METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Evaluate What Is Claimed to Be Confirmed: Initial Version of a Falsification Assessment Form (FAF)

Michael Höfler[1], Anja Kräplin[2], Mahmoud Medhat Elsherif[3,4], Moritz Schepke[1], Maria Montefinese[5], Yashvin Seetahul[6], Bjørn Sætrevik[7], Aaron Peikert[8], Marton A. Varga[9], Lukas Wallrich[10]

[1] *Clinical Psychology and Behavioural Neuroscience, Institute of Clinical Psychology and Psychotherapy, Technische Universität Dresden, Dresden, Germany.* **[2]** *Department of Psychiatry and Psychotherapy, Technische Universität Dresden, Dresden, Germany.* **[3]** *School of Psychology and Vision Sciences, University of Leicester, Leicester, United Kingdom.* **[4]** *School of Psychology, University of Birmingham, Birmingham, United Kingdom.* **[5]** *Department of Developmental Psychology and Socialisation, University of Padua, Padua, Italy.* **[6]** *Institute for Psychology, University of Innsbruck, Innsbruck, Austria.* **[7]** *Department of Psychosocial Science, Faculty of Psychology, University of Bergen, Bergen, Norway.* **[8]** *Max Planck Institute for Human Development Center for Lifespan Psychology, Berlin, Germany.* **[9]** *ELTE Eötvös Loránd University, Institute of Psychology, Budapest, Hungary.* **[10]** *Birkbeck Business School, Birkbeck, University of London, London, United Kingdom.*

Methodology, 2025, Vol. 21(3), 180–196, https://doi.org/10.5964/meth.17705

Received: 2025-04-15 • **Accepted:** 2025-09-01 • **Published (VoR):** 2025-09-30

Handling Editor: Pablo Nájera Álvarez, Universidad Pontificia Comillas, Madrid, Spain

Corresponding Author: Michael Höfler, Clinical Psychology and Behavioural Neuroscience, Institute of Clinical Psychology and Psychotherapy, Technische Universität Dresden, Dresden, Germany. Chemnitzer Straße 46, 01187 Dresden, Germany. Tel: +49-351-46936921. E-mail: michael.hoefler@tu-dresden.de

Supplementary Materials: Materials, Preregistration [see Index of Supplementary Materials]

Abstract

Scientific claims, and the way they are tested, must be unambiguous and flexibility must be disclosed. Grounded in Popper's principle of falsification, we suggest the *Falsification Assessment Form (FAF)*. The form aims to identify ambiguity and undisclosed flexibility in the entire research process with 11 items covering hypothesis formulation, data processing, analysis, and alternative explanations. It also collects information on transparency measures, such as preregistration. The form was developed through consensus among the authors and refined via a collaborative feedback assessment of 19 experts. It is intended for original, quantitative research, it highlights potential issues and requires authors to provide detailed responses. *FAF* is meant to be a structured qualitative audit framework. It can be used to identify concerns in published research, improve the

quality of papers during peer review, or guide rigorous study planning from the outset. We open up further refinement and testing of *FAF* to the scientific community.

Keywords

falsification, peer review, research assessment, research quality, research transparency

Science is commonly defended as a self-correcting enterprise (Ioannidis, 2012). This notion is heavily shaped by Karl Popper's (1959) seminal work on falsification. Falsification is considered by many as the most compelling answer to the epistemic problem of how to learn from data (the particular) about underlying phenomena (the general). The common focus on falsifiability, the capacity to demonstrate false propositions are indeed false, encompasses other, seemingly different approaches to science (Feyerabend, 1993; Gigerenzer & Marewski, 2015; Lakens & DeBruine, 2021; Mayo, 2018; Meehl, 1978; Uygun Tunç & Tunç, 2023). At its core, Popper's principle asks what observation would refute a hypothesis and then seeks that observation. In confirmatory research, the principle translates to a hypothesis or theory predicting a specific result. Observing this result corroborates the hypothesis, while (repeatedly) observing a different result may falsify it. The research process, therefore, must facilitate the ability for false hypotheses to turn out wrong. Despite widespread endorsement of the principle, numerous authors have highlighted its compromise by poor practices throughout the research process. These include vague or ill-defined hypotheses (Devezer & Buzbas, 2021; Oberauer & Lewandowsky, 2019), inadequate or under-reported data processing (Loenneker et al., 2024; Scheel, 2022), and problematic data analysis (Gigerenzer, 2004; Nagy et al., 2024). The key issue is the flexibility with which these procedures are carried out. Without a clear, preregistered prediction, researchers can select results that seemingly support their hypothesis while ignoring others — a practice known as "cherry-picking" (Lakens & DeBruine, 2021).

Over the past decades, valuable tools for evaluating research have been developed. These assess bias (Viswanathan et al., 2018), the confidence in a claim (Alipourfard et al., 2021), its replicability (Fraser et al., 2023), reporting standards (Appelbaum et al., 2018) or provide checklists on general research quality (Héroux et al., 2022; Kerschbaumer et al., 2025; Nosek et al., 2015; Wicherts et al., 2016) and transparency measures (Aczel et al., 2021; Nanyang Technological University Library, 2023).

Here, we introduce the *Falsification Assessment Form (FAF)*, a tool designed to evaluate the falsifiability of a published claim on a hypothesis, and to foster falsifiability in the planning stage. Unlike existing tools, *FAF* is based on a single, unifying principle — falsifiability. It refrains from quantifying the impact of the identified issues, a practice that has been criticized for its arbitrariness and high context-dependence in the absence of specific knowledge on how the identified issues correlate with the quantity of interest (Greenland & O'Rourke, 2001; Herbison et al., 2006).

Thus, *FAF* aims to operationalize Popper's seminal idea through a practical form for behavioural, cognitive, social and health sciences. It serves to inform the assessment of a published paper, study planning, manuscript preparation, or manuscript review.

Method

Approach of the FAF

The form evaluates the falsifiability of a single claim in support of a hypothesis that is likely to be taken to be confirmatory (e.g., "our study suggests…" or "we found evidence that…"; Höfler et al., 2022). It is not suitable for inconclusive (e.g., "our study revealed unclear results…") and exploratory claims (e.g., "we propose the new hypothesis that…"; Höfler et al., 2022). While flexibility in the research process opens the door to the discovery of novelty in exploratory research, flexibility in confirmatory research must be entirely constrained. A confirmatory claim can be typically extracted from the abstract, discussion, or conclusion of a paper. In the planning phase of a study, *FAF* can guide the choice of rigorous research methods, so that once the study is conducted, the claims made would pass the *FAF* items (except possibly Domain 4, see below).

The form is to be filled out once for every claim in a paper and can be used several times for different claims. For composite claims (e.g., "Therapy A is effective, but Therapy B is not"), it is advisable to evaluate each part separately if their implications for theory building or intervention are not the same (Bender & Lange, 2001). *FAF* covers original, quantitative research; it does not apply to meta-analyses, re-analyses, or other study types. Additionally, it does not evaluate the hypothesis's substantive quality and relevance.

Within these limits, *FAF* aims to *uncover as many issues as possible*, using only *a minimum set of straightforward questions*. It returns a list of *potential issues*, which are assumed to be highly context-dependent. The form does not judge these issues but encourages authors and reviewers to address them — ideally before publication. In this sense, *FAF* is meant to be normative. Crucially, it does not constitute a measurement scale or diagnostic instrument requiring high inter-rater reliability, but a structured qualitative audit framework for identifying ambiguities or undisclosed flexibilities that could compromise falsifiability.

Each item contains a broad question paired with an example to illustrate its scope. For example, Item 1.1 asks *"Are there any ambiguities about the meaning of the hypothesis or the conditions under which it is claimed to hold?"* and then explains *"The formulation of a hypothesis must eliminate flexible interpretations. A common instance for flexibility is ambiguity about whether the hypothesis concerns a causal or associational relation. Besides, the hypothesis has to include the conditions under which it holds: the population, materials, stimuli, design procedures and outcomes used."* A broad question, together with the exam-

ple, is intended to prompt more specific considerations that may affect how the question is answered. A potential issue then is raised by endorsing a concern, or by highlighting that information is ambiguous or missing, which is to be described in a free text field. In doing so, we expect a paper or its supplementary material to report anything in the research process where flexibility in procedures could undermine falsifiability. We also expect the reporting to be *verifiable*, wherefore the form ends by evaluating if these procedures have been carried out as reported through a final domain on the transparency measures used (Lakens, 2019; Lakens & Mesquida, 2024).

Although the items on the wording of a hypothesis, data processing, data analysis (Domains 1 – 3) and transparency measures (Domain 5) are designed to elicit a fairly objective assessment, it is not a concern if some users of the form identify more potential issues than others. *FAF* aims to identify as many of them as possible, not to assert that they are all critical. Domain 4, *"Alternative explanations for the claim, not addressed so far"* includes a single free-text item in which any alternative explanations can be listed. This considers the well-known fact that there are always several possibilities for why a hypothesis may be wrong, and a single study can never test all of them (e.g., the effect does not exist, the measurement instrument is not valid; Meehl, 1967; Rakover, 2003). We therefore expect that this item will only remain blank if a paper presents a set of studies that, together, rigorously test all plausible alternative assumptions. Finally, the last domain *"Transparency measures"* asks in detail about preregistration and registered reports, specifically whether they have been timestamped before data access. Other transparency measures are listed in checklist format: the use of open data, open materials, open analysis (code), a reproducibility check, the 21-word solution, or "anything else". For each transparency measure, it is asked whether it was implemented *"in a way that is insufficient to assess falsifiability"* and, if so, a free text explanation of why it is insufficient is requested.

Table 1 provides a full list of each domain's items, i.e., items in FAF (Version 1.0).

Table 1A

Domain 1: Content of the Hypothesis

Item	Content
1.1	What is the hypothesis that the paper claims (concludes) to confirm?
1.2	Are there any ambiguities about the meaning of the hypothesis or the conditions under which it is claimed to hold?

Table 1B

Domain 2: Data Processing, Choice, and Coding of Variables That Entered the Analysis

Item	Content
2.1	Are there undisclosed analytic flexibilities in the data processing, starting from the raw data, selecting measurements, aggregating them into variables (e.g. scales), transforming and categorizing the variables before conducting statistical tests? • Flexibility in data processing from the raw material (e.g., videos, questionnaire items). • Flexibility in the selection or categorization of variables (e.g., dichotomization of actually interval-scaled variables, use of cut-offs, exclusion of items on scales, choice between multiple measures of the same construct). • Flexibility in the exclusion of individuals (e.g. as outliers or inattentive respondents).

Table 1C

Domain 3: Data Analysis and Interpretation

Item	Content
3.1	Is there ambiguity about which of the analyses carried out in the paper relate to the prediction made by the hypothesis and justify the claim?
3.2	Is it unclear how the prediction leads to exactly the analysis (analyses) that were conducted?
3.3	Is it unclear what results, other than those reported, would have led to the opposite conclusion on the hypothesis? What is the decision rule? How could different results lead to supporting instead of rejecting a hypothesis, or conversely, lead to rejecting rather than supporting a hypothesis?
3.4	If multiple analyses were conducted to test the hypothesis, the results may be combined in different ways to decide on the overall confirmation of the hypothesis. How did the paper combine results?
3.5	Among all the assumptions in the statistical test or model, are there any that are clearly inaccurate and favour the claimed hypothesis?
3.6	Are there parts of the interpretation of the results (as stated in the claim) that have not been explicitly tested for? For instance, is the claim on the hypothesis based only on a crude comparison of p-values?

Table 1D

Domain 4: Alternative Explanations for the Claim, Not Addressed So Far

Item	Content
4.1	What are alternative explanations (not addressed in the previous domains), could account for the finding if the hypothesis was false? List explanations that were not sufficiently addressed.

Table 1E

List of Transparency Measures Used

Item
Was a preregistration or Stage 1 registered report done? In case of preregistration, was it before access to data, verifiable by a time-stamp for either preregistration or data-collection?
Preregistration or Stage 1 registered report that mentions:
• a stopping rule for data collection (by date or sample size achieved). • the hypothesis and the conditions under which it is claimed to hold. • raw data processing and computation of all relevant variables. • inclusion and exclusion criteria. • the analyses that exactly inform the decision on the hypothesis. • anything else.
Checklist:
• Open data. • Open materials. • Open analysis/notebook/code. • Reproducibility check. • Authors have confirmed "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study." (the 21-word solution). • Anything else?

After completing the form, *FAF* returns the following results: The extracted claim with paper information (including title and year of publication), a domain-wise listing of "*Identified issues of the claim which might have impaired falsifiability and require consideration*" and a summary of "*transparency measures used*", which might include identified flexibilities, e.g., in the preregistration, that remained undisclosed.

Implementation

FAF was implemented as a Google Sheet (FAF Research Group, 2025). It consists of the sheets *"Background"* (including "*Instructions*"), *"Form"* and *"Results"*. Clicking the form link creates a private copy, ensuring that the data is only stored in one's private Google account. The results sheet can be stored in another spreadsheet or PDF format. (Höfler, Kräplin, Elsherif et al., 2025).

Evolution of the Form

MH had the initial idea in late 2023 to create a form that would shed light on how well a scientific claim could have turned out differently. He drafted a concept and initial items and preregistered the form on the Open Science Framework (Höfler, 2023). AK piloted the items by reviewing a couple of papers. MH presented the idea at the SIPS (Society

for the Improvement of Psychological Science) conference in June 2023, in person and online. Fourteen colleagues joined the project at that point, and a first draft of the *FAF* was agreed upon by consensus in October 2024, after several online meetings. *FAF* was then implemented on Google Sheets by MS, who then joined the project. The ten authors of this paper remained involved in the project ever since.

Collegial Feedback Assessment

After completing the draft, we decided to collect feedback on the form from a wider group of experts. Doing so, we aimed to gauge how the form is received by experts and, above all, to identify any major inaccuracies and incomprehensibilities. The feedback form included the following items:

- *"Have you filled in the Falsification Assessment Form before answering this feedback sheet?"* ("*Yes*"/ "*No answer*"/ "*No*").
- *"What is your general impression of the form, its purpose and approach and structure?"* (Free text assessment).
- *"In what context have you used or would you consider using the form?"* (Multiple options allowed).
  - "*Assessing a published paper.*"
  - "*Reviewing a paper.*"
  - "*Planning research.*"
  - "*Others (specify).*"
- *"Do you have thoughts or suggestions about specific items? Please indicate the number or topic of an item."* (Free text assessment)
- *"Should you be inclined to allow us to contact you for a more in-depth discussion of your suggestions, please leave your email address. This will assist us in following up on particularly noteworthy but unclear ideas."*

We decided to evaluate these questions with a collegial feedback assessment. This anonymous format was chosen to encourage open and honest responses while protecting participant confidentiality. As a result, no identifying information about the respondents was collected. For the development of the initial version of *FAF*, we considered it sufficient that all participants were recognized as experts, as each was personally invited by a member of the *FAF* project. These invitations were made independently to ensure at least a minimal degree of diversity in perspectives. The *FAF* members were instructed to select experts based on predefined criteria: *"You consider somebody to be an expert in meta-science or methods (based on their publications or teaching)."* For group invitations, the criteria included: "*Open science initiatives, meta-science groups, teams working to improve scientific practices and methods, and journal editors."* In addition, members were explicitly instructed not to invite anyone who had previously been involved in the project.

Feedback was provided anonymously and no data beyond the responses to the survey items were collected or stored. However, the final item gave participants the option to leave their email address if they wished to discuss their input further. By doing so, they voluntarily waived their anonymity to ensure their concerns could be properly addressed. Invitations to participate were sent out between 3 and 31 December 2024, and feedback was collected until 31 January 2025. Eighty-five colleagues from psychology, 27 colleagues from other disciplines, and 13 groups were invited by email. For data protection reasons and since each member sent out invitations, we did not create a central database of invitees. Therefore, some overlap in the invitations may have occurred and the number of unique recipients may be slightly lower.

Results

A total of 19 responses were returned. Five respondents confirmed that they had completed the *FAF* before providing feedback. However, as all comments were phrased in a way that suggests familiarity with the *FAF*, no responses were excluded. Free text responses on the item "*general impression*" ranged from full support for the form and its content, to complete disagreement. Eight out of nineteen indicated that the form is too long or complex. Other points of criticism included ambiguity about the form's purpose and, in the "*Suggestions for specific items*" section, ambiguity about how individual items related to falsifiability, as well as concerns about the content and wording of specific items.

When asked what they would consider using the form for, the following frequencies were reported (multiple responses allowed): 5 for "*Reviewing a paper*" and "*planning research*", 4 for "*Evaluating a published paper*", 1 for "*Teaching*", 6 did not respond and 1 stated that she or he would not use it at all.

The authors of this paper reviewed all the comments before deciding how to revise the *FAF*.

No formal criteria were used to decide which items to retain, revise, or remove; instead, we aimed to remain open to all suggestions and evaluated each on its merits. Through several online meetings, we discussed the comments in detail and reached decisions by group consensus. There was unanimous agreement to streamline and condense the form and to clarify its primary objective. Domain 1 ("*Claim content*") was reduced from eight items to three. Domain 3 ("*Data analysis and interpretation")* was shortened from 13 items to six. As Domain 2 (a simple checklist for "*Data processing*") received little feedback, it was only slightly shortened. Domain 4 ("*Alternative assumptions*") and the checklist "*List of Transparency Measures Used*" remained essentially unchanged. In addition, the background text, the instructions and all remaining items were thoroughly revised, condensed, and clarified. The streamlining and rewording was largely guided by one feedback suggestion to focus more on the prediction that a hypothesis makes. For example, in Domain 1, the item *"Is there any doubt about the theoretical foundation of the*

hypothesis?" was deemed no longer necessary. We also removed the item *"Is there any doubt that the hypothesis was fully declared without being affected by the data?"* because we believe this issue is already addressed — more efficiently — in the final domain on transparency measures, with some modifications to the wording in that section. Also, several conceptual terms were hyperlinked to entries in the Framework for Open and Reproducible Research Training (FORRT) glossary (Parsons et al., 2022), where they are explained in more detail.

Following these revisions, the nine out of 19 colleagues who had provided their email address were contacted again and asked if they had any "*further suggestions*" for the revised version of the form. Seven of them responded to the email: three of them expressed support for the revision, three were at least not critical and one still held substantially different views. No further changes were made after this. All previous versions of *FAF* (0.1 to 0.5), along with a summary file documenting the changes from Version 0.5 (pre-feedback) to Version 1.0 (post-collegial feedback assessment), titled "*FAF Changes After Collegial Feedback Assessment*", are available in the *FAF* Open Science Foundation (OSF) repository under the folder "*Versions of FAF*" (Höfler, Kräplin, Varga et al., 2025).

The repository also contains some examples of claims evaluated using *FAF*. These include five evaluations by the authors of this paper of the claim "School bullying results in poor psychological conditions", based on a survey of 95,545 Chinese school students (Zhao et al., 2024). While there was general agreement on most binary items, notable differences emerged regarding whether the hypothesis made a clear prediction (Item 1.2) and whether the results were interpreted beyond what was warranted (Item 3.6). These discrepancies stemmed from differing views on the clarity of the target population and whether the claim implied causality. Differences in free-text assessments offered distinct descriptions of the concerns identified.

These results show that multiple raters identify more potential problems, which naturally and desirably leads to the need for more justification.

Discussion

With *FAF,* we have introduced a tool that seeks to uphold Popper's (1959) longstanding principle of falsifiability in the practices along the research process. The tool scrutinizes the falsifiability of a paper's confirmatory claim about a hypothesis through posing items on 11 sections to be answered. Ideally, the tool informs the study planning, design, and analysis, so that such a claim is later justified if the hypothesis is true and the results turn out as predicted.

Some Illustrative Use Cases

Before discussing *FAF* in detail, some examples of its use are given in Table 2.

Table 2

Use Cases for FAF

Type of use	Example
Assessing a claim in a published paper	A paper claims that "Tai Chi practice enhances self-esteem." While plausible, the reader questions how easily this conclusion might have turned out otherwise. Using *FAF*, she evaluates how robust the evidence is for this claim.
Reviewing a paper	A reviewer notices inconsistent standards in how he evaluates manuscripts. He adopts *FAF* to systematize and standardize his assessments.
Revising a paper	An author uses *FAF* to anticipate reviewer concerns and revise the manuscript accordingly, especially in how the study procedures are reported. While study specifications cannot be changed post hoc, reporting can be clarified.
Planning a study	A researcher aims to generate strong evidence for a hypothesized effect. She uses all *FAF* domains during planning to minimize potential issues, ensuring the hypothesis is clearly formulated (Domain 1), data processing and analysis are predefined (Domains 2 & 3), and the process is fully transparent (last domain). Alternative explanations (Domain 4) are either tested or explicitly acknowledged.
Research methods teaching	Students are asked to formulate a hypothesis and design a study using *FAF* to ensure that it could be falsified. This encourages critical thinking and highlights that while issues can be addressed in all domains, Domain 4 ("Alternative explanations") inherently remains open in single studies.

Qualitative Versus Quantitative Assessment of Falsifiability

FAF is a qualitative tool. It identifies *potential issues* without asserting that a real issue has been brought up. This reflects that the *FAF* items address topics whose impact on falsifiability may depend on context and subjective judgement. For example, the "*meaning of a hypothesis*" may be unclear to one researcher but not to another, or clarity may be achieved by explanation in response to completing the form (e.g., an unpublished claim on a hypothesis can be revised to achieve clarity). Potential issues flagged by *FAF* may be resolved through additional explanation provided in response to filling out the form. While we believe that *FAF* meets its goal if it stimulates the impetus to address such instances, it is natural to ask whether its qualitative approach could be extended towards a quantitative scoring of a claim's trustworthiness.

The quantitative counterpart to falsifiability is the *severity of a test*, the probability that a false hypothesis turns out to be false (Lakens, 2019; Mayo, 2018). This quantity appeals for its clarity of interpretation, but it is extremely difficult to calculate in practice, because it depends on numerous factors whose impact is difficult to determine,

including those covered by *FAF*. Such factors involve bias in analysis — for example, due to unconsidered correlations between observations (see Item 3.2: "*Among all the assumptions in the statistical test or model, are there any that are clearly inaccurate and favour the claimed hypothesis?*"). However, bias is highly context-dependent, and quantifying its extent requires a profound understanding of the mechanisms that produced a particular dataset, like the true magnitude of unconsidered correlations or features of measurement, selection, and unaccounted-for confounders (Greenland, 2005).

Nevertheless, anchoring it in the extremely simple cases in which computing the severity of a test is feasible helps to get an idea on its wider range in more complex, realistic scenarios. Mayo (2018) has elaborated much on severity calculations under the unrealistic assumptions of the absence of any bias (Greenland, 2005) and questionable research practices such as *p*-hacking and HARKing. In this case, and if a claim on a hypothesis is based on a single (frequentist) statistical test, severity is as follows:

- For a statistically significant test result ($p < \alpha$) in favour of, for example, claiming an effect, severity equals $1-\alpha$.
- For a nonsignificant test result in favour of no effect ($p \geq \alpha$), severity is $1-\beta$, where β is actually a (monotonously growing) function of the unknown true effect size.

If the claim relies on multiple tests — as addressed in *FAF* Item 3.4 — and the decision rule requires that *at least one* out of k tests be passed (Item 3.3), severity *decreases*. If the claim is in favor of an effect, severity ranges between $(1 - k \cdot \alpha)$, assuming independent tests, and $(1 - \alpha)$, assuming completely correlated tests. If the claim is against the effect, severity ranges between $(1 - k \cdot \beta)$ and $(1 - \beta)$. Conversely, if the decision rule requires that *all tests* be passed, severity is *higher*. In that case, severity in favour of the effect ranges between $(1 - \alpha)^k$ and $(1 - \alpha)$, depending on the degree of dependence between tests. The analogous range for claims against the effect is between $(1 - \beta)^k$ and $(1 - \beta)$.

Questionable practices reduce severity, and the more ambiguities and flexibilities remain undisclosed (i.e., the more potential issues *FAF* identifies), the more room there is for severity to fall below the above values (Lakens, 2019). In cases of extensive fishing through flexible hypothesis formulation, data processing or analysis, severity can even approach 0 (Head et al., 2015; Simmons et al., 2011). Such a claim is essentially unfalsifiable. When a potential issue is identified by endorsing a *FAF* item — for example, unclear wording of the hypothesis (Item 1.2) or a lack of preregistration or insufficient detail in it (domain on transparency measures) — we do not know which questionable research practices, if any, have been employed, or to what extent. These unknowns determine the degree to which severity is reduced.

While these considerations apply only to specific situations and yield broad severity ranges, we strongly encourage mathematical elaborations of severity in fairly more complex scenarios, particularly regarding the impact of issues addressed by specific *FAF* items.

Strengths and Limitations

A key strength of *FAF* is its foundation in Popper's principle of falsifiability, the most widely accepted scientific standard for hypothesis testing. *FAF's* qualitative approach highlights potential issues without making direct judgments about the validity of a claim. This reduces the risk of arbitrary or overly context-dependent evaluations. Furthermore, since falsifiability is a universal principle of scientific inquiry, *FAF* provides wide usability, allowing researchers to apply it without adjustments in different contexts and across disciplines. Its areas of usage include manuscript preparation, peer review, and post-publication evaluation. While many existing approaches are restricted to post hoc evaluation of compliance with certain reporting standards or methodological benchmarks, *FAF*, when used yet at the planning stage, promotes better research design from the outset and encourages researchers to formulate hypotheses that can be meaningfully tested and potentially refuted. Since *FAF* was developed through a consensus among the authors and refined via a collaborative feedback process involving 19 experts from diverse backgrounds, major weaknesses and inaccuracies in the form should have been identified. The result aims to strike a balance: while it does not identify every possible issue, it captures many and remains concise and accessible to a broad range of researchers.

Despite these strengths, *FAF* has limitations. It is specifically designed for original, quantitative research and does not apply to meta-analyses, re-analyses, or other study types. Assessing falsifiability in such research requires other approaches, though elements of *FAF* may still be useful. For example, in meta-analysis, assessing reproducibility is a major challenge (Maassen et al., 2020), yet flexibilities in the research process must still be disclosed (e.g., predefining the hypothesis, data processing, and analysis before accessing the data). To illustrate *FAF* items, we used examples from the behavioural sciences because they should be familiar to everyone. However, we would like to emphasise that the same items apply to data-intensive research such as neuroimaging or ecological momentary assessment. These fields are not exempt from falsifiability standards simply because they pose greater practical challenges.

As a second limitation, the effectiveness of *FAF* depends on researchers' willingness to engage critically with the identified issues. While the form encourages users to document and reflect on potential ambiguities, there is no mechanism to enforce proper resolutions of these concerns. *FAF* includes a section on transparency measures, but this only addresses whether the reported procedures *could* be verified, not whether they have actually been verified.

Finally, while the expert feedback process contributed valuable insights, the relatively small number of responses (n = 19) means that the revisions to the form were based on a limited set of perspectives. Therefore, the version presented in this paper represents only the initial release — Version 1.0 — of *FAF*. We invite the broader scientific community to contribute to its further development. To facilitate this, we have assigned it a Creative Commons Attribution license (CC BY) and created a GitHub repository

(Höfler, Kräplin, Elsherif et al., 2025), which will allow for the collection of community feedback that can guide future refinements. Further systematic evaluation should aim to achieve broader validation, including large-scale testing across different disciplines. Additionally, interrater agreement should be examined to identify which items — and in which scientific contexts — researchers are most likely to disagree. Such insights could inform improvements to *FAF*'s utility and generalizability.

Conclusion

FAF is beneficial because it avoids the heuristic approach to complexity and bias of multi-criteria assessments by focusing on a single, fundamental principle — falsifiability. This makes it broadly applicable and theoretically grounded, while also encouraging more rigid and less biased research practices. Possible future extensions include training artificial intelligence to automatically extract the relevant information from papers. AI-based large-scale meta-science could analyse thousands of publications to identify not only research fields with particularly high numbers of "not even wrong" hypotheses (Scheel, 2022), but also steps in the research process with particularly frequently undisclosed flexibilities. These insights could then inform the development of targeted research tools and teaching approaches to close the flexibilities. Ultimately, this could encourage the risk-taking that researchers need for conducting studies capable of providing robust evidence.

Funding: The authors have no funding to report.

Acknowledgments: Thanks a lot to the following colleagues who have contributed with their hints and advice: Tao Coll-Martín, Lisa DeBruine, Simona Haasova, Rink Hoekstra, Aaron Peikert, Jill de Ron, Danilo Calero Sequeira, Gilad Feldman, Nicklas Hafiz, Alex Holcombe, Jürgen Hoyer, Amélie Gourdan Kanhukamwe, Daniël Lakens, Niclas Jacobs, Philipp Kanske, Robert Miller, Gerit Pfuhl, Merle-Marie Pittelkow, Priya Silverstein, Anna van 't Veer.

Competing Interests: The authors have declared that no competing interests exist.

Data Availability: The data collected for this study are not publicly available because participants did not consent to data sharing.

Supplementary Materials

Type of supplementary material	Availability/Access
Data	
Data for this study are not publicly available.	—
Preregistration	
Preregistration for study.	Höfler (2023)
Code	
No code was provided for the study.	—
Material	
a) FAF development materials and version history.	Höfler, Kräplin, Varga et al. (2025)
b) Falsification Assessment Form (FAF) Questionnaire.	Höfler, Kräplin, Elsherif et al. (2025)

References

Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, Š., Benjamin, D., Chambers, C. D., Fisher, A., Gelman, A., Gernsbacher, M. A., Ioannidis, J. P., Johnson, E., Jonas, K., Kousta, S., Lilienfeld, S. O., Lindsay, D. S., Morey, C. C., Munafò, M., Newell, B. R., . . .Wagenmakers, E.-J. (2021). A consensus-based transparency checklist. *Nature Human Behaviour, 4*, 4–6. https://doi.org/10.1038/s41562-019-0772-6

Alipourfard, N., Arendt, B., Benjamin, D. M., Benkler, N., Bishop, M. M., Burstein, M., Bush, M., Caverlee, J., Chen, Y., Clark, C., Dreber Almenberg, A., Errington, T. M., Fidler, F., Field, S., Fox, N., Frank, A., Fraser, H., Friedman, S., Gelman, B., . . .Wu, J. (2021, May 4). *Systematizing confidence in open research and evidence (SCORE).* OSF Preprints. https://doi.org/10.31235/osf.io/46mnb

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board Task Force Report. *American Psychologist, 73*(1), 3–25. https://doi.org/10.1037/amp0000191

Bender, R., & Lange, S. (2001). Adjusting for multiple testing — When and how? *Journal of Clinical Epidemiology, 54*(4), 343–349. https://doi.org/10.1016/S0895-4356(00)00314-0

Devezer, B., & Buzbas, E. O. (2021). *Minimum viable experiment to replicate* [Preprint]. PhilSci Archive. https://philsci-archive.pitt.edu/24720/7/Minimum_Viable_Experiment_to_Replicate_preprint.pdf

FAF Research Group. (2025). *Falsification Assessment Form, Version 1.0.* [Spreadsheet] https://docs.google.com/spreadsheets/d/1a1pQ-jQYcBDAZ4p8Tpkq27NYl9p_bxvj-v_3MYq1nOA/copy

Feyerabend, P. (1993). *Against method* (3rd ed.). Verso.

Fraser, H., Bush, M., Wintle, B. C., Mody, F., Smith, E. T., Hanea, A. M., Gould, E., Hemming, V., Hamilton, D. G., & Rumpff, L. (2023). Predicting reliability through structured expert elicitation with the repliCATS (Collaborative Assessments for Trustworthy Science) process. *PLoS One, 18*(1), Article e0274429. https://doi.org/10.1371/journal.pone.0274429

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics, 33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Gigerenzer, G., & Marewski, J. N. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management, 41*(2), 421–440. https://doi.org/10.1177/0149206314547522

Greenland, S. (2005). Multiple-bias modeling for analysis of observational data. *Journal of the Royal Statistical Society. Series A, (Statistics in Society), 168*(2), 267–306. https://doi.org/10.1111/j.1467-985X.2004.00349.x

Greenland, S., & O'Rourke, K. (2001). On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics, 2*(4), 463–471. https://doi.org/10.1093/biostatistics/2.4.463

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of *p*-hacking in science. *PLoS Biology, 13*(3), Article e1002106. https://doi.org/10.1371/journal.pbio.1002106

Herbison, P., Hay-Smith, J., & Gillespie, W. J. (2006). Adjustment of meta-analyses on the basis of quality scores should be abandoned. *Journal of Clinical Epidemiology, 59*(12), 1249.e1–1249.e11. https://doi.org/10.1016/j.jclinepi.2006.03.008

Héroux, M. E., Butler, A. A., Cashin, A. G., McCaughey, E. J., Affleck, A. J., Green, M. A., Cartwright, A., Jones, M., Kiely, K. M., van Schooten, K. S., Menant, J. C., Wewege, M., & Gandevia, S. C. (2022). Quality Output Checklist and Content Assessment (QuOCCA): A new tool for assessing research quality and reproducibility. *BMJ Open, 12*(9), Article e060976. https://doi.org/10.1136/bmjopen-2022-060976

Höfler, M. (2023). *A form to assess severe testing of a paper's claims (Version 1)* [Preregistration]. Open Science Framework. https://osf.io/c8j4w/registrations

Höfler, M., Kräplin, A., Varga, M. A., Wallrich, L., Elsherif, M., Peikert, A., Seetahul, Y., Sætrevik, B., & M., Montefinese (2025). *A falsification assessment form* [FAF development materials and version history]. Open Science Framework. https://osf.io/c8j4w

Höfler, M., Kräplin, A., Elsherif, M. M., Schepke, M., Montefinese, M., Seetahul, Y., Sætrevik, B., Peikert, A., Varga, M. A., & Wallrich, L. (2025). *Falsification Assessment Form* [Questionnaire]. GitHub. https://github.com/FalsificationAssessmentForm

Höfler, M., Scherbaum, S., Kanske, P., McDonald, B., & Miller, R. (2022). Means to valuable exploration: I. The blending of confirmation and exploration and how to resolve it. *Meta-Psychology, 6*. https://doi.org/10.15626/MP.2021.2837

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7*(6), 645–654. https://doi.org/10.1177/1745691612464056

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). https://doi.org/10.1177/25152459241306432

Lakens, D. (2019). *The value of preregistration for psychological science: A conceptual analysis.* PsyArXiv Preprints. https://doi.org/10.31234/osf.io/jbh4w

Lakens, D., & DeBruine, L. M. (2021). Improving transparency, falsifiability, and rigor by making hypothesis tests machine-readable. *Advances in Methods and Practices in Psychological Science, 4*(2). https://doi.org/10.1177/2515245920970949

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

Loenneker, H. D., Buchanan, E. M., Martinovici, A., Primbs, M., Elsherif, M. M., Baker, B. J., Dudda, L., Durdevic, D. F., Misic, K., Peetz, H. K., Roer, J. P., Schulze, L., Wagner, L., Wolska, J., Kuhrt, C., & Pronizius, E. (2024). We don't know what you did last summer. On the importance of transparent reporting of reaction time data pre-processing. *Cortex, 172*, 14–37. https://doi.org/10.1016/j.cortex.2023.11.012

Maassen, E., van Assen, M. A. L. M., Nuijten, M. B., Olsson-Collentine, A., & Wicherts, J. M. (2020). Reproducibility of individual effect sizes in meta-analyses in psychology. *PLoS ONE, 15*(5), Article e0233107. https://doi.org/10.1371/journal.pone.0233107

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars.* Cambridge University Press.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*(2), 103–115. https://doi.org/10.1086/288135

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Nagy, T., Hergert, J., Elsherif, 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. A., Sotola, L. K., Moreau, D., & Rubínová, E. (2024). Bestiary of questionable research practices in psychology. *Advances in Methods and Practices in Psychological Science, 8*(3). https://doi.org/10.1177/25152459251348431

Nanyang Technological University Library. (2023). *Open research checklist (Version 1).* Nanyang Technological University. https://libguides.ntu.edu.sg/openresearchchecklist

Nosek, B. A., Alter, G., Banks, G., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C., Chin, G., Christensen, G., Dumas, T., Ebersole, C., Fidler, F., Hauser, D., Hennessy, S., Hilgard, J., Hogg, M., Humphreys, M., Kaatz, A., & Yarkoni, T. (2015). Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science, 348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review, 26*, 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Parsons, S., Azevedo, F., Elsherif, M. M., Guay, S., Shahim, O. N., Govaart, G. H., Norris, E., O'Mahony, A., Parker, A. J., Todorovic, A., Pennington, C. R., Garcia-Pelegrin, E., Lazić, A., Robertson, O., Middleton, S. L., Valentini, B., McCuaig, J., Baker, B. J., Collins, E., . . .Aczel, B. (2022). A community-sourced glossary of open scholarship terms. *Nature Human Behaviour, 6*(3), 312–318. https://doi.org/10.1038/s41562-021-01269-4

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

Rakover, S. S. (2003). Experimental psychology and Duhem's Problem. *Journal for the Theory of Social Behaviour, 33*(1), 45–66. https://doi.org/10.1111/1468-5914.00205

Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development, 31*(1), Article e2295. https://doi.org/10.1002/icd.2295

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

Uygun Tunç, D., & Tunç, M. N. (2023). A falsificationist treatment of auxiliary hypotheses in social and behavioral sciences: Systematic replications framework. *Meta-Psychology, 7*. https://doi.org/10.15626/MP.2021.2756

Viswanathan, M., Patnode, C. D., Berkman, N. D., Bass, E. B., Chang, S., Hartling, L., Murad, M. H., Treadwell, J. R., & Kane, R. L. (2018). Recommendations for assessing the risk of bias in systematic reviews of health-care interventions. *Journal of Clinical Epidemiology, 97*, 26–34. https://doi.org/10.1016/j.jclinepi.2017.12.004

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-hacking. *Frontiers in Psychology, 7*, Article 1832. https://doi.org/10.3389/fpsyg.2016.01832

Zhao, N., Yang, S., Zhang, Q., Wang, J., Xie, W., Tan, Y., & Zhou, T. (2024). School bullying results in poor psychological conditions: Evidence from a survey of 95,545 subjects. *Frontiers in Psychology, 15*, Article 1279872. https://doi.org/10.3389/fpsyg.2024.1279872

EUROPEAN ASSOCIATION OF METHODOLOGY

Methodology (METH) is the official journal of the European Association of Methodology (EAM).

zpid Leibniz-Institut für Psychologie

PsychOpen GOLD is a publishing service provided by the Leibniz Institute for Psychology (ZPID), Germany.