Addressing the Precision-Breadth-Simplicity Impossible Trinity in Psychological Research: A Comprehensive Exploration Approach

Liqiang Huang*

*Department of Psychology, The Chinese University of Hong Kong, Hong Kong, China

E-mail: lqhuang@cuhk.edu.hk

Abstract

Psychological research faces a fundamental challenge - the Precision-Breadth-Simplicity (PBS) impossible trinity. While experimental findings are often precise and simple, they tend to be narrow in scope. Conversely, broad and simple concepts frequently lack precision. Developing theories that are both precise and broad is scientifically valuable but inevitably introduces complexity, which conflicts with humans' cognitive limitations in processing complexity. To address this impossible trinity, I propose a comprehensive exploration (CE) approach—a data-guided theory-building framework that involves: (1) designing experimental conditions in a stimulus-driven way, with minimal upfront theoretical specification; (2) conducting experiments with tens of millions of observations (e.g., 40 million responses in Huang, 2025a); (3) modeling the results through iterative improvements; and (4) producing the outcome: a moderately complex quantitative information-processing model to integrate diverse empirical findings. Inspired by similar strategies that drove breakthroughs in artificial intelligence (e.g., ImageNet's role in advancing object recognition), the CE approach offers a promising path toward more integrative psychological theories. Initial implementations in visual working memory research demonstrate both its practicality and potential to transform how we study mental processes.

Keywords

Breadth; comprehensive exploration; large-scale experiment; precision; simplicity

A challenge for psychological research

The PBS impossible trinity of psychological research

Psychological research over the past half-century has generated a wealth of experimental findings offering detailed insights into specific phenomena. Yet this knowledge remains fundamentally fragmented (e.g., Almaatouq et al., 2024; Anvari et al, 2025) - like scattered pieces awaiting assembly into a coherent whole. The study of visual working memory (VWM) exemplifies this challenge: while decades of research have identified numerous factors (e.g., Pashler, 1988; Alvarez & Cavanagh, 2004; Zhang & Luck, 2008; Bays & Husain, 2008; Brady & Alvarez, 2011; Brady & Alvarez, 2015; Suchow, Brady, Fournier & Alvarez, 2013; Bae et al., 2015; Huang, 2020; Fougnie, Suchow & Alvarez, 2012; Van den Berg & Ma, 2018; Liesefeld, Liesefeld & Müller, 2019), comprehensive theoretical frameworks (e.g., Oberauer, Farrell, Jarrold & Lewandowsky, 2016; Oberauer et al., 2018; Oberauer, 2023; Suchow, Fougnie, Brady & Alvarez, 2014; Liesefeld & Müller, 2019) have yet to yield a truly unified account that precisely integrates all relevant elements.

Beyond fragmented findings, conceptual vagueness presents a parallel challenge. Core psychological constructs like “attention” and “working memory” frequently operate as overly broad, imprecise categories encompassing diverse phenomena (e.g., Rosenholtz, 2024; Hommel et al., 2019; Anderson, 2011).

This dual challenge of fragmentation and vagueness stands in stark contrast to disciplines like physics, where theories from Newton's laws to Maxwell's equations achieve remarkable explanatory unity through fundamental principles like symmetry and conservation laws. Such exemplary theories successfully achieve three key attributes: precision, breadth and simplicity[1].

Why has psychology failed to develop optimal theories comparable to those in physics? Almaatouq et al. (2024) identify several contributing factors (e.g., experimental incommensurability, theoretical imprecision). While these observations are accurate, I argue they reflect symptoms rather than root causes— resulting from a deeper, unified constraint: the "**PBS impossible trinity**" (**Precision-Breadth-Simplicity**), illustrated in Figure 1. This principle asserts that psychological research can typically satisfy any two of these criteria simultaneously, but achieving all three proves exceptionally challenging.

Three trade-off options

The PBS impossible trinity arises from the inherent complexity of mental mechanisms. As Figure 1 illustrates, this complexity forces three fundamental trade-offs in psychological research: one can either (1) narrow the focus to yield precise yet fragmented findings (**precise-and-simple** approaches); (2) broaden the scope to capture the "big picture" at the cost of precision (**broad-and-simple** approaches);

[1] In this context, precision refers to the accuracy of defining and measuring concepts, breadth describes the scope of a theory's applicability—such as the range of different phenomena or findings it can explain—and simplicity reflects the conciseness of an explanation, exemplified by the number of parameters a model uses to account for the data.

or (3) embrace both breadth and precision while directly confronting the ensuing complexity (**precise-and-broad** approaches).

Most experimental studies exemplify the precise-and-simple approach. While methodologically rigorous, these typically examine isolated phenomena—for instance, demonstrating how VWM performance depends on categorical processing (i.e., how color categories such as red and blue affect memory, see Bae et al., 2015), interactions between items (i.e., how items affect each other's memorized appearance, see Brady & Alvarez, 2011), quality-quantity trade-offs (i.e., the trade-off between memorizing a larger number of imprecise items versus a smaller number of precise ones, see Fougnie et al., 2016), or chunking (i.e., memorizing multiple items as a single unit, see Brady & Tenenbaum, 2013). Such studies provide crucial but piecemeal insights, leaving the broader question ("How are these items memorized?") unanswered.

The broad-and-simple approach dominates theoretical discourse, where umbrella terms like "attention" explain diverse phenomena but often lack operational precision (Rosenholtz, 2024; Hommel et al., 2019; Anderson, 2011). While useful for organizing knowledge, these concepts become vague when applied across different experimental contexts.

The precise-and-broad approach represents the most scientifically valuable direction for psychological research, as it simultaneously expands the scope (breadth) and depth (precision) of our understanding of mental processes. While this approach aligns perfectly with psychology's fundamental goal of comprehensive theoretical understanding, it remains exceptionally rare in practice. Although some studies have begun moving in this direction, a truly ideal example has yet to be realized.

The “Blind Men and the Elephant” parable (Figure 2) captures these trade-offs vividly: precise-and-simple research resembles isolated tactile explorations of individual body parts (Figure 2b), broad-and-simple theories produce oversimplified “spherical elephant in a vacuum” generalizations (Figure 2c), while precise-and-broad synthesis would reveal both components and their interconnections (Figure 2d).

This article investigates why precise-and-broad studies remain elusive despite their scientific value, proposing pathways to overcome these limitations.

Clarifications regarding the PBS impossible trinity

Several clarifications are warranted regarding the PBS impossible trinity. First, while intuitively appealing, the PBS impossible trinity remains challenging to prove rigorously due to the difficulty of quantifying precision, breadth, and simplicity across diverse mental processes. Thus, I present it here as a conjecture.

Second, the severity of the PBS impossible trinity varies across mental processes, with rare exceptions occurring primarily in low-level vision. For instance, Trichromacy theory achieves all three criteria by using just three components to precisely explain diverse retinal color phenomena (for other cases, see also Adelson & Bergen, 1985; Ernst & Banks, 2002; Li, 2002; Zhaoping & Zhe, 2015). These exceptions prove the impossible trinity is not absolute, yet they remain uncommon in psychological research precisely because they involve relatively simple early visual processes - their exceptional status actually reinforces the impossible trinity’s general validity by demonstrating how most psychological phenomena, with their greater complexity, inevitably face these fundamental trade-offs.

Third, while the PBS impossible trinity may seem self-evident, its implications are profound. For instance, it entails that theoretical understanding is better achieved through data-guided theory-building than through theory-driven hypothesis-testing (the *paradox of theorizing* below). Furthermore, it implies that comprehensive theory-driven hypothesis-testing is likely futile (See "challenges in large-scale collaboration" below), pointing to the necessity of fundamental paradigm changes in the field.

Finally, although discussed here in relation to psychology, the PBS impossible trinity likely applies to other behavioral, human, and social sciences. These disciplines may therefore also benefit from the solution proposed below.

Humans' cognitive limitations in processing complexity

Why do precise-and-broad studies remain rare despite their scientific value? This likely stems from a fundamental cognitive bias—the **humans' cognitive limitations in processing complexity** (**HCLPC**)—which systematically prioritizes easily processed information over more complex but scientifically rigorous alternatives. Empirical work demonstrates that both laypeople and experts consistently favor simpler explanations when evaluating competing theories (Lombrozo, 2007; Gigerenzer & Goldstein, 1996; Heath & Heath, 2007), even when more complex alternatives better account for the evidence (Chater & Vitányi, 2003). This preference reflects a basic cognitive limitation—humans struggle with complexity and frequently choose simpler but inferior solutions to reduce mental effort (Gilovich et al., 2002; Kahneman, 2011).

The bias toward simplicity manifests across domains. In perception, we prefer simple visual patterns (Reber et al., 2004); in communication, simplistic messages outperform nuanced arguments (Westen, 2008; Berger, 2013); and in academia, this same preference likely disadvantages complex precise-and-broad research despite its greater scientific value. Faced with limited time and cognitive resources, researchers systematically favor simpler approaches—conducting studies with narrow focus or vague constructs—to avoid the complexity required for precise-and-broad synthesis.

This is not a conscious choice against scientific ideals, but rather an inevitable compromise. In an ideal world, comprehensive precise-and-broad studies would dominate. In reality, cognitive constraints and external pressures (e.g., publishability, visibility) create an ecosystem where simpler, less informative research thrives—even when researchers recognize precise-and-broad work as ultimately more valuable.

Crucially, identifying HCLPC and its consequences is not an indictment of the research community's morals or abilities. Rather, it suggests that pervasive issues like fragmentation and vagueness stem from an innate, irresistible weakness of human cognition—not from practitioners' failures. The solution, therefore, is not to blame researchers but to develop stronger methodological tools that allow us to transcend these limitations. This study presents one candidate for such a tool.

Binary theorizing

Theory-driven hypothesis-testing is a cornerstone of scientific progress, valued for generating testable predictions that incrementally refine theoretical frameworks. In physics, this approach has

yielded remarkable successes—as demonstrated by Poisson's spot, initially conceived as a falsification of wave theory but ultimately becoming decisive evidence for it. One might expect psychological theories to similarly evolve toward precise-and-broad frameworks through either: (1) enriching broad-and-simple theories with mechanistic details, or (2) synthesizing precise-and-simple findings into integrative models.

Yet psychology shows little evidence of this progression, likely due to the aforementioned HCLPC. Rather than converging toward comprehensive accounts, theoretical debates often devolve into binary oppositions: nature vs. nurture, early vs. late attentional selection, slot-based vs. resource-based VWM models. Like the blind men debating whether an elephant resembles a pillar or snake (Figure 2e), these dichotomies oversimplify complex mental phenomena. Their persistence reflects pragmatic realities: binary frameworks are cognitively manageable, easily communicated, and experimentally tractable—qualities that promote publication and impact despite often misrepresenting underlying complexity.

A challenge for psychological research

Figure 3 summarizes the arguments presented so far: the PBS impossible trinity establishes that psychological research cannot simultaneously maximize precision, breadth, and simplicity. Yet the field's fundamental goal—developing comprehensive theoretical understanding of mental processes—demands that we prioritize precise-and-broad approaches despite their inherent complexity. This creates a fundamental tension: the most scientifically valuable work conflicts with HCLPC.

The challenge, therefore, is to develop **a method for precise-and-broad research that reduces complexity-induced burden to a manageable level**. While this does **not** resolve the PBS impossible trinity (which would require precise-and-broad research that is also simple), it represents the best achievable solution.

The proposed solution adapts successful artificial intelligence (AI) approaches by implementing standardized benchmarking—quantifiable evaluation criteria for complex work. Though complexity remains inherent, systematic benchmarking, as demonstrated in AI research, can significantly mitigate its practical challenges while maintaining scientific rigor.

The proposed solution: comprehensive exploration

AI revolution

Psychological mechanisms appear fundamentally more complex than physical systems, creating an inherent PBS impossible trinity. This suggests that psychology's current physics-inspired paradigm—which prioritizes simple, universal theories—may be fundamentally mismatched to its subject matter. Instead, we might look to AI as a more relevant model. AI has achieved remarkable progress in studying the same complex processes that psychology examines (perception, cognition, language), but through a distinctly different, data-driven approach.

The AI revolution offers an instructive case study. Early AI research (pre-1990s) resembled much of contemporary psychology—small-scale, theory-driven, and lacking standardized assessment. The field's

transformation began with the advent of large benchmark datasets like ImageNet, which enabled standardized assessment of model performance through benchmarking. As LeCun et al. (2015) noted, the 2012 ImageNet competition—where AlexNet dramatically outperformed alternatives—marked a turning point in the rise of data-driven AI breakthroughs.

From a PBS impossible trinity perspective, modern AI methods achieve both precision (through quantitative models and metrics) and breadth (via large-scale datasets), while managing the resulting complexity through standardized benchmarking. Benchmarking allows models to be compared fairly using reproducible metrics (e.g., recognition accuracy), reducing reliance on subjective judgment. This creates a virtuous cycle: reliable feedback enables iterative improvement, democratizes innovation, and accelerates progress—all while mitigating complexity-induced burden.

To summarize, AI's solution to the PBS impossible trinity involves: (1) **large datasets** ensuring **breadth**, (2) **quantitative modeling** providing **precision**, and (3) **benchmarking**, which cannot reduce complexity but can **alleviate the complexity-induced burden** by enabling the evaluation of complex studies through simple, relatively objective metrics.

Even in physics, historical precedents like Kepler's analysis of Brahe's astronomical data also show how data-guided theory-building can revolutionize fields by replacing subjective theorizing with empirical pattern-finding. Psychology now stands to benefit similarly by adapting these strategies to mental phenomena.

Adapting AI for Psychology

The AI revolution has inspired growing interest in adapting computational methods for psychological research. Pioneering work by Griffiths (2015) highlighted big data's transformative potential, while Watts (2017) advocated for more solution-focused approaches in social sciences. Yarkoni and Westfall (2017) further argued that psychology should prioritize prediction over explanation, mirroring AI's success. Along these lines, a growing body of arguments, proposals, and empirical findings has emerged, advocating for the application of AI elements to advance psychology (e.g., Awad et al., 2018; Jolly & Chang, 2019; Agrawal, Peterson, & Griffiths, 2020; Schrimpf et al., 2020; Peterson et al., 2021; Bryan, Tipton, & Yeager, 2021; Yarkoni, 2022).

The framework introduced in this work advances beyond prior approaches in three key respects:

1. It is **grounded in the first principle** PBS impossible trinity conjecture, providing a rigorous theoretical foundation.

2. It offers a **complete and operationalizable methodology**, systematically bridging experimental design and model construction.

3. It is **explicitly designed for experimental psychologists**, with theoretical understanding as its primary objective.

This framework's viability is evidenced by multiple successful implementations, including three published studies (Huang, 2023, 2025a, 2025b), numerous ongoing projects, and an ongoing large-scale collaborative initiative (see "large-scale collaboration" below) – all rigorously following the methodological framework illustrated here.

Six methodological dimensions

To develop an optimal framework for integrating AI approaches into psychology, Figure 4 compares typical AI studies (blue) and traditional experimental psychology studies (red) across six dimensions:

Dimension 1: Output Format. AI research typically generates quantitative models, whereas psychological studies often yield verbal interpretations of findings[2]. The optimal approach should adopt AI's quantitative modeling practices to maximize precision, but with a critical modification: employing interpretable information-processing models rather than opaque neural networks, thereby leveraging mechanistic explanations to advance theoretical understanding (Dimension 3).

Dimension 2: Data Scale. Modern AI research leverages large-scale benchmark datasets, whereas psychology traditionally employs smaller, custom-designed experiments. Here, the optimal approach should adopt AI's large-scale paradigm to achieve greater breadth, but with a crucial modification—using large-scale controlled experiments specifically designed for psychological research questions (Dimension 4).

Dimension 3: Research Objective. Whereas AI studies typically prioritize applied solutions, experimental psychology emphasizes theoretical understanding—its fundamental scientific purpose (e.g., Bowers et al., 2023). The optimal approach should therefore use interpretable information-

[2] While exceptions exist in both fields, the discussion focuses on typical cases for simplicity, as with all subsequent dimensions.

processing models, whose mechanistic explanations provide theoretical understanding, rather than the opaque neural networks.

Dimension 4: Data Source. AI research often utilizes large observational datasets, whereas psychology prioritizes controlled experiments. The optimal approach should retain experimental control, as it is essential for isolating causal relationships and minimizing extraneous variables. This methodological rigor remains essential for developing valid theoretical understanding of mental processes.

Dimension 5: Design of Conditions. AI studies typically employ a stimulus-driven design, sampling from stimulus space (e.g., large datasets of diverse faces), whereas psychology traditionally relies on a theory-driven design, sampling from factor space (e.g., computer-generated faces systematically varying specific factors). As noted in Dimension 3, the optimal approach should be theory-oriented, which might seem to suggest that a theory-driven design should be used for consistency.

However, a **paradox of theorizing** arises from the PBS impossible trinity: for gaining theoretical understanding from precise-and-broad studies, **stimulus-driven design is better suited than theory-driven design**. Such studies inevitably grow complex, and theory-driven designs struggle under this complexity due to HCLPC. The most viable strategy is therefore to shift from a theory-driven to a stimulus-driven design, deferring theoretical commitments until empirical evidence can help reduce the burden of complexity (see “Precision, breadth, and standardized benchmarking” below). Simply put, when key factors are unforeseeable and mechanisms are highly complex, it is often more productive to

let the data guide exploration than to impose arbitrary theoretical assumptions in advance (Dubova et al., 2022; Musslick et al., 2023).

Given the paradox of theorizing, stimulus-driven design (i.e., sampling from stimulus space) should be the primary strategy. However, while most uncontrollable factors are best addressed through natural stimulus variations, readily controllable factors should still be explicitly incorporated into the design. This synthesis leads to the optimal approach: **sampling from factor-informed stimulus space** (See below for details).

Dimension 6: Analytical Approach. While psychology traditionally relies on confirmatory analyses (e.g., ANOVA), AI research emphasizes exploratory, iterative model building. The stimulus-driven design of the optimal approach (Dimension 5) deliberately minimizes upfront theoretical specification; consequently, there are no pre-defined hypotheses to test. This necessitates a data-guided theory-building process, using **exploratory, iterative model building** to elucidate the theoretical implications of the experimental conditions.

After introducing all six dimensions, Figure 5 summarizes the underlying rationale for these key methodological choices. The upper cyan region (Dimensions 1-2) reflects selections made to achieve precision, breadth, and standardized benchmarking, which will be further detailed below. The lower purple region (Dimensions 3-4) represents choices prioritizing theoretical understanding, while the central yellow region (Dimensions 5-6) comprises decisions enhancing the feasibility of designing experimental conditions, to be expanded below.

Comprehensive exploration

In summary, the optimal methodology for adapting AI approaches for psychology leverages key AI elements—including stimulus-driven design, large-scale benchmark datasets, quantitative modeling, and exploratory iterative model-building—to advance psychology's core goal of theoretical understanding. This approach, termed **comprehensive exploration (CE)**, employs **stimulus-driven design for large-scale experimentation, followed by iterative modeling to build interpretable quantitative models whose mechanistic explanations yield theoretical understanding**. CE is so named because it diverges from traditional experimental psychology most distinctly in being comprehensive (Dimension 2) and exploratory (Dimension 6).

On one hand, the CE approach differs dramatically from theory-driven methods in psychology, as it constitutes a form of **data-guided theory-building**. This shift does not diminish the importance of theory but reframes its role: in CE, theories emerge from and are refined by empirical evidence, serving as the goal of the research process rather than its starting point.

On the other hand, the CE approach, while data-guided, remains distinct from data-driven AI approaches. In the CE framework, data serves as a GPS, providing guidance while the researcher actively steers toward theoretical understanding. In AI, by contrast, data often acts as a self-driving shuttle bus, delivering outputs automatically without revealing its operational mechanics.

To formalize these distinctions: traditional theory-driven methods in psychology align with **deductive** reasoning, while typical data-driven studies in AI reflect **inductive** reasoning (or inductive

pattern-finding). In contrast, the exploratory, iterative model-building central to the CE approach constitutes **abductive** reasoning—inference to the best explanatory theory based on empirical patterns.

The typical CE study process, illustrated in Figure 6, consists of several steps (details will be elaborated below using a recent example: Huang, 2025a):

Experimental design: Conditions are sampled from a factor-informed stimulus space.

Data collection: A large-scale controlled experiment is conducted to gather data.

Model development: The experimental data undergo exploratory, iterative model building.

Outcome: The above process yields a CE model as the final study output.

Precision, breadth, and standardized benchmarking

Following the successful AI paradigms illustrated above (e.g., ImageNet), the CE approach simultaneously achieves the following:

1. **Precision** through quantitative information-processing models,
2. **Breadth** via large-scale experiments, and
3. **Not simplicity**—since complexity is inevitable in precise-and-broad research—but a **significant reduction in complexity-induced burden** through benchmarking.

In summary, the **CE** directly addresses the fundamental challenge posed above: **methods for precise-and-broad research that reduce complexity-induced burden to a manageable level**.

Several clarifications are needed. First, effective benchmarking requires standardized assessment, which in turn demands both breadth and precision, as demonstrated by foundational psychometric

research (e.g., Crocker & Algina, 2006). Breadth - comprehensive domain coverage - ensures construct validity, while precision - through reliable measurement - ensures consistency; neither dimension alone suffices for standardized assessment. This principle extends to research evaluation, where broad but vague theories produce non-falsifiable claims even when applied to extensive datasets, while precise yet narrow quantitative models generate non-comparable performance metrics lacking generalizability. Consequently, the CE approach necessitates both large-scale experiments (to achieve breadth) and quantitative modeling (to achieve precision) to enable meaningful benchmarking.

Second, benchmarking mitigates complexity-induced burden by enabling the evaluation of complex studies through simple, relatively objective metrics. This also allows researchers to concentrate exclusively on top-performing models—a common practice in AI research. This contrasts sharply with traditional psychological research, where the absence of standardized benchmarking forces scholars to laboriously track all relevant theoretical claims and methodological details across studies. Consequently, the CE approach may yield a net reduction in researchers' cognitive workload despite employing more complex theoretical models, as the field shifts from evaluating every individual study to evaluating performance against shared benchmarks.

Third, while large datasets are commonly associated with breadth and quantitative modeling with precision—relationships we employ for convenient description—these connections are neither automatic nor guaranteed. Poor design compromises dataset breadth, while weak conceptualization undermines model precision. Furthermore, benchmarking reduces but doesn't eliminate subjectivity—researcher biases still influence experimental designs, stimuli, metrics, and interpretations (e.g., Daston & Galison,

2021). Thus, rigorous effort remains just as critical in data-guided CE studies as in traditional theory-driven research.

Details of the CE approach

In this section, I will use Huang (2025a) to illustrate the six methodological dimensions of the CE approach and further justify their current implementation. Although two other published studies (Huang, 2023, 2025b) and numerous ongoing projects further support these methodological choices, they are omitted here for conciseness of presentation.

Large-scale controlled experiment

To achieve comprehensive breadth without compromising experimental rigor, Huang (2025a) implemented a large-scale controlled experiment examining VWM through 10,000 distinct color patterns, collecting 40 million responses (~33,000 hours of data). This design simultaneously fulfills two key dimensions of the CE approach: **Dimension 2 (large-scale data)** and **Dimension 4 (controlled experimentation)**.

Quantitative information-processing model

Huang (2025a) introduced the Quasi-Comprehensive Exploration of VWM (**QCE-VWM**) model, building on established information-processing models of VWM (e.g., Zhang & Luck, 2008; Bays & Husain, 2008; Van den Berg et al., 2014). Information-processing models are formal computational

frameworks that simulate how humans encode, transform, and utilize information, serving as a cornerstone of cognitive psychology. The CE models presented here constitute a specialized class of such models—quantitative, mechanistic, interpretable, and fully specified. Their quantitative nature ensures precise specification of mechanisms, while their interpretability enables meaningful theoretical understanding through clear explanation of these mechanisms. As such, these models simultaneously fulfill two key dimensions of the CE approach: **Dimension 1 (quantitative modeling)** and **Dimension 3 (theoretical research)**.

Specifically, the QCE-VWM takes as input 10,000 randomly generated color patterns (each consisting of 4 colors, totaling 40,000 items) and predicts response distributions for these items. These predictions are then compared against the results from the aforementioned large-scale experiment. Crucially, it balances predictive accuracy with explanatory parsimony, integrating approximately a dozen transparent mechanisms into an integrative framework with just 57 parameters.

For instance, interactions between items are formalized using a combination of normal and Mexican-hat-like functions. Meanwhile, the quality–quantity trade-off is formalized by establishing a unified ratio that governs the co-variation between a representation's precision (standard deviation) and its prevalence (proportion within a mixture). All twelve mechanisms—along with their interrelationships—are explicitly formalized through such algorithmic implementations. This ensures the model remains interpretable, unlike "black-box" neural networks, while still accounting for the

complexity of VWM operations. By maintaining this balance, the QCE-VWM exemplifies the CE approach's ability to reconcile precision[3] and breadth without sacrificing explanatory interpretability.

Sampling from stimulus space

As discussed, the **paradox of theorizing** maintains that stimulus-driven design achieves theory-oriented goals more effectively than theory-driven design in precise-and-broad studies. Huang (2025a)'s design supports this counterintuitive claim.

The use of 10,000 randomly-generated color patterns in Huang (2025a) represents a deliberate departure from traditional theory-driven design in psychology. Typically, researchers carefully select conditions based on specific hypotheses, sometimes using factorial designs to manipulate two variables simultaneously (or three on rare occasions). While this theory-driven design works well for testing focused questions, it becomes untenable for comprehensive investigations of complex phenomena. The challenge lies in operationalizing numerous interacting factors at once—a task that requires not only precisely defining each factor but also creating stimuli that unambiguously represent their myriad combinations. For example, consider designing a stimulus pattern for a VWM study that must simultaneously satisfy:

[3] By their very nature, these models' algorithms are mathematically precise. Yet this computational precision does not preclude divergent theoretical interpretations. A telling example arose in discussions of Huang's (2025a) model, where colleagues interpreted the same mechanisms as supporting either "slot-like" or "resource-like" frameworks. This divergence does not reflect a weakness in the modeling approach but rather highlights its strength: the capacity to expose and clarify vagueness inherent in purely verbal theoretical descriptions.

- Chunking (low)
- Interactions between items (medium)
- Categorical effects (medium)
- Regularity (high)
- Quality-quantity trade-off (high)
- And five other specific criteria…

Now imagine scaling this to 10,000 distinct factor combinations—each requiring peer-defensible implementation. The impossibility speaks for itself[4]. Of course, one may choose to force definitions for these factors, but doing so would likely result in arbitrary, vague, or artificially constrained operationalizations.

This **factor design difficulty** is evident in the VWM literature, which includes three-factor designs (e.g., Van den Berg et al., 2014; Oberauer, 2023) but nothing more complex than that. The absence of higher-dimensional factorial designs underscores the challenge faced by theory-driven designs in precise-and-broad studies.

This factor design difficulty led me to adopt **sampling from stimulus space**—a strategy that **minimizes upfront theoretical specification**, thereby avoiding the challenges of manually designing high-dimensional factor spaces.

[4] This is not a hypothetical example but reflects my actual experience attempting to design a VWM experiment that simultaneously manipulated ten factors. After four weeks of effort, I abandoned not only that specific plan but also my confidence in high-dimensional, theory-driven designs as a whole.

This is inspired by successful AI paradigms like ImageNet. Rather than artificially constructing stimuli by pre-defining and combining specific factors relevant to object recognition—an approach that would be unworkable and counterproductive in ImageNet's context—ImageNet leverages natural variations in images to uncover underlying mechanisms. Similarly, in Huang (2025a), instead of manually designing stimuli based on predetermined factors, I generated 10,000 color patterns, each comprising four colors randomly selected from a color wheel. This allowed natural stimulus variations to reveal the underlying processes of interest.

To understand how theoretical questions can be addressed without upfront theoretical specification, consider various factors in VWM (e.g., interactions between items, chunking, and categorical effects). Traditional sampling from factor space requires explicit, predefined mappings between experimental conditions and these factors—a demanding and inevitably subjective task. In contrast, Huang (2025a) required no upfront theoretical specification regarding these factors. Since these effects are all related to how items are distributed on the color wheel, they were expected to emerge organically from the random variations among the 10,000 patterns. This approach allows the precise nature of these mechanisms to be determined from the data itself, offering a simpler strategy.

In addition to reducing the difficulty of the factor design, sampling from stimulus space offers another key advantage: the **exploration of unanticipated factors**. As Popper (2014) observed, *"Our knowledge can only be finite, while our ignorance must necessarily be infinite."* While sampling from factor space inherently restricts investigations to predefined variables, Sampling from stimulus space allows novel and meaningful factors to emerge organically from natural stimulus variation. For example,

Huang (2025a) identified four new mechanisms: concentration and crosstalk (both of which describe how the color categories of items interact with each other), as well as red advantage and red disadvantage (referring to the enhancement of memory for reddish colors in category-related processing and its reduction in a category-unrelated "unbiased component," respectively).

Sampling from factor-informed stimulus space

Having advocated for sampling from stimulus space, it's crucial to clarify that it should still be theoretically constrained. In Huang (2025a), for example, fixing the number of items at four enabled a focused investigation of multi-item memory phenomena while controlling for the known effects of capacity limitations. Conversely, a CE study specifically designed to examine capacity limitations would instead incorporate systematic variation in number of items. In other words, although the upfront theoretical specification is greatly reduced, some minimal essential ones are still required. This leads to what I term **sampling from factor-informed stimulus space (Dimension 5)**.

Exploratory, iterative model building

Because the CE approach samples from a factor-informed stimulus space—minimizing upfront theoretical specification—the experimental conditions are not derived from pre-defined hypotheses. This makes **exploratory, iterative model building (Dimension 6)** a methodological necessity to elucidate the theoretical implications of the data, as there are no pre-defined theoretical predictions to test.

While this process shares methodological similarities with AI's progressive refinement techniques, it retains a crucial psychological distinction: each iteration of this data-guided theory-building process explicitly tests theoretical hypotheses through rigorous model comparison. For example, Huang (2025a) used effect-size-based criteria (CAD[5]) to evaluate model improvements.

For example, a typical iteration involves refining the "interactions between items" mechanism in the QCE-VWM model. Initially, this mechanism was placed after categorical encoding, reflecting the theoretical view that color interactions depend on category membership (e.g., red, green, blue). However, model comparison revealed significantly better performance when interactions were positioned at a pre-categorical stage, suggesting they operate on early continuous color values rather than later discrete categorical representations. This finding not only revised our understanding of VWM organization but also exemplified how data-guided theory-building can productively challenge theoretical assumptions. Through approximately 1,000 such model comparisons, the development process addressed an equivalent number of fine-grained theoretical questions—a scale of hypothesis-testing impossible in traditional confirmatory frameworks. These theoretical hypotheses originate from two complementary sources (Figure 6): established findings in the literature and empirical patterns emerging from the data itself.

[5] The statistical index CAD (Complexity-Adjusted d) is a fit-parsimony index that evaluates whether each parameter in a model meaningfully contributes to its explanatory power. As an effect-size-based measure, CAD assesses the utility of a parameter independently of dataset scale, ensuring that the effect size of its contribution exceeds a prespecified threshold (e.g., 0.2 or 0.1) across observed patterns. For further details, see Huang (2025a).

This exploratory, iterative model building differs fundamentally from conventional practices in experimental psychology, warranting several important clarifications:

First, this method must be distinguished from the problematic practice of HARKing (Hypothesizing After Results are Known). While both involve post hoc analysis, they differ critically in transparency and intent. HARKing misleadingly presents exploratory findings as confirmatory, whereas the CE explicitly acknowledges its exploratory nature and avoids overinterpretation. The key distinction lies in their treatment of uncertainty: where HARKing obscures it, the CE approach makes it explicit.

Second, while multiple hypothesis-testing does raise legitimate concerns about false positives in traditional small-scale research, the CE approach mitigates this risk through two key features: (1) enormous data scale (e.g., 40 million responses in Huang, 2025a), and (2) strict statistical criteria (e.g., the effect-size-based index CAD in Huang, 2025a). This combination resulted in extraordinarily low false positive rates (e.g., < 1.E-87 in Huang 2025a), rendering the multiple testing problem negligible despite evaluating ~1,000 models.

Third, exhaustive model testing proves impractical given the combinatorial explosion of possibilities. While Van den Berg et al. (2014) tested 32 models—already ambitious by traditional standards—the QCE-VWM's parameter space contains 2^{57} = ~1.4E+17 possible configurations[6]. Exploratory, iterative model building thus represents the only feasible option at this scale.

[6] Strictly speaking, the parameter space is infinite because it includes not only the current 57 parameters but also those that have been tried and discarded, as well as parameters that may be explored in the future.

Fourth, the CE approach explicitly embraces the provisional nature of its models. Like the iterative scientific process described by Popper (2005), initial models may contain inaccuracies that subsequent refinements gradually correct (Figure 2d). Should future work surpass QCE-VWM's performance, this would demonstrate the CE approach's strength—mirroring AI's progressive improvement paradigm—rather than indicating failure[7].

Fifth, iterative exploration outperforms predefined testing by accommodating a wider range of potential solutions. As shown in Figure 7, Huang (2025a)'s model achieved superior fit compared to Van den Berg et al. (2014)'s precisely because its development wasn't constrained to a predetermined solution space.

Sixth, the conventional preference for exhaustive testing may stem less from methodological necessity than from a well-documented cognitive bias. Research on the paradox of choice (Iyengar & Lepper, 2000; Schwartz et al., 2002) demonstrates that decision-makers often favor limited but complete option sets over more expansive but necessarily incomplete ones - even when the latter objectively produces better outcomes. This preference may stem from HCLPC, leading individuals to prioritize ease of evaluation over optimal results. In academia, this manifests as an irrational preference for small, exhaustively testable model spaces despite their limited explanatory scope, while potentially more

[7] A breakthrough—such as unifying the QCE-VWM's dozen mechanisms under a single principle—could challenge the CE approach's foundational assumption of irreducible complexity. Should such unification occur, contemporary complex models would nevertheless serve as necessary precursors to this future theoretical synthesis. Their role would be analogous to that of Ptolemy's system of epicycles: though ultimately incorrect, it provided the essential scaffolding for Kepler's elegant laws of planetary motion. The enduring lesson from Ptolemy is not final accuracy, but the value of systematic engagement with complexity. His model demonstrates that embracing intricacy is often a necessary path to ultimate simplicity.

productive but incompletely testable approaches are avoided due to the discomfort of unexplored alternatives. The CE approach challenges this bias by recognizing that in complex domains, iterative exploration of large solution spaces - while necessarily incomplete - yields superior theoretical and empirical outcomes.

Theory-driven confirmation vs. data-guided exploration

Theory-driven confirmatory analysis is strongly favored over data-guided exploratory analysis in experimental psychology. For instance, Pashler and colleagues, in their analysis of the replication crisis, advocated for solutions like pre-registration and strict adherence to pre-specified plans to ensure studies are truly confirmatory (Pashler & Harris, 2012; Pashler & Wagenmakers, 2012).

At first glance, the CE approach may appear to conflict with this preference for confirmation, given its explicitly exploratory nature. However, this approach can be viewed as an evolution of these methodological reforms—both are driven by the core principle of optimizing scientific methodology. A key insight is that **methodological optimality is scale-dependent**. Issues that are critical in small-scale studies may become negligible in large-scale research, and vice versa.

Figure 8 summarizes four key factors governing the choice between theory-driven confirmation and data-guided exploration, highlighting the divergent trade-offs for small- versus large-scale studies:

Theory-driven confirmation faces two primary challenges:

1. **Lack of exploration**: As discussed above, data-guided exploration allows us to explore unanticipated factors—an advantage lacking in theory-driven confirmation. While this constraint is

harmless in small-scale studies with limited data, it represents a significant opportunity cost in large-scale research, where an abundance of data could allow for more comprehensive exploration.

2. **Factor design difficulty**: As discussed above, small studies can feasibly manipulate one or two factors, but large-scale confirmatory research struggles with the impracticality of manipulating numerous factors.

Data-guided exploration also faces two principal challenges:

1. **False positive risk**: Exploratory analysis traditionally raises concerns about inflated false positive rates due to multiple comparisons. However, as discussed above, in large-scale studies, massive datasets can support the application of exceedingly stringent statistical criteria (e.g., Huang, 2025a's false positive rates of < 1.E-87), effectively mitigating this risk.

2. **Inefficient testing**: Exploratory designs like the CE approach use randomized sampling of stimulus space, which can allocate data points to uninformative or redundant conditions. This inefficiency poses a critical liability in small-scale studies, where data is limited and statistical power is precious. At large scales, however, a certain portion of wasted data becomes easily affordable, as the benefit of broad, unsupervised coverage outweighs the cost.

The shift from small- to large-scale research fundamentally transforms the cost-benefit landscape:

- The limitations of theory-driven confirmatory analysis (lack of exploration & factor design difficulty) become **severely amplified**.
- The drawbacks of data-guided exploratory approaches (false positive risk & inefficient testing) are **greatly reduced**.

Thus, while theory-driven confirmation remains superior for small-scale experiments, data-guided exploration emerges as the more suitable strategy for large-scale investigations.

This methodological shift mirrors the dominant paradigm in AI research, where data-driven exploration prevails—a choice necessitated by the field's reliance on large-scale datasets. This scale-dependent perspective may also explain why such data-intensive fields have largely avoided the replication crises that have plagued psychology, despite being far more exploratory in nature.

Finally, Figure 9 uses a firearms analogy to illustrate three of the four factors discussed above. Theory-driven confirmation in small-scale studies is like a sniper rifle—using a single bullet for precise, narrow targeting. In contrast, data-guided exploration in large-scale studies operates like a shotgun blast, scattering many pellets to achieve broad, comprehensive coverage.

For the sniper, the cost of precise control (analogous to factor design difficulty) is low, and the inability to hit unexpected targets (analogous to a lack of exploration) is acceptable. Conversely, the cost of a missed shot (analogous to inefficient testing) is high, making strict, precise control essential.

This situation reverses for the shotgun. Controlling the trajectory of every pellet is prohibitively difficult, making broad coverage optimal. Here, "missing" with many pellets is inconsequential, while hitting unexpected targets becomes a significant advantage. Ultimately, applying the sniper's precise control to the shotgun's domain—like using theory-driven confirmation in large-scale studies—is both difficult and counterproductive.

An optimal balance between prediction and explanation

Cognitive science is defined by a fundamental trade-off between predictive accuracy and explanatory interpretability. While powerful deep neural networks achieve high accuracy on complex tasks like face recognition, they function as "black boxes" that offer little theoretical understanding (Bowers et al., 2023). Conversely, interpretable psychological models (e.g., featural vs. configural theories of face processing) provide clear explanations but lack robust predictive power (Yarkoni & Westfall, 2017). Consequently, the field remains divided, lacking a unified framework that is both fully predictive and fully explanatory.

The CE approach aims to resolve this dilemma through comprehensive yet interpretable modeling. The QCE-VWM model exemplifies this solution. With only 57 parameters—compared to 30,796 in a benchmark neural network—it achieves superior predictive accuracy while retaining full mechanistic interpretability (Figure 7). This represents an optimal balance between prediction and explanation[8].

Although this success is largely credited to two decades of VWM research that defined the core puzzles, it also proves the value of the CE approach in synthesizing those puzzles into a coherent, complete model.

The utility of exploratory, iterative model building

[8] It should be noted that while Huang (2025a) presents an especially successful case, other CE models (Huang, 2023, 2025b) have not fully matched the effectiveness of benchmark neural networks, though they have come close.

The optimal balance achieved in Huang (2025a) depends not only on large-scale experiments and quantitative models but also—critically—on exploratory, iterative model building. To appreciate this, consider how alternative analytical approaches would have performed if applied to the same data.

The first alternative is traditional confirmatory analysis. As noted above, simultaneously pre-specifying numerous factors is exceptionally challenging. Even if achieved, applying a traditional confirmatory analysis (e.g., a multifactor ANOVA) would fragment QCE-VWM's integrated architecture into statistically significant but theoretically disconnected effects. Such an approach would reduce complex mechanistic relations to simplistic interaction terms, failing to address core questions like whether chunking operates on pre-categorical or categorical information. Moreover, its rigid operationalizations would not adapt to patterns emerging from the data. Consequently, while interpretable, this approach would likely yield inferior predictive accuracy.

This limitation further supports the paradox of theorizing: rigid, pre-specified research plans can stifle discovery by preventing researchers from detecting unexpected yet meaningful patterns (Dubova et al., 2022; Musslick et al., 2023). This suggests that, when facing a large dataset, theoretical understanding may be more effectively achieved by carefully examining the data itself rather than following only pre-determined, theory-driven plans.

The second alternative is predictive modeling using neural networks. While such models excel in predictive accuracy, they function as opaque black boxes, offering minimal explanatory interpretability. This is not to say neural networks lack value; following the scientific regret minimization framework developed by Agrawal, Peterson, & Griffiths (2020), Huang (2025a) employed a "guidance neural

network" to facilitate the development of the QCE-VWM model. However, for psychological science, they should serve as intermediate tools rather than final products, as they cannot provide the mechanistic understanding that defines explanatory progress.

Large-scale collaboration

The need for large-scale collaboration

While Huang (2025a) represents a significant advancement in scale over traditional studies—yielding several hundred times more data—its scope was deliberately restricted to how color variations affect VWM for four items presented simultaneously at fixed locations. This design intentionally excluded extraneous factors such as number of items, temporal order, and stimulus dimension (e.g., color versus object identity). Consequently, a critical future direction for the CE approach is to conduct a more comprehensive investigation of VWM that incorporate these and other dimensions of stimulus variation.

Such an ambitious undertaking demands large-scale collaboration across the field for two reasons. First, the project's unprecedented scope—far exceeding Huang (2025a) in both scale and complexity—necessitates shared resources across multiple laboratories to ensure feasible data collection. Second, and more fundamentally, this collaborative effort could establish a standardized benchmark dataset for the VWM research community—a psychological equivalent to ImageNet in computer vision. Achieving this

vision requires broad participation to ensure the resulting dataset reflects field-wide consensus while serving as a definitive resource for theoretical advancement.

Challenges in large-scale collaboration

Despite its potential benefits, large-scale collaboration presents significant challenges. The Cogitate project (Cogitate et al., 2025), a major adversarial collaboration in consciousness research, illustrates these difficulties. Despite involvement from leading theorists, the project made limited progress (Lenharo, 2024), with critics noting it tested only "idiosyncratic predictions" and "hand-picked auxiliary components already known to be true" rather than core theoretical claims (Fleming et al., 2023).

These difficulties stem from the fundamental constraints of the PBS impossible trinity. The HCLPC makes developing precise-and-broad frameworks challenging even in individual work—a challenge magnified in collaborative settings by self-serving biases. Researchers naturally avoid tests threatening their theoretical positions[9], resulting in designs that focus on safe, known findings rather than critical tests of competing theories[10].

Facilitating collaboration through the CE approach

[9] There have been instances of progress, such as the collaboration on intuitive expertise (Kahneman & Klein, 2009). However, these examples are exceptions rather than the norm. Furthermore, the success of these collaborations often relies on both large datasets and quantitative modeling—two key components of the CE approach.

[10] This tendency resembles the Motte and Bailey fallacy (Shackel, 2005), in which a person advocates for a controversial, ambitious stance (the "Bailey") but retreats to a more defensible, modest position (the "Motte") when challenged.

The CE approach offers a solution to these collaborative challenges. Like adversarial collaborations, a collaborative CE project brings together researchers with diverse perspectives. However, by sampling from stimulus space, it minimizes upfront theoretical specification—collaborators need only agree on relevant stimulus variations, not their theoretical interpretations. This theory-neutral foundation reduces conflict by shifting focus from defending positions to shared empirical exploration. For instance, it's unlikely that any reasonably open-minded researcher would view Huang's (2025a) design (i.e., 10,000 randomly-generated color patterns) as a direct threat to their theoretical positions.

Following the ImageNet model, a collaborative CE project would conclude with an open benchmark dataset, eliminating the need for consensus on theoretical interpretation—a persistent hurdle in traditional adversarial collaborations. Individual researchers could then independently develop and test their models against this standardized benchmark, competing on a more objective basis, as will be explained in "Benchmarking as the engine of change" below.

By reducing these demands—and leveraging the CE approach's inherent resistance to oversimplification— the CE approach enables more viable large-scale cooperation than theory-driven alternatives.

Guided by this design, a large-scale collaborative CE project on VWM is currently underway, co-led by Klaus Oberauer and me. While this project specifically extends Huang (2025a) to study VWM, the advantages of the collaborative CE approach are not domain-specific and can be generalized to other fields of research.

Comparison with Almaatouq et al. (2024)'s integrative experimental design

Similarity between IED and CE

The CE approach and Almaatouq et al. (2024)'s integrative experiment design (IED) share a fundamental critique of psychology's methodological stagnation[11]. Both seek to advance comprehensive theoretical understanding and identify traditional small-scale, hypothesis-driven experiments as a primary obstacle to this goal. Specifically, they argue that such approaches perpetuate theoretical fragmentation, producing either narrow findings or artificial binary debates that fail to capture cognitive complexity. Crucially, they agree that advancing psychological science requires directly engaging with methodological and theoretical complexity rather than circumventing it.

Beyond their shared goals, CE and IED also converge partly in their proposed solutions. Both advocate for larger datasets than those typical in experimental psychology and emphasize more systematic, in-depth analyses that move beyond simple hypothesis-testing. These similarities emerge naturally from their shared intellectual heritage - CE is indeed inspired by earlier work that either influenced IED (e.g., Awad et al., 2018; Jolly & Chang, 2019) or was directly produced by IED's authors themselves (e.g., Griffiths, 2015; Watts, 2017; Agrawal, Peterson, & Griffiths, 2020; Peterson et

[11] Anvari et al. (2025) also share a similar perspective. Their proposed solutions—such as developing more comprehensive datasets, iteratively improving frameworks, and promoting better standardization across the research community—are closely aligned with those of the present work.

al., 2021). Ultimately, the CE approach builds upon the intellectual tradition that led to IED, integrating its key insights as core components.

Theory-driven method vs data-guided theory building

While the CE inherited important insights from this tradition, it also introduced critical modifications.

A primary dimension on which IED and CE differ is their positioning along the theory-driven vs data-guided spectrum. In essence, IED maintains that researchers should use the best available theoretical understanding to structure experimental testing from the outset, thereby ensuring comparability and coherence across studies. In contrast, CE argues that existing theories are usually insufficient to guide such structuring and instead advocates for data-guided theory-building to uncover the underlying structure of phenomena, which can subsequently inform stronger, more robust theories.

Specifically, the IED approach emphasizes the construction of a well-defined "explicit design space" to guide experimental testing and theoretical development. This approach demands even greater upfront theoretical specification than traditional theory-driven methods, placing the formalization of theory at the forefront. While IED acknowledges the importance of data-guided theory-building and suggests that the explicit design space should be updated when data suggest error or inadequacy, such data-guided adjustments remain secondary within this theory-first framework and do not eliminate the challenge of upfront theoretical specification. After all, to be practically useful for the research

community, the explicit design space needs to be stable enough to accommodate ongoing studies and is therefore only updated periodically.

In contrast, CE advocates for a decisive shift toward data-guided theory-building, arguing that excessive upfront theoretical specification is a fundamental limitation of traditional experimental psychology. Therefore, as stated above, CE proposes that upfront theoretical specification should be minimized, and theories should emerge organically through iterative, exploratory model-building. Although CE incorporates some theoretical constraints (i.e., sampling from factor-informed stimulus spaces), these are subordinate to the overarching goal of data-guided theory-building.

For example, Huang (2025a)'s only upfront theoretical constraint is that the number of items is fixed at four, making it a predominantly data-guided approach. In contrast, using the IED to study VWM would require a mainly theory-driven approach. This would involve extensive upfront theoretical specifications such as designing the conditions for "chunking", "interactions between items", "quality-quantity trade-off", and many other factors, as well as their fully factorial combinations.

Iterative experimentation vs. iterative modeling

Another key distinction between IED and CE lies in their approaches to data accumulation. Almaatouq et al. (2024) describe IED as requiring an iterative cycle: first constructing an explicit design space, then progressively generating theories and testing them through experiments systematically sampled from this space. In contrast, CE employs a single comprehensive data collection phase, with subsequent iterations focused exclusively on model refinement without additional experimentation.

Feasibility of IED and CE

Having delineated the key distinctions between CE and IED, it is worth noting that while IED offers a strong theoretical framework for ideal scenarios, its practical implementation can be challenging. In contrast, the CE approach provides a more readily applicable methodological path forward.

CE's emphasis on data-guided theory-building is grounded in the PBS impossible trinity conjecture, which highlights the inherent complexity of precise-and-broad studies and the difficulty of managing this complexity prior to data collection due to HCLPC. Consequently, the explicit design space proposed by IED—though conceptually valuable—faces practical hurdles, as it requires a degree of upfront theoretical specification that may be difficult to achieve in early stages of research. This challenge is compounded when attempting to coordinate such specifications across diverse research teams. In this context, CE's stimulus-driven design offers a more adaptable and feasible alternative.

Additionally, IED's call for iterative experimentation may present greater logistical demands compared to CE's model-focused iteration within a single comprehensive dataset. For example, Huang (2025a) evaluated approximately 1,000 theoretical hypotheses through model comparisons using one dataset—a scale of inquiry that would be difficult to support if each iteration required new experimental data.

These expectations are reflected in current research outcomes. On one hand, CE has been successfully implemented across multiple published studies (Huang, 2023, 2025a, 2025b) and is currently being applied in several ongoing projects and the above-mentioned large-scale collaborative

CE project . On the other hand, while influential, no published study has yet fully implemented the complete IED framework as originally described. Almaatouq et al. (2024) cite Peterson et al. (2021) and Awad et al. (2018) as examples of work aligned with IED principles—and indeed, these studies represent significant advances in scale and systematicity. However, they did not centrally organize around a pre-specified high-dimensional explicit design space, nor did they adopt IED's iterative experimentation protocol. Thus, while they valuably demonstrate the power of large-scale research, they do not yet establish the feasibility of IED's full methodological vision.

IED and CE: Complementary Futures

In summary, IED provides a valuable blueprint for addressing foundational challenges in psychological research, rightly emphasizing the importance of large-scale experimentation and integrative analysis. That said, certain features of its proposal—particularly its reliance on extensive upfront theoretical specification and iterative experimentation—may limit practical implementation in many current research contexts.

The CE approach can be seen as building upon IED's insights while offering a more accessible entry point. It retains IED's core strengths—large-scale data collection and systematic analysis—while shifting emphasis from theoretical pre-specification to data-guided theory-building. This adjustment enhances feasibility, especially in domains where theoretical understanding is still evolving.

It should be emphasized that IED remains a compelling framework for contexts where strong theoretical consensus exists. In such settings, its structured approach could efficiently guide

experimental design. As fields mature through the accumulation of empirical results—for instance, via CE-style research—the more demanding IED framework may become increasingly applicable. For now, however, CE offers a practical and powerful pathway for exploring complex, under-specified domains without requiring premature theoretical convergence.

Advantages of the CE approach

Within the PBS impossible trinity framework, the CE approach's primary advantage is successfully combining breadth and precision – two dimensions traditionally in tension in psychological research. Where conventional studies yield either precise-but-fragmented findings or broad-but-vague theories, CE creates precise linkages between diverse phenomena to form cohesive theoretical frameworks.

Breadth-related advantages

The CE approach's comprehensive scope offers three key advantages for theoretical progress:

First, it provides an integrative, system-level perspective. Traditional studies often examine cognitive mechanisms in isolation, offering fragmented insights akin to possessing scattered puzzle pieces without the picture on the box (Figure 2b). The CE approach, in contrast, assembles this picture (Figure 2d). For instance, while prior research has identified numerous elements of VWM, the QCE-VWM model is the first to offer a direct and comprehensive answer to the broad question: "How are these items memorized?" Crucially, although its constituent mechanisms were largely known, the model's novel contribution lies in its integrated computational architecture, which fully specifies the

relationships *between* these mechanisms. The value of a complete puzzle exceeds that of its scattered pieces; this added value resides in understanding how the pieces connect to form a coherent whole.

Second, it transcends binary theoretical debates. Traditional approaches often force artificial dichotomies (e.g., slot vs. resource models), whereas CE models synthetically incorporate elements from competing theories. The QCE-VWM model exemplifies this by demonstrating how "slot-like" features (e.g., representation precisions that are largely stable against external factors) and "resource-like" features (e.g., quality-quantity trade-offs) coexist within a single architecture. This offers a more nuanced account than either extreme position. Indeed, within the QCE-VWM framework, neither feature serves as the sole guiding principle. Thus, to achieve a precise and broad understanding of VWM, it is necessary to move beyond these debates and embrace a more complex, integrative truth.

Third, it serves as a quantified literature review. This approach combines the systematic coverage of traditional reviews with the empirical rigor of experimental studies (Figure 10). Huang (2025a) exemplifies this by formally integrating decades of VWM findings into a unified computational framework, thereby validating and refining prior insights through data-guided theory-building. Crucially, this method offers unique advantages over conventional reviews: model fitting provides a quantifiable tool for identifying missing or redundant mechanisms. Patterns of poor fit reveal theoretical gaps (e.g., undiscovered processes), while a lack of fit improvement indicates when proposed mechanisms are unnecessary—capabilities beyond the reach of traditional narrative reviews.

Precision-related advantages

While traditional experimental studies are already very precise, CE can enhance the precision even further by utilizing larger datasets and broader coverage of the stimulus space, allowing us to make finer mechanistic distinctions. For example, Huang (2025a) revealed that:

- Bayesian integration follows multiplicative rather than additive rules
- Memory errors follow a truncated normal rather than von Mises distribution

These subtle but theoretically important distinctions often elude smaller-scale studies.

Complexity of the CE approach

Complexity is theoretically necessary

The CE approach's ability to achieve both precision and breadth within the PBS impossible trinity inevitably requires embracing theoretical complexity. Huang (2025a)'s QCE-VWM model exemplifies this necessity, incorporating approximately a dozen distinct mechanisms through 57 parameters while generating 20 theoretical implications about VWM (Table 1, Huang 2025a). This complexity should not be viewed as a methodological flaw, but rather as an epistemically necessary response to the inherent complexity of cognitive systems. Traditional literature reviews face similar challenges when synthesizing diverse findings, suggesting that CE's complexity simply makes explicit what was always implicitly present in comprehensive theoretical accounts.

Complexity is data-justified

The line between scientifically justified simplicity and unjustified simplicity (driven by HCLPC or other biases) must be drawn through evidence-based, statistically rigorous criteria. As demonstrated in previous studies (e.g., Peterson et al., 2021), the optimal complexity of a model scales with data size (see also Figure 11). This justifies the moderate complexity of CE models like the 57-parameter QCE-VWM, which is calibrated to large-scale datasets (40 million responses). Nevertheless, researchers accustomed to small-scale studies may intuitively perceive such models as inherently problematic—a bias stemming from misgeneralized heuristics and HCLPC rather than scientific rationale.

General discussions

A personal journey with theory-driven hypothesis-testing

While incorporating personal experiences into scientific publications is uncommon, I share my own here to clarify the perspectives that shaped this article. My twelve-year effort (2007-2019) to develop the Boolean map theory (Huang & Pashler, 2007) into a comprehensive framework of visual attention proved unexpectedly illuminating through its failures. This odyssey through three distinct phases fundamentally reshaped my understanding of psychological research methods:

Phase 1: Initially, I aimed to develop a theory through traditional hypothesis-testing, striving for an ideal balance of precision, breadth, and simplicity—unaware of the constraints imposed by the PBS impossible trinity. However, years of intensive research led me to conclude this goal may be fundamentally unattainable. A persistent pattern emerged: attempts to simultaneously increase both

precision and breadth inevitably resulted in greater complexity, as additional factors and mechanisms came to light. This impasse forced a difficult realization: the cognitive phenomena I studied resisted elegant simplification.

Phase 2: Upon recognizing that precision, breadth, and simplicity cannot be optimized simultaneously, I shifted toward precise-and-broad (yet complex) frameworks, still relying on one-at-a-time hypothesis-testing. However, progress remained limited. A fundamental barrier was the lack of standardization across experiments, which rendered their results incomparable and thereby necessitated a more integrative and systematic approach.

Phase 3: I adopted a broader, more integrative approach while remaining theory-driven—an approach similar to Almaatouq et al. (2024)'s IED. For instance, building on Huang (2015a)'s findings (2 stimulus types × 3 tasks), I subsequently attempted to standardize attentional processing across a broader range: first to 16 stimulus types × 8 tasks (Huang, 2015b), and ultimately to 16 stimulus types × 26 tasks (Huang, 2022)[12]. While some progress was made within this broad scope, more ambitious attempts consistently failed, compelling a radical methodological shift.

Three key limitations emerged:

[12] Believing it self-evident that broader scope enhanced the work's quality and value, I was shocked when a colleague—initially enthusiastic about Huang's (2015a) focused finding—dismissed the far more comprehensive Huang (2022) as less interesting, overly complex, and lacking an elegant message. This prompted introspection, during which I realized I often react the same way to others' work. This realization led me to believe there is something fundamentally flawed about the so-called "theory-driven" approach. In reality, it is a "simplicity-driven" approach that penalizes rather than rewards the effort toward comprehensive theoretical understanding.

1. **Data constraints**: The largest-scale laboratory projects I can conduct (e.g., Huang 2022) yield only a few thousand hours of data—far too limited for broader ambitions.
2. **Theoretical precision**: Verbal theories proved inadequate for specifying complex cognitive mechanisms, necessitating quantitative information-processing models.
3. **Complexity in design**: Upfront theoretical specification became intractable when handling >5 factors, revealing the need for data-guided theory-building.

These insights—coupled with the empirical successes of AI—motivated the development of the CE approach, which was subsequently validated in Huang (2023, 2025a, 2025b). Ironically, when applied to VWM in Huang (2025a), the QCE-VWM model rejected my initial Boolean map hypothesis. This outcome underscores the reasonable objectivity of the CE approach, demonstrating its ability to mitigate the theoretical confirmation biases inherent in traditional approaches.

Data availability and the next paradigm shift

The rapid growth of data availability and computational power has fueled extraordinary progress in AI over the past two decades. A similar pattern is observable in psychology's history: the field has undergone multiple paradigm shifts—from Freud's psychoanalysis to behaviorism, and later, to cognitive psychology. While these transitions are often attributed to theoretical advancements, the increasing richness and precision of empirical data have consistently played a catalytic role.

For instance, behaviorist research typically relied on studies with hundreds of observations—a scale that constrained investigations to observable behaviors, leaving internal mental processes largely

inaccessible. The introduction of computers enabled studies with tens of thousands of observations, a hundredfold increase that empowered researchers to probe cognitive mechanisms, thereby catalyzing the cognitive revolution.

Today, psychology may be on the verge of another transformative shift. The internet now facilitates studies with tens of millions of observations—a further thousandfold leap—opening unprecedented opportunities for methodological innovation. Yet, data volume alone is not sufficient to drive a paradigm shift. The critical question is: how can this vast new data landscape be harnessed to propel psychological research into its next era? The answer may lie in a practice that has already revolutionized other data-intensive fields: benchmarking.

Benchmarking as the engine of change

Benchmarking provides the essential link between large-scale data and paradigmatic change. As detailed above, benchmarking alleviates complexity-induced burdens. More broadly, benchmarking—powered by standardized assessment—has the potential to fundamentally reorient academic evaluation. By shifting the focus from impression management to the actual content being assessed (Levashina & Campion, 2007), it creates a system that rewards substantive, long-term contributions over superficially impressive but shallow outputs.

The transformative power of standardized benchmarking is irrefutably demonstrated in AI, where it has catalyzed rapid progress by aligning professional incentives with reproducible achievements rather than subjective impressions. Psychology stands poised for a similar revolution. The ongoing large-scale

collaborative CE project on VWM discussed above could serve as a starting point. It would do more than just provide data; it would establish a new arena for scientific progress, replacing cyclical and often inconclusive debates—such as those between slot-based and resource-based models—with rigorous, evidence-based competition. The goal would shift from winning rhetorical arguments to engineering superior models that achieve an optimal, quantifiable balance between prediction and explanation on a shared, standardized stage.

Beyond any single project, the widespread adoption of benchmarking could transform research workflows by making standardized evaluation a central organizing principle[13]. Researchers could shift from perpetually designing isolated small-scale experiments to iteratively refining models against shared benchmark datasets, redirecting energy from generating fragmented findings toward a cumulative, collaborative enterprise of theory-building.

Furthermore, benchmarking presents a powerful systemic solution to the replication crisis by mitigating two of its core drivers: it eliminates selective reporting (as all researchers are evaluated on the same data) and drastically reduces analytical flexibility (through standardized metrics).

Naturalistic tasks

[13] While benchmarking provides a common foundation for communication, an over-reliance on specific benchmarks can stifle methodological diversity and narrow theoretical perspectives. Therefore, it should not be applied so rigidly that it limits intellectual creativity.

The CE approach has demonstrated its ability to achieve an optimal balance between prediction and explanation, as shown in Huang (2025a) using simplified artificial stimuli. An important next step involves extending this success to more complex, naturalistic tasks.

Our lab is currently pursuing two projects in this direction. The first examines aesthetic preferences by collecting ratings for 190,000 abstract artworks, including pieces by established artists, amateur creators, and AI systems. The second project analyzes poetry appreciation through ratings of 95,398 poems spanning various quality levels, from renowned poets to AI-generated verses. Both studies aim to develop interpretable mechanistic CE models that explain and predict human judgments in these domains.

These naturalistic tasks present greater challenges compared to studies using simplified artificial stimuli. The resulting models will likely be more complex and less predictively accurate than that in Huang (2025a). However, they offer significant advantages. Theoretically, they may open up entirely new research domains instead of optimizing existing ones. Practically, they demonstrate psychology's ability to address real-world problems - a crucial consideration as the field faces questions about its relevance in competition with AI approaches.

Observational data

As discussed above, controlled experiments are preferred within the CE approach due to their ability to isolate causal relationships and minimize the influence of extraneous variables. Without such

controls, observational data often reflects more complex and entangled underlying mechanisms, which can hinder the development of accurate and interpretable CE models.

That said, the CE approach is not fundamentally incompatible with observational data. In cases where ethical or practical constraints make experimental manipulation infeasible, the methodology of CE (i.e., data-guided theory-building) may still be applied. If the underlying mechanisms can be reasonably hypothesized or subsequently inferred, this approach remains capable of generating valuable theoretical understanding.

Sampling strategies

Sampling from a factor-informed stimulus space is a general principle that must be adapted to different research contexts.

First, Huang (2025a) illustrates a scenario with focused, artificial stimuli. Here, only a single constraint is imposed (i.e., number of items = 4) to isolate the phenomena of interest (i.e., multi-item memory). The relevant mechanisms are then captured through random variations within that constrained space.

Second, implementation is often more straightforward for naturalistic tasks. As with datasets like ImageNet, these require only a diverse and representative set of real-world examples (e.g., visual artworks or poems, as described above).

Third, more ambitious projects, whether using artificial or naturalistic stimuli, require a more complex stimulus space. For example, the ongoing large-scale collaborative CE project varies stimuli

along several dimensions (e.g., number of items, color, object identity, temporal order). The design is guided by the **principle of minimal sufficient constraints**: the goal is to identify the simplest possible stimulus space that can accommodate the phenomena of interest. Stimuli are randomly generated from this space, ensuring coverage of both known experimental conditions and novel, untested regions that may lead to discovery.

Like traditional design principles, the "minimal sufficient constraints" principle does not prescribe a single correct solution. Just as classic methods require domain knowledge to operationalize variables, constructing an optimal CE stimulus space requires leveraging existing theoretical understanding and, ideally, the collective wisdom of the research community. The key advantage is that a CE design, while sometimes challenging, is invariably more feasible than a theory-driven design of comparable scope, as it focuses on manipulating stimulus dimensions rather than the far more complex task of pre-defining and controlling high-dimensional factor interactions.

Feasibility

While CE studies require larger datasets than traditional psychology research, their associated costs remain manageable through online data collection. Huang (2025a) demonstrated this by obtaining a high-quality dataset of approximately 33,000 participant hours for about $26,000 on a custom-built platform. Subsequent analyses showed that even one-third of this data volume yields reliable parameter estimates, making individual CE projects viable for most researchers.

For more ambitious, community-wide projects—such as the collaborative benchmark study described above—costs scale accordingly but remain practical. An estimated investment of $260,000 for a foundational benchmark dataset becomes feasible when distributed across a consortium of labs. This cost should be framed not merely as an expense, but as a critical one-time infrastructure investment. Unlike individual studies, the resulting dataset would serve as a lasting, open resource for the entire research community, enabling years of cumulative theory-building and providing a high return on investment by eliminating redundant data collection efforts.

Concluding comments

In summary, I argue that psychological research should strive to achieve greater precision and breadth simultaneously. This is accomplished through a data-guided theory-building process that involves:

- leveraging large-scale experimentation
- employing quantitative information-processing models
- minimizing upfront theoretical specification in favor of exploratory, iterative model building
- embracing moderate complexity as a necessary trade-off for more precise and broader understanding.

Acknowledgements

The ChatGPT has been utilized to enhance grammar, spelling, and phrasing of this paper, but it has not been employed to generate any new text.

Funding statement

The work described in this paper was supported by the Research Grants Council of Hong Kong (CUHK 14606622).

Conflicts of interest statement

The author declares no competing interests.

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the optical society of america A, 2*(2), 284-299.

Agrawal, M., Peterson, J. C., & Griffiths, T. L. (2020). Scaling up psychology via scientific regret minimization. *Proceedings of the National Academy of Sciences*, 117(16), 8825-8835.

Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2024). Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behavioral and Brain Sciences, 47*, e33.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science, 15*(2), 106-111. doi:10.1111/j.0963-7214.2004.01502006.x

Anderson, B. (2011). There is no such thing as attention. *Frontiers in psychology, 2*, 246.

Anvari, F., Alsalti, T., Oehler, L. A., Hussey, I., Elson, M., & Arslan, R. C. (2025). Defragmenting psychology. *Nature Human Behaviour*, 1-4.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., . . . Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59-64.

Bae, G.-Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General, 144*(4), 744. doi:10.1037/xge0000076

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science, 321*(5890), 851-854. doi:10.1126/science.1158023

Berger, J. (2013). *Contagious: How to build word of mouth in the digital age*: Simon and Schuster.

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., . . . Heaton, R. F. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences, 46*, e385.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical Encoding in Visual Working Memory: Ensemble Statistics Bias Memory for Individual Items. *Psychological Science, 22*(3), 384-392. doi:10.1177/0956797610397956

Brady, T. F., & Alvarez, G. A. (2015). Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of Vision, 15*(15), 6-6. doi:10.1167/15.15.6

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review, 120*(1), 85. doi:10.1037/a0030779

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour, 5*(8), 980-989.

Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in Cognitive Sciences, 7*(1), 19-22.

Cogitate, C., Ferrante, O., Gorska-Klimowska, U., Henin, S., Hirschhorn, R., Khalaf, A., . . . Vidal, Y. (2025). Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature*, 1-10.

Crocker, L., & Algina, J. (2006). *Introduction to classical and modern test theory*: ERIC.

Daston, L., & Galison, P. L. (2021). *Objectivity*: Princeton University Press.

Dubova, M., Moskvichev, A., & Zollman, K. (2022). Against theory-motivated experimentation in science.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429-433.

Fleming, S., Frith, C., Goodale, M., Lau, H., LeDoux, J. E., Lee, A. L., . . . Slagter, H. A. (2023). The integrated information theory of consciousness as pseudoscience.

Fougnie, D., Cormiea, S. M., Kanabar, A., & Alvarez, G. A. (2016). Strategic trade-offs between quantity and quality in working memory. *Journal of Experimental Psychology Human Perception and Performance, 42*(8), 1231.

Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature Communications, 3*(1), 1229.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review, 103*(4), 650.

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*: Cambridge university press.

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition, 135*, 21-23.

Heath, C., & Heath, D. (2007). *Made to stick: Why some ideas survive and others die*: Random House.

Hommel, B., Chapman, C. S., Cisek, P., Neyedli, H. F., Song, J.-H., & Welsh, T. N. (2019). No one knows what attention is. *Attention, Perception, & Psychophysics, 81*, 2288-2303.

Huang, L. (2015). Color is processed less efficiently than orientation in change detection but more efficiently in visual search. *Psychological Science, 26*(5), 646-652. doi:10.1177/0956797615569577

Huang, L. (2020). Unit of visual working memory: A Boolean map provides a better account than an object does. *Journal of Experimental Psychology: General, 149*(1), 1-30. doi:10.1037/xge0000616

Huang, L. (2022). FVS 2.0: A unifying framework for understanding the factors of visual-attentional processing. *Psychological Review, 129*(4), 696-731. doi:10.1037/rev0000314

Huang, L. (2023). A quasi-comprehensive exploration of the mechanisms of spatial working memory. *Nature Human Behaviour*, 7, 729–739.

Huang, L. (2025a). Comprehensive exploration of visual working memory mechanisms using large-scale behavioral experiment. *Nature Communications*, 16, 1383.

Huang, L. (2025b). Aesthetic preferences among spatial patterns: Large-scale experiment, comprehensive exploration, and a three-component regularity-based model. *Psychology of Aesthetics, Creativity, and the Arts*. Advance online publication.

Huang, L., & Pashler, H. (2007). A Boolean map theory of visual attention. *Psychological Review, 114*(3), 599-631. doi:10.1037/0033-295x.114.3.599

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology, 79*(6), 995.

Jolly, E., & Chang, L. J. (2019). The flatland fallacy: Moving beyond low–dimensional thinking. *Topics in Cognitive Science, 11*(2), 433-454.

Kahneman, D. (2011). Thinking, fast and slow. *Farrar, Straus and Giroux*.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: a failure to disagree. *American psychologist, 64*(6), 515.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444.

Lenharo, M. (2024). The consciousness wars: can scientists ever agree on how the mind works? *Nature, 625*(7995), 438-440.

Levashina, J., & Campion, M. A. (2007). Measuring faking in the employment interview: development and validation of an interview faking behavior scale. *Journal of Applied Psychology, 92*(6), 1638.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences, 6*(1), 9-16.

Liesefeld, H. R., Liesefeld, A. M., & Müller, H. J. (2019). Two good reasons to say 'change!'–ensemble representations as well as item representations impact standard measures of VWM capacity. *British Journal of Psychology, 110*(2), 328-356.

Liesefeld, H. R., & Müller, H. J. (2019). Current directions in visual working memory research: An introduction and emerging insights. *British Journal of Psychology, 110*(2), 193-206.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology, 55*(3), 232-257.

Musslick, S., Hewson, J. T., Andrew, B. W., Strittmatter, Y., Williams, C. C., Dang, G. T., . . . Holland, J. G. (2023). *An evaluation of experimental sampling strategies for autonomous empirical research in cognitive science.* Paper presented at the Proceedings of the annual meeting of the cognitive science society.

Oberauer, K. (2023). Measurement models for visual working memory—A factorial model comparison. *Psychological Review, 130*(3), 841.

Oberauer, K., Farrell, S., Jarrold, C., & Lewandowsky, S. (2016). What limits working memory capacity? *Psychological Bulletin, 142*(7), 758.

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D., Conway, A., Cowan, N., . . . Hurlstone, M. J. (2018). Benchmarks for models of short-term and working memory. *Psychological Bulletin, 144*(9), 885.

Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics, 44*(4), 369-378. doi:10.3758/bf03210419

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*(6), 531-536.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528-530.

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science, 372*(6547), 1209-1214.

Popper, K. (2005). *The logic of scientific discovery*: Routledge.

Popper, K. (2014). *Conjectures and refutations: The growth of scientific knowledge*. Routledge.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and social psychology review, 8*(4), 364-382.

Rosenholtz, R. (2024). Visual Attention in Crisis. *Behavioral and Brain Sciences*, 1-32.

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron, 108*(3), 413-423.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: happiness is a matter of choice. *Journal of Personality and Social Psychology, 83*(5), 1178.

Shackel, N. (2005). The vacuity of postmodernist methodology. *Metaphilosophy, 36*(3), 295-320.

Suchow, J. W., Brady, T. F., Fougnie, D., & Alvarez, G. A. (2013). Modeling visual working memory with the MemToolbox. *Journal of Vision, 13*(10), 9-9.

Suchow, J. W., Fougnie, D., Brady, T. F., & Alvarez, G. A. (2014). Terms of the debate on the format and structure of visual memory. *Attention, Perception, & Psychophysics, 76*, 2071-2079.

Van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial comparison of working memory models. *Psychological Review, 121*(1), 124.

Van den Berg, R., & Ma, W. J. (2018). A resource-rational theory of set size effects in human visual working memory. *ELife, 7*, e34963.

Watts, D. J. (2017). Should social science be more solution-oriented? *Nature Human Behaviour, 1*(1), 1-5.

Westen, D. (2008). *The political brain: The role of emotion in deciding the fate of the nation*: PublicAffairs.

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences, 45*, e1.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100-1122.

Zhang, W. W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*(7192), 233-U213. doi:10.1038/nature06860

Zhaoping, L., & Zhe, L. (2015). Primary visual cortex as a saliency map: a parameter-free prediction and its test by behavioral data. *Plos Computational Biology, 11*(10), e1004375.

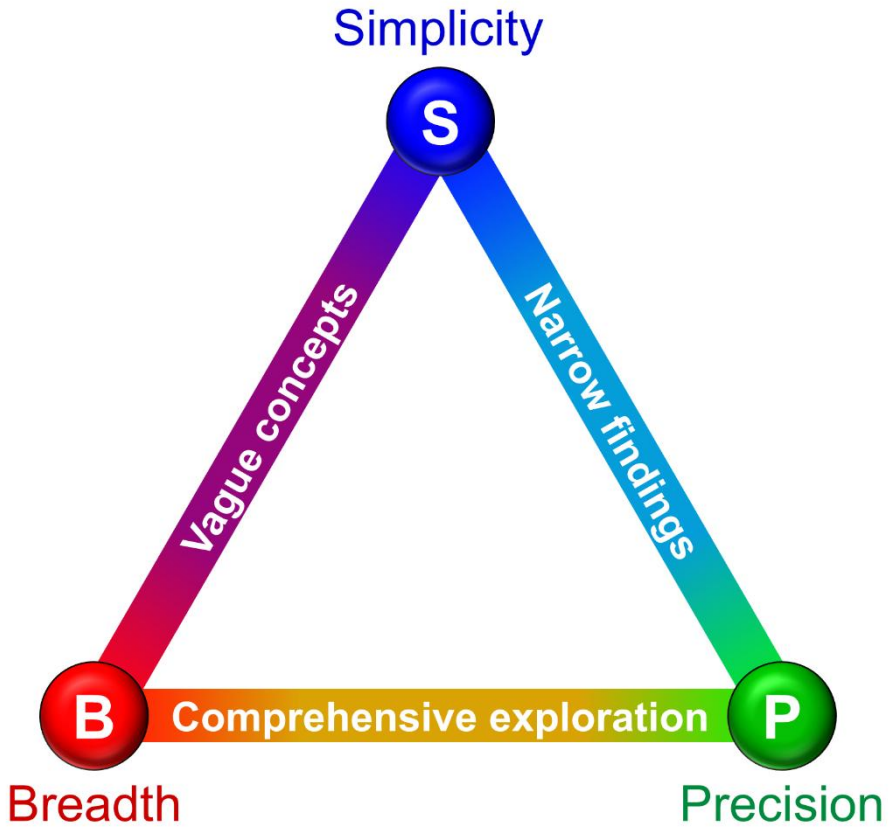

Figure 1. The PBS Impossible trinity in Psychological Research

This figure illustrates the PBS Impossible trinity, a fundamental constraint in psychological research where only two of three criteria can typically be optimized simultaneously: precision, breadth, and simplicity. Experimental studies often yield precise and simple—but narrow—findings, whereas broad theoretical concepts sacrifice precision for generality. The comprehensive exploration (CE) approach proposed here prioritizes precision and breadth, embracing moderate complexity to integrate diverse phenomena into unified information-processing models.

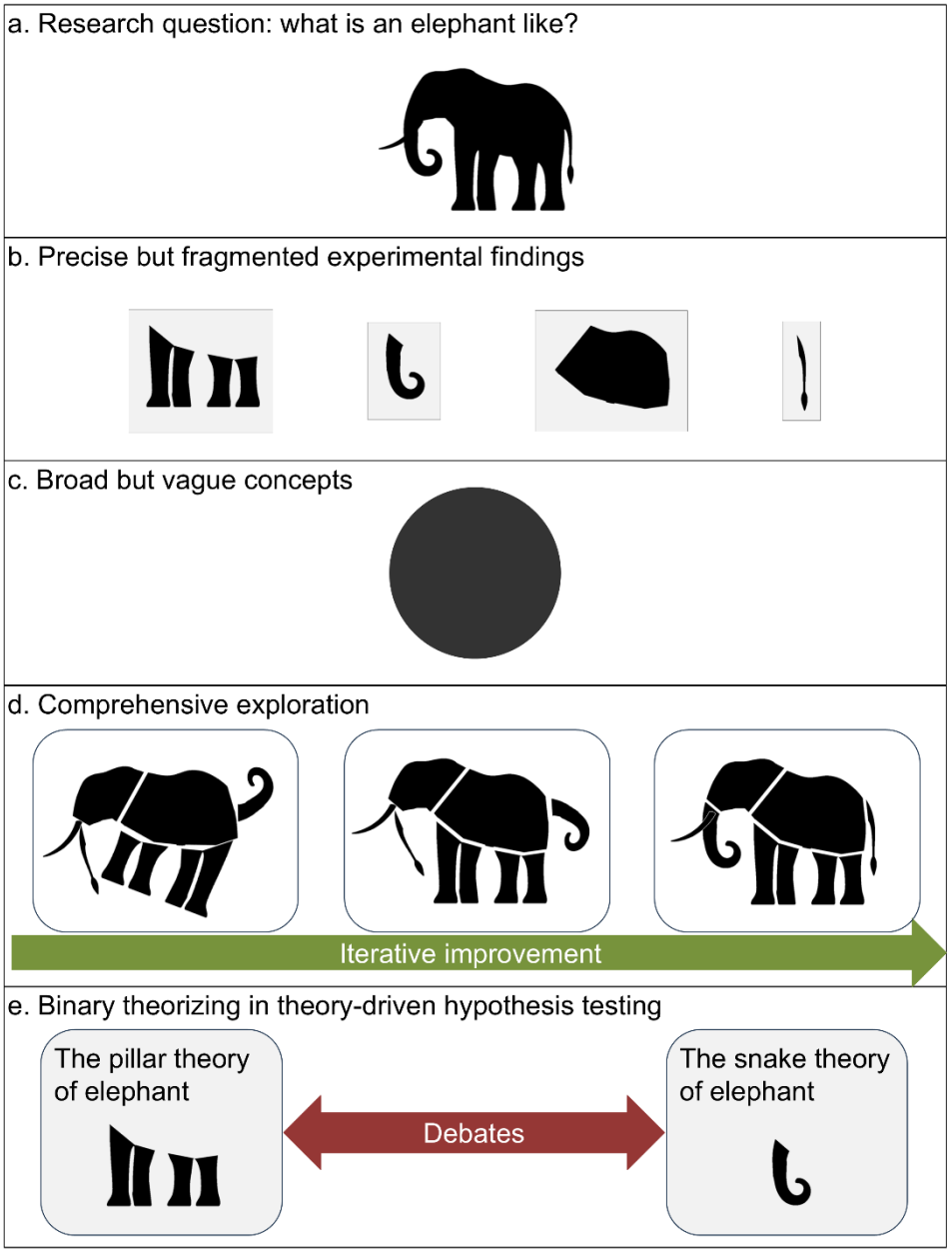

Figure 2. The "Blind Men and the Elephant" Parable

(a) The elephant symbolizes the multifaceted nature of mental mechanisms. (b) Precise-but-fragmented experimental findings examine isolated aspects, much like each blind man perceiving only one part of the elephant. (c) Broad-but-vague theories oversimplify complexity, akin to describing the elephant as a "sphere". (d) Precise-and-broad studies (the CE approach) integrate components, determine their interconnections, and iteratively refine understanding. (e) Binary theorizing reduces complexity to oversimplified dichotomies (e.g., "pillar vs. snake"), forcing artificial opposition.

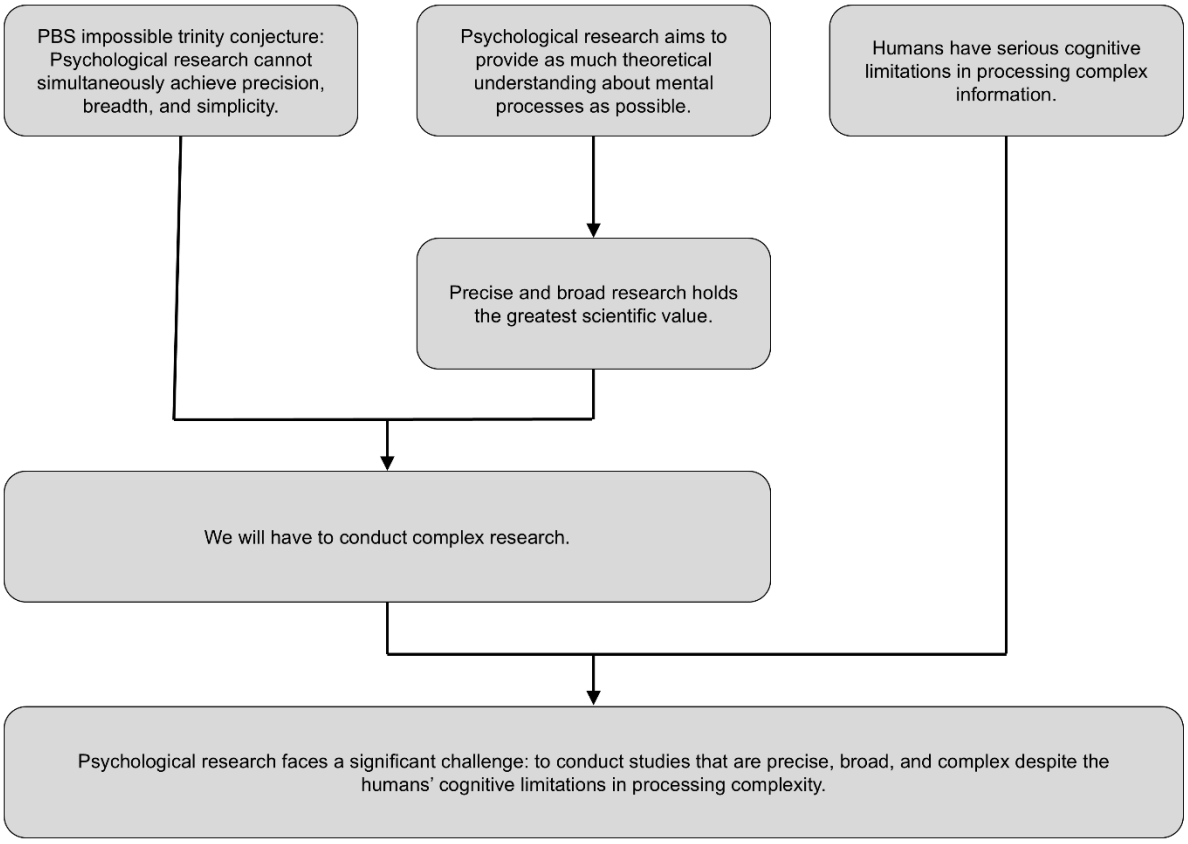

Figure 3. A challenge for psychological research

Human cognition is inherently complex, making it challenging to achieve precision, breadth, and simplicity simultaneously in psychological theories. Although precise-and-broad research holds the greatest scientific value, its inherent complexity conflicts with humans' cognitive limitations in processing complexity (HCLPC). This creates a key challenge: developing methods for precise-and-broad research that reduce complexity-induced burden to a manageable level. The **comprehensive exploration (CE) approach** addresses this challenge by maintaining precision and breadth while mitigating complexity-induced burden through benchmarking.

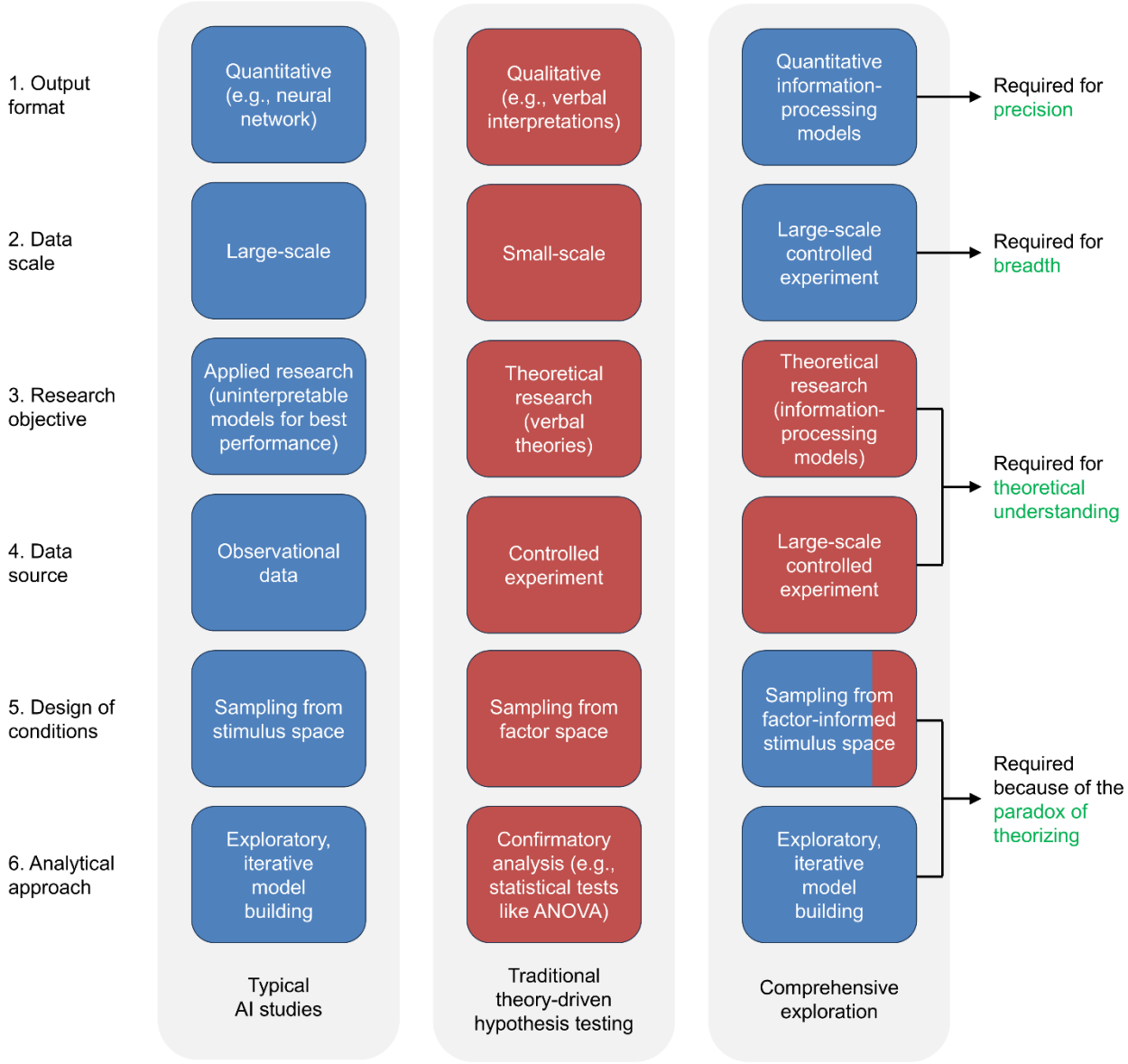

Figure 4. Six Methodological Dimensions

This figure compares six key methodological dimensions in typical AI studies (left column) and traditional theory-driven hypothesis-testing in psychology (middle column). The right column presents the methodological approach of the comprehensive exploration (CE), with blue and red indicating alignment with AI and psychology traditions, respectively.

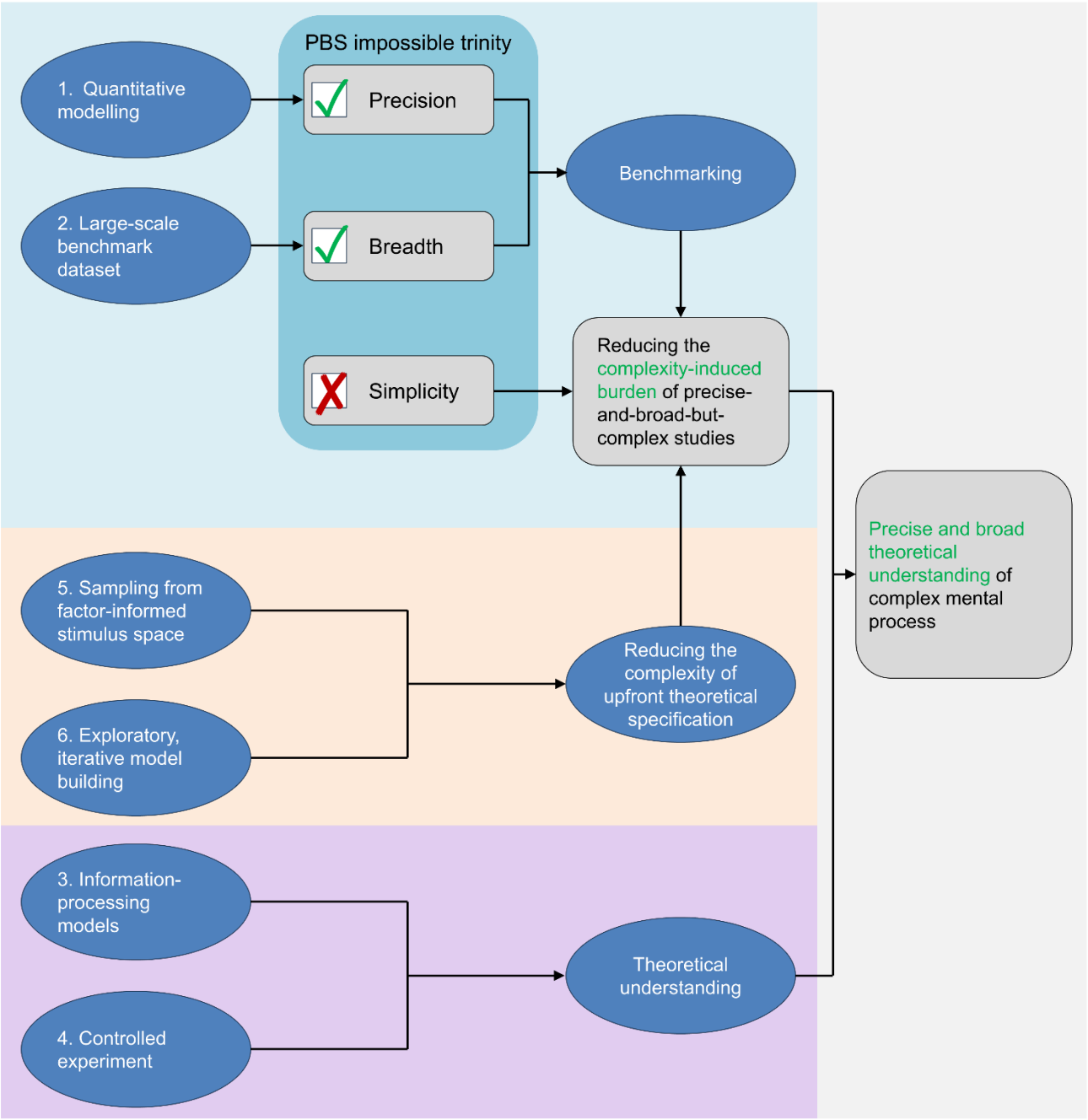

Figure 5. Rationale behind the six methodological choices

The upper cyan region (Dimensions 1-2) reflects choices made to achieve precision, breadth, and benchmarking. The lower purple region (Dimensions 3-4) represents choices made to prioritize theoretical understanding, while the central yellow region (Dimensions 5-6) comprises choices made to enhance the feasibility of experimental design.

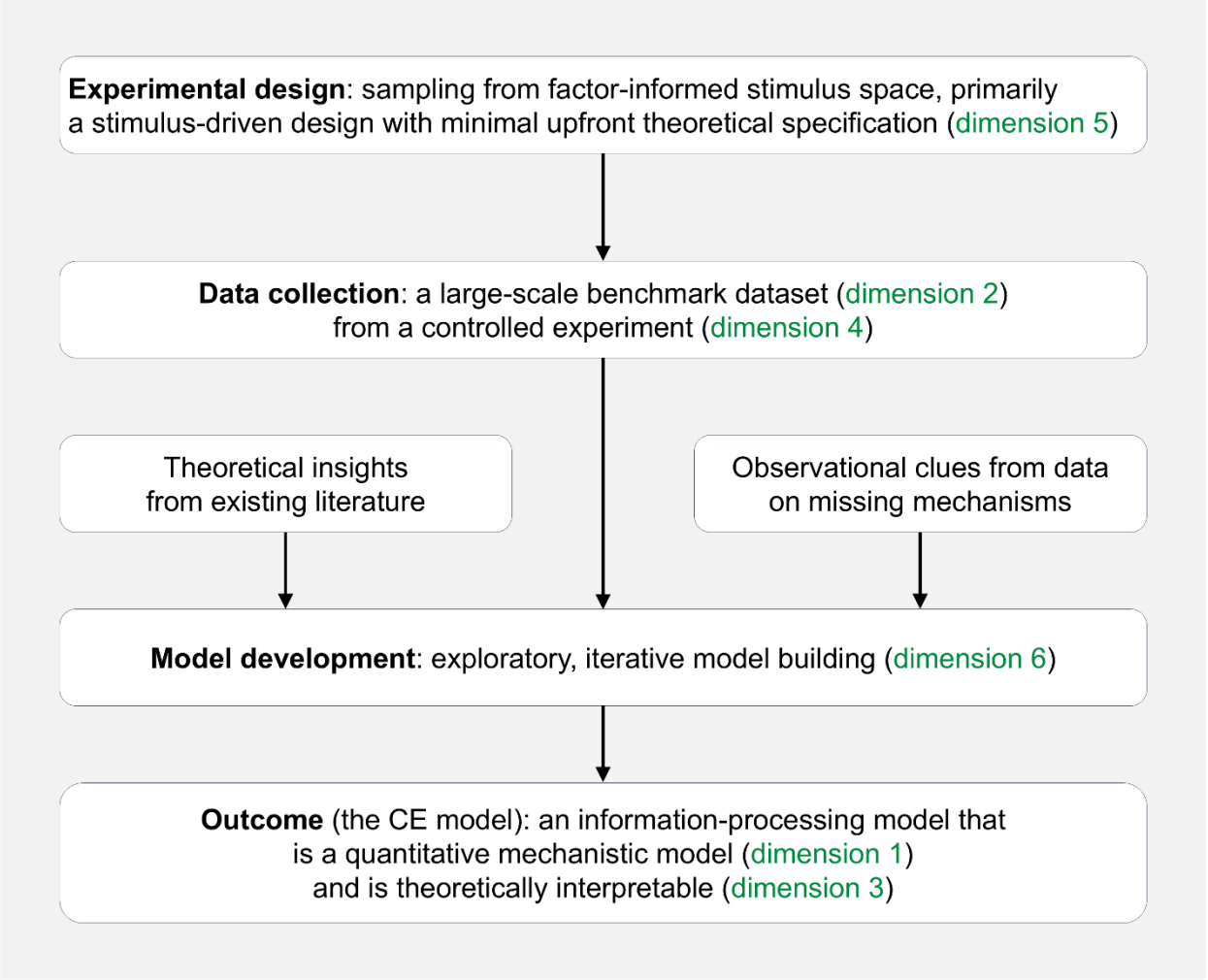

Figure 6. Comprehensive exploration (CE) study process

A typical CE study consists of several steps: experimental design, data collection, and model development, which ultimately produces the outcome of the study—a CE model.

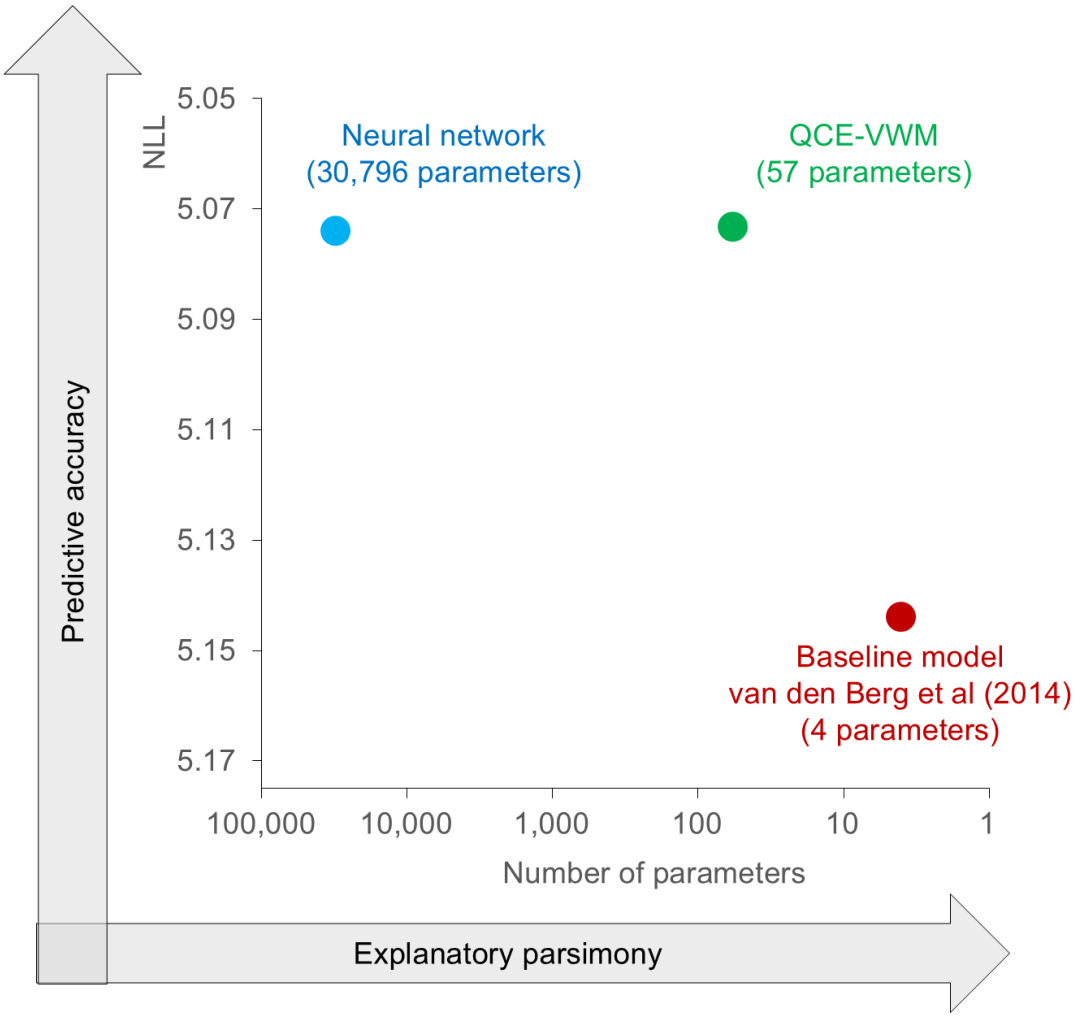

Figure 7. An optimal balance between prediction and explanation

The QCE-VWM model (Huang, 2025a) employs just 57 parameters, all of which are fully interpretable. Despite its explanatory parsimony, it outperforms a benchmark neural network with 30,796 parameters in predictive accuracy. This represents a rare case in which we can assert an optimal balance between prediction and explanation.

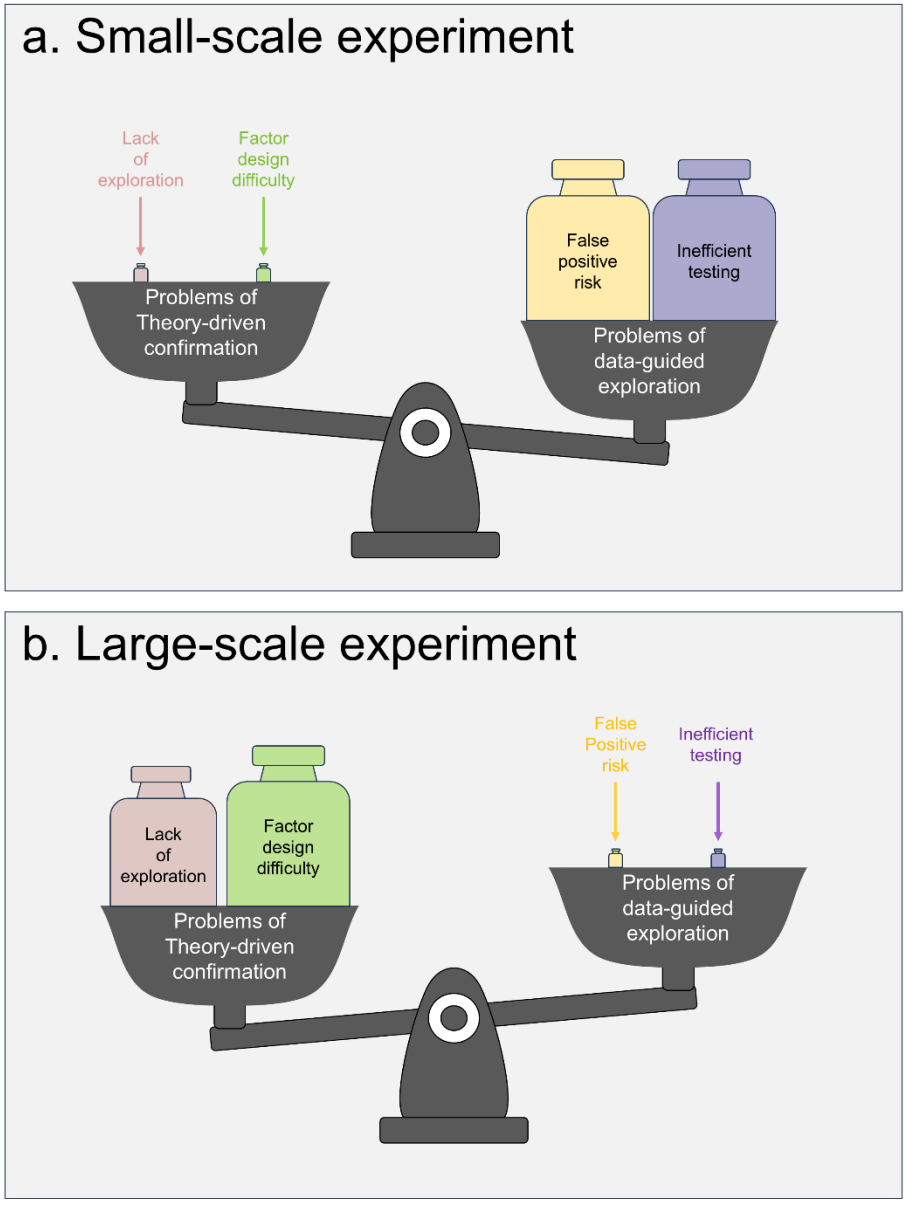

Figure 8. Confirmatory and exploratory analysis

This figure summarizes the factors influencing the choice between theory-driven confirmation and data-guided exploration. When shifting from small-scale to large-scale studies, the problems of confirmatory analysis—such as lack of exploration and factor design difficulty—become significantly more pronounced. In contrast, the drawbacks of exploratory analysis, including false positive risk and inefficient testing, are substantially mitigated. Consequently, while confirmatory methods remain well-suited for small-scale experiments, exploratory analysis emerges as the preferable approach for large-scale investigations.

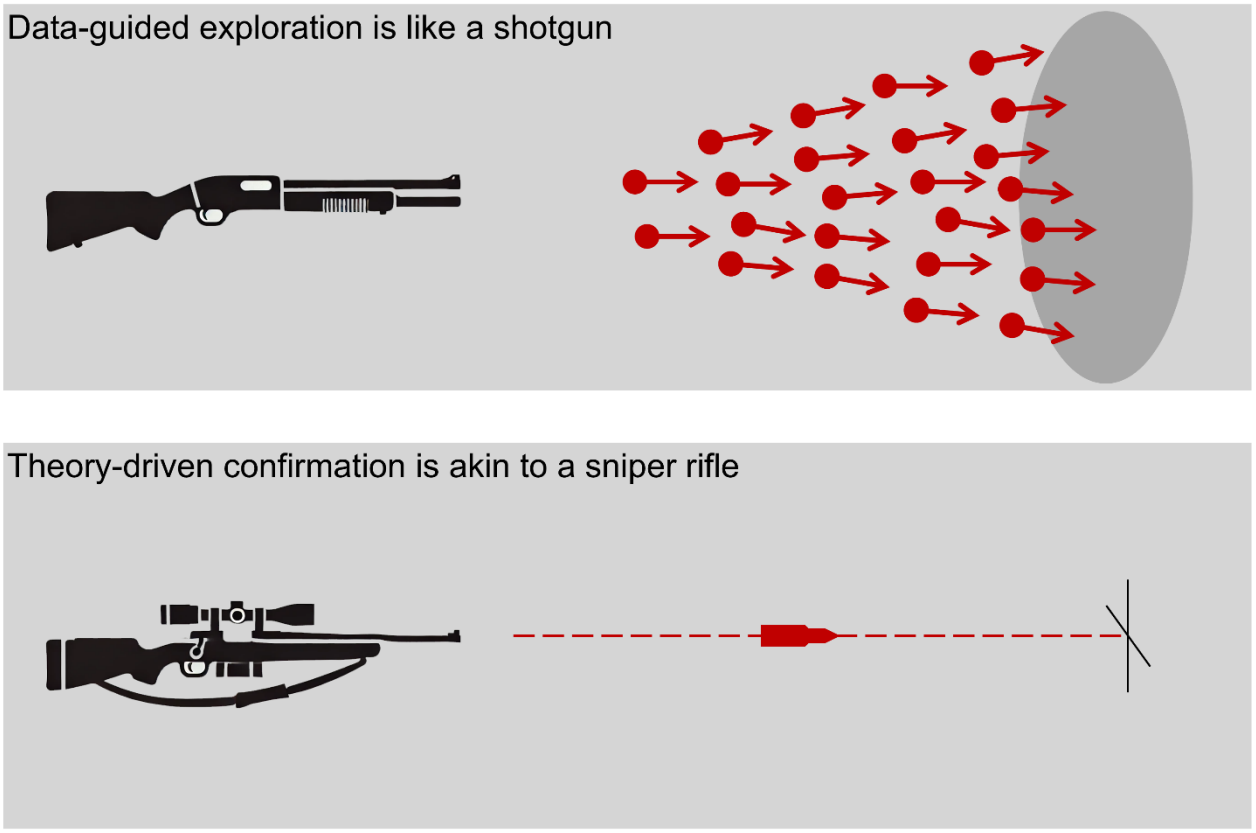

Figure 9. A firearms analogy

Theory-driven confirmation in small-scale studies is like a sniper rifle: a single, precise shot for narrow targeting. The cost of a miss is high, so precise control is essential, while the inability to hit unexpected targets is acceptable. In contrast, data-guided exploration in large-scale studies is like a shotgun blast: many scattered pellets for broad coverage. Controlling the trajectory of every pellet is prohibitively difficult, making broad coverage optimal. Here, "missing" with many pellets is inconsequential, while hitting unexpected targets becomes a significant advantage. Therefore, applying the sniper's precise control to the shotgun's domain—using theory-driven confirmation for large-scale studies—is both difficult and counterproductive.

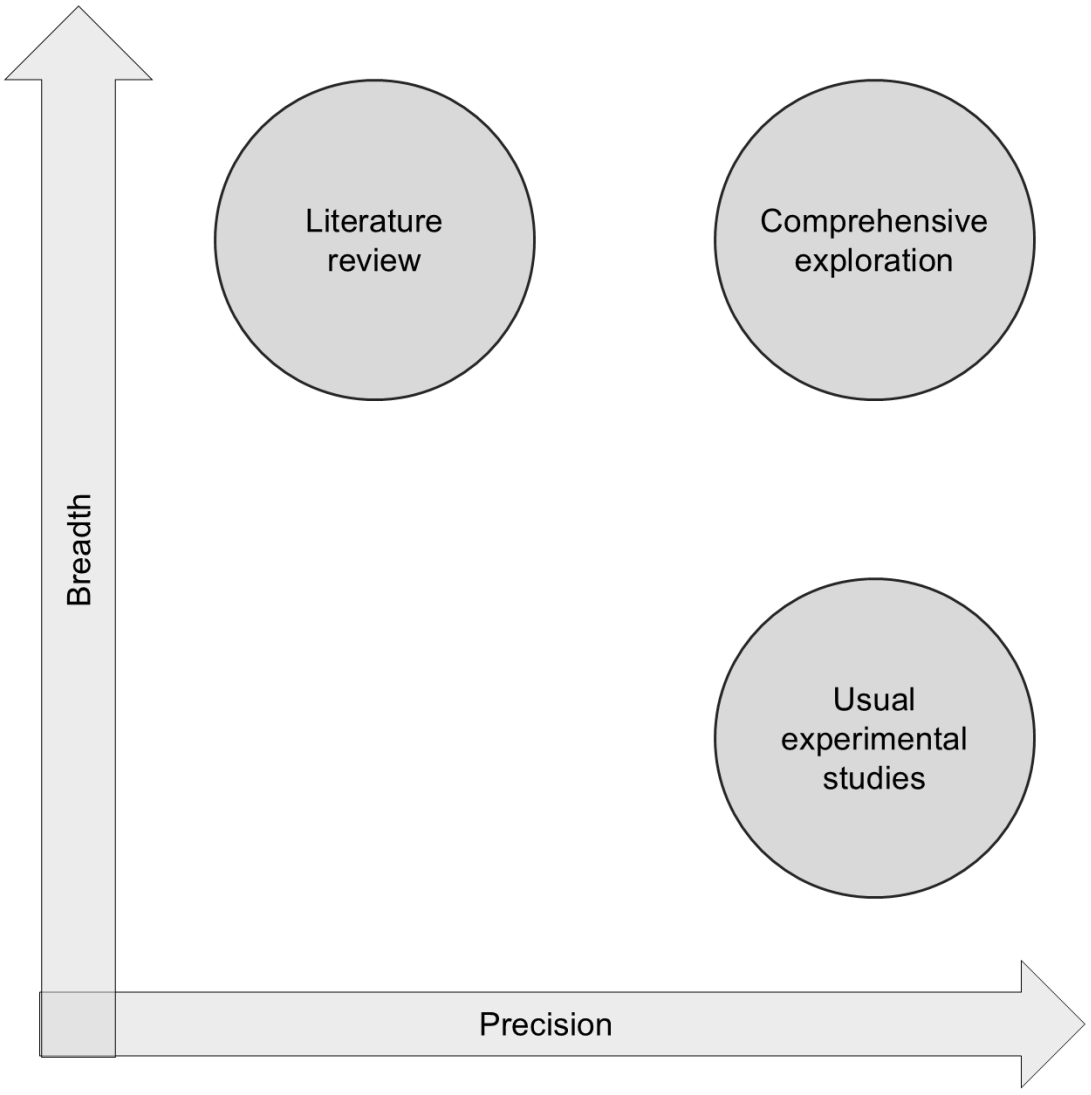

Figure 10. Quantified literature review

The CE approach functions as a quantified literature review, combining the precision and rigor of experimental studies with the integrative scope of traditional reviews.

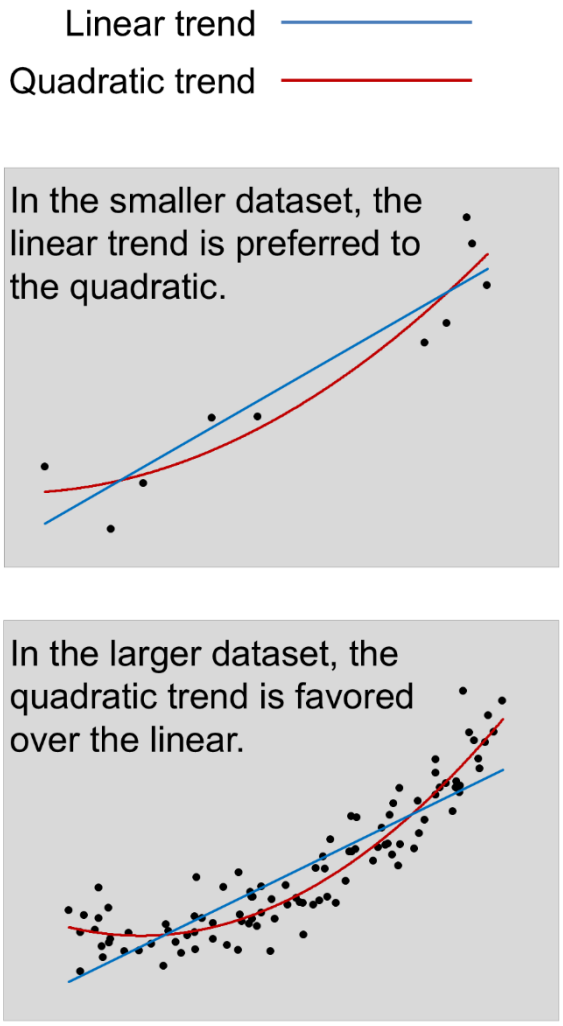

Figure 11. Optimal Complexity and Dataset Scale

The upper panel shows a subset of data points from the lower panel. Statistically, the simpler (linear) model is better suited for smaller datasets, while the more complex (quadratic) model fits larger datasets more effectively. In general, as dataset size increases, the optimal model tends to be more complex— justifying the moderate complexity of CE models.