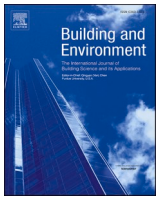
Building and Environment 283 (2025) 113390

Contents lists available at ScienceDirect

Building and Environment

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



The untapped potential of Bayesian region of practical equivalence for assessing null effects in multi-domain research

Jian Pan [a,*], Rania Christoforou [a], Lucy Nives Wiedermann [a], Marcel Schweiker [a,b]

[a] Healthy Living Spaces lab, Institute for Occupational, Social and Environmental Medicine, Medical Faculty, RWTH Aachen University, Pauwelsstr. 30, 52074, Aachen, Germany
[b] Chair of Healthy Living Spaces, Faculty of Architecture, RWTH Aachen University, Germany

ARTICLE INFO

Keywords:
Region of practical equivalence
Null effect
Bayesian statistical inference
Null hypothesis significance testing
Multi-domain
Cross-modal

ABSTRACT

Multi-domain research has become a popular topic in building science. However, studies in this area have not yet paid sufficient attention to the inadequacy of the dominant research practices entrenched in the conventional null hypothesis significance testing (NHST) when dealing with null effects. To address this problem, we first explain why null effects inherently exist in multi-domain research and argue for the importance of null effects. We then highlight numerous limitations of NHST that are particularly relevant to multi-domain research, for example the inability to accept null hypothesis or to refute false theories, the untested and thereby unfalsifiable alternative hypotheses, the underpowered results that mislead cumulative research and aggravate publication bias, the often ignored testing intentions, the fallacious black-and-white thinking, unintuitive interpretation and common misunderstanding, and the asymptotic behavior. Inspired by advances in methodology literature, we introduce Bayesian region of practical equivalence (ROPE) as an alternative approach that allows researchers to decide for, against, or remain undecided regarding whether an effect of research target is practically equivalent to a value of interest (e.g., the null value). The Bayesian ROPE approach has several advantages over NHST, namely the ability to establish practical absence of an effect and to refute false theories, better scalability, no requirement on minimum sample size, intuitive interpretations, and independence from testing intentions. We contend that this alternative approach can promote cumulative multi-domain research and mitigate publication bias.

1. Background

Building science has been paying increasing attention to multi-domain effects among thermal, visual, acoustic, and air quality domains. As extensive reviews (e.g., [1,2]) on previous studies have concluded that existing multi-domain results are often inconsistent or even contradictory, quality criteria have been proposed as guidelines for future multi-domain research [2]. While we recognize the contribution of these quality criteria to improved research practices, we highlight that they are still entrenched in the paradigm of conventional null hypothesis significance testing (NHST), which falls short when it comes to null results (i.e., non-significant results). This limitation seriously hinders the evaluation of null effects that are an inherent part of multi-domain research. In this paper, we understand null effects as either the absence of causal relationships or as practically negligible effects of a trivial size.

1.1. Reasons for the existence of null effects in multi-domain research

Multi-domain research inherently involves null effects for several reasons. First, although researchers have spent much effort in searching for significant multi-domain effects, numerous studies from past decades have yielded abundant null results (e.g., [3–6]). While some of these studies were underpowered and methodologically flawed (see reviews by Schweiker et al., [1] and Chinazzo et al., [2]), so that their null results are insufficient for supporting null effects, the repeated observations of null results by different researchers under diverse settings indicate the potential existence of null effects in multi-domain research.

Moreover, theories and solid arguments are still lacking regarding why considerable multi-domain effects would exist between any combination of multi-domain aspects [1]. One common assumption in the literature is that there are (substantial) multi-domain effects because humans have multiple senses that influence each other (e.g., [2,7–9]).

* Corresponding author.
E-mail address: jpan@ukaachen.de (J. Pan).

https://doi.org/10.1016/j.buildenv.2025.113390
Received 18 November 2024; Received in revised form 20 June 2025; Accepted 4 July 2025
Available online 5 July 2025

While we agree this might be true for certain combinations of specific multi-domain aspects, we do not think this assumption applies to every combination among all possible aspects. Rather, based on common sense and daily experience, we consider null effects as more plausible for many multi-domain combinations. For example, the causal effect of changing CO_2 from 100 ppm to 500 ppm—assuming other environmental factors are properly controlled for—should not have much practical influence on acoustic perception, nor should the causal effect of changing a room from 20 °C to 21 °C noticeably alter our perception of glare.

Furthermore, since the effect size often depends on the magnitude of the difference between conditions, once the difference is small enough, the effect size will inevitably become negligible. For example, assuming temperature does influence our hue perception (i.e., hue-heat-hypothesis; [7,10]), if the difference in manipulation between the thermal conditions is trivial, say 0.1 °C, it is plausible to expect no practical difference in the corresponding hue perception. While a difference of 0.1 °C in this example seems obviously too trivial to be of common research interest, for many (understudied) domains, it can often be unclear whether the difference between conditions is large enough to produce any non-trivial effect sizes. Thus, we expect that at least some research will unconsciously employ conditions with a trivial difference and thereby unintentionally face null effects.

Finally, most empirical research involves an unlimited number of aspects that cannot be controlled or fully randomized (e.g., the characteristics of the equipment, the oxygen level during the experiment, or even the ranking in Bundesliga). Based on scientific evidence, common knowledge, and practical constraints, most of these aspects will be deemed as irrelevant (i.e., null effects) and thus intentionally omitted in research [11]. In this way, null effects are, even if implicitly, an inherent part of research.

1.2. Relevance of null effects (Why should we care about null effects?)

Although the existing research practices generally focus on rejecting null hypotheses and favor non-zero effects (i.e., publication bias; [12–15]), there are several reasons why we should also care about null effects. On the scientific side, knowing the non-existence of an effect is a gain in knowledge. Such knowledge helps us identify irrelevant factors, allowing future research to omit them and focus on (potentially) relevant aspects [11]. Also, evidence supporting null effects can challenge and potentially falsify hypotheses and theories that expect practically meaningful, non-trivial causal effects [16]. By ruling out incorrect theoretical predictions, null effects pave the way toward the discovery of meaningful effects in the sense of cumulative science. Furthermore, since science generally explores the unknown, there is no guarantee regarding whether the effects of research interest are null or not. Thus, it is important for researchers to explicitly consider the possibility of null effects and to be able to appropriately assess them. On the practical side, null effects point out irrelevant aspects to practitioners, thereby informing cost-benefit analysis and contributing to more efficient and effective designs. In this way, null effects facilitate resource allocation and promote sustainability.

1.3. Research gaps and goals of this article

To summarize, we contend that null effects are essential in multi-domain research and that knowledge of null effects is scientifically and practically meaningful. However, the dominant research practices entrenched in NHST have severe limitations, especially when dealing with null effects. Although advances in methodology have proposed alternative practices that overcome these limitations, their potential has remained almost untapped in previous multi-domain studies. Below, we first briefly explain NHST and address its limitations. We then introduce Bayesian region of practical equivalence as an alternative approach, explain its advantages over NHST, and discuss its limitations. Finally, we provide simulated examples to demonstrate application of this alternative approach. All codes and statistical details related to the simulations of this article are presented in supplementary material.

2. Limitations of null hypothesis significance testing

2.1. Brief introduction to null hypothesis significance testing

In existing multi-domain research, the dominant statistical procedure follows the null hypothesis significance testing (NHST). This section briefly explains core concepts regarding NHST, based on the introduction in [17,18].

The NHST procedure starts with a null hypothesis proposing a certain value for a quantity of interest (e.g., the effect of hue on thermal perception is zero), and an alternative hypothesis that is mutually exclusive with the null hypothesis (e.g., the effect of hue on thermal perception is NOT zero). Then data are collected and a test statistic summarizing the observed data is computed (e.g., a *t*-statistic in a *t*-test).

Next, imaginary samples are repeatedly sampled from a hypothetical population that conforms to the null hypothesis. Each imaginary sampling occurs in the same way as the sampling of observed data (i.e., the testing intentions are the same as observed data, more details below). For each imaginary sample, the same summary statistic is calculated.

The summary statistics from all imaginary samples form a sampling distribution (e.g., a *t*-distribution in a *t*-test). This sampling distribution tells us the probability of observing a summary statistic that is at least so extreme as the summary statistic of the data, assuming the entire statistical model—including the null hypothesis, the data collection process, and all other assumptions about data generation—is correct [19]. This probability is known as the *p*-value.

When p is below a pre-defined threshold (i.e., the α-level, conventionally at 5 %), the results are said to be significant, and the null hypothesis is rejected. On the other hand, when p exceeds the threshold, the results are called null results and the null hypothesis cannot be rejected.

Statistical power is the probability of correctly rejecting a false null hypothesis and can be calculated as the probability of rejecting the null hypothesis under the assumption of a certain effect size [20]. A larger sample size increases statistical power, and a greater effect size also enhances power. If the true effect size is smaller than the assumed effect size, then the actual power given the true effect size will be lower than the calculated power based on assumed effect size. If the true effect size is zero, then the concept of power becomes meaningless, since it necessarily assumes a non-zero effect.

In NHST, two types of decision errors are differentiated [20]. A Type I (sometimes referred to as false-positive) error occurs when the null hypothesis is incorrectly rejected. Assuming statistical assumptions are met, the long-run probability of committing Type I errors is controlled by the NHST procedure at the α-level. A Type II (sometimes referred to as false-negative) error occurs when the null hypothesis is incorrectly retained even though a true effect exists. The long-run probability of committing Type II errors, given a specific effect size, can be calculated as one minus the corresponding statistical power. For example, with a power of 80 %, the generally accepted convention value [21], there is still a 20 % chance for false-negative errors in the long run. This means that, across many hypothetical studies with the same effect size and power, approximately 20 % would fail to detect the true effect.

2.2. Null results related limitations

NHST has several limitations when dealing with null results in multi-domain research. To start with, an absence of significant results does not necessarily mean an absence of effects. This fact is primarily due to sampling variability: observed data may deviate from the true effect size to an extent that does not allow the rejection of the null hypothesis, resulting in a false-negative error.

Low statistical power, often resulting from small sample sizes, increases the risk of false-negative errors, making null results unreliable and potentially wasting research resources. Thus, NHST is not appropriate for studies with very small samples [22–26]. Unfortunately, it is not uncommon for multi-domain studies to recruit few participants (e.g., less than 10 subjects [2]). Such small samples can yield statistical power well below the conventional 80 % threshold, especially when the expected effect sizes are small to moderate, as is often the case in real-world occupant-centric research due to the inherent variability in human responses and complex environmental conditions. In such underpowered studies, null results can be unreliable because high false-negative error rates make it difficult to detect true effects. These null results are more likely to reflect noise than a genuine absence of effect.

Importantly, null results from studies with a large sample, although more reliable due to reduced false-negative risks, still do not confirm the null hypothesis, because the power calculation is always based on an a priori assumed effect size. If the true effect size is smaller than assumed, the corresponding power for the true effect size will be smaller, so that the possibility of erroneously keeping the null hypothesis becomes larger. With a small enough true effect size, a limited sample size will eventually end with insufficient power. As many researchers (e.g., [27–29]) have emphasized, conventional NHST can only fail to reject the null hypothesis, but never accept it; significance only provides evidence against the null hypothesis, but not in favor of it.

For decades, statistical literature (e.g., [18,30]) has repeatedly pointed out that the null-results-related limitations of NHST could mislead cumulative research. For example, suppose several underpowered studies with NHST have been independently conducted regarding the multi-domain effect of lighting on sound perception. For demonstration purposes, let's assume the true effect is of practically negligible size. In ideal situations, we should observe many null results. Unfortunately, the individual null results will be inconclusive because of the lack of power. Even worse, because small sample sizes have a greater sampling variability, studies with a very small sample size are more susceptible to producing unstable and potentially exaggerated effect size estimates—particularly when significant results are obtained (Fig. 1). Such exaggerated results can distort the interpretations of the effect and its implication. Considering the prevalent publication bias that favors significant results [12,13,31], such non-zero results will be published more easily than null results. Even if the majority of the literature reports no effect, a few significant results may already suffice to motivate common post-hoc misinterpretations such as that there are other factors influencing the effect/association or that something was wrong with the studies with null results. In this hypothetical example, these interpretations are obviously wrong since the puzzling state of literature is only caused by random variation and the limitation that NHST is not able to appropriately assess null effects.

On the other hand, now we assume some other underpowered studies applying NHST to investigate the multi-domain effect of lighting on visual comfort. This time, suppose the real effect size is moderate. Because of insufficient power, suppose half of the studies observed null results, while the other half found significant results. Again, such a mixed literature state might mislead future research into investigating potential moderating or confounding factors to explain the inconsistencies, while in reality, the significant findings may be false-positives driven by random variation and the limitations of NHST.

Despite these examples being fictional, in reality, the limitations of NHST on null results have led to a "vast graveyard of undead theories" [31]. Psychological theories can often end in a zombie state. That is, theories, even false ones, will often hang around until forgotten [32], because NHST does not provide an informative assessment of evidence for the null hypothesis nor allow for a practical refutation of false theories. Although theory development in multi-domain research still requires much effort, some hypotheses and predicted effects remain similarly in an "undead state" (e.g., the thermo-photometric perception hypothesis [33]) and continue consuming limited research resources.

The above-mentioned limitations are particularly relevant to multi-domain research as this field may observe null results quite often for several reasons. To start with, as mentioned in Section 1.1, previous studies have repeatedly observed numerous null results and at least some multi-domain effects are likely to be of practically neglectable size. Also, because of resource limitations and practical restrictions, it is common for multi-domain experiments to adopt a relatively small sample size, as a previous review [2] showed. The resulting insufficient power can often lead to null results even under the presence of real effects. Furthermore, multi-domain interactions, one of the main research interests of some multi-domain studies, can be methodologically even

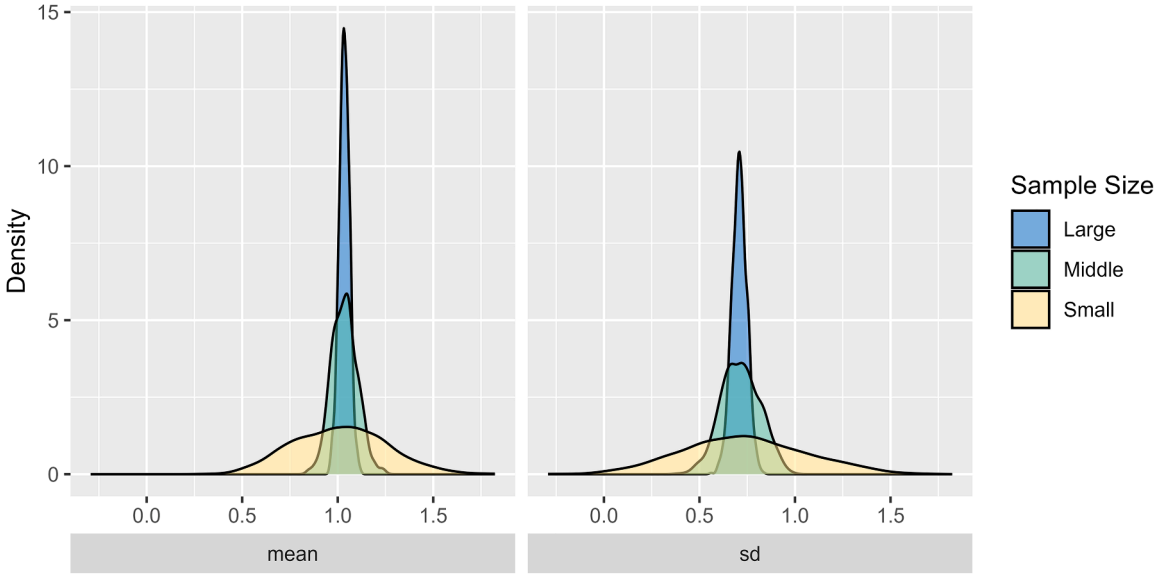


Fig. 1. Density distributions for the mean and the standard deviation (sd) of simulated samples of three sample sizes. Sample sizes of 10, 50, and 200 were repeatedly randomly drawn from a subject pool of 2000. The distributions of means and sds are shown for the three sample sizes. This simulation shows that small samples have a wide density distribution, thereby a larger variance, than large samples. Thus, studies with small sample sizes, common in multi-domain research, are more likely to yield unstable and potentially exaggerated effect size estimates because of greater sampling variability. Such exaggerated results can aggravate publication bias.

harder to detect than main effects. This is because statistical power for interactions is generally lower than for main effects [34,35], due to (a) the typically small standardized effect sizes observed in empirical research studying interactions (e.g., [36]), (b) the increased variability introduced by product terms used to model interactions [35], and (c) the compounded measurement error resulting from combining multiple variables [35]. Finally, the quality of existing research instruments can also restrict the detectable effect size. Based on review results [2], previous multi-domain research often deployed measurements without any psychometric validation to assess psychological constructs, such as perception and comfort. Such measurements can be of questionable reliability and validity, thereby introducing abundant random variance, which can potentially overwhelm the systematic variance from true effects. As a result, studies on effects below a certain size would be more likely non-significant.

2.3. Limitations not related to null results

Beyond the limitations related to null results, diverse statistical literature has also highlighted more general limitations of NHST [18, 37–42]. Based on our experience, these limitations have seldom been considered in multi-domain research. Below, we mention a few that we deem as most relevant for multi-domain research.

First, current analysis practices often ignore the fact that testing intentions influence significance [18]. As introduced earlier, NHST computes the *p*-value with imaginary sampling distributions that have the same testing intentions as observed data. With different testing intentions, the imaginary sampling distributions will differ, causing the *p*-values to change accordingly [18,43].

For example, suppose the testing intention is to test a certain hypothesis multiple times, while the *p*-values are obtained from standard statistical procedures that by default assume a one-time testing intention (e.g., ANOVA or linear regression). If these *p*-values are interpreted at their face value, the probability of committing at least one false-positive error (i.e., family-wise error rate) will be higher than the α-level ([40, 41]; simulation example in Table 1). This issue is known as α-cumulation and can be addressed by adjusting the α-level (e.g., using Bonferroni correction [44]), often with the consequence of reduced power. Unfortunately, we often observe multi-domain studies ignoring their testing intentions and committing α-cumulation, thereby increasing the risk of false-positive errors. This issue is particularly relevant in multidomain research, where multiple dependent variables (e.g., satisfaction, comfort, and sensation from thermal, acoustic, visual, and olfactory domains) across multiple measurement time points can often be analyzed simultaneously under the same overarching hypothesis (e.g., an intervention improves perception).

Another inherent problem of NHST is that the null hypothesis can be wrong from the beginning [42,45–47]. As introduced above, the null hypothesis expects the effect of interest to be EXACTLY equal to a specific value (say absolute zero). However, in reality, the effect of interest may likely have at least some tiny deviation (say 0.000000001 °C of room temperature) from the exact value (e.g., absolute 0 °C) as expected by the null hypothesis. Then, a large enough sample will eventually lead to detection of such non-zero deviation as the power increases with the sample size, and thereby reject the null hypothesis.

A related problem is that NHST generally focuses on rejecting the null hypothesis that represents a null effect, but does not directly test the alternative hypotheses that are of central research interests. When the null hypothesis is a priori wrong, as explained above, then any theory or hypothesis that expects some non-null effect, regardless of its correctness,[1] will be automatically "confirmed" given large enough sample sizes. In this way, the alternative hypotheses are not falsifiable since they are never tested. Although these problems have been pointed out half a century ago [47], most multi-domain research still sets out to reject the null hypothesis, thereby suffering from the same issues.

NHST also induces fallacious black-and-white thinking which has been criticized repeatedly by numerous statisticians (e.g., [18,38,39]). Concretely, not few multi-domain researchers have focused their statistical analyses on making a dichotomous decision between the null and alternative hypotheses (or between "non-significant" and "significant" results). As a result, they (mainly) cared about *p*-values while ignoring further information about the magnitude or the uncertainty of the estimated effect [18]. Such information is critical [48], because significant results may come from an extremely small effect that is practically irrelevant. Also, significant results can greatly differ regarding their range of uncertainty. In extreme cases, the estimation of a significant effect can range from trivial to enormous (Fig. 2). Problematically, although these very different magnitudes of the estimated effect are all plausible, they would have totally different interpretations for the effect of interest and practical implications. Thus, common analysis practices often complement NHST with effect sizes and confidence intervals. Specifically, the effect size indicates the magnitude of an effect, while the confidence interval (CI) tells us the uncertainty of the estimation.

Table 1
False-positive rates from a simulated example. Data were randomly generated with the same sample size for three dependent variables and one independent variable, which has no effect on any of the dependent variables. Linear regressions were conducted respectively on the dependent variables and the percentage of significant results were calculated. The false-positive rates for one-time testing (i.e., the testing intention was to test only one dependent variable) were round the α-level = 0.05. The family-wise false-positive rates (i.e., the testing intention was to test three dependent variables at the same time; significant results were found on at least one of the three dependent variables) became about two times higher than the alpha level.

	Dependent variable A	Dependent variable B	Dependent variable C	Family-wise
False-positive rates	.0488	.0473	.0489	.1382

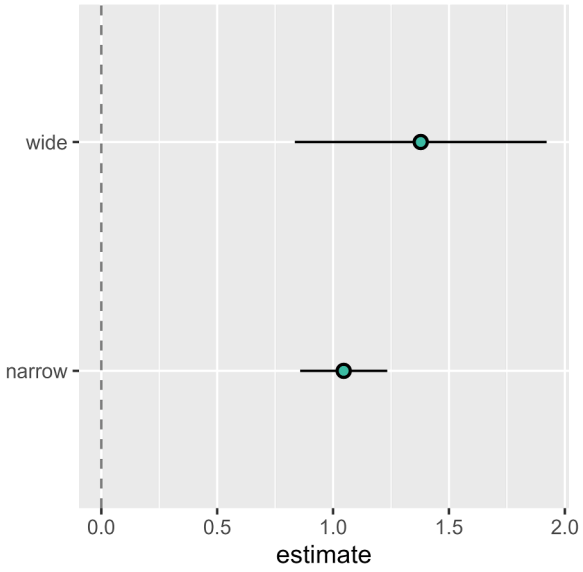


Fig. 2. Two simulated 95 % confidence intervals. Both results are significant as both confidence intervals exclude zero. The upper estimate has a wide confidence interval and thus large uncertainty, while the lower estimate has a narrow confidence interval and small uncertainty.

[1] For example, a hypothesis that predicts a positive effect of size 1 would be wrong if the real effect is negative of size 3.

The 95 % CI is conventionally used in multi-domain research and can be defined as the range of values with corresponding *p*-values not smaller than the α-level [49].

Furthermore, NHST is not intuitive and thus often misunderstood. Studies from diverse fields have repeatedly shown high rates of misunderstanding of basic NHST concepts among researchers, including some with extensive statistical training and research experience (e.g., [37]). For example, researchers often incorrectly interpret *p*-values as the probability of the null hypothesis and 95 % CIs as containing the true value (assuming the statistical model is correct) with a 95 % probability (see previous section for their actual meaning). One reason behind common misinterpretations of NHST concepts is that NHST is designed to answer a question that is often not of interest to researchers. To be specific, the *p*-value, one core output from NHST, tells us the probability of observing the data, given that the null hypothesis is true (e.g., if a coin is fair, what is the probability of observing ten heads in a row). However, most researchers are more interested in the inverse probability question [20]: what is the probability of the null hypothesis, given the observed data (e.g., when we observe ten heads in a row, what is the probability that the coin is fair). These non-intuitive aspects of NHST make it difficult to properly apply it and lead to a multitude of diverse misinterpretations, hindering cumulative multi-domain research.

3. Bayesian region of practical equivalence as alternative approach

In light of the numerous limitations of NHST, various alternative methods have been developed for better data analysis. Unfortunately, their potential remains largely unexplored in multi-domain research. Here, we aim to introduce the Bayesian region of practical equivalence (ROPE) as an alternative to NHST. Most importantly, this alternative approach allows null results to also be informative, so that false theories can be empirically refuted, thereby promoting cumulative science.

3.1. Bayesian analysis

Several publications have introduced Bayesian analysis and related methodological alternatives within the context of building science (e.g., [50–52]). As this article is not intended to serve as a comprehensive tutorial or systematic review of the broader Bayesian inference and its implementation, we deliberately focus our introduction on the essential elements necessary to support the arguments and applications presented in the subsequent sections. Our introduction is based on Kruschke [53]. We refer interested readers to other more extensive yet still accessible materials on beginner-level Bayesian analysis and open-source tools (e.g., [23,54]).

Bayesian analysis starts with a prior (Fig. 3), that is, a distribution representing our a priori belief regarding the respective credibility of candidate values for the target of interest (e.g., a coefficient in a regression model between illuminance and thermal sensation). Then based on data and statistical model, credibility is re-allocated across candidate values following Bayes' rule [55] toward the values that are more consistent with the data. The resulting distribution of the candidate values is the posterior (Fig. 3). A certain percentage of the values with the highest posterior credibility forms the highest density interval (HDI). For example, the values inside the 95 % HDI represent the 95 % most credible values in posterior. In Bayesian literature, the 95 % HDI is one commonly used approach to characterize the uncertainty of the estimation.[2]

3.2. Region of practical equivalence

As an alternative to NHST, Kruschke et al. [18] have proposed a procedure[3] that allows assessing whether some specific value of interest (e.g., the null value representing a null effect between temperature and acoustic satisfaction, or a non-null value of 0.1 K representing a hypothesized effect of illuminance on the skin temperature) is among the most credible values in the posterior. Concretely, a ROPE is first established to serve as a decision threshold. The ROPE includes all values around the value of interest with a trivial difference that is considered too small to be meaningful. In other words, these values in the ROPE are regarded as practically equivalent to the value of interest.

The ROPE should be specified before data collection and transparently justified, for example in a pre-registration [2]. The justification could be based on previous results, background theories, and measurement precision in relation to the expected effect size [18]. Following the recommendations by Lakens et al. [16,59], ideally, a cost-benefit analysis may be conducted to set the ROPE. The main target of the cost-benefit analysis is to clarify how a study with a high probability of rejecting or accepting the defined practical equivalent region will add value to the existing knowledge. It is natural that ROPE varies across studies, researchers, fields, and time as costs and benefits can be very subjective. Importantly, the ROPE should remain independent of the study's own results. For example, researchers should not adjust the ROPE based on their own results to make sure that the ROPE will support the decision they had hoped for. Also, the ROPE should be chosen to ensure that the statistical inferences address important scientific questions. For instance, if the ROPE range is too wide (say equals 100 as in Cohen's *d*), it would likely be meaningless even if a practical equivalence could be established.

Admittedly, under the current research status, substantive information regarding previous results, background theories, and measurement precision can be lacking for many multi-domain research questions, so that it will be hard to develop sound justifications and cost-benefit analysis for setting up the ROPE. In such situations, researchers may follow the suggestion by Kruschke et al. [18] and set the default ROPE as ±0.1 standard deviation of a parameter, which is analogous to a negligible effect size according to Cohen [60].

After the ROPE is determined, it is then compared with the 95 % HDI for making decisions. There are three possible outcomes. If the ROPE entirely includes the 95 % HDI of the posterior (Fig. 4 "Include"), then the 95 % most credible values are deemed as practically equivalent to the value of interest. Thus, the value of interest, which could also represent a null effect, will be accepted. On the other hand, if the ROPE entirely excludes the 95 % HDI (Fig. 4 "Exclude"), the value of interest will be rejected as the 95 % most credible values are not equivalent to the value of interest. If the ROPE partially overlaps with the 95 % HDI (Fig. 4 "Overlap"), then the result from the current data is inconclusive and no decision is made. In this case, more data are required to narrow down the 95 % HDI until a decision can eventually be reached.

It is important to note that this Bayesian ROPE procedure can only confirm the practical absence of effects that are more extreme than the trivially small difference as specified by the decision threshold of the ROPE. In other words, the Bayesian ROPE procedure cannot prove that an effect size is exactly zero. In fact, as statistical literature has pointed out (e.g., [46,61]), no probabilistic methods can ever prove that an effect is absent or present, because there is always random variation, so that rare events can happen.

[2] There are various other types of credible intervals. This article focuses on the 95% HDI following the convention in the classic ROPE approach by Kruschke & Liddell [18].

[3] While various alternatives to NHST exist, such Bayesian hypothesis testing, alternative ROPE-based approaches, and frequentist equivalence testing, we focus on Kruschke et al.'s [18,56] classic Bayesian ROPE approach due to its interpretability, accessibility, and alignment with the inferential needs of our field. Interested readers may refer to broader methodological overviews (e.g., [57,58]).

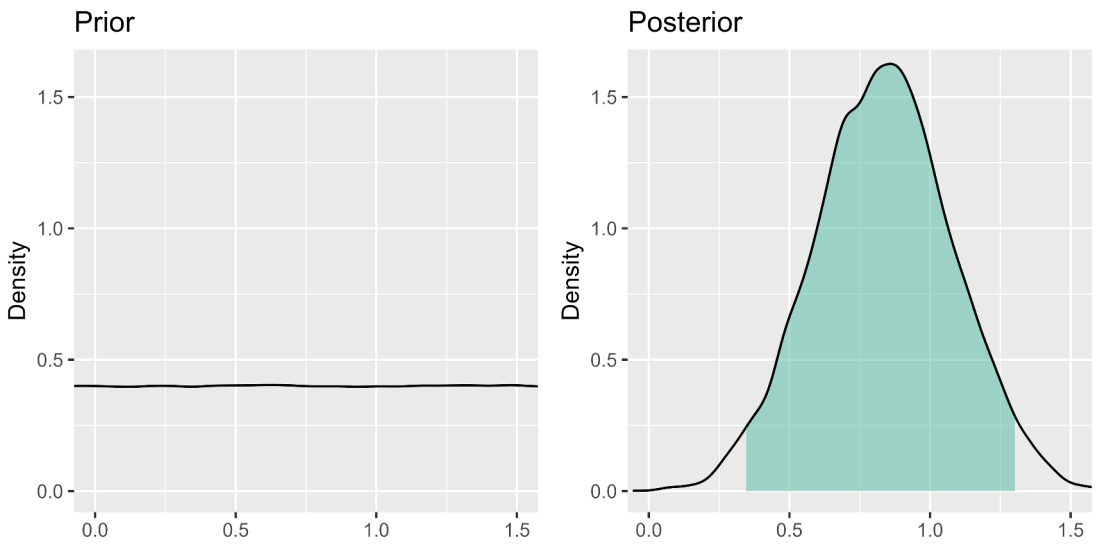


Fig. 3. Simulated example of prior and posterior distributions of credibility (i.e., probability density) for a simulated regression coefficient. The prior follows a uniform distribution. The 95 % highest density interval (HDI) is the green area under the posterior distribution.

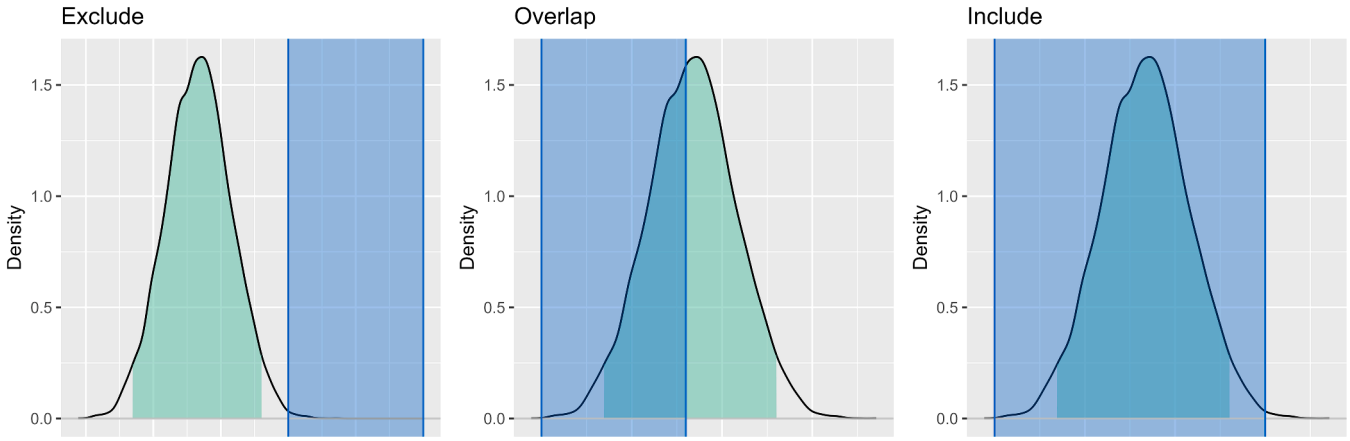


Fig. 4. Illustrations of three possible outcomes when comparing a region of practical equivalence (ROPE; the rectangular region in blue) with a highest density interval (HDI; the green area under the posterior curve). To visually distinguish the outcomes, each panel depicts a separate illustrative scenario with a different ROPE and posterior distribution. The left panel shows a ROPE that entirely excludes the HDI, supporting a decision to reject the value of interest lying within the ROPE. The middle panel shows a ROPE that partially overlaps with the HDI, indicating inconclusive results. The right panel shows a ROPE that entirely includes the HDI, supporting a decision to accept the value of interest.

3.3. Advantages over conventional approach

The Bayesian ROPE procedure has several advantages over NHST. Above all, it allows null results to be informative. As mentioned in previous sections, null results from the conventional NHST are inconclusive and do not provide evidence in favor of the null hypothesis. With Bayesian ROPE, researchers will be able to establish practical absence of an effect based on null results. In this way, alternative hypotheses that predict a presence of effect become falsifiable, thereby also refuting incorrect theories behind the alternative hypotheses. As such, alternative hypotheses will not be automatically confirmed just because the null hypothesis was wrong in the first place due to tiny deviation from the absolute null (cf. Section 2.3.), and theories will not easily end up in a zombie state (cf. Section 2.2.).

In comparison to NHST, Bayesian ROPE will better promote cumulative research. For example, Bayesian ROPE does not aim at finding significant results as the case with NHST; researchers can now make equally important contributions with null results. In this way, the above-mentioned black-and-white thinking regarding significance and publication bias [12–15] will be mitigated. Also, since this alternative procedure is based on the Bayesian approach, it can incorporate existing knowledge, such as previous data, theories, and expert expectations, in the analysis in the form of a prior. In other words, Bayesian results from previous studies (the posterior) can be incorporated into the analysis of later studies (as prior), known as posterior passing [62]. To formulate a prior, original data are not necessary once the posterior is available. Importantly, as the cumulative sample size increases, Bayesian ROPE will eventually converge to the correct decision [14]. In this way, research with Bayesian ROPE is more scalable and could be thought of a relay race.

In comparison, separate NHST results typically cannot be reused in the analysis of another subsequent study if the underlying data are not shared, which is often the case in the existing multi-domain studies.[4] Additionally, even when the data are available, it can be challenging to directly use these data in the subsequent NHST analysis (e.g., because of

[4] Meanwhile, we acknowledge that we can expect to see more data available in the future because of various attempts to improve the situation, such as open data practices and funding requirements.

different experiment designs or data structure). That is, conventional NHST calculations often need to start from scratch and its results cannot be directly reused in future analysis. Suppose we are to perform a conventional ANOVA, access to previous data is necessary to combine it with our own and conduct the analysis on the integrated dataset. If the previous data were not available or the combination with our own data is not possible, our ANOVA results would rely solely on our own data. Such limitations lead to a waste of research resources. Even worse, random variation and lack of power often lead to an unclear state of literature, when part of the results report significant results, while the others do not. Despite a large number of studies, conventional NHST cannot clarify inconsistent results and often leads to misleading interpretations as mentioned in previous section. Moreover, as argued above, NHST can often incorrectly confirm any alternative hypothesis that expects any non-null effect. This issue arises because the null hypothesis can often be a priori wrong. As previously discussed, the null hypothesis expects the quantity of interest to be of an exact value (i.e., the procedure rejects a point-wise hypothesis), but in real world, tiny deviations from this exact value likely exist [42,45,46]. As a result, a large enough sample will necessarily detect such tiny deviations and lead to the rejection of the null hypothesis, thereby automatically "confirming" any alternative hypothesis regardless of its correctness.

Furthermore, Bayesian ROPE does not require a minimum sample size [14,38,45]. One practical issue in multi-domain research is that large (experimental) samples are often not feasible or available. This limitation poses a problem for NHST, because many common NHST procedures, such as linear mixed models, χ^2 tests, and Z-tests, are asymptotic, meaning they rely on large-sample approximations to ensure valid results [25,26]. When sample sizes are small, NHST results can be questionable due to their dependence on large-sample approximations.

In contrast, Bayesian ROPE does not rely on asymptotic approximations and provides valid inference at any sample size (even at $n = 0$), provided the prior and model assumptions are appropriately specified and hold [23,24]. This flexibility in sample size makes Bayesian ROPE especially suitable for studies with limited samples, a common scenario in multi-domain research.

Additionally, Bayesian ROPE does not involve imaginary sampling distributions. As mentioned in the brief introduction to NHST, p-values depend on the imaginary sampling distributions that are directly influenced by the testing intentions. Therefore, with different testing intentions, identical data can result in totally different p-values and corresponding CIs. Based on our experience, much multi-domain research did not consider these distributions, so that the significance would be (unintentionally) biased. In worse cases, testing intentions can easily be intentionally misreported to reach significance in the sense of p-hacking (e.g., [63]) and questionable research practices (e.g., [63–65]). For example, if multiple tests are conducted but only the significant tests are reported, the actual p-values and confidence intervals will be larger than reported because of the discrepant testing intentions. Such unethical practices can be individually hard to detect and bias the literature. In comparison, the issues related to testing intentions are mitigated by Bayesian ROPE, since it does not rely on imaginary sampling distributions to calculate significance, so that its results are invariant to testing intentions. Moreover, since Bayesian ROPE makes null findings more interpretable and valuable, it reduces the pressure to obtain "significant" outcomes, thereby decreasing the incentive for selective reporting. That said, it is important to note that Bayesian ROPE is not entirely immune to researchers' intentions to report only "interesting" results or those that align with their prior beliefs or theoretical preferences.

A further important advantage of Bayesian ROPE is its intuitiveness of basic concepts. As mentioned in Section 2.3., NHST basic concepts are not intuitive and thereby often misinterpreted because it targets the probability of observing the data, given that the null hypothesis is true. In contrast, Bayesian ROPE directly answers the inverse probability question (i.e., the probability of the null hypothesis, given the observed data), thus its results have a more intuitive interpretation. The posterior probability can be interpreted as the probabilities of parameter values. The 95 % HDI contains the true value (assuming the statistical model is correct) with a 95 % probability. When researchers incorrectly understand p-values as the probability of the null hypothesis and 95 % CIs as containing the true value with a 95 % probability (cf. Section 2.3.), they are misinterpreting the unintuitive p-values and CIs as if they were Bayesian posterior probability and HDI.

3.4. Limitations of the alternative approach

The Bayesian ROPE approach, like any statistical approach, has limitations—particularly in the context of complex multi-domain research. A key limitation involves the sensitivity to prior specification. Prior distributions can critically shape the posterior, especially in small-sample contexts where data alone may be insufficient to override prior beliefs. Overly informative or poorly calibrated priors can bias results or impede convergence toward the true parameter value.

Moreover, in multi-domain research, specifying appropriate priors can be particularly challenging. Domains may differ in measurement scales, variability, and underlying mechanisms, making it non-trivial to formulate priors that are both principled and context-sensitive. There is also no universal standard for what constitutes an "appropriate" prior, and the subjectivity inherent in this step may raise concerns about bias or replicability. Rather than prescribing a specific approach, we emphasize the importance of domain expertise and recommend conducting prior predictive checks and sensitivity analyses to evaluate the robustness of conclusions under different plausible priors (cf. [23,53,66]).

The Bayesian approach also relies on the assumption that the statistical model adequately represents the underlying data-generating process. If the model is misspecified—for example, due to an inappropriate likelihood function, unaccounted structural dependencies, or violated assumptions—the resulting inferences may be misleading or invalid. In the context of real-world multidomain research, such misspecifications are especially plausible, given the inherent complexity, heterogeneity, and interdependencies across domains. These points highlight the critical importance of thorough model checking (cf. [23,67,68]), such as posterior predictive checks and convergence diagnostics, to assess model fit and detect potential discrepancies between models and observed data.

From a practical standpoint, Bayesian analyses often require more computational resources than common NHST procedures. Estimating posterior distributions for large datasets or in complex models—particularly with hierarchical structures that are common in multi-domain research—typically involves Markov Chain Monte Carlo or other iterative sampling techniques (cf. [23,53]). These methods can be computationally intensive and require careful monitoring to ensure convergence to prevent unstable or misleading inference. While advances in computational tools have made Bayesian methods more accessible, they still demand technical proficiency and interpretive caution, which may pose a barrier for multi-domain and building science researchers who are unfamiliar with Bayesian statistics. Targeted training and clearer guidance may help mitigate these challenges. For readers seeking further discussion on limitations and best practices, we refer to [23,53,54,69] for additional resources on Bayesian workflow.

4. Simulated application example

In this section, the application of Bayesian ROPE procedure will be demonstrated with simulated examples.[5] To investigate the multi-domain effect of office thermal and visual conditions on body temperature,[6] a 2 × 2 between-subjects experiment was simulated with 100 subjects. The independent variables were the thermal condition (uncomfortable vs. normal) and the visual condition (uncomfortable vs. normal). Each subject was randomly assigned to one of the four resulting conditions. The dependent variable was the standardized mean skin temperature. The true effect of changing the thermal condition from uncomfortable to normal on standardized mean skin temperature was simulated to be 0.8, while the effect of changing the visual condition was set to 0.

Bayesian ROPE procedure was conducted with a Bayesian multiple regression of thermal and visual conditions on skin temperature. 95 % HDIs were respectively calculated for the posterior thermal and visual effects. For demonstration purposes, the ROPE was set to be an interval of ±0.3 around the zero (which represents a null effect), that is, 0.3 of the standard deviation of the standardized skin temperature. This equates to a small effect size following Cohen [60] .[7]

For the thermal condition, the analysis showed that the ROPE entirely excludes the 95 % HDI (Fig. 5 left). Thus, we reject the null effect and conclude a thermal effect on body temperature. For the visual condition, the ROPE partially overlaps with the 95 % HDI (Fig. 5 middle), indicating inconclusive results regarding the visual effect.

To narrow down the 95 % HDIs until a decision can be reached regarding the visual effect, a second experiment with the same setup was replicated with another 150 subjects. By applying posterior passing (i.e., using the posterior from the first experiment as the prior for the second analysis), the new analysis was able to incorporate the results from the first experiment, even without directly using the data from the first experiment (suppose they were somehow unavailable).

The second analysis revealed a ROPE that fully overlaps with the 95 % HDI for the visual condition (Fig. 5 right). Thus, we deem the visual effect on body temperature as practically equivalent to zero (i.e., accepting the null effect).

5. Conclusion

This article addressed a critical deficit underlying multi-domain research in building science, namely the inadequacy of the dominant research practices entrenched in the conventional null hypothesis significance testing (NHST) when dealing with null effects. To this end, we first explained why null effects are inherently involved in multi-domain research and provided several arguments regarding the importance of null effects. We then highlighted numerous limitations of NHST that are particularly relevant to multi-domain research yet have not received enough attention, including the inability to accept null hypothesis or to refute false theories, the untested and thereby unfalsifiable alternative hypotheses, the underpowered results that mislead cumulative research and aggravate publication bias, the often ignored testing intentions that bias significance, the fallacious black-and-white thinking, unintuitive interpretation and common misunderstanding, and the asymptotic behavior.

Inspired by advances in methodology literature, we introduce the untapped potential of Bayesian region of practical equivalence (ROPE) as an alternative approach that allows more appropriate statistical inference. By comparing a ROPE with the highest density interval (HDI) of a Bayesian posterior distribution, researchers will be able to decide for, against, or remain undecided regarding whether an effect of research target is practically equivalent to a certain value of interest (e.g., the null value). The Bayesian ROPE approach has several advantages over NHST, namely the ability to establish practical absence of an effect and to refute false theories, better scalability (e.g., via posterior passing), no requirement on minimum sample size, intuitive interpretations, and independence from testing intentions. As such, we contend that this alternative approach can promote cumulative multi-domain research and mitigate publication bias.

We acknowledge that the Bayesian ROPE approach has its own limitations, such as computational complexity, sensitivity to prior assumptions, and necessity of prior predictive checks and model diagnostics. In addition, a considerably large sample size, often at least in the order of over hundred or thousand, can be necessary for sufficiently narrowing down the HDI until a decision can be reached. Given these limitations, we concede that the Bayesian ROPE approach will not solve all problems that are entrenched in the dominant research practices. Rather, we believe that adopting this alternative approach will be an initial but crucial step toward appropriate evaluation of null effects and cumulative multi-domain research. Investigating and understanding null effects is crucial for advancing cumulative research, as it enables researchers to build upon existing results, both positive and negative, and honestly report their own findings. This alternative approach will promote a more comprehensive understanding of the phenomena under investigation and will help to address inconsistencies in the current literature.

Note that highlighting suboptimal research practices in the field can also reveal remarkable opportunities for future research. The discussed deficits of NHST and advantages of the Bayesian ROPE procedure were also meant to raise the awareness of null effects and methodology in the relevant research community and to encourage reflections and improvements on our established research practices.

More broadly, although we mainly addressed multi-domain research, our criticism on the deficits and recommendations can also benefit other research fields, for example, with regard to occupancy, occupant behavior and single-domain perception in the built environment, since challenges they face are often similar to those in multi-domain research. As such, improved methodology with the Bayesian ROPE procedure is likely to enhance not only multi-domain research, but also general research efforts in occupant-centric building design and operation.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used *ChatGPT* in order to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT authorship contribution statement

Jian Pan: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Conceptualization. **Rania Christoforou:** Writing – review & editing, Conceptualization. **Lucy Nives Wiedermann:** Writing – review & editing, Visualization, Validation, Conceptualization. **Marcel Schweiker:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

[5] The examples are purely hypothetical and illustrative.

[6] This research question may be motivated by the idea that psychological comfort (e.g., perceived relaxation or stress) could influence physiological responses like skin temperature via autonomic nervous system activity. As explored in psychophysiology, subjective states may affect physiological outcomes through mechanisms such as stress-induced alterations in peripheral blood flow.

[7] The illustrative ROPE interval in this example is not intended to serve as a definitive or generalizable threshold. As Cohen noted [60], the terms “small,” “medium,” and “large” are relative and should be interpreted in light of the specific research domain, content, and methodology.

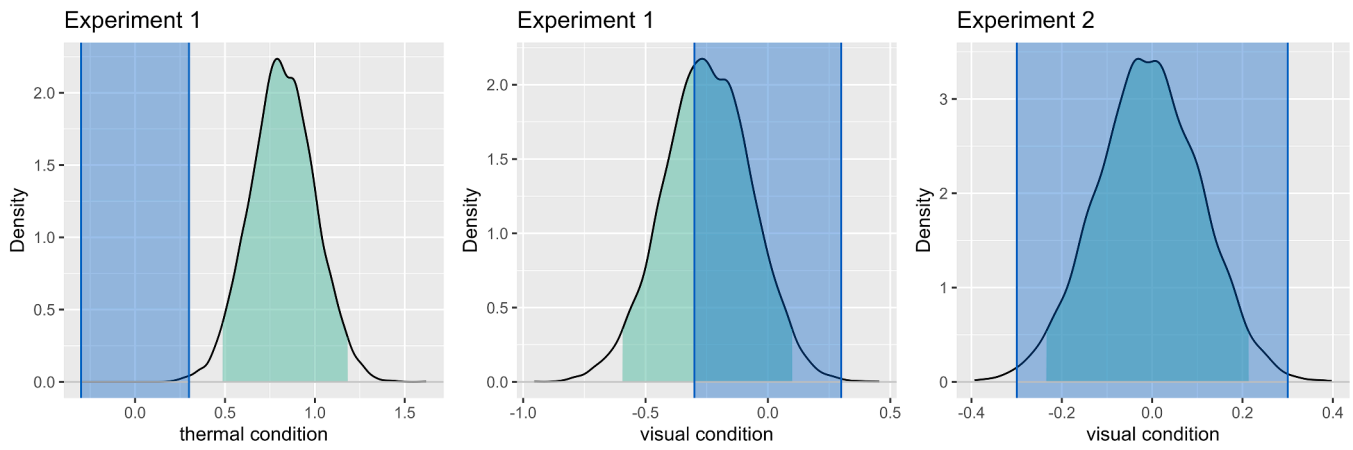


Fig. 5. Illustrations of comparing the region of practical equivalence (ROPE; the rectangular region in blue) with the 95 % highest density intervals (HDI; the green area under the posterior curve) from simulated examples.

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.

Acknowledgement

This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under the project number 498385769. Marcel Schweiker and Rania Christoforou are supported by a research grant (21055) by VILLUM FONDEN. Lucy Nives Wiedermann is supported by a research grant (03EN1081B) by the Bundesministerium für Wirtschaft und Klimaschutz (BMWK, German Federal Ministry for Economic Affairs and Climate Protection). The funding sources provided only financial support for this work. They had no further involvement. We sincerely thank Christina Marie Hofmann for her support in literature searches and constructive feedback.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.buildenv.2025.113390.

Data availability

No data was used for the research described in the article.

Reference

[1] M. Schweiker, E. Ampatzi, M.S. Andargie, R.K. Andersen, E. Azar, V.M. Barthelmes, C. Berger, L. Bourikas, S. Carlucci, G. Chinazzo, L.P. Edappilly, M. Favero, S. Gauthier, A. Jamrozik, M. Kane, A. Mahdavi, C. Piselli, A.L. Pisello, A. Roetzel, A. Rysanek, K. Sharma, S. Zhang, Review of multi-domain approaches to indoor environmental perception and behaviour, Build. Environ. 176 (2020) 106804, https://doi.org/10.1016/j.buildenv.2020.106804.

[2] 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.

[3] L. Bourikas, S. Gauthier, N. Khor Song En, P. Xiong, Effect of thermal, acoustic and air quality perception interactions on the comfort and satisfaction of people in office buildings, Energies 14 (2021) 333, https://doi.org/10.3390/en14020333.

[4] W. Yang, H.J. Moon, Combined effects of acoustic, thermal, and illumination conditions on the comfort of discrete senses and overall indoor environment, Build. Environ. 148 (2019) 623–633, https://doi.org/10.1016/j.buildenv.2018.11.040.

[5] W. Yang, H.J. Moon, M.-J. Kim, Combined effects of short-term noise exposure and hygrothermal conditions on indoor environmental perceptions, Indoor Built Environ 27 (2018) 1119–1133, https://doi.org/10.1177/1420326x17703774.

[6] G. Chinazzo, K. Chamilothori, J. Wienold, M. Andersen, Temperature-color interaction: subjective indoor environmental perception and physiological responses in virtual reality, Hum. Factors 63 (2021) 474–502, https://doi.org/10.1177/0018720819892383.

[7] G. Chinazzo, J. Wienold, M. Andersen, Combined effects of daylight transmitted through coloured glazing and indoor temperature on thermal responses and overall comfort, Build. Environ. 144 (2018) 583–597, https://doi.org/10.1016/j.buildenv.2018.08.045.

[8] A. Gentner, G. Gradinati, C. Favart, K.S. Gyamfi, J. Brusey, Investigating the effects of two fragrances on cabin comfort in an automotive environment, Work 68 (2021) S101–S110, https://doi.org/10.3233/WOR-208009.

[9] S. Lechner, C. Moosmann, A. Wagner, M. Schweiker, Does thermal control improve visual satisfaction? Interactions between occupants' self-perceived control, visual, thermal, and overall satisfaction, Indoor Air 31 (2021) 2329–2349, https://doi.org/10.1111/ina.12851.

[10] M. Ziat, C.A. Balcer, A. Shirtz, T. Rolison, A century later, the hue-heat hypothesis: does color truly affect temperature perception? in: F. Bello, H. Kajimoto, Y. Visell (Eds.), Haptics: Perception, Devices, Control, and Applications Springer International Publishing, Cham, 2016, pp. 273–280, https://doi.org/10.1007/978-3-319-42321-0_25.

[11] 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.

[12] R. Rosenthal, The file drawer problem and tolerance for null results, Psychol. Bull. 86 (1979) 638–641, https://doi.org/10.1037/0033-2909.86.3.638.

[13] G. Cumming, The new statistics: why and how, Psychol. Sci. 25 (2014) 7–29, https://doi.org/10.1177/0956797613504966.

[14] Z. Dienes, How Bayes factors change scientific practice, J. Math. Psychol. 72 (2016) 78–89, https://doi.org/10.1016/j.jmp.2015.10.003.

[15] J. Muradchanian, R. Hoekstra, H. Kiers, D. van Ravenzwaaij, The role of results in deciding to publish: a direct comparison across authors, reviewers, and editors based on an online survey, PLoS One 18 (2023) e0292279, https://doi.org/10.1371/journal.pone.0292279.

[16] D. Lakens, A.M. Scheel, P.M. Isager, Equivalence testing for psychological research: a tutorial, Adv. Methods Pract. Psychol. Sci. 1 (2018) 259–269, https://doi.org/10.1177/2515245918770963.

[17] J.K. Kruschke, Bayesian estimation supersedes the *t*-test, J. Exp. Psychol. Gen. 142 (2013) 573–603, https://doi.org/10.1037/a0029146.

[18] J.K. Kruschke, T.M. Liddell, The Bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective, Psychon. Bull. Rev. 25 (2018) 178–206, https://doi.org/10.3758/s13423-016-1221-4.

[19] 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.

[20] D.S. Moore, G.P. McCabe, B.A. Craig, Introduction to the Practice of Statistics, eighth ed., W.H. Freeman & Company, 2014.

[21] D. Lakens, Sample size justification, Collabra Psychol. 8 (2022) 33267, https://doi.org/10.1525/collabra.33267.

[22] 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.

[23] R. McElreath, Statistical Rethinking: A Bayesian Course with Examples in R and STAN, second ed., CRC Press, London, England, 2020 https://doi.org/10.1201/9780429029608.
[24] S. Depaoli, Bayesian Structural Equation Modeling, first ed., Guilford Press, London, England, 2021.
[25] M. Campo, S.W. Lichtman, Interpretation of research in physical therapy: limitations of null hypothesis significance testing, J. Phys. Ther. Educ. 22 (2008) 43–48, https://doi.org/10.1097/00001416-200801000-00007.
[26] P.M. Sedgwick, A. Hammer, U.S. Kesmodel, L.H. Pedersen, Current controversies: null hypothesis significance testing, Acta Obstet. Gynecol. Scand. 101 (2022) 624–627, https://doi.org/10.1111/aogs.14366.
[27] E.-J. Wagenmakers, A practical solution to the pervasive problems of *p* values, Psychon. Bull. Rev. 14 (2007) 779–804, https://doi.org/10.3758/bf03194105.
[28] H. Campbell, D. Lakens, Can we disregard the whole model? Omnibus non-inferiority testing for R^2 in multi-variable linear regression and η^2 in ANOVA, Br. J. Math. Stat. Psychol. 74 (2021) 64–89, https://doi.org/10.1111/bmsp.12201.
[29] C.R. Gallistel, The importance of proving the null, Psychol. Rev. 116 (2009) 439–453, https://doi.org/10.1037/a0015251.
[30] F. Schmidt, Statistical significance testing and cumulative knowledge in psychology: implications for training of researchers, Psychol. Methods 1 (1996) 115–129, https://doi.org/10.1037/1082-989X.1.2.115.
[31] C.J. Ferguson, M. Heene, A vast graveyard of undead theories: publication bias and psychological science's aversion to the null, Perspect. Psychol. Sci. 7 (2012) 555–561, https://doi.org/10.1177/1745691612459059.
[32] M.I. Eronen, L.F. Bringmann, The theory crisis in psychology: how to move forward, Perspect. Psychol. Sci. 16 (2021) 779–788, https://doi.org/10.1177/1745691620970586.
[33] T. Rakotoarivelo, B. Malet-Damour, Exploring the interplay between thermal and visual perception: a critical review of studies from 1926 to 2022, Buildings 13 (2023) 879, https://doi.org/10.3390/buildings13040879.
[34] G.H. McClelland, C.M. Judd, Statistical difficulties of detecting interactions and moderator effects, Psychol. Bull. 114 (1993) 376–390, https://doi.org/10.1037/0033-2909.114.2.376.
[35] C.P. Durand, Does raising type 1 error rate improve power to detect interactions in linear regression models? A simulation study, PLoS One 8 (2013) e71079, https://doi.org/10.1371/journal.pone.0071079.
[36] H. Aguinis, J.C. Beaty, R.J. Boik, C.A. Pierce, Effect size and power in assessing moderating effects of categorical variables using multiple regression: a 30-year review, J. Appl. Psychol. 90 (2005) 94–107, https://doi.org/10.1037/0021-9010.90.1.94.
[37] R. Hoekstra, R.D. Morey, J.N. Rouder, E.-J. Wagenmakers, Robust misinterpretation of confidence intervals, Psychon. Bull. Rev. 21 (2014) 1157–1164, https://doi.org/10.3758/s13423-013-0572-3.
[38] J.P.A. Ioannidis, Why most published research findings are false, PLoS Med 2 (2005) e124, https://doi.org/10.1371/journal.pmed.0020124.
[39] R.L. Wasserstein, N.A. Lazar, The ASA statement on *p*-values: context, process, and purpose, Am. Stat. 70 (2016) 129–133, https://doi.org/10.1080/00031305.2016.1154108.
[40] H.-M. Hsueh, J.J. Chen, R.L. Kodell, Comparison of methods for estimating the number of true null hypotheses in multiplicity testing, J. Biopharm. Stat. 13 (2003) 675–689, https://doi.org/10.1081/BIP-120024202.
[41] W. Forstmeier, E.-J. Wagenmakers, T.H. Parker, Detecting and avoiding likely false-positive findings - a practical guide: avoiding false-positive findings, Biol. Rev. Camb. Philos. Soc. 92 (2017) 1941–1968, https://doi.org/10.1111/brv.12315.
[42] I.R. Savage, Nonparametric statistics, J. Am. Stat. Assoc. 52 (1957) 331–344, https://doi.org/10.1080/01621459.1957.10501392.
[43] R. Kelter, Bayesian model selection in the M-open setting — Approximate posterior inference and subsampling for efficient large-scale leave-one-out cross-validation via the difference estimator, J. Math. Psychol. 100 (2021) 102474, https://doi.org/10.1016/j.jmp.2020.102474.
[44] O.J. Dunn, Multiple comparisons among means, J. Am. Stat. Assoc. 56 (1961) 52–64, https://doi.org/10.1080/01621459.1961.10482090.
[45] D.R. Anderson, K.P. Burnham, W.L. Thompson, Null hypothesis testing: problems, prevalence, and an alternative, J. Wildl. Manage. 64 (2000) 912–923, https://doi.org/10.2307/3803199.
[46] C. Harms, D. Lakens, Making "null effects" informative: statistical techniques and inferential frameworks, J. Clin. Transl. Res. 3 (2018) 382–393, https://doi.org/10.18053/jctres.03.2017s2.007.
[47] P.E. Meehl, Theory-testing in psychology and physics: a methodological paradox, Philos. Sci. 34 (1967) 103–115, https://doi.org/10.1086/288135.
[48] A.-M. Simundic, Confidence interval, Biochem. Med. 18 (2008) 154–161, https://doi.org/10.11613/bm.2008.015.
[49] D.R. Cox, Principles of Statistical Inference, Cambridge University Press, Cambridge, England, 2006, https://doi.org/10.1017/cbo9780511813559.
[50] M. Favero, S. Carlucci, G. Chinazzo, J.K. Møller, M. Schweiker, M. Vellei, A. Sonta, Ten questions concerning statistical data analysis in human-centric buildings research: a focus on thermal comfort investigations, Build. Environ. 264 (2024) 111903, https://doi.org/10.1016/j.buildenv.2024.111903.
[51] J. Langevin, J. Wen, P.L. Gurian, Modeling thermal comfort holistically: bayesian estimation of thermal sensation, acceptability, and preference distributions for office building occupants, Build. Environ. 69 (2013) 206–226, https://doi.org/10.1016/j.buildenv.2013.07.017.
[52] 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.
[53] J.K. Kruschke (Ed.), Doing Bayesian Data Analysis: A Tutorial With R, JAGS, and Stan, second ed., Academic Press, San Diego, CA, 2014 https://doi.org/10.1016/c2012-0-00477-2.
[54] J.K. Kruschke, T.M. Liddell, Bayesian data analysis for newcomers, Psychon. Bull. Rev. 25 (2018) 155–177, https://doi.org/10.3758/s13423-017-1272-1.
[55] T. Bayes, LII. An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S, Phil. Trans. R. Soc. 53 (1763) 370–418. https://doi.org/10.1098/rstl.1763.0053.
[56] J.N. Tendeiro, R. Hoekstra, T.K. Wong, H.A.L. Kiers, Introduction to the Bayes factor: a Shiny/R app, Teach. Stat. 47 (2024) 5–16, https://doi.org/10.1111/test.12380.
[57] H.A.L. Kiers, J.N. Tendeiro, Bridging null hypothesis testing and estimation: a practical guide to statistical conclusion drawing from research in psychology, PsyArXiv (2024). https://osf.io/preprints/psyarxiv/c7b45_v2 (accessed June 15, 2025).
[58] R. Kelter, Bayesian Hodges-Lehmann tests for statistical equivalence in the two-sample setting: power analysis, type I error rates and equivalence boundary selection in biomedical research, BMC Med. Res. Methodol. 21 (2021) 171, https://doi.org/10.1186/s12874-021-01341-7.
[59] D. Lakens, N. McLatchie, P.M. Isager, A.M. Scheel, Z. Dienes, Improving inferences about null effects with Bayes factors and equivalence tests, J. Gerontol. B. 75 (2020) 45–57, https://doi.org/10.1093/geronb/gby065.
[60] J. Cohen, Statistical Power Analysis For the Behavioral Sciences, second ed., Routledge, London, England, 2013 https://doi.org/10.4324/9780203771587.
[61] K. Sainani, Interpreting "null" results, PM R 5 (2013) 520–523, https://doi.org/10.1016/j.pmrj.2013.05.003.
[62] C.O. Brand, J.P. Ounsley, D.J. Van der Post, T.J.H. Morgan, Cumulative science via bayesian posterior passing, Meta-Psychol. 3 (2019), https://doi.org/10.15626/mp.2017.840. MP.2017.840.
[63] 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.
[64] J.K. Kruschke, Bayesian data analysis, Wiley Interdiscip. Rev. Cogn. Sci. 1 (2010) 658–676, https://doi.org/10.1002/wcs.72.
[65] W. O'Donohue, in: A. Masuda, S. Lilienfeld (Eds.), Avoiding Questionable Research Practices in Applied Psychology, first ed., Springer International Publishing, Cham, Switzerland, 2022 https://doi.org/10.1007/978-3-031-04968-2.
[66] S. Depaoli, S.D. Winter, M. Visser, The importance of prior sensitivity analysis in bayesian statistics: demonstrations using an interactive Shiny App, Front. Psychol. 11 (2020) 608045, https://doi.org/10.3389/fpsyg.2020.608045.
[67] P.B. Conn, D.S. Johnson, P.J. Williams, S.R. Melin, M.B. Hooten, A guide to Bayesian model checking for ecologists, Ecol. Monogr. 88 (2018) 526–542, https://doi.org/10.1002/ecm.1314.
[68] V. Roy, Convergence diagnostics for Markov chain Monte Carlo, Annu. Rev. Stat. Appl. 7 (2020) 387–412, https://doi.org/10.1146/annurev-statistics-031219-041300.
[69] J.K. Kruschke, Bayesian analysis reporting guidelines, Nat. Hum. Behav. 5 (2021) 1282–1291, https://doi.org/10.1038/s41562-021-01177-7.