This is the accepted manuscript version of:

Dolcini, N., Jap, B.A.J., Kun, C.W., & S. Politzer-Ahles. (2025). The pursuitworthiness of experiments in neurolinguistics. *European Journal for Philosophy of Science 15*(65). doi: 10.1007/s13194-025-00691-z

When citing this work, please refer to the published version.

The Pursuitworthiness of Experiments in Neurolinguistics

Nevia Dolcini[1,2*], Bernard A.J. Jap[3], Kun Chan Wa[1], Stephen Politzer-Ahles[4]

[1] Department of Philosophy and Religious Studies, Faculty of Arts and Humanities, University of Macau, Macao, China
[2] Centre for Cognitive and Brain Sciences, University of Macau, Macao, China
[3] Department of Humanities, Language and Translation, Hong Kong Metropolitan University, Hong Kong
[4] Department of Linguistics, University of Kansas, United States
*corresponding author E-mail: ndolcini@um.edu.mo

ABSTRACT

This paper explores the pursuitworthiness of experiments in neurolinguistics by investigating whether criteria for pursuitworthiness can be established for experiments in this field and, if so, which factors are most relevant. Drawing on a detailed analysis of research practices and case studies in neurolinguistics, we propose that pursuitworthiness in this domain should be evaluated along two dimensions: methodological and pragmatic. These dimensions, while both essential, stand in an asymmetrical relationship: methodological criteria primarily serve as non-negotiable thresholds for epistemic adequacy, filtering the space of viable experiments, whereas pragmatic criteria, although context-sensitive, play a role in guiding comparative prioritization *within* that viable space. Additionally, we critically assess Jamie Shaw's distinction between *luxury* and *urgent* science by exploring its utility in mapping the interplay between methodological and pragmatic criteria. We contend that, while the distinction offers valuable theoretical insights, its applicability to neurolinguistics is limited, mainly due to difficulties in predicting which experiments will yield the most impactful outcomes. Ultimately, our analysis demonstrates the value of adopting a structured yet context-sensitive approach to experimental pursuitworthiness in neurolinguistics, grounded in the interplay of methodological integrity and pragmatic viability.

Keywords: experimental research; methodological criteria; neurolinguistics; pragmatic criteria; pursuitworthiness.

1. Introduction

The concept of pursuitworthiness and its prominence in current philosophical debates stem from the need to establish a rational basis for determining which scientific endeavors merit further investigation, hence for guiding research trajectories. This was precisely Larry Laudan's objective in introducing the logic of pursuit (Laudan, 1977): to emphasize that the rational assessment of scientific theories should extend beyond the context of truth, or empirical

adequacy and confirmation, thereby challenging not only Carnap's well-known empiricist emphasis on verification, but also perspectives that narrow evaluation to singular criteria, such as, among others, Lakatos' requirement for progressive theoretical shifts. Instead, scientific assessment, rather than serving as an absolute measure of a theory's empirical and theoretical virtues in a competitive vacuum, is a comparative matter. It should therefore include "[…] judgment as to how a theory stacks up against its known contenders" (Laudan 1977, p. 71). Pursuitworthiness, on this view, is not merely an ancillary concern but a core component of a broader methodology meant to guide decisions about where intellectual and material resources should be allocated. Pursuitworthiness remains a salient ingredient in contemporary discussions about scientific choices, theoretical virtues, and the role of epistemic or non-epistemic values in informing theory assessment, development, and prioritization (Šešelja & Straßer, 2014; Nyrup, 2020; Lichtenstein, 2021; Fleisher, 2022; Shaw, 2022; Han, 2023; Fischer, 2024; Wolf & Duerr, 2024). Yet while much of this debate has traditionally focused on theories, models and research programs, the logic of pursuit—and with it the notion of pursuitworthiness—is just as crucial when applied to experiments.

Experimentation—now understood to be central to scientific inquiry, thanks in no small part to the influential work of Ian Hacking (1983)—plays a vital role in shaping, challenging, and refining theoretical frameworks (Gooding, 1990; Karaca, 2013; Hangel & ChoGlueck, 2023). This suggests the need to treat the pursuitworthiness of *experiments* as a subject of systematic attention in its own right. Key issues in this context include whether pursuitworthiness is primarily an epistemic matter, a pragmatic one, or a combination of both (Kitcher, 2001; Elliot & McKaughan, 2009; Šešelja et al., 2012); which theoretical virtues (e.g., explanatory power, coherence with established theories, and empirical adequacy) are relevant; and what role practical constraints, such as feasibility, technological readiness and funding availability, should play (e.g., Nickels, 2006). Furthermore, it remains unclear whether such criteria should be uniform across the sciences or tailored to the particular demands of specific disciplinary contexts. The urgency of these questions has been increasingly recognized (DiMarco & Khalifa, 2022; Laymon & Franklin, 2022; Fischer, 2024), especially in light of the fact that not all experimental projects can be undertaken simultaneously, which necessitates principled means for assessing their relative merit for pursuit (Elliott & McKaughan, 2014). Nonetheless, despite extensive philosophical work on the role and function of experimentation (Collins, 1985; Shapin & Schaffer 1985; Galison, 1987; Franklin, 1986; Culp, 1995; Burian, 2007; Simons & Vagelli, 2021), the question of what makes an experiment pursuitworthy remains relatively unexplored (Steinle, 2002; Wray, 2024).

This paper contributes to filling that gap by focusing on the specific case of experimental practices in neurolinguistics, a field at the intersection of linguistics and cognitive neuroscience. Neurolinguistic experiments investigate the neural underpinnings of language processing by using techniques that measure brain activity. These

consist of electrophysiological methods that measure synaptic activity—such as electroencephalography (EEG), magnetoencephalography (MEG), and electrocorticography (ECoG)—and hemodynamic methods that measure blood flow in the brain—including functional magnetic resonance imagining (fMRI), positron emission tomography (PET), and near-infrared spectroscopy (NIRS), among others—or techniques that measure individuals' behaviour when their brain function has been affected as in brain damage, or during transcranial magnetic or direct-current stimulation (TMS/TDCS). These methods are highly sensitive to fine-grained linguistic distinctions such as the roles of syntax (sentence structure) and semantics (how linguistic meanings are composed), as well as cognitive variables like memory and attention, making experimental design, execution, and interpretation particularly complex. For this reason, neurolinguistic research provides a compelling case for examining the criteria by which experiments are evaluated and prioritized.

More specifically, two central questions will be addressed: (i) whether criteria of pursuitworthiness for experiments in neurolinguistics can—or should—be established; and if so, (ii) which factors are most relevant and ought to be considered within this specific experimental domain. We argue that such criteria are both feasible and desirable. Based on a detailed analysis of practices and cases in the field, we suggest that the pursuitworthiness of neurolinguistic experiments is best understood as a bidimensional concept, requiring attention to both *methodological* and *pragmatic* considerations. The methodological dimension pertains to an experiment's scientific rigor, interpretability, and significance, while the pragmatic dimension concerns real-world constraints and circumstances, potentially contributing further guiding features for experimental prioritization. Taken together, these two dimensions define the axes of the space in which criteria of pursuitworthiness for experiments in neurolinguistics can be mapped.

At the same time, we also emphasize the importance of parsimony. The broader debate on pursuitworthiness ranges from radical openness—an "anything goes" stance mainly aimed at preserving the potential for scientific discovery, in the spirit of Reichenbach (1938) and Feyerabend (1975)—to an approach that advocates for rule-governed decisions about which research should be pursued. The latter requires identifying the conditions under which a research program can be deemed worthy of pursuit, with various proposals seeking a viable middle path[1].

[1] Multiple accounts seek to strike a middle path between rigid constraints and unfettered exploration. This is evident in Laudan's (1977) notion of a "high rate of progress," which assesses theories not only by their empirical success but also by their capacity to produce progressive problem shifts, thus combining methodological accountability with innovation. Achinstein's (1993) "set of questions" approach similarly aims at balancing methodological rigor (e.g., simplicity, testability) with practical constraints (e.g., time, money). Whitt (1992) reconciles empirical fertility and conceptual viability to avoid both overly rigid empiricism and loose theorizing, and Šešelja and Straßer (2014) propose a framework based on potential explanatory power, inferential density, and consistency, designed to harmonize methodological discipline with exploratory openness.

Notably, the trajectory of this path must safely navigate between two perils: the Scylla of "anything goes" relativism and the Charybdis of narrowly restrictive positivism (Laudan 1996). This balancing act becomes even more delicate in the domain of experimental research: a *laissez-faire* approach risks the proliferation of unreliable or methodologically compromised studies, while excessively rigid standards may prematurely foreclose exploratory work that could yield valuable outputs. The awareness of these two dangers forms the backdrop to our proposal.

The paper proceeds as follows. In section 2, based on the specificities of experimental practices in neurolinguistics, we suggest that pursuitworthiness in this field is best captured by distinguishing between *methodological* and *pragmatic criteria*. The former concern scientific rigor and the reliability of experimental designs, while the latter relate to contextual features, including resources, cost, time, equipment, participant accessibility, expertise, and potential applications. In Section 3, we elaborate on specific methodological criteria, where we examine in detail how they tend to operate in contemporary neurolinguistic research. Importantly, these proposed criteria are not intended as prescriptions of what should or should not be studied, or how it should be studied. Rather, our goal is to critically discuss and identify key methodological components that, more than others, play a role in judgments of pursuitworthiness. In Section 4, we turn to the practical constraints, specifically resource availability limitations and experimental design challenges, involved in conducting neurolinguistic experiments. These constraints serve as important components of pursuitworthiness judgments in neurolinguistic research, and their tension with methodological ideals creates what might be termed the "pursuitworthiness optimization problem". In Section 5, our proposal is tested in light of the distinction between "luxury" and "urgent" science, as introduced by Shaw (2022), where we identify its implications as a potential parameter for assessing pursuitworthiness. The implications of this distinction for prioritization and resource allocation in neurolinguistic research are examined. We argue that, despite the undeniable relevance of the distinction, its practical utility in this context is limited—primarily due to the inherent difficulty of predicting which experiments will ultimately yield clinically (socially, or morally) significant outcomes. We conclude by briefly considering the potential relevance of our approach for experimental research in neuroscience, and by identifying areas where further work is needed to refine experiments' pursuitworthiness criteria, as well as to better understand their complex interplay.

2. The two dimensions of pursuitworthiness in neurolinguistics

Determining which experimental projects are worth pursuing—and thus warrant continued or increased investment—is a fundamental aspect of scientific decision-making. This is particularly crucial in resource-constrained environments, where choices must account not only for funding, but also for the effective use of time, labor, technical

infrastructure, and institutional support. Yet pursuitworthiness involves more than just practical constraints: experiments that fall short of basic methodological standards risk not only wasting resources but also undermining the reliability and credibility of the field as a whole. From this perspective, skepticism about the logic of pursuit—namely, the concern that setting criteria may unduly constrain exploration and restrict scientific discovery (Lynch, 2020)—while theoretically defensible (Stanford, 2006; Shaw, 2022), becomes far less persuasive when applied to experimental research. Unlike theories, which often permit open-ended exploration and greater flexibility in terms of how and when they are developed, experimental work is bound by immediate, and often non-negotiable, material, financial, and logistical constraints. In this context, developing criteria for pursuitworthiness is not only appropriate but necessary. The challenge, then, is not whether such criteria should exist, but what kinds of considerations are relevant, and how they interact in practice.

Developing pursuitworthiness criteria is especially urgent in neurolinguistics due to a combination of four factors. Firstly, the field blends linguistics, psychology, and neuroscience—disciplines that each bring their own methodologies and face different resource limitations (Poeppel & Embick, 2005; Kemmerer, 2015). Secondly, experiments in this area frequently demand substantial resources, such as costly neuroimaging tools (fMRI, MEG, EEG) that are not always available, specialized software, and considerable time from researchers and participants alike (Poline et al., 2012). Thirdly, neuroimaging research often struggles with low statistical power (Button et al., 2013; Nord et al., 2017), and subtle effects sometimes need massive numbers of human participants to be statistically detected (Nieuwland et al., 2018). This warrants further questions about how to best allocate limited resources to the most promising experiments. A fourth factor distinguishing neurolinguistics from many other neuroscientific subfields is its near-complete dependency on human experimentation. While neuroscientists studying visual, motor, or memory systems can conduct invasive experiments on animal models that provide direct access to neural mechanisms (Dehaene & Changeux, 2011; Katz et al., 2025), language's uniquely human nature limits comparative approaches of this nature. Human language involves computational capacities fundamentally different from animal communication systems (Chomsky, 2018; Bolhuis et al., 2018)—which would imply that cross-species generalizations are very limited. This restriction to non-invasive human methods constrains the available methodological toolbox in neurolinguistics (Fedorenko et al., 2010; Poeppel, 2012) and makes it even more important to judiciously select which experiments to pursue with limited resources. Without clear criteria for pursuitworthiness, resources might be allocated inefficiently, and this could lead to favoring experiments of lesser scientific importance over those poised to make more significant contributions (Szucs & Ioannidis, 2017). In what follows, we outline a framework for evaluating pursuitworthiness in neurolinguistics, distinguishing between methodological and pragmatic criteria that reflect both the epistemic and practical dimensions of experimental work.

2.1 Methodological and Pragmatic Criteria

We propose that pursuitworthiness criteria in neurolinguistics can be divided into two main categories: methodological and pragmatic. This distinction builds upon but differs from previous categorizations in the philosophy of science literature. Our proposed methodological/pragmatic distinction broadly aligns with the influential differentiation articulated by McMullin, between epistemic values—those "presumed to promote the truth-like character of science" (McMullin, 1982, p. 19)—and non-epistemic values, which encompass a wide range of values rooted in non-truth oriented human goals. Our bifurcation in methodological and pragmatic criteria shares elements with this influential framework[2] but is tailored specifically to the neurolinguistic experimental context.

The main desiderata for neurolinguistic experiments include scientific rigor and reliability of how an experiment is designed. Accordingly, relevant methodological criteria in this field are identified with reference to their capacity to address the following questions: Is the methodology robust? Is the statistical power adequate? Are the conditions properly controlled? Can we expect interpretable and meaningful results? These standards are drawn from established practices in cognitive neuroscience and linguistics (e.g., Luck, 2014; Kemmerer, 2015; Poldrack et al., 2017). At their core, methodological criteria assess whether an experiment can produce trustworthy knowledge, that is, how well the design can yield evidence that addresses the research questions, avoiding misleading factors (confounds) and statistical issues. Because neurolinguistics investigates complex systems, rigorous methodology is required to separate real findings from statistical or experimental artifacts (de Graaf & Sack, 2011; Luck & Gaspelin, 2017). One major problem in neuroimaging is that the brain signals being measured are often very weak compared to background noise, like eye, heart, and muscle artifacts (Muthukumaraswamy, 2013).

Pragmatic criteria, in contrast, concern the practical feasibility, resource efficiency, and impact of experiments. These include considerations of cost, time requirements, equipment availability, technical expertise

[2] Although subsequent work has challenged the sharpness of McMullin's distinction (e.g., Longino, 1990; Rooney, 1992; Douglas, 2009), it has become widely recognized. Laymon and Franklin (2022), for example, apply a similar distinction to experimental pursuitworthiness, contrasting epistemic values (e.g., predictive accuracy, theoretical coherence) with non-epistemic considerations (e.g., social impact, funding constraints); Šešelja and colleagues distinguish between epistemic and practical notions of pursuitworthiness, where the latter, unlike the former, comprises both epistemic and practical goals (Šešelja et al., 2012, p. 65).

requirements, participant accessibility, availability of stimuli in the language being investigated[3], and potential applications (Merton, 1973; Kitcher, 2001; Firestein, 2016). Pragmatic criteria address questions such as: Is the experiment representing an efficient use of the available yet limited resources? Are there practical barriers to implementation? What is the potential practical impact of the findings? These criteria acknowledge the reality that scientific research operates under resource constraints and within social contexts that influence research priorities (Longino, 1990; Kitcher, 2001; Biddle, 2013). In short, while methodological criteria help assessing the pursuitworthiness of an experiment based on how, and how well[4], it can answer a given research question, pragmatic criteria consider whether the question is worth answering given practical limitations and potential applications (Elliott & McKaughan, 2014).

2.2 Justification of the two kinds of criteria

The bifurcation into methodological and pragmatic criteria offers several advantages for evaluating pursuitworthiness in neurolinguistics. First, it acknowledges the dual nature of scientific inquiry as both an epistemic enterprise seeking knowledge and a practical activity constrained by resources and conducted within social contexts (Douglas, 2009). This recognition aligns with contemporary trends in the philosophy of science that views scientific practice as simultaneously epistemic and social (Longino, 2002; Lusk et al., 2022; Douglas, 2023). Second, this bifurcation corresponds to the types of considerations that arise in practical decision-making about research priorities (e.g., funding). Methodological considerations focus on the internal scientific merit of experiments, while pragmatic considerations situate experiments within broader contexts of resource allocation and potential applications (Elliott & McKaughan, 2014). Third, the methodological-pragmatic distinction allows for a more nuanced evaluation of experiments than would a single-dimensional assessment. An experiment might score highly on methodological criteria yet poorly on pragmatic criteria (e.g., a methodologically rigorous experiment requiring prohibitively expensive

[3] Occasionally, a language simply lacks the right structures, sufficient vocabulary, or other necessary components for conducting a given experiment. For example, if one wanted to compare the processing of common versus rare English irregular verbs with certain characteristics, this may be impossible because English irregular verbs are almost all frequent words (like *is*, *go*, *do*, etc.). Luck (2014) describes some examples of experiments that would require an unfeasible number of trials. In such cases, the experiment is not so much non-pursuitworthy as it is impossible to carry out.

[4] Discipline-specific best practices and standards are essential but do not exhaust methodological considerations. Broader epistemic virtues that transcend particular fields—such as internal validity and reproducibility—as well as meta-methodological judgments about the appropriateness of different techniques for given research questions must also be considered (see Whitt, 1992; Achinstein, 1993; Šešelja & Straßer, 2014; Nyrup, 2015).

equipment), or *vice versa* (e.g., a practically feasible experiment with significant methodological limitations). By evaluating these dimensions separately, researchers (as well as stakeholders who, for example, determine funding allocation) should be able to make more informed judgments about pursuitworthiness (Douglas, 2000; Steel, 2010).

Finally, this bifurcation aligns with institutional structures that often separate methodological and pragmatic evaluations. Grant review panels typically assess both scientific merit (which corresponds to the methodological criteria) and impact (pragmatic criteria), while journal peer review processes focus primarily on methodological criteria (Lee et al., 2013). The rationale for the methodological *vs.* pragmatic distinction can be further supported by showing how it applies to concrete cases. The following examples illustrate how methodological and pragmatic criteria, respectively, can independently guide judgments about pursuitworthiness in neurolinguistic experiments.

Example 1: Methodological criteria in prediction research

Statistical power can affect the methodological pursuitworthiness of neurolinguistic experiments. For instance, consider the large-scale replication study by Nieuwland and colleagues (2018). This experiment examined whether readers predict the phonological forms of words before seeing them. For example, when reading *The day was windy so the boy went outside to fly*…, one may expect the sentence to continue with …*a kite*. If readers predict not just the upcoming meaning (the concept of "a kite") but also predict the word itself (i.e., *kite*, which starts with a [k]), then they should specifically expect the article *a* instead of *an*, since English words beginning with consonant sounds take the article *a*. In this case, brain responses elicited by the article *a* or *an* should vary as a function of the predictability of the following word. Several previous studies with typical sample sizes (several dozen participants) had found conflicting results with respect to whether or not this predictability effect on articles really occurs.

Nieuwland and colleagues (2018) tests this question in a nine-laboratory experiment involving 334 participants. Some of their statistical analyses suggested that there likely *is* a small effect of this nature, but that it was so small it did not reach the traditional threshold of statistical significance even with over 300 participants, an order of magnitude higher than the sample sizes of typical experiments. This finding shows that truly massive sample sizes may be required to reliably detect this effect. Therefore, small-sample ERP studies of this sort of form prediction—at least operationalized through this effect—might not be *methodologically worthwhile* to pursue, as they might have little to no chance to meaningfully measure the effect they intend to investigate (see, e.g., Gelman, 2015, 2017) unless they

substantially increase statistical power either through larger sample sizes or through design changes that increase the size of the effect and/or reduce the variability around it.

Methodological pursuitworthiness criteria in this context include ensuring adequate trial numbers (typically >30 per condition), appropriate participant sample sizes based on power analyses, proper control of linguistic stimuli for confounding variables (e.g., word frequency and cloze probability[5]), robust artifact rejection procedures (Kutas & Federmeier, 2011; Luck, 2014), and proper experiment design and analysis for isolating components of brain activity (Luck, 2014). Experiments meeting these criteria are more methodologically worthy of pursuit because they are more likely to produce results that are both reliable (i.e., can be replicated across labs) and interpretable, thereby advancing our understanding of language processing (Luck & Gaspelin, 2017). Experiments that do not meet these criteria may fail to produce any knowledge. As an extreme example, if a researcher conducts an fMRI experiment but scans participants' shoes instead of their heads, then this experiment will presumably not produce any knowledge about brain function. Similarly, if an experiment has insufficient power or has design or analysis problems that preclude it from accurately measuring brain responses elicited by the linguistic phenomenon of interest, it may also fail to produce any knowledge.

Example 2: Pragmatic criteria in sentence processing

Sentence processing experiments can often face challenges that are not methodological in nature but rather pragmatic. First, creating well-controlled stimulus sets that isolate specific syntactic or semantic features while controlling for confounding variables (word frequency, sentence length, plausibility) requires significant time investment (Kaan, 2007; Kuperberg, 2007). Simply creating stimuli for an experiment can take multiple rounds of pre-tests with additional samples of human participants, in order to norm or rate stimuli on various psycholinguistic dimensions—which is often necessary either to select appropriate stimuli for an experiment, or to measure properties of the stimuli which will eventually be used as independent variables in statistical analysis. Second, these experiments often require equipment like fMRI, EEG, or MEG, which vary in availability and cost across research institutions (Hagoort, 2019). Third, the nature of sentences makes experiments typically requiring very long sessions to obtain

[5] Word frequency is how commonly a word appears in a language, typically quantified based on its rate of occurrence in large text databases/corpora. Cloze probability is how predictable a word is in a specific sentence context. It is measured as the proportion of people who complete a given sentence fragment with that particular word. For example, in "He stirred his coffee with a...", the word "spoon" has a 'high' cloze probability.

sufficient trials across multiple conditions, which may negatively impact a participant's psychological well-being (Jap et al., 2025a). These practical challenges and limitations do not necessarily undermine the scientific rigor of an experiment, but they do influence judgments about whether an experiment is worth pursuing within a given (and often complex) context.

Furthermore, experiments differ in terms of what potential impacts they may have, and how those impacts may take place. For example, studies examining neural markers of syntactic processing in aphasia patients (Thompson et al., 2010; Mack et al., 2013) may have direct clinical applications for assessment and rehabilitation. Meanwhile, basic research investigating garden-path (Novick et al., 2005) or thematic role assignment (Jap et al., 2025b; Jap & Hsu, 2025) might have less immediate clinical relevance but could inform educational practices or computational models of language.

Resource allocation considerations also affect pragmatic pursuitworthiness. Sentence processing experiments using affordable equipment (e.g., behavioral measures like sentence-picture matching [Jap et al., 2016] or EEG [Jap et al., 2024]) may be *more pragmatically worthy* of pursuit in resource-limited contexts (e.g. laboratories in developing areas of the world) than those requiring more expensive technologies (e.g., MEG or fMRI), even when both approaches are methodologically sound (Fedorenko et al., 2010). This pragmatic evaluation shifts, however, if the research question specifically requires the spatial resolution of fMRI. These examples help clarify the proposed distinction between methodological and pragmatic criteria in evaluating the pursuitworthiness of neurolinguistic experiments. Broadly speaking, methodological criteria concern an experiment's capacity to yield reliable knowledge, while pragmatic criteria reflect contextual factors such as resource constraints and potential applications.

Before turning to a more detailed elaboration of these criteria, it is worth addressing potential objections to the distinction itself. It might be argued that the two dimensions are often intertwined—for example, a methodological choice (e.g., using EEG vs. fMRI) can have pragmatic implications (e.g., cost, accessibility), and vice versa. Furthermore, it might be objected that the classification of a concern as either methodological or pragmatic may depend on the context; sample size, for instance, could be seen as a methodological issue affecting statistical power or a pragmatic one affecting time and feasibility. Yet this overlap does not undermine the distinction; rather, it highlights its value as an analytical tool. By differentiating between the two dimensions of pursuitworthiness, the distinction proves to be heuristically useful in guiding research decisions.

3. Methodological Factors

The bifurcation of pursuitworthiness criteria distinguishes between methodological considerations—addressing the scientific rigor and epistemic reliability of experimental designs—and pragmatic considerations related to feasibility and resource efficiency. This section elaborates on methodological factors, with particular attention to Event-Related Potential (ERP) research as a case study, exploring how methodological rigor constitutes (what we deem) a *non-negotiable* criterion for experimental pursuitworthiness and examining current methodological challenges in neurolinguistic research.

3.1 Methodological criteria for ERP research

ERP research represents a cornerstone methodology in neurolinguistics. ERP is commonly used to provide temporally precise measurements of neural activity associated with language processing (Kutas & Federmeier, 2011). The methodological quality of ERP experiments—their capacity to generate reliable, interpretable, and valid results—depends on adherence to established standards and best practices across multiple dimensions of experimental design, data collection, analysis, and appropriate sample size (Picton et al., 2000; Luck, 2014).

At the experimental design level, several parameters determine methodological rigor. First, stimulus design must *isolate* the linguistic variable of interest while controlling for the various potential confounding factors which include (but are not limited to): word frequency, semantic relatedness, and syntactic complexity (Kutas et al., 2011). Second, experiments must include sufficient trials per condition to achieve adequate signal-to-noise ratio (Luck, 2014). Third, appropriate counterbalancing and randomization procedures must be implemented to mitigate order effects and strategic participant responses (Picton et al., 2000).

Data collection standards similarly affect methodological quality. Proper electrode placement, reference selection, and impedance maintenance are prerequisites for valid data (Kappenman & Luck, 2016). Recording parameters such as sampling rate and filter settings must be selected to prevent signal distortion or aliasing (Widmann, et al., 2015). Experimental protocols must minimize participant fatigue and consider the participants' attention because these factors can significantly impact the data quality and ERP components of interest (Woodman, 2010).

Analysis procedures make up the third dimension of methodological rigor. Preprocessing pipelines must include appropriate filtering, artifact detection and rejection/correction, segmentation, and baseline correction (Luck, 2014; Keil et al., 2014). Ideally, these procedures *should* be determined prior to data analysis to prevent data-dependent decisions that inflate Type I error[6] (Simmons, et al., 2011). Statistical analyses should account for multiple comparisons across time points and electrodes (Groppe et al., 2011; Luck & Gaspelin, 2017). This issue is especially important in neuroscience experiments because the data tend to be multidimensional—a difference between brain responses to two different linguistic stimuli may occur over one of several dozen scalp sites or brain regions, and at one of several thousand points in time, and may or may not interact with other factors of interest. Conducting several thousand statistical tests almost guarantees that some spurious significant differences will be found, even when there is no plausible difference in reality (Bennett et al., 2009). Therefore, statistical control for multiple comparisons is crucial for ensuring the epistemic adequacy of an experiment. Finally, interpretations of the data must also distinguish between statistical significance and effect size magnitude (Sullivan & Feinn, 2012).

Methodologically rigorous ERP experiments in neurolinguistics also require appropriate sample sizes based on a priori power analyses (Button et al., 2013; Larson & Carbine, 2017). While earlier studies often featured small samples (e.g., fewer than twenty participants), current understanding of statistical power and effect sizes discourages this practice, particularly for subtle effects (effects that are generated by smaller or more subtle differences in stimuli) or smaller ERP components (Boudewyn et al., 2018; Nieuwland et al., 2018). Beyond technical and analytical concerns, experiments may be limited by their theoretical and design assumptions. An experiment that aims to use some particular pattern of brain activity to measure something that it is simply incapable of measuring may be doomed to fail. Poeppel and Embick (2005) describe the "ontological incommensurability problem", by which they mean the fact that linguistic primitives (like sentences, phonemes, etc.) are not the primitives that the brain uses, and thus any experiments attempting to identify brain responses to those sorts of things are unlikely to meaningfully contribute to knowledge about the neuroscience of language (see also Poeppel, 2012). This means that experiments motivated by and couched in an inaccurate understanding of what certain components of brain activity represent may fail to meet the methodological adequacy criterion for pursuitworthiness.

[6] Type I error occurs in statistical hypothesis testing when a true null hypothesis is incorrectly rejected. In simpler terms, it is concluding there is an effect or relationship when, in reality, there is none (often referred to as a "false positive"). Type II error is when a false null hypothesis is incorrectly not rejected. In simpler terms, it is failing to detect an effect or relationship that actually exists ("false negative").

3.2 Methodological rigor as a primary criterion for pursuitworthiness

Methodological rigor serves as a primary criterion for experimental pursuitworthiness in neurolinguistics for a fundamental reason: it determines the epistemic value of experimental outcomes. This operates on two related levels. At a basic level, methodologically compromised experiments produce questionable, potentially misleading results that fail to advance scientific understanding despite consuming resources (Szucs & Ioannidis, 2017). As Curran (2004) stated, methodological decisions in ERP research directly influence the probability of both Type I and Type II errors, ultimately determining whether reported effects reflect genuine neurocognitive processes or experimental artifacts. Beyond merely avoiding false results, methodological rigor also enhances the positive epistemic contribution of experiments. Studies adhering to rigorous standards generate knowledge with higher evidential value—results that are more likely to be replicated, contribute to cumulative knowledge, and provide trustworthy foundations for both theoretical development and subsequent research (Open Science Collaboration, 2015; Munafò et al., 2017). The epistemic quality of experimental outcomes directly relates to whether an experiment merits pursuit given limited resources (Laymon & Franklin, 2022).

Second, methodological rigor relates to the 'traveling capacity' of scientific knowledge—its ability to inform and constrain future research across contexts. Methodologically sound experiments produce results that can be meaningfully integrated with existing literature, compared across studies, and incorporated into meta-analyses (Maumet et al., 2016; Poldrack et al., 2017). By contrast, methodologically questionable experiments generate "orphaned" findings that remain isolated from the broader scientific discourse. The replication crisis in psychology and neuroscience has highlighted how methodological shortcomings undermine scientific progress (Button et al., 2013; Open Science Collaboration, 2015). Failed replications are often taken as indicators of underlying methodological flaws, such as weaknesses in experimental design that may be fatal to the credibility of the original findings. This concern reflects Karl Popper's view that "non-reproducible single occurrences are of no significance to science" (Popper, 1959, p. 86). While reproducibility is arguably an overarching value for research—given its sensitivity to environmental variances (Leonelli, 2018)—in most cases, irreproducible findings can consume resources that could otherwise support more promising research avenues, creating scientific 'dead ends' that divert attention from more productive questions (Spellman, 2015). In neurolinguistics specifically, where experiments often require substantial resource investments, pursuing methodologically questionable experiments imposes opportunity costs on the field as a whole (Szucs & Ioannidis, 2017). Incidentally, it is worth noting that the methodological criteria emphasized above

(and to be discussed in the forthcoming paragraphs) include both explicit discipline-specific norms and standards shared within the research community, as well as implicit rules that are, to varying degrees, recognized and accepted.

Methodological rigor represents not merely an epistemic ideal but a concrete prerequisite for experimental pursuitworthiness. An experiment lacking methodological rigor cannot generate the knowledge it purports to seek, regardless of its conceptual merits or theoretical promise. As such, methodological standards serve as necessary (though perhaps not sufficient) conditions for experimental pursuitworthiness in neurolinguistics.

3.3 Current methodological challenges in neurolinguistic research

Despite widespread recognition of methodological best practices, neurolinguistic research—particularly ERP studies—continues to face significant methodological challenges that impact pursuitworthiness judgments. A systematic review by Larson and Carbine (2017) examining 100 human electrophysiology papers revealed concerning patterns: no studies (0%) reported sample size calculations, and while 77% utilized repeated-measures designs, none provided the necessary variance and correlation information required for accurate power calculations in future studies. Only 40% reported effect sizes, and less than 60% reported basic descriptive statistics such as means and measures of variability. This opacity not only undermines the evidential value of published research but also complicates decisions about which experimental avenues merit pursuit.

3.3.1 Statistical issues and inferential practices

Statistical practices in neurolinguistic research often fail to address the complex inferential challenges inherent to high-dimensional ERP data. Luck and Gaspelin (2017) identify several problematic practices that inflate false positive rates in ERP research, including: (1) selective analysis of subsets of electrodes or time windows identified post-hoc from data visualization; (2) flexible definitions of measurement windows that maximize statistical significance; (3) inconsistent application of correction for multiple comparisons; and (4) publication bias favoring results that include statistically significant differences over those that do not. These practices represent variations of what Simmons et al. (2011) term "researcher degrees of freedom"—undisclosed analytical flexibility that increases the likelihood of finding statistically significant results irrespective of underlying effects. In ERP research specifically, the high dimensionality of data (multiple electrodes × multiple time points × multiple conditions) creates numerous

opportunities for such flexibility—that is, multiple ways to potentially analyze the data - particularly when analysis decisions are not preregistered (Nosek et al., 2018). For example, without a pre-specified plan, researchers might choose to focus statistical tests only on a subset of electrodes or within a specific time window after observing where effects appear largest in the data, rather than relying on *a priori* hypotheses or standard practices.

The consequences of these statistical issues extend beyond individual studies to shape the pursuitworthiness landscape in neurolinguistics. Experimental paradigms appearing productive based on published literature may actually represent statistical artifacts rather than genuine neurocognitive phenomena (Ioannidis, 2005). This distortion undermines rational resource allocation by diverting attention and funding toward experimental avenues with limited scientific value—precisely the outcome that pursuitworthiness criteria aim to prevent.

3.3.2 Power considerations and effect size interpretation

Insufficient statistical power represents a second critical methodological challenge in neurolinguistic research. Button et al. (2013) demonstrate that low-powered studies not only have reduced chances of detecting true effects but also produce inflated effect size estimates when significant results are found. In neurolinguistics, where many effects of interest are subtle compared to measurement noise, low power creates a particularly pernicious combination of missed effects and exaggerated effect sizes in the published literature. Nieuwland and colleagues' (2018) large-scale replication study, discussed above, illustrates that some ERP effects of theoretical interest may be so small that they can only be accurately measured with massive sample sizes.

The power problem directly impacts pursuitworthiness judgments in two ways. First, small-sample experiments in areas characterized by subtle effects may simply not merit pursuit because they cannot reasonably be expected to provide informative results (even if the underlying theory is correct). Second, the established pattern of power limitations necessitates resource-intensive larger-sample studies to address questions that initially appeared answerable through smaller-scale research (Boudewyn et al., 2018). This recalibration alters the cost-benefit calculus for experimental pursuitworthiness across the field (Yarkoni, 2009). This problem compounds itself because low-powered studies overestimate effect sizes. When a study has small power, it can only find significant differences when those 'effects' are very large. This means that studies may zero in on noise (i.e., spurious effects) that is large in magnitude, while missing real but smaller-magnitude effects. A result of this is that a literature can become rife with reports of large effects (because the studies were only capable of finding large effects), missing small effects, leading

to an inaccurate view of how large or small certain ERP effects really are, and hindering the ability of future researchers to correctly estimate the sample sizes they will need in their own experiments to achieve sufficient statistical power.

3.3.3 The case for interval estimates

A third methodological issue is the overreliance on Null Hypothesis Significance Testing (NHST) at the expense of interval estimation and effect size reporting. Significant tests are statements about whether some effect (usually a difference between brain responses elicited by two kinds of stimuli) is *statistically significant*, whereas interval estimates are statements about how large that difference probably is. This practice often leads to excessive focus on whether a p-value crosses an arbitrary threshold (e.g., $p < .05$) to declare a result 'significant', while neglecting information provided by interval estimation (e.g., confidence intervals), which indicates the precision of the findings, and effect size reporting, which quantifies the magnitude or practical importance of an effect. Cumming (2014) and Wasserstein and Lazar (2016) argue that point estimates accompanied by confidence intervals provide more informative evidence than dichotomous significance testing. In neurolinguistics, where effect sizes vary substantially across paradigms and populations, confidence intervals around effect estimates provide important information for evaluating both theoretical implications and practical significance (Sullivan & Feinn, 2012).

Interval estimates are difficult to provide, however, in some of the statistical methods commonly used for ERP research. For example, the cluster-based permutation method (Maris & Oostenveld, 2007), which is a valuable technique because it addresses the multiple comparison problem discussed above, relies on dichotomizing ERP data into bins of datapoints which exceed a certain significance criterion and bins of datapoints that do not; there is no way to perform interval estimation within this approach. Interval estimation also facilitates cumulative knowledge building by making these studies robust enough to be included in meta-analysis (Cumming, 2012). This cumulative perspective aligns with the 'epistemic landscape' view of pursuitworthiness—which evaluate experiments not in isolation but as contributions to broader epistemic projects. In this context, experiments providing precise interval estimates may merit pursuit even when they cannot reject null hypotheses in isolation.

The dominance of NHST over interval estimation has implications for pursuitworthiness judgments, as it places the focus on *whether* there is an effect (significance), rather than *how big* the effect is (magnitude). This can be

misleading in small-sample studies: significant p-values in small-sample studies may suggest promising research avenues that later prove disappointing when larger studies reveal the true (smaller) effect sizes (Gelman & Carlin, 2014). On the other hand, nonsignificant findings may prematurely terminate pursuit of genuinely promising experimental directions when point estimates and confidence intervals would reveal meaningful effects that require larger samples to achieve statistical significance (Maxwell et al., 2015).

The balance of value between significance testing and interval estimation also differs between different fields and research questions. For example, in clinical contexts, interval estimates of effect size are important: if a clinical intervention is found to 'significantly' improve patient outcomes but only by a tiny amount, then this intervention might not be clinically meaningful or not worth pursuing as a treatment option. On the other hand, in basic science research, the mere existence of an effect may be important for theory-building, and thus may matter more than the size of the effect; in such cases, interval estimation is important for researchers to decide how confident they should be that an effect exists, but the ultimate question of interest is whether the effect exists rather than how large it is.

3.4 Toward methodologically justified pursuitworthiness judgments

The methodological factors examined above suggest several principles for making pursuitworthiness judgments in neurolinguistic research. First, experiments adhering to established methodological standards merit higher priority than those with methodological limitations, regardless of their theoretical attractiveness or novelty (Munafò et al., 2017). Second, sample size and power considerations should feature prominently in pursuitworthiness judgments, with preference given to adequately powered studies capable of detecting effects of theoretical interest (Button et al., 2013; Nieuwland et al., 2018). Third, experimental designs facilitating precise interval estimation and effect size interpretation deserve priority over those yielding only dichotomous significance judgments (Cumming, 2014). These principles are in line with the current emphasis on methodological rigor in science more broadly, as reflected in initiatives like preregistration, registered reports, and open science practices (Chambers, 2013; Nosek et al., 2018). Such methodological reforms address not just how research is conducted but also which research merits pursuit in the first place.

Finally, while we have focused here on methodological issues common to all neurolinguistic research (such as statistical power), we note that methodological pursuitworthiness also depends on adherence to technique-specific methodological best practices—for example, appropriate methods for component isolation in ERP research (Luck,

2014), appropriate methods for head movement correction in fMRI research, etc. In conclusion, methodological factors constitute essential, non-negotiable criteria for evaluating experimental pursuitworthiness in neurolinguistics. Through the case study of ERP research, we have identified specific methodological standards and challenges that inform judgments about which experimental avenues may be more or less worthy of pursuit, especially under resource constraints. These criteria do not constitute rigid requirements, but they can support a more structured and graded assessment of pursuitworthiness—one that accommodates degrees of methodological robustness.

4. Practical constraints as pursuitworthiness criteria

The factors put forward here interact with the methodological considerations discussed in Section 3. An experiment may be methodologically sound yet pragmatically impractical, or vice versa. The tension between methodological rigor and practical constraints requires pursuitworthiness judgments that weigh both dimensions rather than applying rigid criteria (Elliott & McKaughan, 2014). Nevertheless, methodological factors represent non-negotiable constraints on experimental pursuitworthiness in neurolinguistics—experiments failing to meet basic methodological standards cannot justify resource allocation regardless of their pragmatic merits. While methodological factors represent necessary conditions for experimental pursuitworthiness, pragmatic considerations constitute equally important dimensions of the pursuitworthiness calculus. This section examines how practical constraints—specifically resource availability limitations and experimental design challenges—function as important components of pursuitworthiness judgments in neurolinguistic research.

4.1 Resource availability constraints

Resource-intensive neuroimaging technologies create limitations on who can conduct neurolinguistic research and which questions can be feasibly investigated. There are three dimensions of resource availability constraints: equipment costs, expertise requirements, and funding limitations. Neuroimaging equipment requires a large financial investment that impacts pursuitworthiness considerations (Poldrack, 2012). De Haas (2018) reported that MRI machine booking costs approximately €150 per hour in Germany—and in some locations can be an order of magnitude higher. Given the high cost per participant and the large number of participants sometimes necessary (see section 3.3.2 above), some experiments may be methodologically adequate but simply too expensive to run. These

high costs create a hierarchical structure of pursuitworthiness across methods, with behavioral studies (requiring minimal equipment) being most accessible, electrophysiological methods in the intermediate position, and fMRI and MEG representing the most resource-intensive methodologies (Bishop, 2013).[7]

Beyond equipment costs, neuroimaging methodologies impose knowledge barriers. Clayson et al. (2019) document the high level of technical expertise required for ERP data processing and analysis, while Gorgolewski et al. (2016) identify even steeper technical requirements for fMRI. These expertise requirements create entry barriers, significant training time costs, and collaborations that may introduce coordination challenges (Bishop, 2013). The technical complexity of neuroimaging data analysis also introduces quality control concerns, as inadequate methodological reporting often reflects insufficient technical expertise (Clayson et al., 2019).

Financial constraints are the third dimension of resource availability. Researchers often make compromises in study design due to cost factors; for example, in survey research, cost pressures frequently lead to the adoption of non-probability sampling methods, which, while more affordable, can introduce biases and limit the generalizability (Baker et al., 2013). Hayasaka et al. (2007) provide a quantitative framework for cost-benefit calculations in neuroimaging sample sizes. These financial constraints can create systemic biases in neurolinguistic research, which may potentially discourage certain research questions—particularly those requiring larger samples or longitudinal designs (Poldrack, 2012).

4.2 Experimental design challenges

The feasibility of experimental protocols constitutes another constraint on pursuitworthiness in neurolinguistic research. There are three dimensions of experimental design challenges: participant recruitment, protocol feasibility, and cross-institutional disparities.

[7] The nature of the costs can also differ between methods. For example, fMRI researchers often have to pay a cost for every hour of scan time. MEG, on the other hand, has a regular cost to keep the facilities running (as MEG sensors require super-cooled liquid helium that needs to be regularly replaced), but do not always require additional costs per each run. EEG costs less to maintain and often has little to no costs for each run, but each run does use up some (relatively cheap) consumable supplies, like conductive gel, supplies for cleaning the EEG cap after the experiment, and (in some laboratories) disposable combs and disposable gel applicators.

Participant recruitment represents the first experimental design challenge. Button et al. (2013) demonstrate how underpowered studies waste resources and produce unreliable results—connecting recruitment challenges to both methodological rigor and resource efficiency. The severity of this constraint is evident in empirical assessments, with Szűcs and Ioannidis (2020) reporting that the median sample size for experimental fMRI studies was only 12 participants—although this may also be limited due to cost/resources. These recruitment challenges vary systematically across populations (e.g. some clinical conditions are very rare, thus making this group a very difficult sample to recruit) and research questions (Clayson et al., 2019).

The second experimental design issue would be protocol feasibility. Luck (2014) documents practical limitations on trial numbers and participant fatigue in ERP research that necessitate deliberate protocol design to maintain data quality. Jap et al. (2025a) demonstrate how prolonged experimental sessions reduce positive affect among participants, which raises ethical concerns about participant welfare. Protocol feasibility creates trade-offs between competing methodological ideals, as Chen et al. (2022) show that trial sample size has nearly the same impact on statistical efficiency as subject sample size in some contexts. These feasibility issues vary across components and paradigms, with some requiring more 'intensive' protocols than others to produce methodologically sound results (Jensen & MacDonald, 2023).

ERP research in particular requires large numbers of stimuli in order to adequately separate the signal (the brain response of interest) from the noise (everything else that affects the data recorded from the scalp) (Luck, 2014). For example, research on sentence processing needs to include several dozen sentences of each type tested—if an experiment will compare brain responses elicited by grammatical sentences and ungrammatical sentences, a participant may need to read, say, 60 grammatical sentences and 60 more ungrammatical ones. Many such experiments need to include multiple types of sentences to rule out confounding variables or test nuanced hypotheses; it is not uncommon for experiments to use 2×2 designs (with four types of sentences) or 2×2×2 designs (with eight types of sentences), meaning that each participant reads or listens to hundreds of sentences in an experiment. This can cause challenges for trying to investigate phenomena that might only happen in some sentences. For example, a growing area of neurolinguistics seeks to examine brain responses related to pragmatic phenomena—essentially, how individuals figure out nonliteral interpretations of sentences beyond their literal semantic meanings (Politzer-Ahles, 2020). If the realization of nonliteral meaning requires effortful deployment of some cognitive and inferential resources, it is not outlandish to imagine that readers might only need to do this the first one or two times they read such sentences. After that, if they read 300 more sentences with similar nonliteral interpretations, they might not have to work so hard.

This means that ERP experiments, which rely on presenting individuals with huge numbers of sentences or other stimuli, may struggle to detect brain activity that only occurs on the first few sentences those individuals read.

Finally, resource availability varies substantially across research institutions and countries, creating systematic disparities in the capacity to pursue certain neurolinguistic questions, and affecting what questions can be pursued in different institutional contexts. These disparities, manifesting in both infrastructure differences and expertise availability, result in what might be described as the 'geographical contingency' of pursuitworthiness, introducing knowledge-based constraints that operate independently of equipment or funding availability (Harris, 2022).

4.3 Balancing methodological ideals against pragmatic realities

The tension between methodological ideals and practical constraints creates what might be termed the "pursuitworthiness optimization problem" in neurolinguistics. Simmons et al. (2011) show that limited resources may lead to questionable research practices, which leads to the need for pursuitworthiness judgments that realistically reflect practical limitations rather than strictly adhering to methodological ideals beyond available means. The issue of sample size and statistical power is a good example of this tension. Button et al. (2013) emphasize the value of sufficient statistical power, yet achieving these standards is often difficult in neurolinguistic research due to recruitment and resource challenges. This creates a pursuitworthiness dilemma: should researchers pursue imperfectly powered studies that may yield preliminary evidence, or restrict research to questions answerable with adequate power given available resources?

Rather than applying universal criteria, pursuitworthiness judgments require calibration to specific research environments. An experiment that merits pursuit in a well-resourced environment may fail to meet pursuitworthiness thresholds in resource-limited contexts, despite *identical* methodological merits. This calibration addresses the problem of 'implementation gap' in scientific planning—the disconnect between theoretical research priorities and practical feasibility. Practical constraints should not only be seen as limiting factors but also as opportunities for innovation that can improve pursuitworthiness. Nikolaidis et al. (2023) discussed how implementing open science practices can improve inclusiveness and collaboration while increasing research efficiency, thereby broadening the range of feasible research questions. Moreover, these practical constrains cannot be easily separable from the epistemic goals of an experiment; for example, technical realizability, despite its pragmatic character, often directly interacts with methodological concerns (Nickels 2006; Šešelja et al., 2012). By considering both methodological standards and

practical realities, such integrated judgments offer a realistic framework for distributing limited resources among competing priorities in neurolinguistics.

5. Luxury vs. urgent science and pursuitworthiness judgments in neurolinguistics

In the previous sections, we suggested that pursuitworthiness judgments in neurolinguistics experiments should integrate both methodological and pragmatic aspects, with methodological criteria—unlike pragmatic ones—being non-negotiable. This hierarchy reflects both the variability of pragmatic constraints, shaped by the specific context, and the foundational role of methodological rigor in safeguarding scientific integrity. Without minimal methodological adequacy, an experiment cannot generate reliable data or contribute meaningfully to the field, regardless of its feasibility. By contrast, pragmatic considerations (e.g., cost, technological feasibility, infrastructure) can be assessed only in light of the particular context in which a study is conducted. Although the interplay between methodological and pragmatic criteria is hard to pin down in abstraction from specific experimental contexts, it remains important to consider how they might interact in practice. In the following, we outline a provisional framework that can help guide decisions, especially regarding the role of pragmatic criteria in promoting more principled forms of experimental prioritization.

5.1 Mapping the Interplay

A helpful starting point for examining the interplay between methodological and pragmatic criteria is offered by a scenario proposed by Jamie Shaw (2022), in which we are asked to imagine the case in which science is driven solely by epistemic goals. In such a scenario, he argues, there would be no purely epistemic reason to prioritize one line of scientific inquiry over another, since all would, in principle, contribute equally to the expansion of knowledge. By analogy, we can imagine the case in which pursuitworthiness judgments for experiments in neurolinguistics are guided exclusively by methodological considerations. In such an ideal scenario, as long as a set of experiments equally comply with the same methodological criteria, there would be no purely methodological grounds for ranking them in terms of priority[8]. Methodological criteria would function primarily as a threshold of adequacy—they can disqualify

[8] This scenario, by virtue of being ideal, assumes that all experiments equally comply with the methodological criteria. Outside of this ideal scenario, however, experiments can be comparatively assessed based on the extent to which they exemplify the relevant features previously discussed (e.g., analysis procedures, pool size, etc.; see Section 3).

poorly designed or unreliable studies, but they cannot, on their own, determine which among the remaining candidates ought to be pursued first or with greater intensity. Shaw's scenario reveals an important asymmetry: methodological criteria, while essential for securing the epistemic integrity of an experiment, are comparatively weak tools for making finer-grained decisions about resource allocation. Once the threshold of methodological adequacy is crossed, decisions about pursuitworthiness—such as which experiment to prioritize, when, and with what level of support—depend on pragmatic considerations, including cost, feasibility, time constraints, expected impact, and broader institutional or societal priorities.

In this light, we attempt to clarify the interplay between methodological and pragmatic criteria. First, methodological criteria set the floor by providing the conditions under which an experiment counts as (more or less) epistemically viable, and thus worth pursuing to varying degrees. Given the methodological complexities of neurolinguistic experiments, these criteria may also guide comparative assessments of their methodological strengths (see, Section 3). While methodological criteria filter the space of possibilities of pursuit, pragmatic criteria structure comparative judgments by guiding selection within that space. One way to explore this selection mechanism is by critically applying Shaw's luxury vs. urgent science distinction, developed in the debate on scientific pursuitworthiness.

Let us return to Shaw's scenario of a science governed exclusively by epistemic goals. From this it follows that decisions about *when* a discovery should be made are independent of epistemic values and must instead rely on non-epistemic ones—especially the value of knowing "*when* that knowledge can be put into action" (Shaw, 2022, p. 108). Timing matters, for example, in curing diseases, addressing climate change, or advancing social justice. On this basis, Shaw argues that (short-term) pursuitworthiness criteria are necessary only for what he calls *urgent science*, driven by practical or moral reasons to obtain results within a set timeframe[9], while *luxury science* lacks such urgency and is therefore free from such constrains. Although our proposal departs from Shaw's in several respects[10], his distinction remains philosophically valuable, offering a useful perspective on cases where pragmatic criteria are especially salient. It also invites us to ask whether certain experiments in neurolinguistics could plausibly be treated as *urgent*[11], and

[9] Timeframe is the core of Shaw's urgent/luxury distinction: "A research proposal is urgent if there is a practical or moral reason to need a result within a specified timeline and the research can realistically be carried out within that timeline. […] is luxurious iff it has no expected timeline for returning particular results" (Shaw, 2022, p. 106).

[10] Unlike Shaw, our focus is on the pursuitworthiness of experiments. More importantly, we do not treat the urgent/luxury distinction as setting the boundaries within which pursuitworthiness criteria apply; rather, we assume that some form of evaluation is required in *all* experimental contexts.

[11] Note that Shaw's notion of *urgent science*, though related to Stegenga's (2024) concept of *fast science*, should not be conflated with it. Both are grounded in the idea that practical or moral reasons demand timely results, but Stegenga's 'fast

whether urgency so defined might guide prioritization among methodologically sound studies. If workable in this field, the distinction would add an axis along which pragmatic factors are structured, refining the temporal and comparative dimensions of pursuitworthiness. At the same time, however, its application to neurolinguistic experimentation reveals important limitations: while it highlights the societal relevance of research and supports prioritization when resources are scarce, many neurolinguistic studies not originally framed as urgent have later yielded unanticipated but significant clinical applications, making urgency difficult to assess prospectively. In what follows, we present paradigmatic cases of research initially motivated by theoretical curiosity that ultimately proved instrumental for clinical diagnostics and intervention—cases that caution against using the luxury/urgent distinction as a strict prioritization framework.

5.2 Historical cases challenging the luxury/urgent distinction

The history of neurolinguistics provides numerous examples that challenge the luxury/urgent dichotomy. Research initially pursued for purely theoretical interests—seemingly "luxury" by Shaw's definition—has repeatedly yielded unexpected clinical and practical applications. One of the most paradigmatic cases is Kutas and Hillyard's (1980) discovery of the N400, a negative deflection elicited by semantic anomalies. Conceived as basic research on language comprehension, the N400 later became a key diagnostic tool for conditions such as aphasia, schizophrenia, and developmental language disorders (Hirano et al., 2020), its clinical value emerging gradually through subsequent studies (Kutas & Federmeier, 2011). A similar trajectory characterizes Osterhout and Holcomb's (1992) identification of the P600—a positive deflection about 600ms after syntactic anomalies—originally aimed at understanding the neural bases of grammatical processing. This work later proved crucial for treating syntactic deficits in aphasia, with its theoretical and methodological contributions informing targeted interventions for agrammatic aphasia (Raymer et al., 2008). Both cases undermine the luxury/urgent distinction, showing that research with "no practical or moral reasons to need a result within a specific timeline" (Shaw, 2022, p. 108) can later generate significant clinical, and arguably practical and moral, value recognized only in retrospect.

A further component that followed this course is the Mismatch Negativity (MMN), a brain response to violations of an auditory pattern (e.g., hearing *pa* in a sequence of *ba*s), even without conscious attention (Näätänen,

science' pertains specifically to scenarios of supreme emergency, where accelerated research is justified by the exceptional stakes involved.

2007). Initially investigated in basic auditory processing research, it has since provided insights into autism (O'Connor, 2012) and shown promise as a biomarker for abnormal brain function in vegetative-state patients (Zarza-Luciáñez, 2007; Schall, 2016). Moreover, the P300, a component with similar functional properties, has been used to develop brain-computer interfaces enabling paralyzed and locked-in patients to communicate without muscle movement (Birbaumer, 2006). Pulvermüller et al.'s (2001) constraint-induced therapy for aphasia provides another clear challenge to the luxury/urgent dichotomy. Now widely used in rehabilitation, this intervention grew out of basic research on cortical lateralization, neural plasticity, and syntax processing—initially without clinical aims—and exemplifies what García et al. (2023) call the "unanticipated clinical potential" of theoretically motivated work.

Taken together, these cases show how research deemed "luxury science" under Shaw's framework has repeatedly acquired unexpected clinical significance, showing that in neurolinguistics the luxury/urgent dichotomy fails to capture the intricate ties between basic research and application (Thompson & Shapiro, 2007).

5.3 The unpredictability problem

The cases above exemplify what may be termed the "unpredictability problem" in research prioritization: the systematic difficulty of predicting which studies will ultimately yield significant practical applications. In neurolinguistics, this unpredictability challenges to the luxury/urgent science framework, which presupposes an ability to assess research by its eventual applications and their capacity to meet (either perceived or factual) salient societal needs. Friedman and Šešelja (2023) analyse this issue through the lens of scientific disagreement and "fast science" pressures, noting that the time lag between basic research and applications can bias decisions against valuable work lacking foreseeable 'fast' payoffs. This temporal dimension of scientific research challenges the categorization underlying the luxury/urgent framework, complicating the application of practical criteria in selecting among epistemically viable possibilities. Raymer et al. (2008) offer a potential response, proposing a "translational continuum" in aphasia research that connects basic neuroscience, cognitive neuropsychology, and clinical application. Their model acknowledges time lags and recognizes that translation occurs across multiple timeframes and pathways, often through unpredictable connections between unrelated areas of investigation.

In neurolinguistics, the luxury/urgent distinction also overlooks forms of urgency which extend beyond immediate societal applications to include other time-sensitive research imperatives. One is the temporal urgency of endangered language documentation: languages on the brink of extinction represent irreplaceable sources of

information about linguistic diversity and cognitive processing (Whalen & McDonough, 2015), even with no immediate application. Another arises from critical developmental windows in language acquisition; some research questions must be pursued within these periods to capture time-sensitive phenomena, regardless of their immediate payoffs (Friederici & Gierhan, 2013). A further form of urgency stems from equity considerations in clinical research: García et al. (2023) highlight the need for cross-linguistic studies to establish reliable neurodegenerative markers for non-English-speaking populations. The gap driven by the dominance of English-language research risks perpetuating clinical inequities for the majority of the world's speakers. This perspective identifies urgency based on equity considerations rather than temporal constraints again providing evidence of the flaws of a simple luxury/urgent dichotomy.

The historical and conceptual considerations surveyed above collectively point to a central insight: while methodological and pragmatic criteria serve distinct roles in pursuitworthiness judgments—filtering vs. selecting—they are not strictly separable in practice. The interplay is often intricate, context-dependent, and temporally extended. In neurolinguistics, the luxury/urgent distinction, though heuristically useful, cannot serve as a definitive guide to prioritization among methodologically sound candidates. Instead, what is needed is a more flexible approach—one that acknowledges the unpredictability problem and resists overconfident prescriptions.

6. Implications and open questions

Our analysis reveals a structured, yet flexible framework grounded in the interplay between methodological and pragmatic criteria. While centred on neurolinguistic experiments, our case study raises broader issues for the philosophical discussion on pursuitworthiness. A first concerns the generalizability of the hierarchical relationship we identified—methodological criteria functioning as thresholds, pragmatic criteria as comparative guides. This hierarchy appears to be broadly applicable towards cognitive sciences, especially subfields that rely on experimental methods. Subfields like cognitive or developmental psychology share with neurolinguistics the reliance on resource-intensive methods (neuroimaging as well as the difficulty in recruitment of atypical populations) where the reliability of findings is critical (Munafò et al., 2017). In these fields, the epistemic cost of methodological failure is universally high; experiments that cannot produce trustworthy results are poor candidates regardless of the pragmatic appeal. The logic that one cannot prioritize what is not first viable seems a general principle of rational resource allocation in science. While the specific *content* of methodological criteria is necessarily domain-specific, their primary gatekeeping *function* should hold across similar domains.

The question then arises: should methodological criteria in pursuitworthiness judgments *always* serve as non-negotiable thresholds for epistemic adequacy, filtering the space of viable experiments, while pragmatic criteria guide comparative prioritization within that viable space? While we have defended this hierarchy, its application requires nuance. The framework is most robustly applied to confirmatory experimental research, where the goal is to test well-defined hypotheses. In other contexts, however, the hierarchy may be more flexible. In highly exploratory research, where the aim is to generate novel hypotheses rather than test existing ones, some methodological standards may be relaxed to foster discovery (Burian, 2007; Steinle, 2002; Mizrahi, 2022). Similarly, experiments focused on methodological innovation—developing a new technique—are pursued precisely because the method's properties are not yet fully understood; one example of this is the 'advent' of functional MRI (fMRI), or the use of BOLD signals as a proxy of neural activity (Kwong et al., 1992). In such cases, pursuitworthiness lies exactly in the potential of the method itself, not in the substantive knowledge generated by its initial applications. These cases show not that methodological criteria lose their gatekeeping role, but that their content shifts: in exploratory or innovation-driven contexts, what counts as 'adequate' is defined in relation to the specific epistemic aims of the experiment.

The relationship between methodological and pragmatic criteria may also shift depending on the urgency of practical applications. For example, in contexts of genuine emergency—such as public health crises[12]—pragmatic considerations may temporarily override some methodological constraints. More generally, the salience of certain pragmatic factors on prioritization is more evident in some fields than in others. In biomedical research, for instance, factors such as disease prevalence and treatment accessibility play a particularly prominent role in guiding which lines of inquiry are pursued. Importantly, however, this should be understood as a calculated *trade-off* reflecting the heightened stakes of such situations[13] rather than an abandonment of methodological standards (Douglas, 2009; Steel, 2010; Stegenga, 2024).

[12] The COVID-19 crisis exemplifies how pragmatic considerations can prevail over methodological ones. This dynamic has been also described within the *Post-Normal Science* framework (Funtowicz & Ravetz, 1993), which characterizes contexts of high uncertainty, urgent decisions, and contested values. In such situations, public communication becomes a key central pragmatic factor, often amplifying pressures to deliver rapid answers (Amoretti & Lalumera, 2023; Dolcini & Valle, 2024), thereby contributing to a *de facto* relaxation of methodological safeguards (Jung et al., 2021).

[13] As Elliot (2017, p. 95) observes, " if one thought that public health should be valued particularly highly, one might call for lowering the statistical significance level for individual studies." Nevertheless, our account aligns with the view that long-term credibility and utility of science hinge on maintaining robust methodological standards, even in areas subject to pressing social demands (Hudson, 2021).

What follows, then, is that methodological thresholds, although not dogmas, remain the necessary ground on which pragmatically informed prioritization can occur. Without them, the risk of privileging expedient but epistemically fragile research increases, especially under the "fast science" pressures described by Friedman and Šešelja (2023). This way of framing the hierarchy also invites comparison with the virtue-economic account of pursuitworthiness proposed by Duerr & Fischer (2025). While their approach does not assign methodological criteria a distinct gatekeeping role, it similarly treats pursuitworthiness at a meta-methodological level. Yet insisting on methodological thresholds does not reduce pragmatic criteria to a secondary role. Once the space of epistemically viable experiments has been established, pragmatic considerations—including, urgency, equity, societal relevance, feasibility—legitimately guide comparative prioritization. Furthermore, the cases in neurolinguistics here discussed, suggest that pragmatic criteria must be expanded beyond immediate societal needs to encompass temporal and normative urgencies, and that context-sensitivity is intrinsic to pursuitworthiness judgments more generally.

Our discussion points to several unresolved questions for the body of work on pursuitworthiness of experiments. First, our analysis of Shaw's luxury/urgent distinction revealed the difficulty of predicting which research will yield valuable applications, but this raises broader questions about how temporal considerations should influence decisions. Should pursuitworthiness evaluations prioritize short-term over long-term potential? How should researchers balance immediate applications against theoretical advances with uncertain timelines? Should ethical values—including considerations related to participant well-being (Jap et al., 2025a)—require a separate ethical evaluation, and perhaps be treated as a third dimension of pursuitworthiness? And to what extent should equity-based urgencies—such as addressing the underrepresentation of non-English languages in neurolinguistic diagnostics (cf. García et al., 2023)—be weighted relative to other pragmatic factors in research prioritization?

In neurolinguistics specifically, several questions remain open. First, the field's heavy reliance on neuroimaging technologies raises questions about epistemic justice: how can pursuitworthiness criteria avoid systematically disadvantaging researchers in under-resourced institutions or developing countries? Additionally, the field's focus on neurotypical populations in well-studied languages creates biases in research priorities. Should pursuitworthiness criteria actively promote diversity in participant populations and linguistic phenomena, even when such studies face greater practical challenges? A pressing challenge is the trade-off between methodological rigor and ecological validity. Highly controlled experiments excel at isolating cognitive processes, but often at the cost of tasks far removed from

natural language use. For example, Barbet and Thierry (2016) used an ERP oddball paradigm to demonstrate that people do not automatically interpret "*some*" as meaning "not all". Participants saw quantity words (e.g., *all*, *two*, *none*) presented with varying mixes of uppercase and lowercase letters, and clicked a button whenever the number of uppercase letters matches the meaning of the word: e.g., they should click a button when they see *ALL*, *tWO*, or *none*, and they should not click a button when they see *aLL*, *tWo*, or *nONE*. Participants were instructed to interpret the word "*some*" pragmatically, i.e., as meaning "at least one, but not all". Nevertheless, the ERP data revealed that when participants saw *SOME*, they still showed neural signs of preparing to respond—indicating that the pragmatic "not all" interpretation was not applied automatically. This experiment thus provided powerful evidence, from a highly controlled manipulation, about an issue that had been challenging to address with other methods (see Politzer-Ahles, 2020, for discussion of the benefits of this approach to ERP research). A limitation of the experiment, however, was that it involved highly unnatural processing: participants watched long series of isolated words, were engaged in a metalinguistic task that is very different from natural language use, and were instructed to interpret the word "*some*" in a particular way. Perhaps even more concerning, experiments like this require large numbers of stimuli (Barbet & Thierry's participants saw 1560 words), but processes like inferencing about nonliteral meaning may occur only on the first few exposures and then be applied heuristically thereafter. Other studies take a more naturalistic approach; for example, several studies have recorded brain signals in participants while they read stories like *The Little Prince* or listen to an audiobook (Li et al., 2022; Stehwien et al., 2020; Zhang et al., 2022). Absent experimental manipulations, it can be difficult to isolate processes of interest and test specific hypotheses in a dataset like this. Recent research, however, has begun to use statistical and computational methods to extract meaning from large datasets like these. These approaches, including EEG decoding and computational investigations of cognitive dynamics like entropy (e.g., Salicchi et al., 2025) and multivariate pattern analysis (Zhang et al., 2022), offer a promising way forward, though their success in balancing experimental control with ecological validity remains a key open question for the field.

The more difficult challenge arises at a higher level, namely how to adjudicate pursuitworthiness across disciplinary domains that operate with different methodological norms. Here our framework aligns with accounts of methodological pluralism (e.g., Longino, 2002). The open problem is how to accommodate such pluralism while retaining evaluative criteria robust enough to support cross-disciplinary assessment. One promising direction is the development of bridging criteria—principles that connect domain-specific thresholds of adequacy to more general standards. Our proposal can be seen as a step in this direction: it structures pluralism without collapsing into relativism. Clarifying how such bridging principles might function across heterogeneous fields marks an important task for future work on the pursuitworthiness of experiments.

7. Concluding Remarks

We proposed a framework for the pursuitworthiness of experiments in neurolinguistics, grounded in a distinction between two key dimensions: methodological and pragmatic. We argued that methodological and pragmatic criteria, although both essential, stand in a relation of internal hierarchy due to the context-sensitivity of pragmatic constraints, and the foundational role of methodological rigor. Methodological criteria primarily serve as non-negotiable thresholds for epistemic adequacy, filtering the space of viable experiments, while pragmatic criteria play a role in guiding comparative prioritization *within* that viable space. As such, our proposal emphasizes parsimony to avoid prescriptive overload. On the one hand, by placing methodological criteria as foundational, we establish a baseline for epistemic adequacy. However, these criteria are not intended to prescribe specific methods or designs; rather, they provide a framework for identifying what makes an experiment viable from a scientific perspective, without mandating a single fixed approach. On the other hand, pragmatic criteria, while acknowledging and addressing constraints of resources, time, and feasibility, remain inherently context-sensitive. The interplay between these two kinds of criteria, though central to the framework, is structurally complex and context-dependent, and a more systematic exploration of this relationship would be desirable. In this paper, we limited the analysis of this interplay to the application of Jamie Shaw's luxury-urgent distinction, aiming to evaluate its potential utility for guiding experimental prioritization. Based on historical experimental cases, we showed that its applicability in neurolinguistics is constrained by what we have termed the "unpredictability problem," which complicates decisions about experiment prioritization. Finally, although the proposed framework is specifically tailored to neurolinguistics, it also offers a model that could inform pursuitworthiness evaluations for experiments in other scientific domains. We recognize that the distinctive features of neurolinguistics may limit direct generalization; different fields, with their own methodologies and practical challenges, may require domain-specific adaptations. Nonetheless, our framework suggests a general strategy for balancing epistemic and practical considerations—one that could prove valuable across neuroscience and other fields facing similar methodological and practical challenges.

References

Achinstein, P. (1993). How to defend a theory without testing it: Niels Bohr and the 'Logic of Pursuit'. *Midwest Studies in Philosophy*, *18*, 90–120. https://doi.org/10.1111/j.1475-4975.1993.tb00259.x

Amoretti M.C. & Lalumera E. (2023). Unveiling the interplay between evidence, values and cognitive biases. The case of the failure of the AstraZeneca COVID-19 vaccine. *Journal of Evaluation in Clinical Practice*, *29*, 1294–1301. https://doi.org/10.1111/jep.13903

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, *1*(2), 90–143. https://doi.org/10.1093/jssam/smt008

Barbet, C., & Thierry, G. (2016). Some alternatives? Event-related potential investigation of literal and pragmatic interpretations of some presented in isolation. *Frontiers in Psychology, 7*, 1479.

Bennett, C., Miller, M., & Wolford, G. (2009). Neural correlates of interspecies perspective taking in the post-mortem Atlantic salmon: an argument for multiple comparisons correction. *NeuroImage (Orlando, Fla.)*, *47*, S125–S125. https://doi.org/10.1016/S1053-8119(09)71202-9

Biddle, J. (2013). State of the field: Transient underdetermination and values in science. *Studies in History and Philosophy of Science Part A*, *44*(1), 124–133. https://doi.org/10.1016/j.shpsa.2012.09.003

Birbaumer, N. (2006). Breaking the silence: Brain–computer interfaces (BCI) for communication and motor control. *Psychophysiology*, *43*(6), 517–532. https://doi.org/10.1111/j.1469-8986.2006.00456.x

Bishop, D. V. M. (2013). Research review: Emanuel Miller Memorial Lecture 2012 – Neuroscientific studies of intervention for language impairment in children: interpretive and methodological problems. *Journal of Child Psychology and Psychiatry*, *54*(3), 247–259. https://doi.org/10.1111/jcpp.12034

Bolhuis, J. J., Beckers, G. J. L., Huybregts, M. A. C., Berwick, R. C., & Everaert, M. B. H. (2018). Meaningful syntactic structure in songbird vocalizations? *PLoS Biology*, *16*(6), e2005157–e2005157. https://doi.org/10.1371/journal.pbio.2005157

Boudewyn, M. A., Luck, S. J., Farrens, J. L., & Kappenman, E. S. (2018). How many trials does it take to get a significant ERP effect? It depends. *Psychophysiology*, *55*(6), e13049. https://doi.org/10.1111/psyp.13049

Burian, R. M. (2007). On MicroRNA and the need for exploratory experimentation in post-genomic molecular biology. *History and Philosophy of the Life Sciences*, *29*(3), 285–311.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafó, M. R. (2013). Erratum: Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(6), 451–451. https://doi.org/10.1038/nrn3502

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex*, *49*(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

Chen, G., Pine, D. S., Brotman, M. A., Smith, A. R., Cox, R. W., Taylor, P. A., & Haller, S. P. (2022). Hyperbolic trade-off: The importance of balancing trial and subject sample sizes in neuroimaging. *NeuroImage*, *247*, 118786–118786. https://doi.org/10.1016/j.neuroimage.2021.118786

Chomsky, N. (2018). Two notions of modularity. In R. G. de Almeida & L. R. Gleitman (Eds.), *On concepts, modules, and language: Cognitive science at its core* (pp. 25–40). Oxford University Press.

Clayson, P. E., Carbine, K. A., Baldwin, S. A., & Larson, M. J. (2019). Methodological reporting behavior, sample sizes, and statistical power in studies of event-related potentials: Barriers to reproducibility and replicability. *Psychophysiology*, *56*(11), e13437. https://doi.org/10.1111/psyp.13437

Collins, H. (1985). *Changing order: Replication and induction in scientific practice*. Sage Publications.

Culp, S. (1995). Objectivity in experimental inquiry: Breaking data-technique circles. *Philosophy of Science*, *62*(3), 438–458. https://doi.org/10.1086/289877

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge, Taylor & Francis Group.

Cumming, G. (2014). The new statistics: why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Curran, T. (2004). Effects of attention and confidence on the hypothesized ERP correlates of recollection and familiarity. *Neuropsychologia*, *42*(8), 1088–1106. https://doi.org/10.1016/j.neuropsychologia.2003.12.011

de Graaf, T. A., & Sack, A. T. (2011). Null results in TMS: From absence of evidence to evidence of absence. *Neuroscience & Biobehavioral Reviews*, *35*(3), 871–877. https://doi.org/10.1016/j.neubiorev.2010.10.006

de Haas, B. (2018). How to enhance the power to detect brain-behavior correlations with limited resources. *Frontiers in Human Neuroscience*, *12*, 421–421. https://doi.org/10.3389/fnhum.2018.00421

Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70*(2), 200–227. https://doi.org/10.1016/j.neuron.2011.03.018

DiMarco, M., & Khalifa, K. (2019). Inquiry tickets: Values, pursuit, and underdetermination. *Philosophy of Science*, *86*(5), 1016–1028. https://doi.org/10.1086/705446

DiMarco, M., & Khalifa, K. (2022). Sins of inquiry: How to criticize scientific pursuits. *Studies in History and Philosophy of Science*, *92*, 86–96. https://doi.org/10.1016/j.shpsa.2021.12.008

Dolcini, N. & Valle, V. (2024). Heroines in the East and the West: a comparative semiotic study of female imagery in the anti-pandemic discourse of Italy and China. *Chinese Semiotic Studies*, 20(4), 647–666. https://doi.org/10.1515/css-2024-2031

Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science*, *67*(4), 559–579. https://doi.org/10.1086/392855

Douglas, H. E. (2009). *Science, policy, and the value-free ideal.* University of Pittsburgh Press.

Douglas, H. (2023). The importance of values for science. *Interdisciplinary Science Reviews*, *48*(2), 251–263. https://doi.org/10.1080/03080188.2023.2191559

Duerr, P. M. & Fischer, E. (2025). Rationally warranted promise: the virtue-economic account of pursuit-worthiness. *Synthese*, *206*(2), 1–33. https://doi.org/10.1007/s11229-025-05077-5

Elliott, K. C., & McKaughan, D. J. (2009). How values in scientific discovery and pursuit alter theory appraisal. *Philosophy of Science*, *76*(5), 598–611. https://doi.org/10.1086/605807

Elliott, K. C., & McKaughan, D. J. (2014). Nonepistemic values and the multiple goals of science. *Philosophy of Science*, *81*(1), 1–21. https://doi.org/10.1086/674345

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, *104*(2), 1177–1194. https://doi.org/10.1152/jn.00032.2010

Feyerabend, P. (1975). *Against method.* Verso.

Firestein, S. (2016). *Failure: Why science is so successful.* Oxford University Press.

Fischer, E. (2024). The promise of supersymmetry. *Synthese*, *203*(1), 6. https://doi.org/10.1007/s11229-023-04447-1

Fleisher, W. (2022). Pursuit and inquisitive reasons. *Studies in History and Philosophy of Science*, *94*, 17–30. https://doi.org/10.1016/j.shpsa.2022.04.009

Franklin, A. (1986). *The neglect of experiment*. Cambridge University Press.

Friederici, A. D., & Gierhan, S. M. (2013). The language network. *Current Opinion in Neurobiology*, *23*(2), 250–254. https://doi.org/10.1016/j.conb.2012.10.002

Friedman, D. C., & Šešelja, D. (2023). Scientific disagreements, fast science and higher-order evidence. *Philosophy of Science*, *90*(4), 937–957. https://doi.org/10.1017/psa.2023.83

Funtowicz, S.O. & Ravetz, R.J. (1993). Science for the post-normal age. *Futures*, *25*(7), 739–755. https://doi.org/10.1016/0016-3287(93)90022-L.

Galison, P. (1987). *How experiments end*. University of Chicago Press.

García, A. M., De Leon, J., Tee, B. L., Blasi, D. E., & Gorno-Tempini, M. L. (2023). Speech and language markers of neurodegeneration: A call for global equity. *Brain*, *146*(12), 4870–4879. https://doi.org/10.1093/brain/awad253

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (Sign) and type M (Magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/10.1177/1745691614551642

Gelman, A. (2015, April 21). The feather, the bathroom scale, and the kangaroo. *Statistical Modeling, Causal Inference, and Social Science*. https://statmodeling.stat.columbia.edu/2015/04/21/feather-bathroom-scale-kangaroo/

Gelman, A. (2017). Ethics and statistics: Honesty and transparency are not enough. *CHANCE*, *30*(1), 37–39. https://doi.org/10.1080/09332480.2017.1302720

Gooding, D. (1990). *Experiment and the making of meaning: Human agency in scientific observation and experiment*. Kluwer Academic Publishers.

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*(1), 160044–160044. https://doi.org/10.1038/sdata.2016.44

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, *48*(12), 1711–1725. https://doi.org/10.1111/j.1469-8986.2011.01273.x

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science*. Cambridge University Press.

Hagoort, P. (2019). The neurobiology of language beyond single-word processing. *Science*, *366*(6461), 55–58. https://doi.org/10.1126/science.aax0289

Han, H. (2023). Taking model pursuit seriously. *European Journal for Philosophy of Science*, *13*(2), 22. https://doi.org/10.1007/s13194-023-00524-x

Hangel, N., & ChoGlueck, C. (2023). On the pursuitworthiness of qualitative methods in empirical philosophy of science. *Studies in History and Philosophy of Science*, *98*, 29–39. https://doi.org/10.1016/j.shpsa.2022.12.009

Harris, J. (2022). Mixed methods research in developing country contexts: Lessons from field research in six countries across Africa and the Caribbean. *Journal of Mixed Methods Research*, *16*(2), 165–182. https://doi.org/10.1177/15586898211032825

Hayasaka, S., Peiffer, A. M., Hugenschmidt, C. E., & Laurienti, P. J. (2007). Power and sample size calculation for neuroimaging studies by non-central random field theory. *NeuroImage*, *37*(3), 721–730. https://doi.org/10.1016/j.neuroimage.2007.06.009

Hirano, S., Spencer, K. M., Onitsuka, T., & Hirano, Y. (2020). Language-related neurophysiological deficits in schizophrenia. *Clinical EEG and Neuroscience*, *51*(4), 222–233. https://doi.org/10.1177/1550059419886686

Hudson, R. (2021). Should we strive to make science bias-free? A philosophical assessment of the reproducibility crisis. *Journal for General Philosophy of Science, 52*(3), 389–405. https://doi.org/10.1007/s10838-020-09548-w

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124–e124. https://doi.org/10.1371/journal.pmed.0020124

Jap, B.A., Alimu, S., & Dolcini, N. (2025a). Affect and human electrophysiological research. *Neuroethics*, *18*(1), 2. https://doi.org/10.1007/s12152-024-09572-3

Jap, B.A., Hsu, Y.Y., & Politzer-Ahles, S. (2025b). Registered Report: Neural correlates of thematic role assignment for passives in Standard Indonesian. *PloS ONE, 20*(5), e0322341. https://doi.org/10.1371/journal.pone.0322341

Jap, B.A., & Hsu, Y.Y. (2025). An ERP study on verb bias and thematic role assignment in standard Indonesian. *Scientific Reports, 15*, 11847. https://doi.org/10.1038/s41598-025-96240-y

Jap, B.A., Martinez-Ferreiro, S., & Bastiaanse, R. (2016). The effect of syntactic frequency on sentence comprehension in standard Indonesian Broca's aphasia. *Aphasiology, 30*(11), 1325-1340. https://doi.org/10.1080/02687038.2016.1148902

Jap, B.A., Hsu, Y.Y., & Politzer-Ahles, S. (2024). Are cleft sentence structures more difficult to process? *Neuroscience Letters*, Article 138029. https://doi.org/10.1016/j.neulet.2024.138029

Jensen, K.M., & MacDonald, J.A. (2023). Towards thoughtful planning of ERP studies: How participants, trials, and effect magnitude interact to influence statistical power across seven ERP components. *Psychophysiology*, *60*(7), e14245. https://doi.org/10.1111/psyp.14245

Jung, R.G., Di Santo, P., Clifford, C., Prosperi-Porta, G., Skanes, S., Hung, A., Parlow, S., Visintini S., Ramirez F.D., Simard T., Hibbert B. (2021). Methodological quality of COVID-19 clinical research. *Nature Communications*, *12*(1): 943. https:// doi.org/10.1038/s41467-021-21220-5.

Kaan, E. (2007). Event-related potentials and language processing: A brief overview. *Language and Linguistics Compass*, *1*(6), 571–591. https://doi.org/10.1111/j.1749-818X.2007.00037.x

Kappenman, E. S., & Luck, S. J. (2016). Best practices for event-related potential research in clinical populations. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *1*(2), 110–115. https://doi.org/10.1016/j.bpsc.2015.11.007

Karaca, K. (2013). The strong and weak senses of theory-ladenness of experimentation: Theory-driven versus exploratory experiments in the history of high-energy particle physics. *Science in Context*, *26*(1), 93–136. https://doi.org/10.1017/S0269889712000300

Katz, L. N., Bohlen, M. O., Yu, G., Mejias-Aponte, C., Sommer, M. A., & Krauzlis, R. J. (2025). Optogenetic manipulation of covert attention in the nonhuman primate. *Journal of Cognitive Neuroscience*, *37*(2), 266-285. https://doi.org/10.1162/jocn_a_02274

Keil, A., Debener, S., Gratton, G., Junghöfer, M., Kappenman, E. S., Luck, S. J., Luu, P., Miller, G. A., & Yee, C. M. (2014). Committee report: Publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. *Psychophysiology*, *51*(1), 1–21. https://doi.org/10.1111/psyp.12147

Kemmerer, D. (2015). *Cognitive neuroscience of language*. Psychology Press, Taylor & Francis Group.

Kitcher, P. (2001). *Science, truth, and democracy*. Oxford University Press.

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, *1146*, 23–49. https://doi.org/10.1016/j.brainres.2006.12.063

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205. https://doi.org/10.1126/science.7350657

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. In M. Bar (Ed.), *Predictions in the brain: Using our past to generate a future* (pp. 190–207). Oxford University Press.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology*, *62*(1), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kwong, K.K., Belliveau, J.W., Chesler, D.A., Goldberg, I.E., Weisskoff, R.M., Poncelet, B.P., Kennedy, D.N., Hoppel, B.E., Cohen, M.S., Turner, R., Cheng, H.-M., Brady, T. J., & Rosen, B.R. (1992). Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. *Proceedings of the National Academy of Sciences*, *89*(12), 5675-5679.

Larson, M. J., & Carbine, K. A. (2017). Sample size calculations in human electrophysiology (EEG and ERP) studies: A systematic review and recommendations for increased rigor. *International Journal of Psychophysiology*, *111*, 33–41. https://doi.org/10.1016/j.ijpsycho.2016.06.015

Laudan, L. (1977). *Progress and its problems: Towards a theory of scientific growth*. University of California Press.

Laudan, L. (1996). *Beyond positivism and relativism: Theory, method, and evidence*. Westview Press.

Laymon, R., & Franklin, A. (2022). *Case studies in experimental physics: Why scientists pursue investigation* (1st ed.). Springer.

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, *64*(1), 2–17. https://doi.org/10.1002/asi.22784

Leonelli, S. (2018). Rethinking reproducibility as a criterion for research quality. In *Including a Symposium on Mary Morgan: Curiosity, imagination, and surprise* (Vol. 36B, pp. 129–146). Emerald Publishing Limited. https://doi.org/10.1108/S0743-41542018000036B009

Li, J., Bhattasali, S., Zhang, S., Franzluebbers, B., Luh, W.-M., Spreng, R.N., Brennen, J., Yang, Y., Pallier, C., & Hale, J. (2022). Le Petit Prince multilingual naturalistic fMRI corpus. *Scientific Data, 9*, 530. https://doi.org/10.1038/s41597-022-01625-7

Lichtenstein, E. I. (2021). (Mis)Understanding scientific disagreement: Success versus pursuit-worthiness in theory choice. *Studies in History and Philosophy of Science Part A*, *85*, 166–175. https://doi.org/10.1016/j.shpsa.2020.10.005

Longino, H. E. (1990). *Science as social knowledge: Values and objectivity in scientific inquiry*. Princeton University Press.

Longino, H. E. (2002). *The fate of knowledge*. Princeton University Press.

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). MIT Press.

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, *54*(1), 146–157. https://doi.org/10.1111/psyp.12639

Lusk, G., & Elliott, K. C. (2022). Non-epistemic values and scientific assessment: An adequacy-for-purpose view. *European Journal for Philosophy of Science*, *12*(2), 35. https://doi.org/10.1007/s13194-022-00458-w

Lynch, W. (2020). *Minority report: Dissent and diversity in science*. Rowman & Littlefield Publishing Group.

Mack, J., Meltzer-Asscher, A., Barbieri, E., & Thompson, C. (2013). Neural correlates of processing passive sentences. *Brain Sciences*, *3*(3), 1198–1214. https://doi.org/10.3390/brainsci3031198

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Maumet, C., Auer, T., Bowring, A., Chen, G., Das, S., Flandin, G., Ghosh, S., Glatard, T., Gorgolewski, K. J., Helmer, K. G., Jenkinson, M., Keator, D. B., Nichols, B. N., Poline, J.-B., Reynolds, R., Sochat, V., Turner, J., & Nichols, T. E. (2016). Sharing brain mapping statistical results with the neuroimaging data model. *Scientific Data*, *3*(1), 160102–160102. https://doi.org/10.1038/sdata.2016.102

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

McMullin, E. (1982). Values in Science. *PSA: Proceedings of the biennial meeting of the Philosophy of Science Association*, *1982*, 3–28. https://doi.org/10.1111/j.1467-9744.2012.01298.x

Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations* (4th ed.). University of Chicago Press.

Mizrahi, M. (2022). Theoretical virtues in scientific practice: An empirical study. *British Journal for the Philosophy of Science*, *73*(4), 879–902.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

Muthukumaraswamy, S. D. (2013). High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations. *Frontiers in Human Neuroscience*, *7*, 138–138. https://doi.org/10.3389/fnhum.2013.00138

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology*, *118*(12), 2544–2590. https://doi.org/10.1016/j.clinph.2007.04.026

Nickles, T., Schickore, J., & Steinle, F. (2006). Heuristic appraisal: Context of discovery or justification. In *Revisiting Discovery and Justification* (Vol. 14, pp. 159–182). Springer Netherlands. https://doi.org/10.1007/1-4020-4251-5_10

Nieuwland, M. S., Politzer-Ahles, S., Heyselaar, E., Segaert, K., Darley, E., Kazanina, N., Von Grebmer Zu Wolfsthurn, S., Bartolozzi, F., Kogan, V., Ito, A., Mézière, D., Barr, D. J., Rousselet, G. A., Ferguson, H. J., Busch-Moreno, S., Fu, X., Tuomainen, J., Kulakova, E., Husband, E. M., … Huettig, F. (2018). Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *eLife*, *7*, e33468. https://doi.org/10.7554/eLife.33468

Nikolaidis, A., Manchini, M., Auer, T., L. Bottenhorn, K., Alonso-Ortiz, E., Gonzalez-Escamilla, G., Valk, S., Glatard, T., Selim Atay, M., M.M. Bayer, J., Bijsterbosch, J., Algermissen, J., Beck, N., Bermudez, P., Poyraz Bilgin, I., Bollmann, S., Bradley, C., E.J. Campbell, M., Caron, B., … P. Zwiers, M. (2023). Proceedings of the OHBM Brainhack 2021. *Aperture Neuro*, 87. https://doi.org/10.52294/258801b4-a9a9-4d30-a468-c43646391211

Nord, C. L., Valton, V., Wood, J., & Roiser, J. P. (2017). Power-up: A reanalysis of 'power failure' in neuroscience using mixture modeling. *The Journal of Neuroscience*, *37*(34), 8051–8061. https://doi.org/10.1523/JNEUROSCI.3592-16.2017

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Novick, J. M., Trueswell, J. C., & Thompson-Schill, S. L. (2005). Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, *5*(3), 263–281. https://doi.org/10.3758/CABN.5.3.263

Nyrup, R. (2015). How explanatory reasoning justifies pursuit: A Peircean view of IBE. *Philosophy of Science*, *82*(5), 749–760. https://doi.org/10.1086/683262

Nyrup, R. (2020). Of water drops and atomic nuclei: Analogies and pursuit worthiness in science. *The British Journal for the Philosophy of Science*, *71*(3), 881–903. https://doi.org/10.1093/bjps/axy036

O'Connor, K. (2012). Auditory processing in autism spectrum disorder: A review. *Neuroscience & Biobehavioral Reviews*, *36*(2), 836–854. https://doi.org/10.1016/j.neubiorev.2011.11.008

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Osterhout, L., & Holcomb, P. J. (1992). Event-related brain potentials elicited by syntactic anomaly. *Journal of Memory and Language*, *31*(6), 785–806. https://doi.org/10.1016/0749-596X(92)90039-Z

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R., Miller, G. A., Ritter, W., Ruchkin, D. S., Rugg, M. D., & Taylor, M. J. (2000). Guidelines for using human event-related potentials to study cognition: Recording standards and publication criteria. *Psychophysiology*, *37*(2), 127–152. https://doi.org/10.1111/1469-8986.3720127

Plante, E., & Gómez, R. L. (2018). Learning without trying: The clinical relevance of statistical learning. *Language, Speech, and Hearing Services in Schools*, *49*(3S), 710–722. https://doi.org/10.1044/2018_LSHSS-STLT1-17-0131

Poeppel, D. (2012). The maps problem and the mapping problem: Two challenges for a cognitive neuroscience of speech and language. *Cognitive Neuropsychology*, *29*(1–2), 34–55. https://doi.org/10.1080/02643294.2012.710600

Poeppel, D., & Embick, D. (2005). Defining the relation between linguistics and neuroscience. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 103–118). Lawrence Erlbaum Associates.

Poline, J.-B., Breeze, J. L., Ghosh, S., Gorgolewski, K., Halchenko, Y. O., Hanke, M., Haselgrove, C., Helmer, K. G., Keator, D. B., Marcus, D. S., Poldrack, R. A., Schwartz, Y., Ashburner, J., & Kennedy, D. N. (2012). Data sharing in neuroimaging research. *Frontiers in Neuroinformatics*, *6*, 9–9. https://doi.org/10.3389/fninf.2012.00009

Poldrack, R. A. (2012). The future of fMRI in cognitive neuroscience. *NeuroImage*, *62*(2), 1216–1220. https://doi.org/10.1016/j.neuroimage.2011.08.007

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., Nichols, T. E., Poline, J.-B., Vul, E., & Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, *18*(2), 115–126. https://doi.org/10.1038/nrn.2016.167

Politzer-Ahles, S. (2020). What can electrophysiology tell us about the cognitive processing of scalar implicatures? *Language and Linguistics Compass*, *14*(10), 1–22. https://doi.org/10.1111/lnc3.12401

Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson & CO.

Pulvermüller, F., Neininger, B., Elbert, T., Mohr, B., Rockstroh, B., Koebbel, P., & Taub, E. (2001). Constraint-induced therapy of chronic aphasia after stroke. *Stroke*, *32*(7), 1621–1626. https://doi.org/10.1161/01.STR.32.7.1621

Raymer, A. M., Beeson, P., Holland, A., Kendall, D., Maher, L. M., Martin, N., Murray, L., Rose, M., Thompson, C. K., Turkstra, L., Altmann, L., Boyle, M., Conway, T., Hula, W., Kearns, K., Rapp, B., Simmons-Mackie, N., & Gonzalez Rothi, L. J. (2008). Translational research in aphasia: From neuroscience to neurorehabilitation. *Journal of Speech, Language, and Hearing Research*, *51*(1), S259–S275. https://doi.org/10.1044/1092-4388(2008/020)

Reichenbach, H. (1938). *Experience and prediction: An analysis of the foundations and the structure of knowledge*. University of Notre Dame press.

Rooney, P. (1992). On values in science: Is the epistemic/non-epistemic distinction useful? In D. Hull, M. Forbes, & K. Okruhlik (Eds.), *Proceedings of the 1992 Biennial Meeting of the Philosophy of Science Association*. Philosophy of Science Association.

Salicchi, L., Hsu, Y. Y., & Jap, B. A. J. (2025, April). Sequential prediction and semantic evaluation, but alongside entropy: A computational investigation of N400 and P600 cognitive dynamics. In *The 5th International Conference on Theoretical East Asian Psycholinguistics*.

Schall, U. (2016). Is it time to move mismatch negativity into the clinic? *Biological Psychology*, *116*, 41–46. https://doi.org/10.1016/j.biopsycho.2015.09.001

Šešelja, D., & Straßer, C. (2012). The rationality of scientific reasoning in the context of pursuit: Drawing appropriate distinctions. *Philosophica*, *86*(3). https://doi.org/10.21825/philosophica.82146

Šešelja, D., & Straßer, C. (2014). Epistemic justification in the context of pursuit: A coherentist approach. *Synthese*, *191*(13), 3111–3141. https://doi.org/10.1007/s11229-014-0476-4

Shapin, S., Schaffer, S., & Hobbes, T. (1985). *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life*. Princeton University Press.

Shaw, J. (2022). On the very idea of pursuitworthiness. *Studies in History and Philosophy of Science*, *91*, 103–112. https://doi.org/10.1016/j.shpsa.2021.11.016

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simons, M., & Vagelli, M. (2021). Were experiments ever neglected? Ian Hacking and the history of the philosophy of experiments. *Philosophical Inquiries*, *9*(1), 167–188. https://doi.org/10.4454/philinq.v9i1.339

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, *10*(6), 886–899. https://doi.org/10.1177/1745691615609918

Stanford, P. K. (2006). *Exceeding our grasp: Science, history, and the problem of unconceived alternatives*. Oxford University Press, Incorporated.

Steel, D. (2010). Epistemic values and the argument from inductive risk. *Philosophy of Science*, *77*(1), 14–34. https://doi.org/10.1086/650206

Stehwien, S., Henke, L., Hale, J., Brennan, J., & Meyer, L. (2020). The Little Prince in 26 languages: Towards a multilingual neuro-cognitive corpus. *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources,* 43–49.

Stegenga, J. (2024). Fast science. *British Journal for the Philosophy of Science*. Advance online publication. https://doi.org/10.1086/729617

Steinle, F. (2002). Experiments in history and philosophy of science. *Perspectives on Science*, *10*(4), 408–432. https://doi.org/10.1162/106361402322288048

Sullivan, G. M., & Feinn, R. (2012). Using effect size—Or why the *P* value is not enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *15*(3), e2000797–e2000797. https://doi.org/10.1371/journal.pbio.2000797

Szucs, D., & Ioannidis, J. P. (2020). Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals. *NeuroImage (Orlando, Fla.)*, *221*, 117164–117164. https://doi.org/10.1016/j.neuroimage.2020.117164

Thompson, C. K., & Shapiro, L. P. (2007). Complexity in treatment of syntactic deficits. *American Journal of Speech-Language Pathology*, *16*(1), 30–42. https://doi.org/10.1044/1058-0360(2007/005)

Thompson, C. K., Den Ouden, D.-B., Bonakdarpour, B., Garibaldi, K., & Parrish, T. B. (2010). Neural plasticity and treatment-induced recovery of sentence processing in agrammatism. *Neuropsychologia*, *48*(11), 3211–3227. https://doi.org/10.1016/j.neuropsychologia.2010.06.036

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on *p* -values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Whalen, D. H., & McDonough, J. (2015). Taking the laboratory into the field. *Annual Review of Linguistics*, *1*(1), 395–415. https://doi.org/10.1146/annurev-linguist-030514-124915

Whitt, L. A. (1992). Indices of theory promise. *Philosophy of Science*, *59*(4), 612–634. https://doi.org/10.1086/289698

Widmann, A., Schröger, E., & Maess, B. (2015). Digital filter design for electrophysiological data – a practical approach. *Journal of Neuroscience Methods*, *250*, 34–46. https://doi.org/10.1016/j.jneumeth.2014.08.002

Wolf, W. J., & Duerr, P. M. (2024). Promising stabs in the Dark: theory virtues and pursuit-worthiness in the Dark Energy problem. *Synthese (Dordrecht)*, *204*(6), 155. https://doi.org/10.1007/s11229-024-04796-5

Woodman, G. F. (2010). A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception, & Psychophysics*, *72*(8), 2031–2046. https://doi.org/10.3758/BF03196680

Wray, K. B. (Ed.). (2024). *Kuhn's* The structure of scientific revolutions *at 60* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781009122696

Yarkoni, T. (2009). Big correlations in little studies: Inflated fMRI correlations reflect low statistical power—Commentary on Vul et al. (2009). *Perspectives on Psychological Science*, *4*(3), 294–298. https://doi.org/10.1111/j.1745-6924.2009.01127.x

Zarza-Luciáñez, D., Arce Arce, S., Bhathal Guede, H., & Sanjuán Martín, F. (2007). Mismatch negativity y nivel de conciencia en el traumatismo craneoencefálico grave. *Revista de Neurología*, *44*(08), 465–468. https://doi.org/10.33588/rn.4408.2006306

Zhang, S., Li, J., Yang, Y., & Hale, J. (2022). Decoding the silence: Neural bases of zero pronoun resolution in Chinese. *Brain and Language, 224*, 105050. https://doi.org/10.1016/j.bandl.2021.105050