Khanie, D. Khovalyg, C. Lingua, A. Luna-Navarro, A. Mahdavi, C. Miller, I. Mino-Rodriguez, I. Pigliautile, A. L. Pisello, R.F. Rupp, A.-M. Sadick, F. Salamone, M. Schweiker, M. Syndicus, G. Spigliantini, N.G. Vasquez, D. Vakalis, M. Vellei, S. Wei, Quality criteria for multi-domain studies in the indoor environment: critical review towards research guidelines and recommendations, Build. Environ. 226 (2022) 109719, https://doi.org/10.1016/j.buildenv.2022.109719.
[120] A. Gelman, E. Loken, The statistical crisis in science, Am. Sci. 102 (2014) 460–465.

Matteo Favero is a Postdoctoral Research Fellow at the École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. He has a background in Building Engineering from Politecnico di Milano (Italy) and holds a PhD in Civil and Environmental Engineering from the Norges teknisk-naturvitenskapelige universitet (NTNU), Norway. His research interests and expertise include (dynamic) thermal comfort, occupant-centric models for thermal comfort in buildings and statistical thinking.

Salvatore Carlucci is a Full Professor at the Department of Theoretical and Applied Sciences of the Insubria University in Varese, Italy. He received MSc and PhD degrees, both with honours, from Politecnico di Milano in Italy. His research interests include building physics, occupant-centric building design and operation, indoor environmental quality (thermal, visual and acoustic comfort, and indoor air quality), climate change and microclimate simulation, high-performance buildings and smart building technologies, and building performance simulation and optimisation.

Giorgia Chinazzo is an Assistant Professor of Instruction in the Department of Civil and Environmental Engineering at Northwestern University, USA. She has a background in Architectural Engineering from Politecnico di Torino and Milano (Italy) and holds a Ph.D. in Civil Engineering from the École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. Her research focuses on the influence of multi-domain exposures in the built environment on different human responses, spanning from perceptual to physiological.

Jan Kloppenborg Møller is an Associate Professor in stochastic dynamical systems at the Technical University of Denmark (DTU). He holds a MSc in Applied Mathematics from DTU (2006) and a PhD in Engineering from DTU (2011). His research concentrates on modelling and forecasting (continuous or discrete time) stochastic dynamical systems. He has been involved in projects and worked on many different systems like ecosystems, urban drainage, water treatment, wind and solar power forecast, and occupancy behaviour. In all cases focus has been on models that have a clear interpretation of states and parameters.

Marcel Schweiker is a Full Professor heading the Healthy Living Spaces lab at the Institute for Occupational, Social and Environmental Medicine at the Medical Faculty of RWTH Aachen University. He has a background in Architecture from the University of Kassel, Germany, and a PhD in Environmental and Information Sciences from Tokyo City University, Japan. His core interests are crossing disciplines in improving our understanding of those indoor environmental conditions promoting well-being and health within the built environment.

Marika Vellei is a permanent Researcher at the French National Centre for Scientific Research (CNRS) and she is based at the University of Bordeaux, France. She has a background in Energy Engineering and holds a PhD in Building Engineering from the University of Bath, UK. Her current research focuses on better understanding and predicting thermal comfort from physiological signals.

Andrew Sonta is an Assistant Professor of Civil Engineering at the École Polytechnique Fédérale de Lausanne (EPFL), Switzerland and director of the ETHOS Lab: Engineering and Technology for Human Oriented Sustainability. He holds a PhD from the Sustainable Design and Construction program of the Department of Civil and Environmental Engineering at Stanford University and was previously a postdoctoral fellow at the Data Science Institute at Columbia University. His research focuses on human-building interaction across the multiple scales of the built environment.