UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

White, M., & Stovner, R. (2024). Breakdowns in Scientific Practices: How and Why Some Accepted Scientific Claims May Have Little Actual Support. *Collabra: Psychology*, *10*(1). https://doi.org/10.1525/collabra.121436

Methodology and Research Practice

Breakdowns in Scientific Practices: How and Why Some Accepted Scientific Claims May Have Little Actual Support

Mark White[1][a], Roar Stovner[2][b]

[1] Department of Teacher Education and School Research, University of Oslo, Oslo, Norway, [2] Department of Primary and Secondary Teacher Education , OsloMet – Oslo Metropolitan University, Oslo, Norway

Keywords: quantitative methods, institutional theory, validity theory, significance testing, model fit, null hypothesis, meta-research

https://doi.org/10.1525/collabra.121436

Collabra: Psychology

Vol. 10, Issue 1, 2024

This paper considers the reasonableness of claims made in empirical psychological science. Drawing on validity and institutional theories, our conceptual model views research methods as institutionalized approaches to supporting the (implicit) inferential argument that is used to validate conclusions. Breakdowns occur when researchers falsely believe that a method strongly supports the inferential argument, but where little support is provided. We identify two characteristics of methods that promote breakdowns and show that these characteristics explain breakdowns of two common methods, null hypothesis significance testing and cutoffs for fit indices. Last, we discuss broadly how to reduce breakdowns in scientific practice.

This paper responds to recent critical reflections on the problematic state of social sciences (e.g., over-belief in unreplicated findings from small studies, p-hacking, HARKing; Hedges, 2018; Hoekstra & Vazire, 2021; Munafò et al., 2017; Renkewitz & Heene, 2019; Simmons et al., 2011). We provide a framework for thinking about the validity of study findings that could be helpful to most social scientists but is of particular interest to the field of meta-science and those working to improve the state of the social sciences. We understand the problematic state of the social sciences as arising from the rise, spread, and continuation of research methods that promote breakdowns, which is when a method is believed to provide support for a scientific claim when little support for the claim is justified. For example, p-hacking refers to a set of practices (e.g., testing multiple outcomes and interactions) that make finding a statistically significant result probable even in the absence of "true" effects (Simmons et al., 2011). When p-hacking, statistical significance testing provides little evidence to support study conclusions (Mayo & Spanos, 2006). P-hacking (and other related practices), then, represents a condition under which there is a breakdown in the practice of significance testing (i.e., conditions under which researchers believe that study conclusions are strongly supported when little support exists).

This paper provides a three-step, theoretically-based argument for how methods that promote breakdowns remain in use, sometimes despite a widespread understanding of a method's problematic nature (e.g., p-hacking). First, we conceptualize social science as the practice of building, typically implicit, inferential arguments to support the reasonableness of specific claims and define research methods as approaches to support key inferences in inferential arguments. Second, we argue that researchers' are motivated to reduce the uncertainty of claims and that the same, inductive inferences arise repeatedly across studies (Cronbach, 1982), which leads to the institutionalization of research methods. This sets the stage for the last step, where the application of a research method becomes a substitute for providing support to key inferences, leading the reasonableness of a claim to be judged by the proper application of research methods rather than the actual support provided by the (implicit) inferential argument. This leads to breakdowns by making it difficult to fully evaluate the support provided by the inferential argument. We propose two conditions that facilitate this last step. The remainder of the paper develops and expands this argument, providing two selected examples, and discussing the implications this argument has for improving social science research.

The First Step: Claims, Validity and Research Methods

Broadly speaking, we characterize the research field as many individual researchers or research groups that strive to put forth specific claims that are taken up by other researchers or groups. Claims are any conclusion, recommendation, or suggestion stemming from published or un-

a Correspondence concerning this article should be addressed to Mark White, mark.white@ils.uio.no.

b robast@oslomet.no

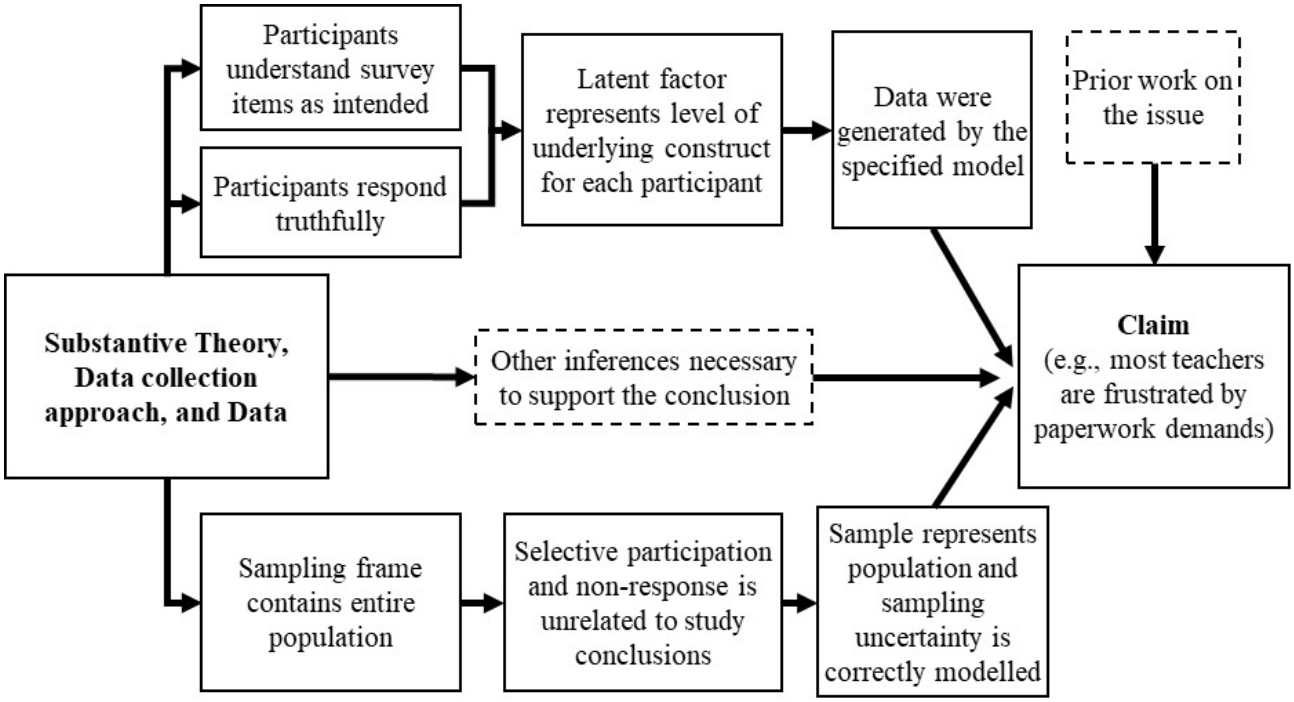

Figure 1. Example of an inferential argument to support a research claim. Only some inferences are shown.

published articles or presentations (e.g., construct X exists or is important, intervention Y is effective)[1]. Researchers are cognitively, emotionally, and professionally motivated to make claims that are reasonable, accurate, and important (e.g., through gaining professional satisfaction/prestige) and institutionally rewarded by doing so (e.g., through promotions or job hiring). We focus here on the reasonableness of empirical claims since a claim's accuracy is often unknowable and its importance is evaluable only within the context of a specific field. Drawing on modern validity theory (Kane, 2006), we view empirical claims as proposed interpretations or uses of data (based on theory and analyses), and a claim is accepted as reasonable when there is a well-supported, possibly implicit, inferential argument that lays out why a combination of theory, analyses, and data supports the claim, see Figure 1. From this perspective, breakdowns occur when researchers broadly believe that a validity argument provides strong evidence for a claim, but when one or more inferences in that argument have limited support. That is, when researchers have incorrect beliefs about the support for a claim.

From this perspective, the key challenge for identifying breakdowns becomes identifying how researchers support the (implicit) inferential argument and the scenarios under which researchers might have more faith in the inferential argument than is justified, given the evidence. We define a research method as a systematized procedure used to support one or more inferences. For example, the method of random sampling is used to support the inference that the sample represents the population (in expectation). The method of systematic test development procedures is used to support the inference that a test measures the intended construct. The method of regression modelling is used to support inferences about structural relationships between variables, and so on. Figure 2 shows a representation of this argument, highlighting a single "focal inference" within the inferential argument for expository clarity. A full version of Figure 2 might show many claims arising from a single study, each with a unique inferential argument, and many methods being used to support the inferences that compose the inferential arguments. As shown in Figure 2, it is the set of inferences (i.e., the inferential argument) and strength of their support that determines the reasonableness of claims. Research methods provide support for specific inferences within the inferential argument and the overall level of support provided to the inferential argument determines the reasonableness of a claim.

Note that we only seek to provide a theoretically-based, descriptive accounting of the practice of research. We take no normative stand on scientific reasoning or methodology. The inferential argument could be built on hypothetico-deductive/falsificationist (e.g., Tunç et al., 2023) or abductive (Haig, 2018) lines of reasoning, or combinations. The research aims could be confirmatory or exploratory. We further take no stand on the accuracy of claims. Claims with little support could still be accurate. Last, we do not assume that inferential arguments are explicit, simply that researchers use methods to convince themselves that claims are reasonable and believe that the claim's audience will do the same, and this process of making claims that the

1 One might replace the word claim with "contribution to the field", but we find the word claim to be more neutral and concrete.

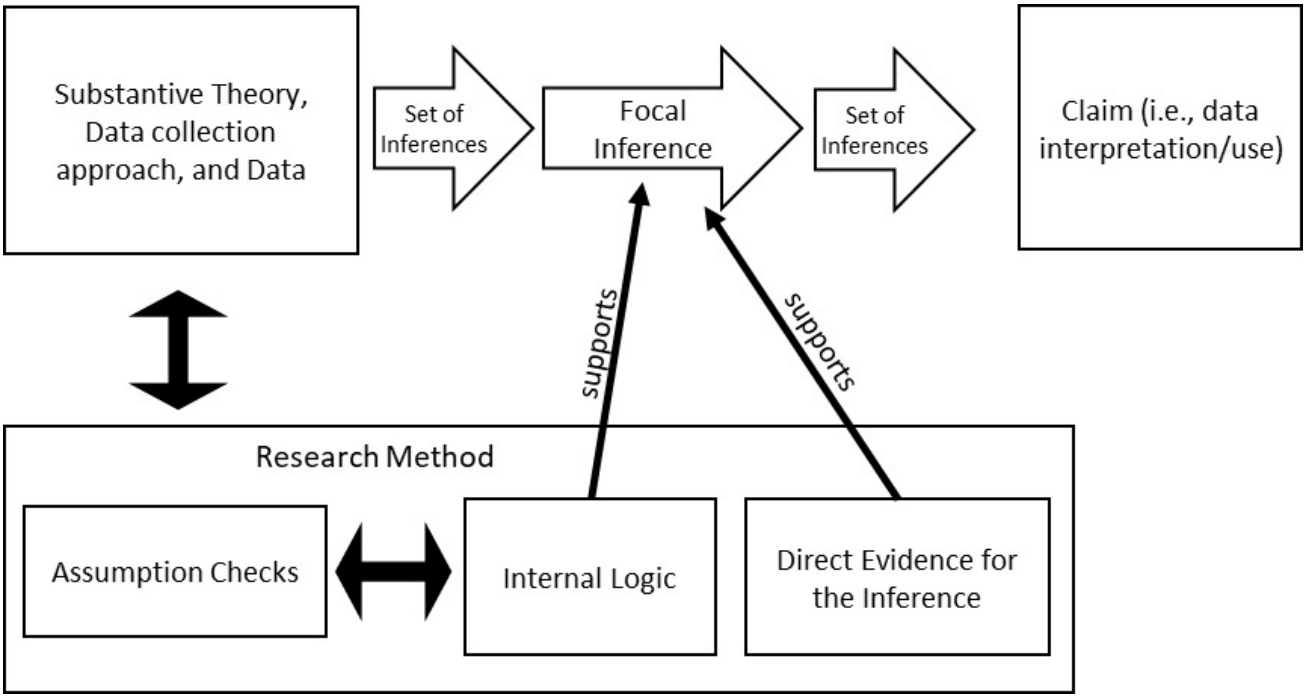

Figure 2. Conceptual Model

intended audience will find reasonable can be represented as an inferential argument. Our goal is simply to characterize the practice of research.

Figure 2 additionally highlights three aspects of a research method: the Internal Logic, Assumptions, and Direct Evidence for the Inference. The Internal Logic of a research method is the implicit or explicit logic by which a research method provides support for the focal inference. Importantly, the internal logic connects directly to the inference since it is this logic that provides support for the inference. For example, Null hypothesis significance testing (NHST; Gigerenzer, 2018) is used to support the inference that the data is inconsistent with the associated statistical model (including the corresponding null hypothesis). Support for this claim follows from NHST's probabilistic *modus tollens* logic (Cohen, 1994; Tunç et al., 2023). More specifically, the premises are (1) if the statistical model and Null hypothesis were accurate descriptions of the data-generating function, then the observed statistic follows the calculated distribution and (2) given the distribution, it is unlikely to have calculated a statistic at least as extreme as what was observed. The conclusion is that the data is inconsistent with the proposed statistical model (which includes the Null hypothesis; Wasserstein & Lazar, 2016). In this way, the NHST logic serves as an Internal Logic that directly supports the "data-model inconsistency" inference.

The second aspect of a research method highlighted by Figure 1 is the Assumption Checks, which are checks on whether preconditions and assumptions necessary for using a method are met. They usually become part of routines for properly using a research method. Importantly, assumption checks do not directly support an inference, but only check conditions under which the internal logic could not be correctly applied. For example, using NHST often requires checking for (or assuming) normality of residuals (Cohen et al., 2003, ch. 2). Non-normality in regression residuals does not provide information about the size of a regression parameter, but solely provides information about the applicability of NHST's internal logic (i.e., is important for calculating and interpreting p-values). When Assumption Checks fail, the method's internal logic is undermined and may not provide support for inferences, though some methods can be robust to some assumptions. Assumption checks can, at best, denote 'necessary, but not sufficient' conditions for applying the research method (St. Clair, 2005). For example, parametric correlations assume linear relationships between variables (Cohen et al., 2003, ch. 2), but while clear instances of non-linearity in the relationship can be found, one can generally not prove that the true relationship between two variables is linear (i.e., identifying an apparently linear relationship is necessary, but not sufficient for applying the method).

The last part of a method in Figure 1 is Direct Evidence for the Inference. Consider the method of random sampling (Kish, 1965). Random sampling supports the generalization inference (i.e., an effect in the sample is present in the population) through the logic of randomization, as randomization creates a representative sample, on average (i.e., in expectation). A representative sample allows for generalizing from sample to population, and random sampling creates a sample that is, on average, representative of the population. Additionally, the sample's representativeness can be *partially* empirically examined by comparing the sample and population on observed variables. That the sample and population are similar on observed variables provides direct empirical support for the generalization inference independent of the logic of random sampling. This empirical evidence does not replace the logic of random sampling,

but it provides important, direct empirical evidence to supplement that logic. Only some research methods provide such direct evidence while others rely solely on their internal logic. For example, a p-value cannot directly support an inference but provides information only when interpreted through the logic of a statistical significance testing paradigm.

Our argument thus far can be summarized as: Researchers seek to make claims that intended readers will accept as reasonable, using their own (or colleagues) judgement to make this determination. A claim's reasonableness depends on the extent to which a well-supported inferential argument underlies that claim and this support comes from research methods. Breakdowns occur when researchers believe that a research method provides strong support for one or more inferences when little support is actually provided (i.e., when researchers strongly believe claims that have weak support). We turn now to discuss how institutional theories can help illuminate when and why breakdowns might arise, spread, and be sustained.

The Second Step: Institutionalization of Research Methods

This section draws on institutional theory (DiMaggio & Powell, 1983; Meyer & Rowan, 1977) to characterize the spread of research methods. Institutionalization is a process where groups of actors (traditionally, organizations) that experience similar challenges with no clear, optimal solution develop common solutions that become widely rationalized as adequate or even ideal solutions for the shared challenges. Three conditions facilitate the institutionalization process: (1) uncertainty in either goals or approaches to meet goals, which prevents purely rational approaches from developing; (2) desire to reduce this uncertainty, which motivates institutional forces; and (3) groups of organizations working towards similar goals, as drivers of institutionalization work across groups.

Here, we envision individual researchers, or research groups, as "organizations" that share knowledge resources (e.g., theories, past claims) and have a shared interest in maintaining a positive public perception of science while competing for resources and seeking to gain prestige by making claims. We assume that the same basic inferences arise repeatedly across studies, driven by the fact that researchers study specific units (e.g., individuals), treatments, observations (e.g., measurement tools), and settings (i.e., utos) and seek to make claims about broader populations of units, treatments, observations, and settings (i.e., UTOS; Cronbach, 1982). This creates the need to support the inductive inferences that, e.g., specific samples represent populations or specific measures generate scores that represent constructs. Then, many inferences that arise repeatedly across studies are inductive and so inherently uncertain (Haig, 2018). We assume that researchers, in trying to support claims, are highly motivated to reduce this uncertainty, providing the strongest possible defense for a claim. In this way, researchers (1) face uncertainty in determining the validity of claims[2], (2) have a desire to reduce this uncertainty (i.e., ensure that claims are accurate), and (3) face similar challenges in supporting specific inferences across studies and research questions, setting the stage for the forces of institutionalization. We propose, then, that research methods arise, spread, and become institutionalized as attempts to manage and minimize the inherent uncertainty arising from the need to support inferences.

Three pressures lead to the institutionalization of research methods: mimetic, coercive, and normative pressures (DiMaggio & Powell, 1983). Applying institutional theory to characterize the development of the research field, the following story emerges. Researchers struggle to support key inferences when making claims, leading them to identify approaches from similar past work that seemingly successfully supported claims. Importantly, at the start of this process, the adopted method need not be highly effective at reducing the uncertainty of claims, researchers must simply believe it reduces the uncertainty more than existing approaches. Mimetic pressure, then, leads other researchers to copy previous methods, citing the successful past work as justification for the approach. As methods are copied, they are systematized, spreading through both formal courses and informal social networks (i.e., normative pressures). As a method spreads, normative pressures to adopt the method arise (i.e., others grow to expect the approach's usage). Eventually, gatekeepers, including peer reviewers, editors, and funding agencies, can begin requiring the method, creating coercive pressures.

As a method spreads, it becomes systematized with clear rules and routines for enactment, including Assumption Checks (i.e., the conditions for using the approach are formalized), and shared understandings of the method's affordances and limitations develop (i.e., the internal logics are formalized and the inferences a method might support are agreed upon). This systematization of the method results in shared understandings of when, how, and why the method should be used, as well as how assumptions of the method should be checked. However, there may not ever develop universal agreement over all aspects of a method, both because experts disagree on complex subtleties associated with the method and because applied researchers may only be motivated to develop the knowledge they believe to be necessary to apply the method.

A note about the institutionalization process is important. It is driven by pressures that promote conformity and

2 It may be possible that some claims are supportable only through deductive logics (e.g., falsificationist approaches), though philosophers of science have argued that while deductive logics can play an important role in science, there is an unavoidable need to rely on abductive logics, as deductive logics cannot create new knowledge (see Haig, 2018). The only point necessary for the arguments in this paper is that many claims have some uncertainty associated with them, as this would be enough to drive the institutionalizing forces that we discuss.

reduces uncertainty in claims through this conformity (i.e., claims are perceived as less uncertain when supported by methods that the field accepts as being sufficient to support a claim), making it inherently conservative. There is no guarantee that effective methods will result from this process, though we assume that clearly ineffective methods will be deselected at the start of the institutionalization process. However, the conservative nature of institutionalization may sustain methods that already are institutionalized after flaws are identified (e.g., commonly critiqued approaches like NHST have remained remarkably robust despite long-standing critiques; Gigerenzer, 2018). Similarly, there is likely a "good enough" principle, where methods that are understood as non-ideal, but that are better than alternatives, may become institutionalized and so remain self-propagating even after better methods arise.

The Third Step: The Emergence of Breakdowns in Scientific Practices

Overall, the institutionalization of research methods is neither solely negative nor positive. It has several positive impacts. Institutionalized methods are those that have been vetted and approved by the field, so they are likely to be more effective and fully developed than methods individual researchers might develop on their own. For example, regression models (and many other classes of statistical models) have standards for use (e.g., guides for evaluating the quality of a regression analysis; Hancock et al., 2019, ch. 23), established computer software that includes built-in assumption checks (e.g., the *lm* package in R; R Core Team, 2024), and various fitting algorithms that relax assumptions (e.g., OLS versus least absolute deviation; Cohen et al., 2003). These features represent a systematization of regression modelling that can, ideally, reduce misuse of the method by ensuring that core procedures are enacted and key assumptions tested. This is not a perfect fix, as individual researchers may not apply methods as intended and/or engage in the expected assumption checks, due either to a lack of knowledge, cognitive biases (e.g., confirmation bias), or institutional pressures (e.g., lack of time, pressure to publish). However, the existence of institutionalized methods, such as regression modelling, prevents researchers from having to start from scratch in every study, allowing the methodology of science to be cumulative.

On the other hand, institutionalization has potential negative impacts. The institutionalization process shifts the focus from building and supporting inferential arguments to the proper enactment of research methods. This reduces the uncertainty of claims because inferential arguments (especially inductive inferences) are inherently uncertain while the proper application of a research method can be evaluated with little uncertainty (albeit typically only using information not available in published research). Research methods, then, can serve to buffer claims from true scrutiny, as researchers (and peer reviewers and the public) often focus on the enactment of specific research practices when judging the validity of conclusions rather than evaluating the overall support of the (implicit) inferential argument linking data to conclusions (i.e., ceremonial inspection; Meyer & Rowan, 1977). For example, factor analysis often supports the inference that one has measured the intended, theoretical construct, but fully supporting this inference is highly complex, requiring more than statistical modelling (Maul, 2017; McGrane & Maul, 2020).

While we emphasize the desire to reduce uncertainty in claims as the driver here, other cognitive and/or social biases could also play a role here. For example, the emphasis on applying methods facilitates confirmation bias by allowing appeals to the limitations of methods when outcomes do not confirm to expectations while allowing researchers to minimize a method's limitations (by appealing to the method's common use) when outcomes confirm expectations. Further, social conflicts may be minimized by allowing researchers to focus on the application of the method (i.e., something highly concrete) rather than much more complex questions related to inferential arguments (e.g., how much support for an inference is enough? What epistemological/ontological assumptions are necessary for building a convincing inferential argument?). In this way, the shift in focus from the inferential argument to the application of methods might facilitate a wide range of cognitive and social biases that promote or sustain further breakdowns.

A further consequence of the focus on methods is the resultant ambiguity in inferential arguments. It can be difficult to determine what inferences are necessary to support a claim and how precisely research methods support the inferential argument. This promotes breakdowns in at least two ways. First, it allows difficult to support inferences to go unnoticed (typically unintentionally, but also potentially intentionally by those with poor motives). Second, and relatedly, ambiguity in the inferential argument allows researchers to substitute difficult inferences with similar, but easier to support inferences. For example, NHST is often inappropriately used to support the inference that a specific parameter is non-zero, when it, at best, can be used to support the broader data-model inconsistency inference, implying that one or more parameters or assumptions is wrong; Wasserstein & Lazar, 2016).

Beyond these two potential sources of breakdowns, we propose two characteristics of research methods that might promote breakdowns or allow breakdowns to persist across time: The Logic of Confidence condition and the Deductive-Style Reasoning condition. We discuss these conditions next.

The Logic of Confidence Condition

The Logic of Confidence condition occurs when the internal logic of the research method is the sole basis for supporting an inference (i.e., when no direct evidence is generated; see Figure 2). NHST serves as an example for this condition. While enacting NHST might involve many Assumption Checks (e.g., distributional tests), these checks provide no direct evidence for the "data-model consistency" inference (i.e., the typical focal inference for NHST; Wasserstein & Lazar, 2016), but simply identify if the conditions necessary to apply NHST exist. However, one can never fully specify or check the complete set of auxiliary

assumptions necessary for applying a method (c.f., Gershman, 2019; St. Clair, 2005). That is, assumptions denote at best necessary, but not sufficient conditions for using a method. Then, one can never guarantee that the research method can be applied in each case. For example, the basis of NHST involves severe testing (i.e., the claim that a test result is highly unlikely to have occurred given the model; Mayo & Spanos, 2006), but the necessary conditions for a statistical test to be a severe test are still under-specified (though pre-registration of analyses seems like a possible pre-condition; Gehlbach & Robinson, 2017).

Relying solely on a method's internal logic increases researchers' confidence in the appropriateness of a method, especially when a series of assumption checks show no problems, while shielding researchers from confronting any direct evidence about whether an inference is supported. The result is that enactment of the method becomes the only standard through which researchers themselves and peer reviewers can judge the appropriateness of the focal inference, allowing the very use of methods (according to agreed upon standards) to gain a self-perpetuating momentum, regardless of whether that use promotes breakdowns (see, e.g., Gigerenzer, 2018). In this way, the internal logic of a research method becomes a logic of confidence that is accepted as a matter of faith while avoiding true tests of the adequacy with which the method supports claims (i.e., ceremonial inspection; Meyer & Rowan, 1977). We hypothesize, then, that research methods that fit the logic of confidence condition likely promote breakdowns, yet still persist in usage despite these breakdowns.

The Deductive-Style Reasoning Condition

The second characteristic of research methods that promotes breakdowns is the Deductive-Style Reasoning condition. Here, we define deductive-style reasoning as reasoning where the conclusion is guaranteed by a set of premises, contrasting this with inductive reasoning, which involves tentatively suggesting a broad conclusion based on evidence from specific examples. The contrast here is both in terms of the certainty of conclusions (i.e., only deductive-style reasoning gives certainty) and the nature of the reasoning (i.e., only inductive reasoning argues from specific example to a general principle; Haig, 2014).

We proposed that many shared inferences are inherently inductive (i.e., focused on generalizing beyond the specific units, treatments, measures, and settings studied) and so uncertain (c.f., Cronbach, 1982; Haig, 2018). Researchers are motivated to reduce this uncertainty and so expected to prefer deductive-style logics that promise more certainty in conclusions. For example, problematic applications of NHST ritualistically interpret low p-values as proof that a parameter is non-zero (Gigerenzer, 2018). That is, after accepting the premises of (a) assumption checks fail to identify problems and (b) the p-value is below the cut-off, the validity of the focal inference is taken as a foregone conclusion. We might contradict this with Fisher's original interpretation of p-values as a source of inductive evidence (Halpin & Stam, 2006; see also Wasserstein et al., 2019). This certainty, however, can lead to overconfidence in the focal inference, especially for inductive inferences. That is, the apparent certainty of deductive logics makes it appear to be unnecessary or a waste of time to conduct the sorts of replications and checks necessary to detect breakdowns, discouraging such checks (see Amrhein et al., 2019). Thus, we contend that the use of deductive-style reasoning when applying research methods to support inductive inferences could hinder the detection and correction of breakdowns.

This section has argued that breakdowns will persist across time if research methods meet the Logic of Confidence condition, because researchers will never be faced with direct empirical evidence that exposes the minimal support provided for the focal inference, or the Deductive-Style Reasoning condition, because certainty in conclusions discourages the sorts of checks necessary to detect breakdowns. This same logic points towards two desired properties of research methods: they should generate Direct Evidence to support the focal inference and, at least when the focal inference is inductive, they should emphasize inductive-style reasoning. In the following, we apply the theory to highlight potential problems with two common methods.

Application of the Theory to Two Research Methods

To this point, the paper has put forth a theoretical account that explains the institutionalization of research practices, or how research practices spread and become formalized into widely adopted methods and hypothesized when and why some research methods promote breakdowns. This section discusses two specific sets of practices in light of this account.

Null Hypothesis Significance Testing (NHST)

The first method we discuss is NHST, which was briefly discussed in several places above (see Gigerenzer, 2018 for broader discussions of NHST). While NHST arguably can be appropriately applied, the widespread usage of p-values to justify scientific hypotheses that arises from NHST has been heavily criticized (see Wasserstein & Lazar, 2016 for a formal statement from the American Statistical Association on problems with p-values), leading us to confidently point to NHST as a case where a research method has led to a breakdown. We follow Gigerenzer's (2018) characterization of problematic forms of NHST in what follows here. Importantly, in this characterization, the inference supported by NHST is the generalization inference, or the inference that an effect found in a sample is present in the population, which is a subtly broader inference than that discussed above in the context of (the appropriate use of) NHST. This highlights our point above that the institutionalization process leads to an emphasis on research methods that can make it difficult to follow precisely what parts of the inferential argument are being supported by a given method. This development and persistence of breakdowns related to NHST is predicted by our theory, see Figure 3.

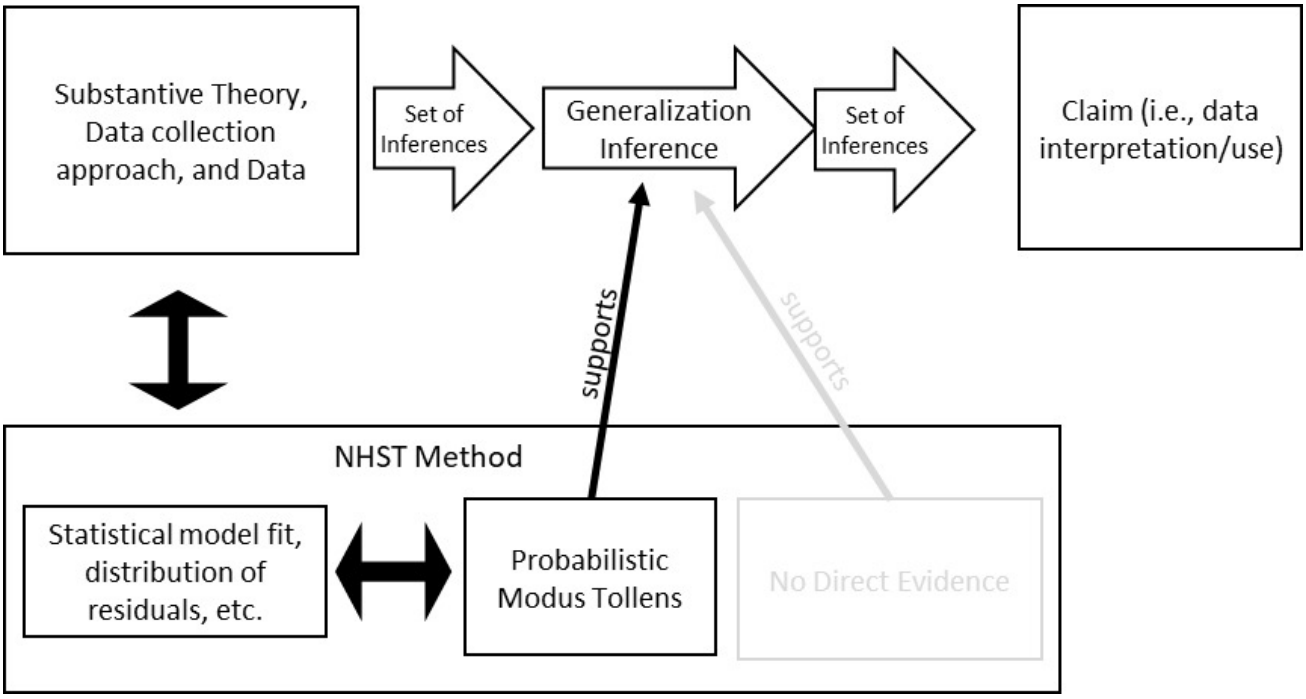

Figure 3. The conceptual framework as applied to the NHST method

NHST fulfills the logic of confidence condition

We discussed the logic of confidence condition as it relates to NHST previously, but shortly stated, the application of NHST generates only p-values, which can only be interpreted within the logic of NHST. No direct evidence that can independently lead to an evaluation of the generalization inference is generated, leading researchers to rely solely on the NHST logic when supporting the generalization inference.

Consider the usage of NHST from the perspective of the average researcher. The typical researcher may routinely support the generalization inference using NHST. In doing so, the researchers rely on the logic of confidence, never having to directly confront evidence about whether the generalization inference is accurate. This insulates researchers from theoretical or conceptual critiques by ensuring personal, positive experiences with publishing and obtaining peer recognition using NHST, allowing researchers to minimize any critiques they might come across and explaining the persistence of NHST-related breakdowns across time. Researchers can go their whole career without having to directly confront whether the generalization inferences that they make are empirically supported (c.f., Gershman, 2019).

NHST fulfills the deductive-style reasoning condition

Further, while the probabilistic *modus tollens* logic of NHST (Cohen, 1994) is properly understood as providing uncertain conclusions, problematic interpretations of NHST effectively interpret NHST as invoking the deductive *modus tollens* logic (i.e., they reason deductively when drawing conclusions from NHST; Gigerenzer, 2018). This leads researchers to be overconfident in single studies, undermining the likelihood that they will engage in further efforts to validate the generalization inference (Amrhein et al., 2019; Nelder, 1999).

The deductive-style reasoning underlying problematic applications of NHST undermines systematic attempts to externally validate the generalization inference by creating certainty in conclusions (Nelder, 1999). This, in turn, creates the impression that direct replications are unoriginal, simply confirm already known facts, or are a waste of time, reducing the demand for replications (at least until recently; Amrhein et al., 2019; Makel & Plucker, 2014). This reinforces the logic of confidence by reducing the likelihood of direct replications, which is the main possible source of direct evidence that a researcher might use to empirically test the generalization inference.

Discussion of breakdowns when using NHST

Despite decades of discussions of flaws in NHST (e.g., Cohen, 1994; Hubbard, 2004), it is still widely used in problematic ways (Gigerenzer, 2018), which we argue is perpetuated by the two conditions. If the critiques were accurate and common uses of NHST were questionable, one would predict that disruptions to NHST's logic of confidence or deductive-style reasoning would lead to growing demands for changes to NHST. This is exactly what happened. The "replication crisis" began as empirical evidence mounted that NHST led to breakdowns (i.e., studies were far less likely to replicate than assumed; Amrhein et al., 2019). That is, replication studies provided direct empirical evidence about the generalization inference, and when this was found to be discrepant with results from the NHST logic, many researchers reconsidered how they conducted research. The result is widespread reconsideration of when NHST supports the generalization inference and even ef-

forts to ban NHST (e.g., Wasserstein et al., 2019). Additionally, there is a growing push for so-called "many lab studies" that explicitly and directly provide direct empirical evidence to test the generalization inference (e.g., Makel et al., 2019), replacing NHST's internal logic with direct empirical evidence to support the generalization inference. While problematic usage of NHST has not disappeared, disruptions of the logic of confidence and deductive-style reasoning conditions have led to a wide scale recognition of the breakdowns resulting from problematic applications of NHST and efforts to reform the use of this method.

What are the prospects of reducing the field's dependence on the questionable usage of NHST? Structural changes, such as better methods training, reforming textbooks, changing publication incentives, and ensuring gatekeepers (i.e., peer reviewers, editors, funding agencies) understand the limitations of NHST, can be an important tool in promoting change, a common conclusion to meta-research studies (e.g., Hoekstra & Vazire, 2021; Munafò et al., 2017; Renkewitz & Heene, 2019). However, these structural changes do not address researchers' need to support the generalization inference. They also do not force individual researchers to confront the limitations of NHST within their own work, though the direct challenges to the NHST logic that were discussed in the previous paragraph could accomplish this. Without addressing such needs, NHST may continue or be replaced by similarly questionable practices (e.g., Morey et al., 2016).

It remains to be seen if any of the many suggested replacements for NHST will avoid its problems (see Wasserstein et al., 2019 for a brief overview of several such approaches). Unfortunately, while replication may be the current gold standard for supporting the generalization inference, the emphasis currently placed on drawing clear policy and practice recommendations from single studies undermines the role of replication by emphasizing the need to draw conclusions from individual studies (Nelder, 1999). Efforts to combat this viewpoint, then, are important (e.g., Pashler & Ruiter, 2017; Robinson et al., 2013), as is providing alternative, breakdown-resistant approaches to support the generalization inference.

The use of pre-registration has been suggested as a potential solution to some of the problems stemming from problematic usage of NHST (Gehlbach & Robinson, 2017). We support the goals of pre-registration and believe that it can help with some of the issues around NHST, namely those related to p-hacking and the garden of forking paths (Simmons et al., 2011). However, the key problem in this paper is how NHST is used to support the wrong inference (driven by institutionalization's focus on evaluating methods rather than inferences) while leading researchers to be overly confident in the inference (i.e., the deductive logics condition) and never confronting researchers with direct evidence regarding the inference's accuracy (i.e., the logic of confidence condition). We do not believe that this key problem is likely to be impacted by pre-registration.

Using Cutoffs for Model Fit Statistics

Next, we apply our conceptual framework to the practice of using cutoffs, or rules of thumb, on summary statistics to make decisions. This practice occurs in a range of areas when a continuous summary statistic is found that provides information that could be used to support an inference. The difficulty of determining whether a specific value on the summary statistic provides support for an inference has led to the institutionalization of using a specific range of values as supportive of the inference and another range of values as unsupportive. The origin or basis for specific cutoff values is not always known (e.g., for Cohen's Kappa; Wilhelm et al., 2018), but a small number of cutoff values often become widely adopted. We focus our discussion on the use of cutoffs in the specific case of generating model fit statistics within confirmatory factor analysis or structural equation modelling (e.g., Hu & Bentler, 1999). Here, the focal inference is that the data were generated by the specified model, which we call the model fit inference, see Figure 4. That is, researchers use model fit statistics to imply that the specified model is correct, and the observed data come from that model. Note that this is subtly, but importantly different from the inference that model fit statistics should support, namely that the data are consistent with the model. Here, again subtle shifts in the inference being supported are a starting place for breakdowns. The difference in inferences comes from the fact that any data set is consistent with an infinite number of models (MacCallum et al., 1993), so the focal inference (i.e., that the model is correct) is much stronger than the reasonable inference (i.e., that the data are consistent with the model).

The use of model fit cutoffs fulfills the deductive-style reasoning condition

This practice highlights the deductive-style reasoning condition, as the approach of establishing a cutoff takes an inherently uncertain situation, where the continuous statistic denotes varying levels of support for an inference, and creates an accept/reject conclusion (i.e., it creates certainty from uncertainty by introducing the deductive-style rule of 'if and only if you exceed the threshold, the inference is supported'). Note that the use of such cutoffs fits well within the institutional framework described by this paper: The uncertainty of supporting an inference with a continuous measure is reduced by the broad scale agreement that specific cutoffs denote meaningful ranges of values.

We are not proposing that researchers have a black and white view of cutoffs, as we have found most researchers are far more nuanced in their understanding when engaged in discussion. Further, we note that model fit statistics (and most other uses of cutoffs) often propose many levels of cutoffs to denote bad, moderate, and good levels of fit, which can help to nuance findings. Rather, we propose that, within the text of written scientific articles, a deductive-style reasoning is applied whereby exceeding specific cutoffs is presented as sufficient evidence for model fit and failing to exceed specific minimal cutoffs is taken as sufficient evidence for a lack of fit. Fit statistics falling within

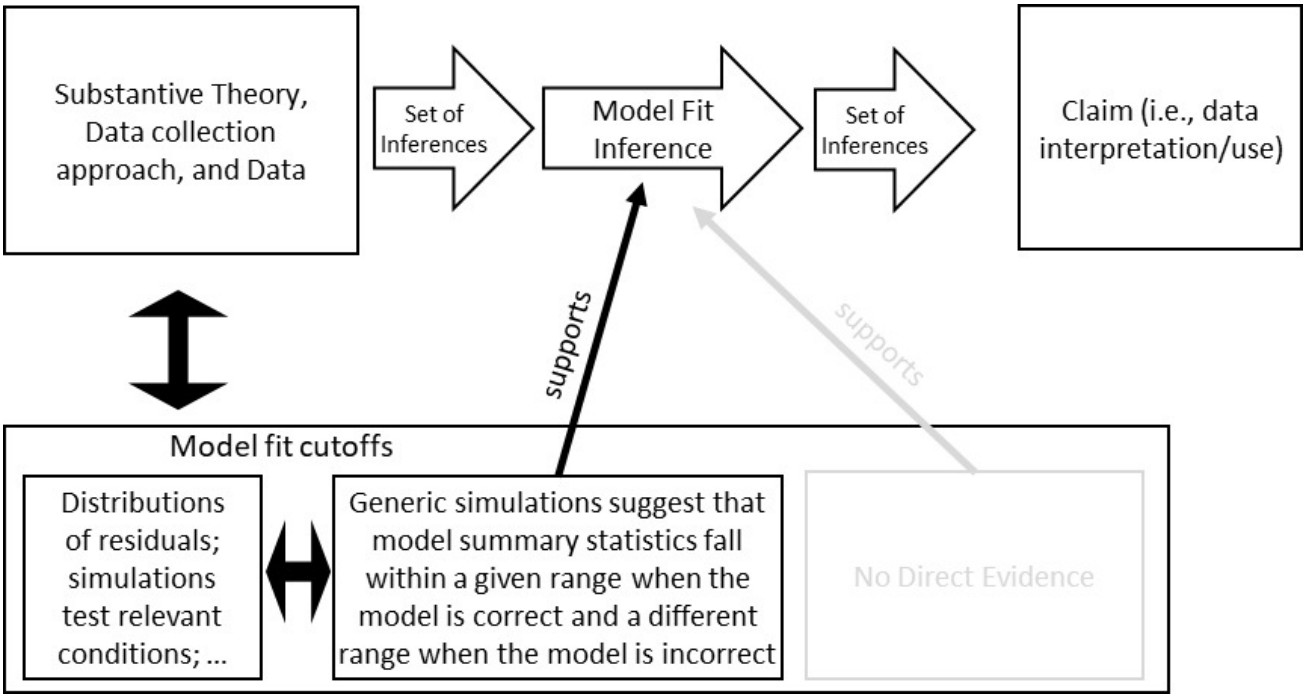

Figure 4. The conceptual framework for the use of cutoffs in model fit

a middle range can be accepted, but only with explicit justification for the moderate levels of fit or explicitly noting moderate fit as a limitation of the paper. See van de Grift and colleagues (2019) for what we consider a typical example of this sort of argumentation. Note also that this reasoning is supported by official APA guidelines (*APA Style JARS*, n.d.). We propose that this deductive-style reasoning seeks to minimize the perceived uncertainty of interpreting fit statistics by reference to commonly accepted cutoffs.

The use of model fit cutoffs fulfills the logic of confidence condition

For any given paper, there is often little clear evidence that a cutoff validly identifies when the inference is supported (McNeish & Wolf, 2021). Further, since meeting the cutoff value is often the only evidence used to support the inference, researchers engaging in this practice never face empirical evidence regarding whether meeting the cutoff actually provides support for the inference it is purported to support (e.g., van de Grift et al., 2019)—the logic of confidence condition is met.

Discussion of breakdowns when using model fit cutoffs

There is a growing case to be made that this process of using cutoffs is fundamentally flawed because simulation-validated cutoffs become over-generalized to apply to cases or problems not tested by the simulation and because specific test statistics or cutoffs cannot account for all threats to the support of key inferences (e.g., Marsh et al., 2004). Interestingly, suggested solutions for this problem include using simulations to generate direct evidence for the model fit inference (or at least the limited inference that the data is consistent with the model; i.e., reforming the method to add direct evidence; McNeish & Wolf, 2021). This involves simulating both (1) data of the same size and type as study data that was generated using the model to be tested in order to show that a certain range of values of model-fit statistics are likely if the model were accurate and (2) data of the same size and type of study data with known misspecifications to show that, if specific misspecifications were present, the model-fit statistics would likely take on a different range of values. These range of fit values found in simulations for the correct and misspecified simulations conditions provide direct evidence, in the context of a specific study, of whether specific ranges of fit-statistic values provide support for the inference. This solution fits well with what our framework would suggest. In fact, even in ideal cases, fit statistics are a necessary, but not sufficient condition to supporting the model fit inference since an infinite number of models may provide equivalent fit to any given data set (Fried, 2020; Yarkoni, 2020).

Other Possible Applications of our Framework

A reviewer suggested that we provide further examples of applying our framework to understand research methods to help better elucidate the framework. We do this in Table 1, which is based on methods volunteered by colleagues or the reviewer and/or that the authors happened to have recently read about. We chose to retain some methods that did not fit perfectly into our framework, as we see our attempt to make sense of methods from the perspective of our framework as potentially illuminating. Note that space limitations provide restrictions on the level of detail that we can provide.

Table 1. Interpretations of other methods based on the paper's framework

Method (reference)	Basic Description	(Hypothesized) Institutionalization process & focal inference supported	Proposed Model-Based Factors	Notes
Magnitude Based Inference (MBI; Sainani, 2018)	When evaluating the impact of an intervention, define areas of negative, trivial, and positive impacts. Use estimated confidence intervals from a study to examine how likely it is for the intervention to have a negative, trivial, or positive effect.	Researchers evaluating interventions with moderate sample sizes struggled to interpret large, but non-significant results given low power and the threat of Type II errors. MBI provided an apparently reasonable approach to incorporating effect sizes into conclusions. **Inference**: Is the intervention effective?	**Logic of Confidence**: No data beyond the confidence interval appears to be used. **Deductive Logic**: Fixed proposed interpretations are established based on confidence interval levels that overlap with the boundary between negative and trivial effects and the boundary between trivial and positive effects.	This approach effectively reparametrizes confidence intervals into statements of intervention effectiveness in ways that significantly increase false positives.
Reflective Measurement (White, 2024)	Measurement approach where items are assumed to share a common cause and associated statistical models (e.g., factor analysis, item-response theory) for modelling item scores under this assumption	Researchers struggle to determine if they are measuring the intended construct, a necessary assumption for many research questions. The reflective modelling framework provides established procedures for establishing measurement of the intended construct. **Inferences**: 1. Measurement of the intended construct has occurred. 2. Scores/estimates are adjusted for measurement error. 3. Measurements are comparable across groups.	**Logic of Confidence**: No evidence beyond model fit is generally provided for these inferences. Response processes governing how item-scores are created are generally considered independently of reflective modelling. **Deductive Logic**: See the "Using cutoffs for Model Fit Statistics" section, as the same principles apply.	Contradictions inherent in the building of measurement scales and the application of reflective measurement models leads to questions about the effectiveness of these models in supporting inferences (see White, 2024).
Sample standardized effect sizes (Baguley, 2009)	Effects are standardized based on sample-level standard deviations to facilitate interpretations of scores.	Researchers often measure variables on difficult to interpret scales (e.g., survey responses) and want to be able to compare estimates across studies and populations. Such comparisons are difficult due to challenges in making sense of measurement scales. This leads to standardizing scales, which is believed to make effects more comparable. **Inference**: Two parameters are directly comparable.	**Logic of Confidence**: No direct evidence is generated. **Deductive Logic**: standardized effects are comparable?	Note that such standardized scores are a function of sample standard deviations so may be dependent on sampling procedures, which could impact standardized effect sizes by impacting sample standard deviations (which are theoretically irrelevant to comparisons of parameters). There is a case to be made then that sample-standardized effect sizes are not inherently more comparable than unstandardized effect sizes.
Randomization in very small samples	Treatment assignment or other variables are randomly assigned, even in small samples that cannot take advantage of the equality in expectation that results from random sampling	Researchers are (appropriately) trained that randomization is the best way to study causal impacts, leading them to use random assignment even in cases of very small samples. Since random assignment ensures only equal groups in expectation, the impact of	**Logic of Confidence**: Comparisons of group characteristics may be unreliable in small samples, making it hard to use direct evidence to determine group	This method is included to highlight how actual good practices that should be encouraged can become less useful when used to support inferences that are not well-supported by the method in new contexts. In small studies, (matched) randomization should be encouraged, but largely because of its impact on reviews and/or meta-analyses than its impact on results from individual

Method (reference)	Basic Description	(Hypothesized) Institutionalization process & focal inference supported	Proposed Model-Based Factors	Notes
		randomization is minimal in (very) small studies. **Inference**: Groups being compared are equivalent.	equivalence, especially across many variables. **Deductive Logic**: Randomization leads to unbiased estimates, ignoring the probability of this statement?	studies since random errors with greatly trump non-random errors.
Rater reliability statistics (Wilhelm et al., 2018)	Calculation of statistics such as Cohen's Kappa to determine whether raters agree sufficiently on assigned scores.	Researchers using scores from raters typically find that raters agree often, but not always. There arises the question of how much agreement is enough in order to justify that raters are scoring the same construct and chance-corrected agreement statistics propose to provide answers to this question. **Inference**: Study finding is not rater dependent.	**Logic of Confidence**: No data other than the statistics are provided. **Deductive Logic**: Cut-offs are determined to collectively decide how much agreement is enough, just as the cutoff example in the paper.	Note that specific study findings could be more or less robust to rater error so rater agreement statistics tend to be a poor source of information for the focal inference, but those statistics (when calculated to be high) are often viewed as convincing evidence for the inference.
Positionality Statements (King, 2024)	Public, explicit reflections by researchers about their own biases and how these might have influenced interpretations and conclusions	Researchers always face the risk of biased or motivated reasoning leading them to draw biased conclusions. Positionality statements are argued to surface internal biases of researchers to reduce the influence of this motivated reasoning. **Inference**: Conclusions are not driven by motivated reasoning of the researcher.	**Logic of Confidence**: Existence of positionality statement reduces bias in findings? **Deductive Logic**: Positionality statements reduce bias by surfacing that bias?	We include this method because we know of several cases where peer reviewers asked researchers to draft positionality statements after conclusions were made (i.e., after such a statement could have helped to reduce biased reasoning), suggesting that adherence to institutionalized norms may often drive this method rather than the methods affordances.
Binning continuous variables (e.g., median splits; Maxwell & Delaney, 1993)	Researchers transform continuous variables into categorical variables in order to support interpretation of complex models	The simplicity and clarity of stories told when using binned data likely motivates later researchers to engage in this practice as they work to create a narrative to present results. **Inference**: Unclear, but potentially something related to how binning does not impact results while simplifying the presentation of results	**Logic of Confidence**: repeating analyses with an without binned data could serve as direct evidence for an inference that binning has no impact on results, but results are typically only presented binned **Deductive Logic**: not generally applicable	This approach can increase false-positive and/or false-negative rates depending on specific details of distributions.

Discussion

We have presented a three-step argument characterizing how and why scientific breakdowns occur, or why researchers might have far more faith in scientific claims than is deserved. The first step views social science as the practice of making claims supported by an (implicit) inferential argument with research methods serving to support the inferential argument. Second, we argued that researchers' motivation to reduce the uncertainty of claims, combined with the same inferences arising across different studies mobilizes forces that lead to the institutionalization of research methods. The third step occurs when institutionalized research methods come to serve as substitutes for building and explicitly supporting inferential arguments. On the positive side, this allows researchers to maintain collective professional control over the practices that become widespread in their field and those practices ensure that researchers use methods that provide strong support for the implicit inferential argument.

However, this creates a risk of breakdowns. The implicitness of the inferential argument along with the ambiguity in how research methods link to specific parts of the inferential argument can make it difficult for researchers themselves, along with readers of research articles and peer reviewers, to evaluate both the completeness of the inferential argument and its support. This, in turn, can promote breakdowns as researchers over-estimate the support provided for a claim. Note that some part of this risk is inherent in the institutionalization process, as substituting the evaluation of how research methods are used for the evaluation of the inferential argument is driven by efforts to minimize the uncertainty in claims (i.e., the key driver of institutionalization). Some part of this risk is driven by how research methods are defined and used. We argue that two characteristics of methods are especially likely to promote breakdowns.

The first is the logic of confidence condition, or when research methods provide support for a specific inference only through their own internal logic. In this case, breakdowns could conceivably persist forever since the empirical evidence that would be needed to identify and correct breakdowns is never generated. The second is the deductive-style reasoning condition, or when research methods are built on deductive-style reasoning that takes a conclusion as a fact (i.e., as certain) conditional on establishing some premises. Then, the logic of confidence condition ensures research methods do not inherently generate the evidence needed to identify breakdowns while the deductive logics condition demotivates the active search for evidence that could detect and address breakdowns.

Research methods that meet either of these conditions, then, may be more likely to result in breakdowns that persist across time. Further, methods that meet these conditions allow individual researchers to successfully apply those methods without being aware of breakdowns in their own work. This allows the usage of such methods to persist despite critiques in the methodological literature. That is, such methods may remain self-perpetuating even after it is discovered that they promote breakdowns (e.g., NHST).

The arguments put forth in this paper provide a broad scale critique of how using specific research methods can lead to problematic breakdowns in scientific practice. We discuss three implications of this framework: viewing of research as an inferential argument, possible changes to research methodology training, and the need to adapt existing methods or develop new methods that avoid the two key conditions.

Research as an Inferential Argument

The framework highlights limitations that can manifest when evaluating the application of methods substitutes for evaluating inferential arguments. These limitations were shown in the presented examples. For example, when discussing NHST, we noted that NHST was designed to support the inference that the data is inconsistent with a model (one part of which is the Null hypothesis; Wasserstein et al., 2019), but problematic versions of NHST use NHST to support the stronger inference that a given parameter is nonzero in the population. Similarly, the use of model cutoffs is used to support the inference that the specified model is true, but only supports the simpler inference that the data is consistent with the specified model. When inferential arguments are not laid out explicitly, research methods can inadvertently be used to support inferences that are far stronger than appropriate. That is, statistical reasoning (or methods reasoning more broadly) can replace scientific reasoning (Hubbard et al., 2019). This can make it difficult for researchers themselves, peer reviewers, and readers of scientific papers to judge the reasonableness of claims. This almost certainly interacts with various cognitive biases, such as confirmation bias, increasing the likelihood that such biases impact reasoning.

Beyond challenges arising from this subtle shifting of inferences, there is the problem of hidden inferences that do not get properly evaluated. For example, in our experience, measurement models typically lead to many comments or critiques by peer reviewers. However, when single items are used as variables, there are rarely comments about those items, even though both measurement models and single items must support the inference that the obtained score represents the intended construct. When evaluating the application of methods substitutes for evaluating the strength of an inferential argument, inferences not associated directly with a specific method (e.g., using a single item to represent a construct) may get overlooked. Such failure to consider all necessary inferences in an inferential argument could easily lead to overconfidence in claims being made (i.e., breakdowns). In most cases, we believe that this likely results from an oversight, but it is not hard to imagine that bad actors could leverage this point to avoid appropriate scrutiny.

Based on these expanded points, we recommend making claims, inferential arguments, and the connection between methods and inferential arguments explicit in scientific writing. This would require some shifts in norms for writing articles, but would ensure that researchers themselves, as

well as peer reviewers and readers, would be able to directly evaluate the inferential argument for themselves. This would hinder the sort of shifting and obscuring of inferences that were just discussed and are likely to lead to breakdowns.

This would involve some shifts to how researchers describe their methods. For example, Instead of describing the sample, authors would lay out explicitly why they believe the given conclusions can be made from the given sample (e.g., what the sample is representative of). Instead of describing measurement tools, authors describe why specific measures should be interpreted as capturing the target constructs. Instead of describing the analytic models used, authors describe how the analytic method supports drawing specific conclusions. The point here would be to explicitly surface the implicit assumptions and inferences necessary to support a claim and to show how each step in the research process supported that claim. Note this suggestion is in line with several other recent calls for change, including calls to more explicitly justify the generalizability of findings (e.g., Simons et al., 2017), calls to be more explicit about uncertainties in conclusions ("Tell It like It Is," 2020; Wasserstein et al., 2019), and calls to avoid over-generalizing findings (e.g., Robinson et al., 2013).

Changes in Research Methodology Training

A second recommendation stemming from this work would be a call to improve the teaching of research methods. Such a call is not new (e.g., Aiken et al., 2008). Beyond the call for better and more methods training, our framework would suggest a different style of methods instruction, one that connects more directly to training in scientific reasoning and how research methods support specific, concrete inferences. Over the course of the past few months, the first author has taken to asking graduate students and colleagues to explicitly state the claims that are being made in a paper/presentation and how methodological choices allowed them to support those claims. This is a very difficult task for many researchers. Reflecting on our own methodological training, the reason for this difficulty may be that there is often very little emphasis on scientific reasoning and how to combine methods to support complex claims (except perhaps in the domain of causal reasoning, e.g., Shadish et al., 2002), but methods courses typically emphasize the proper application and interpretation of a specified method. It is, of course, important that researchers learn how to properly apply and interpret methods. However, being able to properly leverage a wide range of appropriate methods to support claims is just as important.

Designing new classes that emphasize building inferential arguments to support complex claims and how a wide range of research methods support such inferential arguments may help avoid breakdowns. Additionally, adapting existing classes on specific research methods to emphasize the specific inferences that methods are used to support may also help to avoid breakdowns. These shifts in methodology training would be facilitated by explicitly cataloguing the set of inferences necessary to support typical claims. Some examples of such cataloguing exist. For example, Cronbach (1982) emphasized the common need to generalize from specific units, treatments, observations/measures, and settings to broader units, treatments, observations/measures, and settings. Similarly, the classic book by Shadish, Cook, and Campbell (2002) catalogues a broad set of inferences that are necessary for making causal claims.

While these changes in methods training could help, we have emphasized the resilience of some methods that promote breakdowns despite knowledge of problems in a method, a resilience that is driven by individual researchers' personal success in applying a method. One way to combat this resilience would be to highlight the prevalence of breakdowns in specific fields, which, we argued, is a driver behind the strong recent efforts to reconsider the use of NHST. Individual researchers could work towards this goal by selecting highly-cited articles and decomposing the implicit inferential argument and the ways that described methods appear to support that inferential argument. This could, if done well, force specific fields to reevaluate the reasonableness of claims and confront researchers with the limitations of chosen methods within the context of their own (or similar) work. Care must be taken, though, to respect the complexity of trying to build and support inferential arguments and the impossibility of providing strong support to every possible inference within an inferential argument. The goal of such work should not be to embarrass individual researchers, but to provide a lens for self-reflection in a field.

Adaptation of Existing Methods

A third takeaway from our framework is the importance of adapting existing methods to remove the logic of confidence and/or deductive-style logics, or developing new methods that avoid these two conditions. Earlier in this paper, we discussed regression analysis as a positive example of an institutionalized method, noting that software and routines make it easy to conduct regressions and check regression model assumptions. One important inference supported by regression analysis is the functional-form inference (i.e., that the functional form of the modelled relationship between independent and dependent variables is correct). Diagnostic plots of residuals provide direct evidence for this inference, as violations of the inference often show up in such plots. However, it is generally not routine or expected that such plots are presented within papers nor do reviewers, in our experience, expect such plots to be presented during the peer review process. It is unclear, consequently, how often such diagnostic plots are examined. For methods like regression analysis, then, there may simply be a need to shift reporting norms, encouraging peer reviewers and editors to demand direct evidence for inferences to be presented, either in papers or appendices. In other cases, where no way of presenting direct evidence exists, methods may have to be adapted or developed so that direct evidence can be provided or deductive logics avoided. Still in other cases, there will likely be a need to develop new methods

that avoid the logic of confidence and/or deductive logics conditions.

Conclusion

We have created a theoretical explanation for why breakdowns both occur and can be sustained across time. Breakdowns are likely to occur when the evaluation of how research methods are applied substitutes for the evaluation of inferential arguments. They are sustained by research methods that rely on (a) logics of confidence that provide no empirical validation for important inferences and (b) deductive-style reasoning that leads researchers to be overly confident in conclusions. These two conditions allow individual researchers to "successfully" use research methods without ever becoming aware of breakdowns that exist within their own work, which gives research methods a self-perpetuating momentum that is difficult to combat through methodological critiques and criticism. Addressing breakdowns, then, requires being more explicit about claims, inferential arguments, and the ways that methods support inferential arguments, as well as adaptations to methods to avoid the two identified conditions and training that helps researchers to adjust to the shift towards more making more explicit inferential arguments.

Contributions

Contributed to conception and design: MW, RB
Drafted the article: MW, RB
Revised the article: MW
Approved the submitted version for publication: MW

Funding Information

The first author's time was supported by Nordforsk, grant number: 87663

Competing Interests

The authors declare they have no competing interests.

Data accessibility statement

There are no data associated with this paper.

Submitted: February 27, 2024 PDT, Accepted: July 12, 2024 PDT

References

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist*, *63*(1), 32–50. https://doi.org/10.1037/0003-066X.63.1.32

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication. *The American Statistician*, *73*(sup1), 262–270. https://doi.org/10.1080/00031305.2018.1543137

APA Style Journal Article Reporting Standards. (n.d.). Https://Apastyle.Apa.Org. Retrieved May 23, 2023, from https://apastyle.apa.org/jars

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617. https://doi.org/10.1348/000712608X377117

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

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.). L. Erlbaum Associates.

Cronbach, L. J. (1982). *Designing Evaluations of Educational and Social Programs* (First Edition). Jossey-Bass Inc Pub.

DiMaggio, P. J., & Powell, W. W. (1983). The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields. *American Sociological Review*, *48*(2), 147–160. https://doi.org/10.2307/2095101

Fried, E. I. (2020). Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature. *Psychological Inquiry*, *31*(4), 271–288. https://doi.org/10.1080/1047840X.2020.1853461

Gehlbach, H., & Robinson, C. D. (2017). Mitigating Illusory Results through Preregistration in Education. *Journal of Research on Educational Effectiveness*, *11*(0), 1–20. https://doi.org/10.1080/19345747.2017.1387950

Gershman, S. J. (2019). How to never be wrong. *Psychonomic Bulletin & Review*, *26*(1), 13–28. https://doi.org/10.3758/s13423-018-1488-8

Gigerenzer, G. (2018). Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science*, *1*(2), 198–218. https://doi.org/10.1177/2515245918771329

Haig, B. D. (2014). *Investigating the Psychological World: Scientific Method in the Behavioral Sciences*. MIT press. https://doi.org/10.7551/mitpress/9780262027366.001.0001

Haig, B. D. (2018). An Abductive Theory of Scientific Method. In B. D. Haig (Ed.), *Method Matters in Psychology: Essays in Applied Philosophy of Science* (pp. 35–64). Springer International Publishing. https://doi.org/10.1007/978-3-030-01051-5_3

Halpin, P. F., & Stam, H. J. (2006). Inductive Inference or Inductive Behavior: Fisher and Neyman: Pearson Approaches to Statistical Testing in Psychological Research (1940-1960). *The American Journal of Psychology*, *119*(4), 625–653. https://doi.org/10.2307/20445367

Hancock, G. R., Stapleton, L. M., & Mueller, R. O. (Eds.). (2019). *The reviewer's guide to quantitative methods in the social sciences* (Second Edition). Routledge.

Hedges, L. V. (2018). Challenges in Building Usable Knowledge in Education. *Journal of Research on Educational Effectiveness*, *11*(1), 1–21. https://doi.org/10.1080/19345747.2017.1375583

Hoekstra, R., & Vazire, S. (2021). Aspiring to greater intellectual humility in science. *Nature Human Behaviour*, *5*(12), 1602–1607. https://doi.org/10.1038/s41562-021-01203-8

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hubbard, R. (2004). Alphabet Soup: Blurring the Distinctions Between p's and a's in Psychological Research. *Theory & Psychology*, *14*(3), 295–327. https://doi.org/10.1177/0959354304043638

Hubbard, R., Haig, B. D., & Parsa, R. A. (2019). The Limited Role of Formal Statistical Inference in Scientific Inference. *The American Statistician*, *73*(sup1), 91–98. https://doi.org/10.1080/00031305.2018.1464947

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Praeger Publishers.

King, K. A. (2024). Promises and perils of positionality statements. *Annual Review of Applied Linguistics*, 1–8. https://doi.org/10.1017/S0267190524000035

Kish, L. (1965). *Survey sampling*. J. Wiley.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, *114*(1), 185. https://doi.org/10.1037/0033-2909.114.1.185

Makel, M. C., & Plucker, J. A. (2014). Facts Are More Important Than Novelty Replication in the Education Sciences. *Educational Researcher*, *43*(6), 304–316. https://doi.org/10.3102/0013189X14545513

Makel, M. C., Smith, K. N., McBee, M. T., Peters, S. J., & Miller, E. M. (2019). A Path to Greater Credibility: Large-Scale Collaborative Education Research. *AERA Open*, *5*(4). https://doi.org/10.1177/2332858419891963

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In Search of Golden Rules: Comment on Hypothesis-Testing Approaches to Setting Cutoff Values for Fit Indexes and Dangers in Overgeneralizing Hu and Bentler's (1999) Findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

Maul, A. (2017). Rethinking Traditional Methods of Survey Validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*(2), 51–69. https://doi.org/10.1080/15366367.2017.1348108

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, *113*(1), 181. https://doi.org/10.1037/0033-2909.113.1.181

Mayo, D. G., & Spanos, A. (2006). Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction. *The British Journal for the Philosophy of Science*, *57*(2), 323–357. https://doi.org/10.1093/bjps/axl003

McGrane, J. A., & Maul, A. (2020). The human sciences, models and metrological mythology. *Measurement*, *152*, 107346. https://doi.org/10.1016/j.measurement.2019.107346

McNeish, D., & Wolf, M. G. (2021). Dynamic fit index cutoffs for confirmatory factor analysis models. *Psychological Methods*, NoPaginationSpecified-NoPaginationSpecified. https://doi.org/10.1037/met0000425

Meyer, J. W., & Rowan, B. (1977). Institutionalized Organizations: Formal Structure as Myth and Ceremony. *American Journal of Sociology*, *83*(2), 340–363. https://doi.org/10.1086/226550

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123. https://doi.org/10.3758/s13423-015-0947-8

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

Nelder, J. A. (1999). From Statistics to Statistical Science. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *48*(2), 257–269.

Pashler, H., & Ruiter, J. P. de. (2017). Taking Responsibility for Our Field's Reputation. *APS Observer*, *30*. https://www.psychologicalscience.org/observer/taking-responsibility-for-our-fields-reputation

R Core Team. (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Renkewitz, F., & Heene, M. (2019). The replication crisis and open science in psychology: Methodological challenges and developments. *Zeitschrift Fur Psychologie*, *227*(4), 233–236. https://doi.org/10.1027/2151-2604/a000389

Robinson, D. H., Levin, J. R., Schraw, G., Patall, E. A., & Hunt, E. B. (2013). On Going (Way) Beyond One's Data: A Proposal to Restrict Recommendations for Practice in Primary Educational Research Journals. *Educational Psychology Review*, *25*(2), 291–302. https://doi.org/10.1007/s10648-013-9223-5

Sainani, K. L. (2018). The Problem with "Magnitude-based Inference." *Medicine & Science in Sports & Exercise*, *50*(10), 2166. https://doi.org/10.1249/MSS.0000000000001645

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

St. Clair, R. (2005). Similarity and Superunknowns: An Essay on the Challenges of Educational Research. *Harvard Educational Review*, *75*(4), 435–453. https://doi.org/10.17763/haer.75.4.a263u5q535658h41

Tell it like it is. (2020). *Nature Human Behaviour*, *4*(1), Article 1. https://doi.org/10.1038/s41562-020-0818-9

Tunç, D. U., Tunç, M. N., & Lakens, D. (2023). The epistemic and pragmatic function of dichotomous claims based on statistical hypothesis tests. *Theory & Psychology*, 09593543231160112. https://doi.org/10.1177/09593543231160112

van de Grift, W. J. C. M., Houtveen, T. A. M., Hurk, H. T. G. van den, & Terpstra, O. (2019). Measuring teaching skills in elementary education using the Rasch model. *School Effectiveness and School Improvement*, *30*(4), 455–486. https://doi.org/10.1080/09243453.2019.1577743

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

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond "p < 0.05." *The American Statistician*, *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

White, M. (2024, February 8). *A Peculiarity in Educational Measurement Practices*. https://doi.org/10.31219/osf.io/qzynh

Wilhelm, A. G., Rouse, A. G., & Jones, F. (2018). Exploring Differences in Measurement and Reporting of Classroom Observation Inter-Rater Reliability. *Practical Assessment, Research & Evaluation*, *23*(4), 16.

Yarkoni, T. (2020). Implicit Realism Impedes Progress in Psychology: Comment on Fried (2020). *Psychological Inquiry*, *31*(4), 326–333. https://doi.org/10.1080/1047840X.2020.1853478

Supplementary Materials

Peer Review Communication

Download: https://collabra.scholasticahq.com/article/121436-breakdowns-in-scientific-practices-how-and-why-some-accepted-scientific-claims-may-have-little-actual-support/attachment/237082.docx?auth_token=JPf-6ICt6QI-Jnq6D2ta