Illusions of replication, illusions of truth

Clintin P. Davis-Stober[1,2,*], Konstantina Sokratous[1], and Joachim Vandekerckhove[3]

This is the author final copy of:
Davis-Stober, C. P., Sokratous, K., & Vandekerckhove, J. (2026). Illusions of replication, illusions of truth. *Computational Brain & Behavior*.

Shiffrin et al. argue that scientific practice often produces illusions of understanding; situations in which familiar inferential tools generate misplaced confidence. We describe a related illusion that arises not from statistical misapplications, but from ordinary theory testing itself. Here we highlight a different source of epistemic uncertainty or overconfidence; one that emerges even when statistical tools are used correctly, effect sizes are precisely estimated, findings replicate without controversy, and heterogeneity is explicitly acknowledged within hierarchical models.

paradox of converging evidence | replication | philosophy of science | scientific uncertainty

Shiffrin, Stigler, and Keil (2026) argue that common scientific practice may produce illusions of understanding: familiar inferential tools and models can encourage misplaced confidence, and prediction is not causation. They catalog a number of illusion types—linked to frequent practices best avoided—and discuss levels of understanding, but they do not examine how evidence accumulates when theories are tested across many studies, something we believe to be core to the practice of science. Here, we describe a structural illusion that arises from the accumulative process, using a common scenario that involves statistical tools that are used correctly, effect sizes that are precisely estimated, and findings that replicate without controversy.

This illusion is not primarily about misuse of statistics—indeed it occurs when inferential statistics are implemented in entirely conventional ways—but follows from how theories are routinely implicitly operationalized and tested, with predictions being checked one by one at the population level, while theory satisfaction is intended to be conjunctive at the individual level.

To see the illusion, consider a familiar pattern of scientific practice. A theory $\mathcal{T}$ implies multiple predictions. Each prediction is translated into a treatment-control contrast. Separate experiments are conducted, each well powered, preregistered, independently replicated. If every predicted effect emerges in the expected direction, confidence in $\mathcal{T}$ increases. Additional entailments, especially those yielding larger effects, appear to strengthen the evidential base.

When science is conducted in this manner, it is constrained by at least three structural features:

1. Predictions accumulate conjunctively at the individual level; to satisfy the theory an individual must satisfy all predictions.
2. Evidence accumulates additively at the population; each positive effect contributes incremental support.
3. Observed effects aggregate across heterogeneous individuals; population means can conceal substantial variation in individual response patterns.

These features are not statistical mistakes – they are ordinary features of how researchers test theories. Yet, they create a subtle limitation: *population level confirmation does not imply that many, or even* any*, individuals satisfy all predictions of a theory simultaneously*. This limitation gives rise to what Davis-Stober and Regenwetter (2019) called the "paradox of converging evidence." As more predictions are confirmed and more effects are replicated, the appearance of evidential convergence strengthens. At the same time, the proportion of individuals who satisfy all predictions can shrink or even collapse, depending on the degree of interindividual heterogeneity.

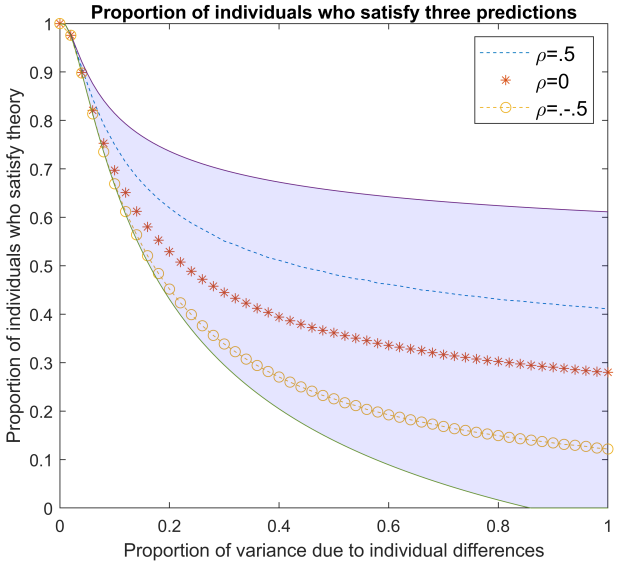
Under their framework, there are three key ingredients when determining the proportion of participants in the population who satisfy a theory $\mathcal{T}$: (1) the number of predictions entailed by $\mathcal{T}$, (2) the size of the corresponding effects, and (3) the proportion of variance in the population that is attributable to true individual differences as compared to 'noise'. As an illustration, suppose that $\mathcal{T}$ consisted of three predictions, with each prediction operationalized as a population mean on some dependent variable being greater in value under a treatment condition compared to a control condition (i.e., $\mu_c^i < \mu_t^i$, for the i^{th} prediction). Effect size is calculated via $\delta_i = \frac{\mu_t^i - \mu_c^i}{\sigma}, i \in \{1, 2, 3\}$, where σ is the population variance of the dependent variable. Suppose three effect sizes, corresponding to the three predictions, were perfectly estimated to be $\delta_1 = .4, \delta_2 = .6$, and $\delta_3 = .7$. The top graph of Figure 1 displays the proportion of individuals who could satisfy all three predictions of $\mathcal{T}$ as a function of the proportion of σ that is attributable to true individual differences. When the x-axis is zero, σ reflects only noise, and no true heterogeneity, so all participants are impacted by each treatment condition, in each experiment, in exactly the same way. When the x-axis equals 1, there is no 'noise' to observe, all variation reflects actual differences in true scores on the dependent variable. The top curve reflects the upper bound on the proportion of individuals who could satisfy $\mathcal{T}$, while the bottom curve denotes the lower bound. The shaded region reflects all possible values of this proportion.

As can be seen in the top graph of Figure 1, once even a small amount of true individual differences appear, the lower bound quickly drops. If 40% of σ reflects true individual differences, at best, 70% of the population satisfy the theory, at worst, only 20% do. Perfect replication, even knowledge of true effect sizes, does not make this uncertainty go away. Suppose that a new prediction from $\mathcal{T}$ was tested and an effect size of $\delta_4 = .35$ was

[1]Department of Psychological Sciences, University of Missouri, USA; [2]MU Institute for Data Science and Informatics, University of Missouri, USA; [3]Department of Cognitive Sciences, University of California, Irvine, USA
All authors contributed to the final draft.
To whom correspondence should be addressed. E-mail: stoberc@missouri.edu.
The authors declare no conflicts of interest.

perfectly estimated, replicated, etc. As it can be seen by examining the bottom graph of Figure 1, this additional effect dramatically increases this uncertainty. Unless the amount of true heterogeneity is very small, we cannot be certain that even a single individual satisfies $\mathcal{T}$.

Heck (2021) refined the model of Davis-Stober and Regenwetter (2019) by explicitly modeling the correlations among true difference scores across the different predictions of $\mathcal{T}$. For simplicity, suppose that this correlation, denoted ρ, is equal across all prediction pairs. Under Heck's formulation, the upper curves in Figure 1 are attainable only when $\rho = 1$, indicating that if a randomly drawn participant satisfies one prediction they will necessarily satisfy all other predictions. When the correlations are maximally negative, a somewhat less pessimistic lower bound than that obtained by Davis-Stober and Regenwetter (2019) is attained – see also Davis-Stober and Regenwetter (2022). The 'paradox' of converging evidence, however, does not disappear. Even a 'reasonable' value of $\rho = .5$ demonstrates that the proportion of individuals who satisfy $\mathcal{T}$, even for these perfectly estimated and replicated effects, falls below a majority, especially when four predictions are considered.



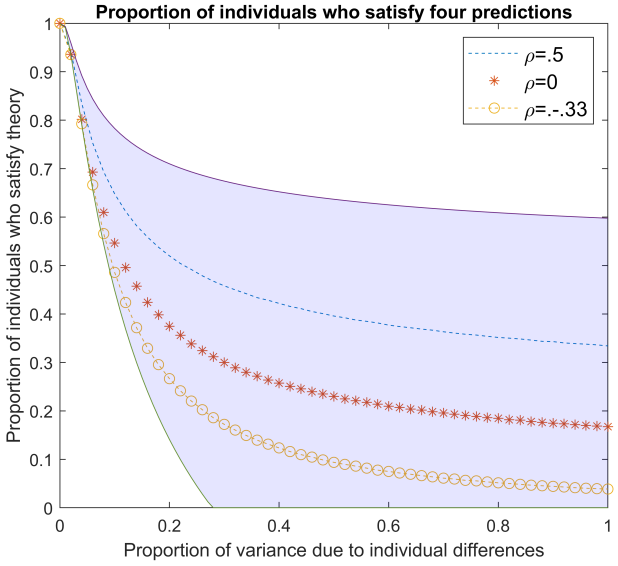


Fig. 1. The top figure shows the proportion of individuals who satisfy theory $\mathcal{T}$ considering four effect sizes:$\delta_1 = .4, \delta_2 = .6, \delta_3 = .7$. The bottom graph considers size effect sizes:$\delta_1 = .4, \delta_2 = .6, \delta_3 = .7, \delta_4 = .35$.

Linking back to Shiffrin et al., this is another example of the need for better bridges between various types of uncertainty. When theories are evaluated through multiple population-level effects, evidence accumulates additively across studies, while theory satisfaction remains conjunctive at the individual level. As heterogeneity increases, the space of individuals who satisfy all predictions can contract even as the number of successful replications grows.

The resulting illusion is consequential: Converging evidence at the population level can create the impression that a theory is broadly true, when in fact its individual level applicability may be limited, underdefined, indeterminate, or even null. Statistical tools exist that directly model individual level constraints (Rouder, Haaf, Davis-Stober, & Hilgard, 2019; Haaf & Rouder, 2019; Schnuerch, Nadarevic, & Rouder, 2021) and alternative probabilistic formulations of theory satisfaction may mitigate the paradox (Davis-Stober & Regenwetter, 2019; Kellen, Pedersen, & Klauer, 2023). However, they do not solve the conceptual problem of how we operationalize theories. Only the way we define what it means for a theory to be satisfied can place inherent limits on what replication alone can establish. Recognizing these limits does not weaken science. Instead, it clarifies what our evidence can and cannot warrant. Without explicitly modeling heterogeneity and joint structure, converging evidence may converge only in appearance – and in that appearance lurks another illusion of understanding.

References

Davis-Stober, C. P., & Regenwetter, M. (2019). The 'paradox'of converging evidence. *Psychological review*, *126*(6), 865-879.

Davis-Stober, C. P., & Regenwetter, M. (2022). A re-examination of the lower bounds on the 'paradox'of converging evidence.

Haaf, J. M., & Rouder, J. N. (2019). Some do and some don't? accounting for variability of individual difference structures. *Psychonomic Bulletin & Review*, *26*(3), 772–789.

Heck, D. (2021). Assessing the "paradox" of converging evidence by modeling the joint distribution of individual differences: Comment on davis-stober and regenwetter (2019). *Psychological review*, *128*(8), 1187-1196.

Kellen, D., Pedersen, A. P., & Klauer, K. C. (2023). On piecemeal testing and the possibility of psychological science.

Rouder, J. N., Haaf, J. M., Davis-Stober, C. P., & Hilgard, J. (2019). Beyond overall effects: A bayesian approach to finding constraints in meta-analysis. *Psychological Methods*, *24*(5), 606.

Schnuerch, M., Nadarevic, L., & Rouder, J. N. (2021). The truth revisited: Bayesian analysis of individual differences in the truth effect. *Psychonomic Bulletin & Review*, *28*(3), 750–765.

Shiffrin, R., Stigler, S., & Keil, F. (2026). Illusions of understanding in the sciences. *Computational Brain & Behavior*.