What Pilot Studies Can (and Cannot) Do for Validity in Psychological Research

Seetahul, Y. (Department of Psychology, University of Innsbruck, Austria)*, **Elsherif, M. M.** (University of Leicester, University of Birmingham, UK), **Zygar-Hoffmann, C.** (Charlotte Fresenius Hochschule, Germany; LMU Munich, Germany), **Wallrich, L.** (Birkbeck, University of London, UK), **Silverstein, P.** (University of Coimbra, Portugal; Institute for Globally Distributed Open Research and Education), **Sætrevik, B.** (University of Bergen, Norway), **Pit, I. L.** (Institute of Human Sciences, University of Oxford, UK; Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany, Calleva Research Centre for Evolution and Human Sciences, Magdalen College, Oxford, UK), **Loenneker, H. D.** (University of Tübingen, Tübingen, Germany), **Heiser, N. H.** (University of Geneva, Switzerland), **Handley-Miner, I. J.** (Boston College, USA), **Graham, C. J.** (Royal College of Physicians of Edinburgh, Scotland), **Chou, Y. Y.** (King's College London, UK), **Buttliere, B.** (University of Warsaw, Poland), **Bochynska, A.** (University of Oslo, Norway), **Beitner, J.** (Central Institute of Mental Health, Germany), & **Neff, M.B.** (University of Oslo, Norway)*

Note: Middle authors contributed equally and are listed in reverse alphabetical order.

*__Corresponding authors__:

Yashvin Seetahul (Yashvin.Seetahul@uibk.ac.at)
Mary Beth Neff (marybethneff@gmail.com)

What Pilot Studies Can (and Cannot) Do for Validity in Psychological Research

Abstract

Recent debates about psychology's credibility have renewed attention to validity threats that are often difficult to evaluate from standard research reports. We argue that pilot studies can function as methodological due diligence by improving early decisions about design, measurement, manipulation, sampling, and analysis prior to a main study. We examine four types of validity: construct, internal, external, and statistical-conclusion validity. Across these domains, we show how piloting can help researchers assess whether materials and procedures function as intended, identify potential validity threats, and provide key design parameters for planning and model choice. However, we also caution that piloting can reduce validity when treated as a consequence-free space for undocumented tinkering, or when small, noisy pilots are used to justify overconfident inferences. Drawing on recent methodological debates, we distinguish two problematic uses of piloting: fragile use and opportunistic use. We discuss how hidden selection and analytic flexibility can distort the evidential value of confirmatory tests. We, then, conclude with practical recommendations for transparent pilot reporting. Greater transparency would enable readers to evaluate how pilot evidence shaped the final design of a study and consequently, how to interpret a study's findings.

Keywords: *validity, transparency, open science, pilot study, feasibility study, preliminary study*

1. Introduction

For over a decade, "a crisis of confidence" has challenged the credibility of psychological findings (Anvari & Lakens, 2018; Nosek et al., 2022; Pashler & Wagenmakers, 2012). There has been growing recognition that research reports often lack sufficient information to systematically evaluate construct, internal, external, and statistical-conclusion validities (Kerschbaumer et al., 2025; Schiavone et al., 2023; Vazire et al., 2022). In this article, we focus on piloting—an iterative process of preparatory investigations conducted before the full-scale or "main" study (In, 2017)—as a practical means of generating some of the missing information required to assess and improve validity.

Initial discussions of the crisis primarily addressed statistical-conclusion validity. False positives in confirmatory studies were attributed to research practices, such as small sample sizes and opportunistic decision-making, that made results in support of a hypothesis easier to obtain than they should have been (Lakens, 2019, 2025; Nagy et al., 2025; Spellman, 2015; Vazire, 2018). More recently, the "credibility revolution" has recognized that credibility relies on the integrity of the entire research process, from theory development to operationalization, design, and analysis, rather than solely on controlling false positives (Vazire, 2018).

This broader notion of integrity is commonly discussed in terms of validity, i.e., the degree to which each stage of the research process supports the intended inferences. This includes whether measures accurately capture the targeted constructs (Flake & Fried, 2020), experimental manipulations genuinely operationalize the intended constructs (Chester, 2018, 2020; Chester & Lasko, 2021), designs support causal inference (Rohrer, 2018), and effects generalize across people, stimuli, and settings

(Nosek & Errington, 2020; Yarkoni, 2022). This renewed focus on validity also highlights an ongoing practical challenge. Researchers often have to plan, justify, and communicate a wide range of design and analysis decisions before the consequences of those choices are known (or knowable).

In principle, many decisions can be specified in advance through pre-registration (Lakens et al., 2024) or Registered Reports (Chambers & Tzavella, 2022). However, researchers often find detailed study planning difficult. As Nosek et al. (2019) note, pre-registering a study requires extensive research and contingency planning, and many researchers struggle to make design and analysis decisions without concrete information about how the study and data will unfold. Sarafoglou et al. (2022) present similar qualitative evidence: respondents describe situations in which they find it "hard or sometimes impossible" (p. 14) to know how to analyze data before seeing its structure, and they emphasize that they cannot anticipate every eventuality.

In practice, study plans typically fall somewhere between two extremes. On one end, are vague protocols that leave numerous analytic paths open, allowing researchers to obtain significant results only weakly related to the original hypothesis (Claesen et al., 2022; Rubin, 2025). On the other end, are highly detailed but arbitrary plans that lock in decisions that may prove ill-suited when unexpected problems arise (Sarafoglou et al., 2022). When these issues are identified only after data collection, researchers face a dilemma. Revising the design or analysis may address validity concerns but introduces data-driven flexibility, thereby increasing the risk of false positives. Conversely, adhering to the original, flawed plan avoids this flexibility but leaves the validity problems unresolved.

One way to address this dilemma is to gather information on validity-relevant aspects of the study *before* the main data collection. Piloting allows researchers to assess key features of measurement, manipulation, design, sampling, and analysis in advance. Pilot studies[1] can therefore offer a practical alternative to both vague planning and arbitrarily rigid pre-specification. Despite its potential, piloting is rarely reported, seldom taught systematically, and only weakly supported by concrete guidance (Handley-Miner et al., 2025; Pilot Reporting Task Force, 2024). In this paper, we argue that when implemented carefully, piloting can generate information relevant to all major forms of validity and improve the quality of design and analysis decisions without inflating the error rate of the main confirmatory test. We then discuss how to prevent piloting from undermining the very credibility it is meant to support.

2. Piloting Can Strengthen Validity

Piloting practices vary widely across studies and subfields. Nonetheless, pilot studies typically support the main study in two broad ways (for a comprehensive list of pilot study use-cases, see Pilot Reporting Task Force, 2024; see also Thabane et al., 2010). First, pilots inform decisions that must be finalized before the main data collection begins, such as materials, instructions, recruitment strategies, and task structure. Second, they guide decisions that take effect after data collection but must be specified in advance to remain credible, including exclusion criteria, model selection, stopping rules, and outlier handling. Because piloting practices vary widely, we do not presume a single "correct" approach. Instead, we outline general ways piloting can enhance validity and discuss strategies for preventing it from undermining credibility.

---

[1] For simplicity, this paper uses the terms "pilot studies", "pilots", and "piloting." However, pilot studies are also commonly referred to as "preliminary studies," "feasibility studies," "pretests," or "pre-tests." In some contexts, preliminary studies may also be described more specifically as "pre-study manipulation checks" or "norming studies".

We organize the following discussion using Shadish, Cook, and Campbell's (2002) framework with four types of validity, as adopted by tools such as Seaboat (Schiavone et al., 2023) and VALID (Kerschbaumer et al., 2025). These tools catalog common threats to validity and highlight that many published studies do not report enough information to systematically evaluate them. Building on this work, we illustrate for each of the four validities: (a) representative threats, (b) how pilots can address these threats, and (c) how pilots themselves can introduce new validity problems.

2.1. Piloting for Construct Validity

Construct validity refers to whether a study's operationalizations provide a defensible link between abstract theory and participants' experiences and responses (Cronbach & Meehl, 1955; Flake & Fried, 2020; Kenny, 2019; Shadish et al., 2002). Many threats arise from disconnects between theory and practice. For example, researchers may use construct labels that are clearer to experts than to participants, rely on measures that behave differently than anticipated, or employ manipulations that shape responses through demand cues rather than through the intended psychological process (Kerschbaumer et al., 2025; Schiavone et al., 2023).

Piloting can help determine whether proposed measures and procedures function as the theory assumes (e.g., Baumeister et al., 1998; Brohmer et al., 2024; Cramer, 1967; Eisele et al., 2025; Hundley et al., 2000; Le, 2025; Loenneker et al., 2024; McDougal et al., 2024; Nagel et al., 2024; Neff, 2024; Oľhová et al., 2023; Schäfer et al., 2022; Schönbrodt et al., 2022; Zygar-Hoffmann et al., 2022; see also Table 1). First, qualitative and mixed-methods piloting, such as think-aloud protocols (Ericsson & Crutcher, 1991; Foddy, 1993; Wolcott & Lobczowski, 2021) or open-ended feedback, can assess whether participants interpret construct labels, interview questions, item content, and instructions as intended (Beatty & Willis, 2007; Borsboom et al., 2004; Flake & Fried, 2020; Fried & Flake, 2018; Irwin et al., 2010; Padilla & Benitez, 2014; Wood et al., 2021). Second, pilots allow for an initial evaluation of measurement quality. Researchers can estimate internal consistency and test-retest reliability, examine item distributions including floor and ceiling effects, detect patterns of non-response, and inspect response patterns for inattentive responding or demand characteristics (Coles et al., 2025; Corneille & Lush, 2023; Flake et al., 2017). Third, pilots can validate and calibrate manipulations. By measuring the intended construct alongside plausible alternatives, researchers can compare the resulting profile to theoretical expectations and identify unintended side effects (Chester, 2018, 2020; Chester & Lasko, 2021; Fabrigar et al., 2020; Schiavone et al., 2023). Pilots involving manipulation checks can assess whether a manipulation remains effective in the intended setting and primarily affects the target construct rather than related but unintended constructs (e.g., arousal instead of anger; Hauser et al., 2018). Finally, pilots can test translations and cultural appropriateness (Baldassarri & Abascal, 2017; Kenny, 2021; Pit et al., 2024).

However, piloting focused on construct validity can itself create problems if used incautiously. Very small pilot samples can give unreliable estimates of reliability (Bonett, 2002; Charter, 2003) or factor structure (MacCallum et al., 1999), leading to false reassurance or unjustified item pruning (Clark & Watson, 1995). Measures or manipulations can be “over-tuned” to produce “clean” effects. This can narrow the construct, distort its meaning, or introduce demand effects by making the design too obvious to the participant.

Table 1. **Examples of how piloting can inform and mislead construct validity inferences**

Construct validity concern	Information gained from piloting	Recommended uses of piloting	Cautions for interpretation and common misuses of piloting
Vague or inconsistently defined constructs	- Identification of ambiguities in definitions, instructions, and items; - Evidence of how participants spontaneously describe the state/trait in their own words; - Clarification or refinement of construct boundaries as understood in the planned context.	- Use cognitive interviews, think-aloud protocols, qualitative interviews, and open-ended feedback to assess how participants interpret definitions, instructions, and items; - Examine how participants spontaneously describe the state/trait; - Refine construct boundaries based on systematic patterns in pilot feedback.	- Over-interpreting idiosyncratic pilot feedback as definitive evidence about the construct; - Narrowing the construct to what a small convenience sample understands; - Making conceptual refinements without documenting changes.
Inadequate description or mismatch of operationalizations	- Confirmation that items, scoring, manipulations, and administration procedures are interpretable and feasible in planned context; - Preliminary evidence that observed responses align with theoretical expectations in the planned context.	- Document items, scoring, manipulations, and administration procedures in detail; - Use pilot data to check whether responses align with theoretical expectations in the planned context.	- Treating insufficient pilot checks as full validation of construct validity; - Overestimating construct validity based on feasibility or face validity alone.

Misinterpreting reliability and scale functioning	- Early estimates of internal consistency and short-term stability; - Identification of item distributions, floor or ceiling effects, and response styles in the target population.	- Use pilot data to obtain preliminary estimates of internal consistency and short-term stability; - Inspect item distributions, floor or ceiling effects, and response styles in the target population.	- Drawing strong conclusions from unstable reliability estimates in very small samples; - Pruning items solely to maximize internal consistency, thereby narrowing the construct; - Ignoring uncertainty of estimates (e.g., wide confidence intervals) around reliability estimates.
Measures/manipulations introduce bias or demand	- Detection of demand characteristics, cues, social desirability pressures, and inattentive responding; - Evidence on whether manipulations selectively affect intended versus collateral constructs across outcomes or settings.	- Use pilots to identify demand cues, social desirability pressures, and inattentive responding; - Evaluate whether manipulations selectively affect intended rather than collateral constructs across outcomes and settings.	- Informal "tuning" of manipulations until a large effect is observed, without documenting failed variants ("manipulation hacking" or "stimulus hacking"); - Allowing pilots to drift into overt hypothesis testing; - Underestimating remaining bias in the main study.

2.2. Piloting for Internal Validity

Internal validity refers to whether observed effects can be attributed to the intended causal variable rather than to alternative explanations, such as confounding factors, differential attrition, or implementation differences (Campbell et al., 1963; Kenny, 2019; Shadish et al., 2002). Threats to causal interpretation arise when participants are not comparable between conditions, procedures are inconsistently implemented, blinding fails, or other unintended differences exist between conditions (Kerschbaumer et al., 2025; Schiavone et al., 2023). In observational designs, threats include unmeasured confounding, selection bias, and order effects (Rohrer, 2018, Shadish et al., 2002).

Pilot studies can assess internal validity (e.g., Baumeister et al., 1998; Brohmer et al., 2024; David et al., 2021; Greenwald et al., 1998; McDougal et al., 2024; Nagel et al., 2024; Neff, 2024; Zygar-Hoffmann et al., 2022; see also Table 2). Piloting can screen for unintended confounds, reveal whether assignment and recruitment operate as intended, assess whether participants remain blind to conditions, and check whether dropout rates or procedural details differ across conditions. Piloting can test randomization by running the planned assignment process and examining whether resulting groups are comparable on key background variables. Researchers can then adjust recruitment strategies, introduce balancing procedures, or pre-plan statistical adjustments if imbalances emerge. Piloting can probe participant blinding and sensitivity to demand effects or experimenter effects by asking participants (and even researchers when double blinding is necessary) what they think the study is about and whether they can guess their condition (Schiavone et al., 2023; Slack & Draugalis, 2001). Pilots can also evaluate counterbalancing schemes and task length to reduce fatigue effects, and can map attrition dynamics by tracking dropout rates and reasons across conditions.

However, while piloting can help with the assessment, it cannot guarantee internal validity, and poorly designed or selectively used pilots can mislead. Small-sample pilots may lack power to detect modest but important imbalances or demand effects, and not detecting problems in a pilot does not imply their absence in the main study (Cooper et al., 2018). Moreover, running many variants during piloting and selecting the version that produces the largest effect, without disclosing this search, shifts internal validity threats from the main study to the pilot stage rather than resolving them. Finally, when the main study departs from pilot conditions (e.g., different recruitment sources, instructions, or experimenters), pilot-based assurances about internal validity may no longer hold.

Table 2. **Examples of how piloting can inform and mislead internal validity inferences**

Internal validity concern	Information gained from piloting	Recommended uses of piloting	Cautions for interpretation and common misuses of piloting
Non-equivalent groups and unaccounted confounders	- Empirical evidence that the randomization pipeline functions as intended; - Preliminary assessment of baseline balance on key covariates across conditions; - Feasibility of blocking or stratification strategies in the planned design.	- Use pilot data to empirically check the randomization pipeline; - Examine baseline balance on key covariates across conditions; - Assess the feasibility of blocking or stratification strategies.	- Pilot N too small to detect relevant imbalances; - Assuming balanced pilots guarantee balanced main samples; - Quietly adjusting assignments to "fix" problems without documenting changes.
Selective or differential attrition	- Preliminary estimates of dropout rates; - Information about reasons for attrition, and condition- or trait-related patterns; - Input for refining incentives, reminders, task burden and missing-data plans.	- Estimate overall dropout rates and commons reasons for attrition; - Examine whether attrition varies by condition- or trait-related patterns; - Use pilot information to refine incentives, reminders, task burden, and missing-data plans.	- Treating the absence of clear patterns in a small-sample pilot as evidence of ignorable attrition; - Using post-hoc pilot-based exclusions tailored to produce desired patterns.
Lack of blinding and demand/expectancy effects	- Indications that participant (or experimenter) makes assumptions about experimental condition or research hypothesis; - Identification of qualitative feedback on cues that reveal condition or desired	- Assess participant and experimenter guesses about condition and hypothesis; - Use qualitative feedback to identify cues that reveal condition or desired responses; - Improve masking procedures and experimental scripts based on pilot data.	- “Piloting away” obvious demand cues while leaving untested subtler expectations; - Failing to implement the same masking procedures in the main study (or switching to new masking procedures that were not

	responses; - Basis for improving masking procedures and experimental scripts.		piloted); - Ignoring evidence that blinding is not feasible.
Order and carryover effects in within-person designs	- Preliminary estimates of practice and fatigue trends across trial order; - Comparison of alternative counterbalancing schemes; - Identification of problematic trial sequences.	- Use pilot data to estimate practice and fatigue trends across trial order; - Compare alternative counterbalancing schemes; - Identify problematic trial sequences.	- Underpowered pilots that miss subtle but systematic order effects; - Selecting a trial structure based on inflated false positive rates in pilots rather than actual differences; - Selecting a trial structure that happens to yield a desirable pilot pattern, inflating Type I error.
Poorly matched control conditions	- Participant ratings of time-on-task, workload, stress, valence, interest, and expectancy across treatment and control conditions; - Qualitative feedback on perceived similarities and differences between conditions.	- Use participant ratings to compare time-on-task, interest, and expectancy across treatment and control conditions; - Gather qualitative feedback on perceived similarities and differences between conditions.	- Iteratively tweaking control conditions until the largest difference is observed, without documenting prior versions; - Creating overly "inert" controls that sacrifice psychological realism.

2.3. Piloting for External Validity

External validity refers to the extent to which findings generalize beyond the specific people, settings, measures, treatments or manipulations, and time points studied (Campbell, 1957; Kenny, 2019; Shadish et al., 2002). Common threats include vague generalization targets, narrow or poorly described samples, and measures that may rely on specific sample characteristics (Kerschbaumer et al., 2025; Schiavone et al., 2023).

Piloting can help researchers clarify the likely scope of their conclusions (e.g., Cramer, 1967; David et al., 2021; Le, 2025; Loenneker et al., 2024; McDougal et al., 2024; Oľhová et al., 2023; Zygar-Hoffmann et al., 2022; see also Table 3). Pilots can compare recruitment channels and sampling frames. Researchers might recruit small pilot samples from different sources and compare them on demographics, key covariates, and outcome distributions. These comparisons can inform eligibility criteria and sampling strategies that better approximate the theoretical population of interest (Dyrvig et al., 2014; Schiavone et al., 2023; Slack & Draugalis, 2001). Stimulus sets can be piloted more broadly to estimate between-stimulus variance and calibrate difficulty before selecting a subset. Pilots can assess procedural robustness across contexts and devices by running the same task online and in the lab, on mobile and desktop, or at different times to reveal systematic differences that may warrant standardizing procedures. Finally, piloting can test cultural and contextual fit by checking whether translations and examples are interpretable across subgroups (Baldassarri & Abascal, 2017; Kenny, 2021; Pit et al., 2024).

However, piloting alone cannot guarantee generalizability. Moreover, pilots based on small or convenience samples may give an incomplete picture of which populations are reachable or how heterogeneous responses will be. Stimulus piloting can become "stimulus hacking" (Experimental Philosophy, 2025; Jaeger, 2023) if researchers test a large pool of stimuli and retain only those producing large effects without reporting the selection process. Similarly, comparing multiple recruitment channels and choosing the one yielding the largest pilot effect can bias the main study.

Table 3. **Examples of how piloting can inform and mislead external validity inferences**

External validity concern	Information gained from piloting	Recommended uses of piloting	Cautions for interpretation and common misuses of piloting
Unclear or ill-defined target population	- Feedback and evidence on who is actually reached through different recruitment channels; - Descriptive comparisons between pilot samples and the intended target population on key variables; - Information to refine inclusion/exclusion criteria.	- Identify who is actually reached through different recruitment channels and adjust strategy; - Compare pilot samples descriptively with the intended target population on key variables; - Gather information to refine inclusion/exclusion criteria.	- Treating the most accessible pilot sample as the target population by default; - Failing to revise theory or scope of claims when pilot sampling suggests recruitment diverges from target population.
Non-transparent or non-representative samples.	- Detailed demographic and contextual information about and from pilot participants (e.g., age, culture, language, context of participation); - Preliminary evidence about the feasibility of recruiting more diverse samples.	- Collect and report detailed demographic and contextual information from pilot participants (e.g., age, culture, language, context of participation); - Use pilots to assess the feasibility of recruiting more diverse populations.	- Over-generalizing from small or homogeneous pilots (e.g., assuming a manipulation that works in the pilot will work outside that sample); - Using pilot results to justify broad population claims without adequate sampling; - Failing to report recruitment limitations or sampling failures.
Over-generalizing beyond specific measures, settings, or times	- Differences in task performance and responses across devices, contexts, or time points; - Preliminary estimates of temporal	- Compare task performance and responses across devices, contexts, or time points; - Use pilot data to obtain preliminary estimates of temporal stability or context	- Conducting pilots only in the most convenient context (e.g., university lab) while generalizing to broader contexts (e.g., everyday life, cross-cultural

	stability or context-dependent moderation.	moderation.	contexts); - Selecting contexts that yield the largest pilot effect.
Ignoring contextual and cultural fit	- Qualitative feedback on scenarios, translations, and norms; - Identification of content that is confusing, inappropriate or poorly aligned with particular subgroups.	- Use qualitative feedback to evaluate scenarios, translations, and norms; - Identify content that is confusing, inappropriate or poorly aligned with particular subgroups.	- Treating negative or complex feedback from participant subgroups as sufficient grounds for excluding them from the main study.

2.4. Piloting for Statistical-Conclusion Validity

Statistical-conclusion validity refers to the appropriateness of statistical analyses and inferences given the study's design, underlying assumptions, and data quality (Cook & Campbell, 1979; Shadish et al., 2002; Kenny, 2019). Common threats include weakly justified sample sizes, unclear stopping or exclusion rules, mismatches between hypotheses and statistical models, overly rigid pre-analysis plans based on assumptions that do not match the observed data, and opaque reporting (Kerschbaumer et al., 2025; Schiavone et al., 2023).

Piloting can provide design parameters and modeling insights that are difficult to obtain before data collection (e.g., Brohmer et al., 2024; Greenwald et al., 1998; Hundley et al., 2000; Loenneker et al., 2024; McDougal et al., 2024; Schäfer et al., 2022; Schönbrodt et al., 2022; Zygar-Hoffmann et al., 2022; see also Table 4). Pilot data can estimate outcome variance (Teare et al., 2014), intraclass correlations (Snijders & Bosker, 2012), baseline-outcome correlations (Borm et al., 2007), rates of missing data (Heo, 2014), and typical numbers of valid trials per participant (Miller, 2024). These estimates can support more realistic sample-size justifications and help determine cluster sizes and expected data loss (Kerschbaumer et al., 2025). Pilot analyses can help identify appropriate model families. For example, pilots can check whether count outcomes show overdispersion or zero inflation, residuals are strongly skewed, scales are constrained by floor or ceiling effects, or proposed random-effects structures are sufficiently stable (De Boeck & Jeon, 2018; Dyrvig et al., 2014). Pilots can inform data-handling rules by showcasing realistic response time ranges and plausible thresholds for excluding trials or participants (Fabrigar et al., 2020; Thomas & Clifford, 2017). Finally, pilot experience can clarify which outcomes and contrasts are substantively relevant, helping anticipate multiplicity and distinguish confirmatory from exploratory analyses (Ditroilo et al., 2025; Li et al., 2017).

However, statistical-conclusion piloting is especially prone to misuse. A central risk is treating effect size estimates from small-sample pilots as accurate inputs for power analyses or for specifying the smallest effect size of interest (Lakens et al., 2018). Simulations show that basing main-study designs on effect sizes from small-sample pilots tends to produce underpowered studies (Albers & Lakens, 2018; Lakens, 2022). Although pilot effect size estimates are noisy in both directions, they become systematically inflated when researchers base decisions on "promising" pilot outcomes (e.g., statistical significance or feasibility). This approach selects from a truncated subset of the sampling distribution, inflating the effect size used in power analyses and leading to underpowered main studies (Albers & Lakens, 2018). Consequently, pilot-based effect size estimates should be treated as rough guides and calibrated against effect sizes in previous studies and field norms.

Another risk is using pilots to try many analytic models or transformations and then selecting the one that yields the most favorable results, treating it as if it had been specified in advance, without reporting the pilot. When this selection is driven by which specification yields the smallest p-value or largest effect, the pilot becomes a hidden model-selection stage. Consequently, the confirmatory analysis is no longer independent of the data that motivated it, and the resulting evidence is easier to overstate. Finally, if pilot data are included in the main analysis without proper adjustments, or if the boundary between "pilot" and "main" samples is unclear, the probability of falsely confirming a hypothesis may be increased.

***Table 4.* Examples of how piloting can inform and mislead statistical-conclusion validity inferences**

Statistical-conclusion validity concern	Information gained from piloting	Recommended uses of piloting	Cautions for interpretation and common misuses of piloting
Low power, low precision, and under-justified sample sizes	- Approximate estimates of variance components, intraclass correlations, autocorrelations, baseline-outcome correlations; - Estimates of usable trial rates and data loss; - Input for precision- or power-based sample size justification that does not rely solely on published effect sizes.	- Use approximate estimates of variances, intraclass correlations, autocorrelations, baseline-outcome correlations; - Estimate rates of usable trials and data loss; - Inform precision- or power-based sample size justification without relying solely on published effect sizes.	- Treating small sample pilot effect sizes as if they were unbiased or precise; - Selecting only "promising" pilots to justify main study sample sizes; - Defining the smallest effect size of interest directly from small or noisy sample pilot estimates.
Flexible stopping and exclusion rules	- Realistic estimates of attrition rates and data quality; - Typical ranges of response times, and prevalence of low-quality data; - Information to define feasible stopping rules and exclusion thresholds in advance of the main study.	- Use pilot data to estimate attrition rates and typical ranges of response times; - Assess the prevalence of low-quality data; - Define feasible stopping rules and exclusion thresholds in advance, ahead of the main study.	- Tailoring stopping or exclusion rules post-hoc based on pilot plus main data; - Treating ad hoc thresholds as if they were pre-planned; - Failing to report that pilot-informed decisions were based on multiple unreported pilot variants.
Inappropriate models or ignored dependence	- Evidence on outcome distributions (e.g., skew, zero inflation) and variance components; - Identification of clustering or dependence structures;	- Use pilot data to examine outcome distributions (e.g., skew, zero inflation) and variance components; - Identify clustering or dependence structures;	- Overfitting complex models to small pilot datasets and locking it in; - Choosing models primarily because they yield desirable statistical significance patterns in the pilot;

	- Basis for choosing appropriate model families, random-effects structures, or robust methods.	- Select appropriate model families, random-effects structures, or robust methods.	- Ignoring that model assumptions may differ in larger or more heterogeneous main samples.
Opaque reporting and fragile results	- Experience conducting multiverse or robustness checks in pilot data; - Early identification of ambiguous or unstable patterns that motivate clearer hypotheses or designs; - Practice in documenting analysis code and decisions.	- Experience with multiverse or robustness checks in pilot data; - Identify early ambiguous or unstable patterns that motivate clearer hypotheses or improved designs; - Practice documenting analysis code and decisions.	- Treating exploratory pilot analyses as confirmatory; - Failing to disclose that the final analysis pipeline was tuned using pilot data; - Using pilot-based robustness checks to select favorable analyses without reporting alternatives.

2.5. Summary of Piloting's Role for Validity

Recent developments have brought renewed attention to validity. We argue that piloting can provide early, validity-relevant information that reduces guesswork in study design. However, while piloting can strengthen study validity, it is not foolproof. The flexibility that makes it useful can also create risks if pilot work is poorly designed or selectively interpreted.

Given these risks, it is reasonable to ask how much value piloting can add. We advocate piloting as a form of methodological due diligence rather than a guarantor of validity. Ultimately, piloting offers a structured way to identify and address validity threats before committing to a full-scale study, though understanding what piloting can and cannot achieve is essential for using it effectively.

When pilot studies are used to justify decisions about measures, designs, or analyses, they function as scientific evidence. The same principles of validity and inference therefore apply to piloting as they do to the main study. In the next section, we examine cases where piloting can backfire when these principles are not followed or misused.

3. When Piloting Goes Wrong

Piloting can go wrong in at least two broad ways. The first is methodological overextension, where well-intentioned researchers treat low-precision pilot data as a stable foundation for consequential design choices, despite inherent uncertainty. The second is opportunistic misuse, where pilots become a vehicle for selective exploration and outcome-contingent decision-making. This introduces hidden degrees of freedom and distorts evidential value, whether intentionally or through undocumented iteration, especially when the processes that shaped downstream decisions are not transparently documented.

3.1. Fragile Inference Problems Due to Methodological Overextension

As outlined earlier, many pitfalls arise when pilots are treated as exempt from basic statistical constraints. And although piloting practices vary, pilots are typically conducted with small samples. With very small samples, these judgments are inherently unstable (Lewis et al., 2021). Observed effects are noisy estimates, and sampling error alone can generate large apparent effects that will not replicate or obscure meaningful effects that do not appear in the pilot.

This problem is well-documented for power analysis. Albers and Lakens (2018) show that using small-sample pilots to estimate effect sizes and decide which path to pursue leads to biased and imprecise estimates, often resulting in underpowered main studies even when planning is "by the book." The same logic applies when pilot estimates determine what effect sizes to expect, as any inference based on small pilot samples must contend with substantial uncertainty.

Similar issues arise when pilots inform exclusion rules, device restrictions, or control conditions. For example, if a small-sample pilot suggests that mobile participants produce noisier data, it may be tempting to exclude them from the main study. Yet such a decision also depends on the magnitude and uncertainty of the observed difference (between mobile and non-mobile participants), and the recruitment constraints of the pilot. Treating a small-sample difference as decisive risks overfitting the design to pilot idiosyncrasies.

These examples illustrate how well-intentioned uses of piloting can lead to fragile inferences when statistical uncertainty and evidential limits are not fully acknowledged. Hence, when pilot data are used as evidence about how measures, manipulations, samples, or models behave, the pilot should be designed and interpreted with that evidential purpose in mind.

Pilot studies may, and perhaps often will, be smaller than main studies because their aims are narrower and more diagnostic. However, labeling a study a pilot does not exempt it from the usual constraints of scientific inference. Limited sample size, selective recruitment, and contextual specificity still shape what can be learned. Accordingly, a "pilot sample" should not be treated as synonymous with a "small convenience sample," and there is no principle by which a study stops being a pilot once its sample becomes large. The distinction between a pilot and a main study is functional: pilots are conducted to inform and refine subsequent decisions, whereas main studies are conducted to adjudicate the focal claims (e.g., Bell, 2018).

This does not mean that pilots must be large. Pilots should be tailored to the specific information they are meant to provide and to the researcher's broader scientific approach. From a Bayesian perspective (see Wagenmakers et al., 2018), pilot data can be valuable even when small if the goal is learning and refining understanding. Any amount of information, even if very limited, can update beliefs about key features of the study and its data-generating process. But from a Neyman–Pearson decision-oriented perspective (see Lehmann & Romano, 2005; Neyman & Pearson, 1933), the central question is whether the pilot provides enough information to justify acting on it. In some cases, it may be better to make no pilot-based decision at all rather than base consequential choices on a pilot that cannot adequately support those decisions. Importantly, this perspective does not require pilots to systematically have large samples either. For example, consider a manipulation check where the manipulation is only worth using if it produces a clearly noticeable change. In such cases, even a pilot with a relatively small sample could provide enough information to judge whether the manipulation is capable of producing an effect of that order of magnitude.

3.2. Hidden Flexibility and Epistemic Distortion

Recent methodological debates have also highlighted opportunistic misuses, in which preliminary work is concealed or selectively used to steer downstream decisions and repeatedly tested until it produces a desired outcome.. For example, critics of Bem's (2011) precognition experiments have suggested that pilot-based p-hacking might explain the unusually high rate of significant findings (Schimmack, 2018).

In response, Simonsohn et al. (2018) formalized and evaluated this strategy, which they called "pilot-dropping." They describe pilot-dropping as a procedure in which researchers run an initial pilot, inspect its p-value, drop pilots with unpromising p-values, and proceed when results look like they will reject the null-hypothesis. They contrast this with "pilot-hacking," where researchers try multiple analyses until a small p-value is achieved, after which data collection continues, and the combined dataset is analyzed using the tuned specifications. The simulation by Simonsohn et al. suggests that pilot-hacking would be far more effective than pilot-dropping at producing significant results. Precisely because it blurs the line between exploratory piloting and confirmatory main studies, these practices represent a serious distortion of evidential value (Simonsohn et al., 2018). However, both cases show how pilots can enable hidden flexibility and selective exploration.

A related concern is cycling through stimuli, operationalizations, or manipulations until a desired result appears, then presenting the final configuration as if it were planned from the start (Fiedler, 2011; Handley-Miner et al., 2025). Yarkoni's (2022) analysis of the generalizability crisis emphasizes that stimulus selection can strongly shape observed effects, yet the processes behind these decisions are under-documented. When such iterative search processes are treated as piloting, and only the final "successful" main study is reported, the main study is easily over-interpreted as rigorously demonstrating a general phenomenon rather than one that emerges only in specific conditions.

To be clear, exploratory method development is not the issue. Instead, the problem is misallocation of evidential credit. As Popper (1959) emphasized, strong tests expose hypotheses to genuine risk. Hidden pilot selection changes this calculus. If many versions are tried but only those that reject the null are reported, some of the apparent "success" reflects exploratory search over designs, materials, and analyses rather than the reported hypothesis alone. The reported p-value for the main study may still be valid given its stated design and analysis, yet the overall evidential situation is weaker than it appears because the broader process that produced that final test has not been disclosed. Consequently, the problem is not exploration itself, but *undisclosed* exploration, as readers cannot assess what the main study truly tested when the pilot history that shaped design or analytic choices is obscured.

3.3. The Case for Transparency

Across both inferential approaches discussed above, the core requirement is that pilot evidence be used in a principled manner. For this reason, transparency is a solution for both fragile inferences and strategic misuse. When substantive design or analytic choices are justified on the basis of pilot findings, readers and reviewers need sufficient information to evaluate whether those inferences are warranted. This entails being explicit about what the pilot was designed to reveal, the inferential limits it faces, which data will be part of the pilot versus the main study, and the kinds of updates or downstream decisions it can reasonably support. Without such information, pilot-informed choices become effectively opaque: the main study can appear carefully planned while the evidential basis for its design decisions remains largely invisible.

Given the loose definition of piloting and variation across subfields, one might intuitively advocate for strict boundaries around what counts as piloting. This is not the approach we take. The defining feature of a pilot is functional: it is conducted to obtain information needed to support a subsequent main study, and that informational target can vary widely. Accordingly, many limitations emphasized here are not limitations of piloting per se, but of the scientific and statistical methods deployed within pilots. Fully formalizing "piloting" would require drawing boundaries around an open-ended set of preparatory practices, which is unlikely to be feasible or valuable. Instead, we propose a singular boundary: whenever a pilot study is used as evidence to motivate substantive choices in a main study, it should be treated as part of the evidential record and reported transparently.

Handley-Miner et al. (2025) argue that routine disclosure of piloting should become standard and outline concrete steps to facilitate this reporting. For example, they suggest including a brief pilot-transparency statement in the manuscript that summarizes whether pilots were conducted and how they informed the main study, and sharing materials and de-identified pilot data when feasible. More generally, these goals require reporting templates and explicit journal guidance so authors are not penalized for disclosing pilots. Such practices would make visible the selection processes that shape main studies, reduce file-drawering of pilot outcomes, and help readers gauge how informative a study is. Most importantly, transparent reporting preserves the link between validity concerns and methodological adjustments, as it allows readers to see not only that piloting occurred but how it was used to address specific threats to validity.

4. Limitations and Open Questions

Our argument for piloting as a tool for strengthening validity has limitations. First, there is relatively little systematic evidence on how piloting is practiced across psychological subfields.

Survey work suggests that piloting is rarely reported, and that what counts as a pilot varies widely (Pilot Reporting Task Force, 2024). However, a lack of reported pilots does not necessarily imply that piloting occurred behind the scenes. In some areas, pilots may be uncommon due to resource constraints or field norms and lack of perceived value. Moreover, the practice of piloting ranges from brief technical checks to substantial preliminary studies with different evidential weights. More descriptive work and publication audits are needed to map the frequency, design, decision role, and reporting quality of pilot studies in psychology.

Second, calls for more and better piloting have resource and incentive implications. Well-designed pilots require time, funding, and participant access, and may be particularly infeasible in costly paradigms or hard-to-reach populations. Pilots are also less visible and rewarded than main studies, so increasing reliance on piloting without changes in incentives could widen inequalities by favoring better-resourced labs. Addressing this likely requires policy shifts beyond individual practice (Gärtner et al., 2024), including support from funders and journals and potentially dedicated outlets for methods and materials.

Third, piloting is only one tool among several for addressing validity problems. For some questions, methods such as simulations (e.g., Pfaffel et al., 2016), meta-analyses (e.g., Farrar et al., 2020; Roth et al., 2005), or using existing large-scale datasets may more efficiently provide relevant information. Our goal is to encourage the uptake of piloting and its transparent reporting, but not to prescribe it as a single workflow.

Finally, open questions remain about how piloting should be defined and reported. Because piloting derives its value from versatility, this looseness is not a problem in itself. For example, checking whether a validated in-person study works online is necessarily different from assessing whether a novel questionnaire is understandable to lay audiences. Rather than standardizing a single approach, we advocate clearly operationalizing and reporting how piloting was conducted. A pragmatic starting point is a brief pilot-transparency statement indicating whether pilot data informed the reported study, how they shaped key decisions, and, when feasible, where to find the pilot's materials and data (Handley-Miner et al., 2025). Future work could categorize pilot types across subfields and test which reporting levels best balance informativeness and reader needs.

5. Conclusion

Concerns about validity have spurred the proliferation of tools for identifying threats, but researchers still need concrete ways to test out their planned study in time to make adjustments. We propose that piloting be treated as methodological due diligence: a structured way to learn about one's own study in advance, and a form of scientific evidence when used to justify substantive design or analytic choices. When piloting is conducted with disciplined aims and reported transparently, it can help researchers identify validity threats early, reduce guesswork in study planning, and avoid common pitfalls in practice. Piloting, understood in this way, should play a central role in psychologists' methodological toolkit.

References

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, *74*, 187–195. https://doi.org/10.1016/j.jesp.2017.09.004

Anvari, F., & Lakens, D. (2018). The replicability crisis and public trust in psychological science. *Comprehensive Results in Social Psychology*, *3*(3), 266–286. https://doi.org/10.1080/23743603.2019.1684822

Baldassarri, D., & Abascal, M. (2017). Field experiments across the social sciences. *Annual Review of Sociology*, *43*, 41–73. https://doi.org/10.1146/annurev-soc-073014-112445

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource? *Journal of Personality and Social Psychology*, *74*(5), 1252–1265. https://doi.org/10.1037/0022-3514.74.5.1252

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, *71*(2), 287–311. https://doi.org/10.1093/poq/nfm006

Bell, M. L., Whitehead, A. L., & Julious, S. A. (2018). Guidance for using pilot studies to inform the design of intervention trials with continuous outcomes. *Clinical Epidemiology*, *10*, 153–157. https://doi.org/10.2147/CLEP.S146397

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. https://doi.org/10.1037/a0021524

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, *27*(4), 335–340. https://doi.org/10.3102/10769986027004335

Borm, G. F., Fransen, J., & Lemmens, W. A. J. G. (2007). A simple sample size formula for analysis of covariance in randomized clinical trials. *Journal of Clinical Epidemiology*, *60*(12), 1234–1238. https://doi.org/10.1016/j.jclinepi.2007.02.006

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Brohmer, H., Hofer, G., Bauch, S. A., Beitner, J., Berkessel, J. B., Corcoran, K., ... & Athenstaedt, U. (2024). Effects of the generic masculine and its alternatives in Germanophone countries: A multi-lab replication and extension of Stahlberg, Sczesny, and Braun (2001). *International Review of Social Psychology*, *37*, 17. https://doi.org/10.5334/irsp.522

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297–312. https://doi.org/10.1037/h0040950

Campbell, D. T., Stanley, J. C., & Gage, N. L. (1963). *Experimental and quasi-experimental designs for research*. MA: Houghton, Mifflin and Company.

Chambers, C. D., & Tzavella, L. (2022). The past, present and future of registered reports. *Nature Human Behaviour*, *6*(1), 29–42. https://doi.org/10.1038/s41562-021-01193-7

Charter R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *The Journal of general psychology*, *130*(3), 290–304. https://doi.org/10.1080/00221300309601160

Chester, D. (2018, December 3). *Questionable MAnipulation Practices (QMAPs)*. David Chester Personal Webpage. http://davidchester.weebly.com/1/post/2018/12/questionable-manipulation-practices-qmaps.html

Chester, D. (2020, June 24). *Validating Experimental Manipulations w/ Passive Control Conditions*. David Chester Personal Webpage. http://davidchester.weebly.com/1/post/2020/06/validating-experimental-manipulations-w-passive-control-conditions.html

Chester, D. S., & Lasko, E. N. (2021). Construct validation of experimental manipulations in Social Psychology: Current Practices and Recommendations for the Future. *Perspectives on Psychological Science*, *16*(2), 377–395. https://doi.org/10.1177/1745691620950684

Claesen, A., Lakens, D., Vanpaemel, W., & van Dongen, N. N. N. (2022). Severity and crises in science: Are we getting it right when we're right and wrong when we're wrong? *PsyArXiv Preprint*. https://doi.org/10.31234/osf.io/ekhc8

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*(12), 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Coles, N. A., Wyatt, M., & Frank, M. C. (2025). A meta-analysis of the impact and heterogeneity of explicit demand characteristics. *Collabra: Psychology*, *11*(1), 143005. https://doi.org/10.1525/collabra.143005

Cooper, C. L., Whitehead, A., Pottrill, E., Julious, S. A., & Walters, S. J. (2018). Are pilot trials useful for predicting randomisation and attrition rates in definitive studies: a review of publicly funded trials. *Clinical Trials*, *15*(2), 189-196.

Corneille, O., & Lush, P. (2023). Sixty years after Orne's *American Psychologist* article: A conceptual framework for subjective experiences elicited by demand characteristics. *Personality and Social Psychology Review*, *27*(1), 83–101. https://doi.org/10.1177/10888683221104368

Cramer, P. (1967). The Stroop effect in preschool aged children: A preliminary study. *Journal of Genetic Psychology*, *111*, 9–12

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

David, E. J., Beitner, J., & Võ, M. L. H. (2021). The importance of peripheral vision when searching 3D real-world scenes: A gaze-contingent study in virtual reality. *Journal of Vision*, *21*(7), 3. https://doi.org/10.1167/jov.21.7.3

De Boeck, P., & Jeon, M. (2018). Perceived crisis and reforms: Issues, explanations, and remedies. *Psychological Bulletin*, *144*(7), 757–777. https://doi.org/10.1037/bul0000154

Ditroilo, M., Mesquida, C., Abt, G., & Lakens, D. (2025). Exploratory research in sport and exercise science: Perceptions, challenges, and recommendations. *Journal of sports sciences*, *43*(12), 1108–1120. https://doi.org/10.1080/02640414.2025.2486871

Dyrvig, A.-K., Kidholm, K., Gerke, O., & Vondeling, H. (2014). Checklists for external validity: A systematic review. *Journal of Evaluation in Clinical Practice*, *20*(6), 857–864. https://doi.org/10.1111/jep.12166

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., Lazarević, L. B., Rabagliati, H., Ropovik, I., Aczel, B., Aeschbach, L. F., Andrighetto, L., Arnal, J. D., Arrow, H., Babincak, P., … et al. (2020). Many Labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science*, *3*(3), 309–331. https://doi.org/10.1177/2515245920958687

Eisele, G., Hiekkaranta, A., Kunkels, Y. K., van Ballegooijen, W., Bartels, S. L., Bastiaansen, J. A., ... & Kirtley, O. J. (2025). ESM-Q: A consensus-based quality assessment tool for experience sampling method items. *Behavior Research Methods*, *57*(4), 1–20.

Enserink, M. (2012). Final Report on Stapel Also Blames Field As a Whole. *Science*, *338*(6112), 1270–1271. https://doi.org/10.1126/science.338.6112.1270

Ericsson, K. A., & Crutcher, R. J. (1991). Introspection and verbal reports on cognitive processes—two approaches to the study of thought processes: A response to Howe. *New Ideas in Psychology*, *9*,57-71. https://doi.org/10.1016/0732-118X(91)90041-

Experimental Philosophy. (2025). Submissions. Retrieved January 27, 2026, from https://journals.ub.uni-koeln.de/index.php/xphi/about/submissions

Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). A validity-based framework for understanding replication in psychology. *Personality and Social Psychology Review*, *24*(4), 316–344. https://doi.org/10.1177/1088868320931366

Fanelli, D. (2010). "Positive" results increase down the Hierarchy of the Sciences. *PLoS One*, *5*(4), e10068. https://doi.org/10.1371/journal.pone.0010068

Farrar, B. G., Altschul, D. M., Fischer, J., van der Mescht, J., Placì, S., Troisi, C. A., Vernouillet, A., Clayton, N. S., & Ostojić, L. (2020). Trialling meta-research in comparative cognition: Claims and statistical inference in animal physical cognition. *Animal Behavior and Cognition*, *7*(3), 419–444. https://doi.org/10.26451/abc.07.03.09.2020

Fiedler, K. (2011). Voodoo correlations are everywhere—not only in neuroscience. *Perspectives on Psychological Science, 6*(2), 163-171. https://doi.org/10.1177/1745691611400237

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current Practice and Recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Fried, E. I., & Flake, J. K. (2018). Measurement matters. *APS Observer*, *31*. https://www.psychologicalscience.org/observer/measurement-matters

Foddy, W. (1993). Constructing questions for interviews and questionnaires. *Theory and Practice in Social Research*. London: Cambridge University Press.10.1017/CBO9780511518201

Gigerenzer, G. (1998). Surrogates for theories. *Theory & Psychology*, *8*(2), 195–204. https://doi.org/10.1177/0959354398082006

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Handley-Miner, I., Bochynska, A., Buttliere, B., Chung, K. L., Elsherif, M., Graham, C., Heiser, N., Loenneker, H., Sætrevik, B., Seetahul, Y., Sulik, J., Tomczak, J., Zygar-Hoffmann, C., & Neff, M. B. (2025). A call for greater transparency in piloting. *PsyArXiv Preprint*. https://osf.io/q95zn_v1

Hauser, D. J., Ellsworth, P. C., & Gonzalez, R. (2018). Are manipulation checks necessary? *Frontiers in Psychology*, 9, Article 998. https://doi.org/10.3389/fpsyg.2018.00998

Heo, M. (2014). Impact of subject attrition on sample size determinations for longitudinal cluster randomized clinical trials. *Journal of Biopharmaceutical Statistics*, *24*(3), 507–522. https://doi.org/10.1080/10543406.2014.888442

Hundley, V, Milne, J, Leighton-Beck, L. et al. (2000). Raising research awareness among midwives and nurses: does it work? *Journal of Advanced Nursing*, *31*(1), 78–88. https://pubmed.ncbi.nlm.nih.gov/10632796/

In, J. (2017). Introduction of a pilot study. *Korean Journal of Anesthesiology*, 70(6), 601–605. https://doi.org/10.4097/kjae.2017.70.6.601

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

Irwin, D. E., Varni, J. W., Yeatts, K., & DeWalt, D. A. (2009). Cognitive interviewing methodology in the development of a pediatric item bank: a patient reported outcomes measurement information system (PROMIS) study. *Health and quality of life outcomes*, 7, 3. https://doi.org/10.1186/1477-7525-7-3

Jaeger, B. (2023, April 19). *Stimulus-hacking: Stimulus variability increases the need for preregistration* [Conference presentation]. *Plymouth Meeting*, Experimental Psychology Society, Plymouth, England. https://eps.ac.uk/wp-content/uploads/2023/04/EPS-Plymouth-Meeting-Programme-April-v2.pdf

Kenny, D. A. (2019). Enhancing validity in psychological research. *American Psychologist*, *74*(9), 1018–1028. https://doi.org/10.1037/amp0000531

Kenny, A. R. (2021). Commentary on the beyond WEIRD special issue: The importance of open research practices to empirical research in the evolutionary social sciences. *Evolution and Human Behavior*, *42*(3), 268–270. https://doi.org/10.1016/j.evolhumbehav.2021.02.008

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Kerschbaumer, S., Voracek, M., Aczél, B., Anderson, S. F., Booth, B. M., Buchanan, E. M., Carlsson, R., Heck, D. W., Hiekkaranta, A. P., Hoekstra, R., Karch, J. D., Lafit, G., Lin, Z., Liu, S., MacKinnon, D. P., McGorray, E. L., Moreau, D., Papadatou-Pastou, M., Paterson, H., … Tran, U. S. (2025). VALID: A checklist-based approach for improving validity in psychological research. *Advances in Methods and Practices in Psychological Science*, *8*(1), 25152459241306432. https://doi.org/10.1177/25152459241306432

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr., R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating variation in replicability: A “Many Labs” replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr., R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … et al. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods* *and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Kunselman, A. R. (2024). A brief overview of pilot studies and their sample size justification. *Fertility and Sterility*, *121*(6), 899–901. https://doi.org/10.1016/j.fertnstert.2024.01.040

Lakens, D. (2019). The value of preregistration for psychological science: A conceptual analysis. 心理学評論 (*Japanese Psychological Review*), *62*(3), 221–230. https://doi.org/10.24602/sjpr.62.3_221

Lakens, D. (2025). Concerns about replicability across two crises in social psychology. *International Review of Social Psychology*, *38*(1). https://doi.org/10.5334/irsp.1036

Lakens, D., Mesquida, C., Rasti, S., & Ditroilo, M. (2024). The benefits of preregistration and Registered Reports. *Evidence-Based Toxicology*, *2*(1), Article 2376046. https://doi.org/10.1080/2833373X.2024.2376046

Lakens D., Scheel A. M., & Isager P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Le, N. (2025, August 4). Development and validation of ecological momentary interventions to increase the current closeness motivation in romantic relationships (Version 1) [Open-ended registration]. *OSF Registries*. https://doi.org/10.17605/OSF.IO/SEJ64

Lehmann, E. L., & Romano, J. P. (2005). The general decision problem. In: *Testing statistical hypotheses* (3rd ed., pp. 3–28). Springer. https://doi.org/10.1007/0-387-27605-X

Lewis, M., Bromley, K., Sutton, C. J., McCray, G., Myers, H. L., & Lancaster, G. A. (2021). Determining sample size for progression criteria for pragmatic pilot RCTs: The hypothesis test strikes back! *Pilot and Feasibility Studies*, *7*(1), Article 40. https://doi.org/10.1186/s40814-021-00770-x

Li, G., Taljaard, M., Van den Heuvel, E. R., Levine, M. A. H., Cook, D. J., Wells, G. A., Devereaux, P. J., & Thabane, L. (2017). An introduction to multiplicity issues in clinical trials: The what, why, when and how. *International Journal of Epidemiology*, *46*(2), 746–755. https://doi.org/10.1093/ije/dyw320

Loenneker, H. D., Artemenko, C., Willmes, K., Liepelt-Scarfone, I., & Nuerk, H.-C. (2024). Deficits in or preservation of basic number processing in Parkinson's disease? A registered report. *Journal of Neuroscience Research*, *102*, e25397. https://doi.org/10.1002/jnr.25397

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

McDougal, E., Silverstein, P., Treleaven, O., Jerrom, L., Gilligan-Lee, K., Gilmore, C., & Farran, E. K. (2024). Assessing the impact of LEGO® construction training on spatial and mathematical skills. *Developmental Science*, *27*(2), e13432. https://doi.org/10.1111/desc.13432

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244. https://doi.org/10.2466/PR0.66.1.195-244

Miller, J. (2024). How many participants? How many trials? Maximizing the power of reaction time studies. *Behavior Research Methods*, *56*, 2398–2421. https://doi.org/10.3758/s13428-023-02155-9

Nagel, J., Morgan, D. P., Gürsoy, N. Ç., Sander, S., Kern, S., & Feld, G. B. (2024). Memory for rewards guides retrieval. *Communications Psychology*, *2*(1), 31. https://doi.org/10.1038/s44271-024-00074-9

Nagy, T., Hergert, J., Elsherif, M. M., Wallrich, L., Schmidt, K., Waltzer, T., Payne, J. W., Gjoneska, B., Seetahul, Y., Wang, Y. A., Scharfenberg, D., Tyson, G., Yang, Y.-F., Skvortsova, A., Alarie, S., Graves, K., Sotola, L. K., Moreau, D., & Rubínová, E. (2025). Bestiary of questionable research practices in psychology. *Advances in Methods and Practices in Psychological Science*, *8*(3), 25152459251348431. https://doi.org/10.1177/25152459251348431

Neff, M. B. (2024). *Challenging the Metaphor Deficit: Unpacking Children's 'Literally Biased' Communicative Development*. [Doctoral dissertation, University of Oslo]. 10.13140/RG.2.2.36584.17921

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*. Series A, Containing Papers of a Mathematical or Physical Character, *231*(694–706), 289–337. https://doi.org/10.1098/rsta.1933.0009

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS Biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*(Volume 73, 2022), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Oľhová, S., Lášticová, B., Kundrát, J., & Kanovský, M. (2023). Using fiction to improve intergroup attitudes: Testing indirect contact interventions in a school context. *Social Psychology of Education*, *26*(1), 81–105. https://d-nb.info/1278797513/34

Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136–144. https://doi.org/10.7334/psicothema2013.259

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Pfaffel, A., Kollmayer, M., Schober, B., & Spiel, C. (2016). A missing data approach to correct for direct and indirect range restrictions with a dichotomous criterion: A simulation study. *PLOS ONE*, *11*(3), e0152330. https://doi.org/10.1371/journal.pone.0152330

Pilot Reporting Task Force. (2024). Piloting practices across psychological sub-disciplines. *PsyArXiv Preprint*. https://doi.org/10.17605/OSF.IO/3QDY2

Pit, I. L., Kenny, A. R., & Fortunato, L. (2024). Validation of materials for replication of the field-based experiments in Maass et al. (1989): A mixed-methods pilot study. *OSF Preprint*. https://osf.io/hdrx6/files/u2b58

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Roth, P. L., Bobko, P., & McFarland, L. A. (2005). A meta-analysis of work sample test validity: Updating and integrating some classic literature. *Personnel Psychology*, *58*(4), 1009–1037. https://doi.org/10.1111/j.1744-6570.2005.00714.x

Rubin, M. (2025). Preregistration does not improve the transparent evaluation of severity in Popper's philosophy of science or when deviations are allowed. *Synthese*, *206*(111), 1–25. https://doi.org/10.1007/s11229-025-05191-4

Sarafoglou, A., Kovacs, M., Bakos, B., Wagenmakers, E.-J., & Aczel, B. (2022). A survey on how preregistration affects the research workflow: Better science but more work. *Royal Society Open Science*, *9*(7), 211997. https://doi.org/10.1098/rsos.211997

Schäfer, S. J., Simsek, M., Jaspers, E., Kros, M., Hewstone, M., Schmid, K., ... & Christ, O. (2022). Dynamic contact effects: Individuals' positive and negative contact history influences intergroup contact effects in a behavioral game. *Journal of Personality and Social Psychology*, *123*(1), 107. https://pubmed.ncbi.nlm.nih.gov/34582243/

Schiavone, S. R., Quinn, K., & Vazire, S. (2023). A consensus-based tool for evaluating threats to the validity of empirical research. *PsyArXiv Preprint*. https://osf.io/fc8v3_v1

Schimmack, U. (2018, January 5). *Why the Journal of Personality and Social Psychology Should Retract Article DOI: 10.1037/a0021524 "Feeling the Future: Experimental evidence for anomalous retroactive influences on cognition and affect" by Daryl J. Bem*. https://replicationindex.com/2018/01/05/bem-retraction/

Schönbrodt, F. D., Zygar-Hoffmann, C., Nestler, S., Pusch, S., & Hagemeyer, B. (2022). Measuring motivational relationship processes in experience sampling: A reliability model for moments, days, and persons nested in couples. *Behavior Research Methods*, *54*(4), 1869–1888.

Shadish, W. R., Cook, T., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. MA: Houghton Mifflin Boston.

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

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2018, January 25). [68] Pilot-dropping backfires (So Daryl Bem Probably Did Not Do It). *Data Colada*. https://datacolada.org/68

Slack, M. K., & Draugalis, J. R. (2001). Establishing the internal and external validity of experimental studies. *American Journal of Health-System Pharmacy: AJHP: Official Journal of the American Society of Health-System Pharmacists*, *58*(22), 2173–2181.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). SAGE Publications.

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

Teare, M. D., Dimairo, M., Shephard, N., Hayman, A., Whitehead, A., & Walters, S. J. (2014). Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: A simulation study. *Trials*, 15, 264. https://doi.org/10.1186/1745-6215-15-264

Thabane, L., Ma, J., Chu, R., Cheng, J., Ismaila, A., Rios, L. P., Robson, R., Thabane, M., Giangregorio, L., & Goldsmith, C. H. (2010). A tutorial on pilot studies: The what, why and how. *BMC Medical Research Methodology*, 10, Article 1. https://doi.org/10.1186/1471-2288-10-1

Thomas, K. A., & Clifford, S. (2017). Validity and Mechanical Turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, *77*, 184–197. https://doi.org/10.1016/j.chb.2017.08.038

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science*, *13*(4), 411–417. https://doi.org/10.1177/1745691617751884

Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: improving the four validities in psychological science. *Current Directions in Psychological Science*, *31*(2), 162–168. https://doi.org/10.1177/09637214211067779

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wicherts, J. M. (2011). Psychology must learn a lesson from fraud case. *Nature*, *480*(7375), 7. https://doi.org/10.1038/480007a

Wolcott, M. D., & Lobczowski, N. G. (2021). Using cognitive interviews and think-aloud protocols to understand thought processes. *Currents in Pharmacy Teaching and Learning*, *13*(2), 181-188.

Wood, J. K., Anglim, J., & Horwood, S. (2021). A less evaluative measure of Big Five personality: Comparison of structure and criterion validity. *European Journal of Personality*, *36*(5), 809-824. https://doi.org/10.1177/08902070211012920

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685

Zygar-Hoffmann, C., Cristoforo, L., Wolf, L., & Schönbrodt, F. D. (2022). Eliciting short-term closeness in couple relationships with ecological momentary interventions. *Collabra: Psychology*, *8*(1), 38599.

Author Contributions:

Seetahul, Y. – Conceptualization, Writing, Drafting, Editing, Administration
Elsherif, M. M. – Conceptualization, Writing, Editing, Administration
Zygar-Hoffmann, C. – Conceptualization, Writing, Editing
Wallrich, L. – Conceptualization, Writing, Editing
Silverstein, P. – Conceptualization, Writing, Editing
Sætrevik, B. – Conceptualization, Writing, Editing
Pit, I. L. – Conceptualization, Writing, Editing
Lönneker, H. D. – Conceptualization, Writing, Editing
Heiser, N. H. – Conceptualization, Writing, Editing
Handley-Miner, I. J. – Conceptualization, Writing, Editing
Graham, C. J. – Conceptualization, Writing, Editing
Chou, Y. Y.– Conceptualization, Writing, Editing
Buttliere, B. – Conceptualization, Writing, Editing
Bochynska, A. – Conceptualization, Writing, Editing
Beitner, J. – Conceptualization, Writing, Editing
Neff, M.B. – Conceptualization, Writing, Drafting, Editing, Administration, Supervision

Acknowledgements:

We thank Justin Sulik and other members of the Pilot Reporting Task Force (https://pilotreportingtf.github.io) for their early feedback on the conceptualization of this work, as well as Claudia Dörrler for helpful comments on an earlier draft of the manuscript.

Funding:

The authors received the following financial support for the publication of this article: M.M.E. was funded by the Leverhulme Early Career Research Fellow RM56G0344/44143; P.S. was co-funded by Horizon Europe (101087416), the EU Recovery and Resilience Facility, and Portuguese national funds via FCT – Fundação para a Ciência e a Tecnologia, under projects, LA/P/0058/2020 [DOI: 10.54499/LA/P/0058/2020], UID/PRR/04539/2025 [DOI: 10.54499/UID/PRR/04539/2025] and UID/04539/2025; I.J.H.M was funded by the Alfred P. Sloan Foundation (G-2024-22462).

Conflict of Interest:

The authors have no conflict of interest to disclose.