What can we expect of MINFLUX, and more broadly, superresolution and single-molecule imaging? Because MINFLUX can now reach a resolution less than 5 nm, single-molecule fluorescence resonance-energy transfer, which can determine distances of up to ~7 nm at 0.3-nm resolution with only about 100 photons (*11*), may be combined to obtain dynamic structural information continuously covering from the length scale of single amino acids to the cellular scale or larger. A considerable challenge would be to extend the molecular resolution to three-dimensional imaging, which most certainly would require interferometric methods (*12*). Moving toward multicolor imaging is likely to be more straightforward because the precision in position determination is largely wavelength-independent in MINFLUX and because more fluorescent reporters become eligible because of the reduced photon budget.

> ***"The minimal photon flux... is particularly advantageous for single-molecule tracking experiments..."***

Ultimately, the true spatial resolution of an image is going to be limited by how densely the sample can be labeled, However, the greater resolving power achieved at molecular distances that has been enabled by MINFLUX is likely to stimulate further developments in probe and labeling technologies. MINFLUX also requires more hardware engineering as compared with other localization-based nanoscopy. Nevertheless, rapid commercialization, pending further developments necessary for cellular imaging, may make it available to biologists in the not-too-distant future. ■

REFERENCES

1. C. Eggeling, K. I. Willig, S. J. Sahl, S. W. Hell, *Q. Rev. Biophys.* **48**, 178 (2015).
2. F. Balzarotti *et al.*, *Science* **355**, 606 (2017).
3. A. Yildiz *et al.*, *Science* **300**, 2061 (2003).
4. E. Betzig *et al.*, *Science* **313**, 1642 (2006).
5. M. J. Rust, M. Bates, X. Zhuang, *Nat. Methods* **3**, 793 (2006).
6. T. A. Klar, S. Jakobs, M. Dyba, A. Egner, S. W. Hell, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 8206 (2000).
7. M. G. Gustafsson, *J. Microsc.* **198**, 82 (2000).
8. M. Dai, R. Jungmann, P. Yin, *Nat. Nanotechnol.* **11**, 798 (2016).
9. S. W. Hell, J. Wichmann, *Opt. Lett.* **19**, 780 (1994).
10. E. Betzig, *Opt. Lett.* **20**, 237 (1995).
11. T. Ha *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 6264 (1996).
12. G. Shtengel *et al.*, *Proc. Natl. Acad.Sci. U.S.A.* **106**, 3125 (2009).

10.1126/science.aam5409

STATISTICS

Measurement error and the replication crisis

The assumption that measurement error always reduces effect sizes is false

By **Eric Loken**[1] *and* **Andrew Gelman**[2]

Measurement error adds noise to predictions, increases uncertainty in parameter estimates, and makes it more difficult to discover new phenomena or to distinguish among competing theories. A common view is that any study finding an effect under noisy conditions provides evidence that the underlying effect is particularly strong and robust. Yet, statistical significance conveys very little information when measurements are noisy. In noisy research settings, poor measurement can contribute to exaggerated estimates of effect size. This problem and related misunderstandings are key components in a feedback loop that perpetuates the replication crisis in science.

It seems intuitive that producing a result under challenging circumstances makes it all the more impressive. If you learned that a friend had run a mile in 5 minutes, you would be respectful; if you learned she had done it while carrying a heavy backpack, you would be awed. The obvious inference is that she would have been even faster without the backpack. But should the same intuition always be applied to research findings? Should we assume that if statistical significance is achieved in the presence of measurement error, the associated effects would have been stronger without noise? We caution against the fallacy of assuming that that which does not kill statistical significance makes it stronger.

Measurement error can be defined as random variation, of some distributional form, that produces a difference between observed and true values (*1*). Measurement error and other sources of uncontrolled variation in scientific research therefore add noise. The latter is typically an attenuating factor, as acknowledged in various scientific disciplines. Spearman (*2*) famously derived a formula for the attenuation of observed correlations due to unreliable measurement. In epidemiology, it is textbook knowledge that nondifferential misclassification tends to bias relative risk estimates toward the null (*3*). According to Hausman's "iron law" of econometrics, effect sizes in simple regression models are underestimated when the predictors contain error variance (*4*).

It is understandable, then, that many researchers have the intuition that if they manage to achieve statistical significance under noisy conditions, the observed effect would have been even larger in the absence of noise. As with the runner, they assume that without the burden—that is, uncontrolled variation—their effects would have been even larger (*5*–*7*).

The reasoning about the runner with the backpack fails in noisy research for two reasons. First, researchers typically have so many "researcher degrees of freedom"—unacknowledged choices in how they prepare, analyze, and report their data—that statistical significance is easily found even in the absence of underlying effects (*8*) and even without multiple hypothesis testing by researchers (*9*). In settings with uncontrolled researcher degrees of freedom, the attainment of statistical significance in the presence of noise is not an impressive feat.

The second, related issue is that in noisy research settings, statistical significance provides very weak evidence for either the sign or the magnitude of any underlying effect. Statistically significant estimates are, roughly speaking, at least two standard errors from zero. In a study with noisy measurements and small or moderate sample size, standard errors will be high and statistically significant estimates will therefore be large, even if the underlying effects are small. This is known as the statistical significance filter and can be a severe upward bias in the magnitude of effects; as one of us has shown, reported estimates can be an order-of-magnitude larger than any plausible underlying effects (*10*).

In a low-noise setting, the theoretical results of Hausman and others correctly show that measurement error will attenuate coefficient estimates. But we can demonstrate with a simple exercise that the opposite oc-

[1] Department of Educational Psychology, University of Connecticut, Storrs, CT 06269-3815, USA. [2] Department of Statistics and Department of Political Science, Columbia University, New York, NY 10027-6902, USA. Email: gelman@stat.columbia.edu

curs in the presence of high noise and selection on statistical significance.

Suppose we measure x and y in a setting where the underlying truth is that there is a small effect of x on y. Imagine four conditions based on changes in two factors. First, we might have either a high-powered study (sample size $N = 3000$) or a low-powered study ($N = 50$). Second, we might have measurements on x and y that are high quality, or have some degree of additional measurement error. In the large-N scenario, adding measurement error will almost always reduce the observed correlation between x and y (see the figure, left panel). But in the small-N setting, this will not hold; the observed correlation can easily be larger in the presence of measurement error (see the figure, middle panel).

Take these scenarios and now add selection on statistical significance. We can track the proportion of studies, as a function of sample size, where the observed effect is larger than the original error-free effect. For the largest samples, the observed effect is always smaller than the original. But for smaller N, a fraction of the observed effects exceeds the original. If we were to condition on whether or not the observed effect was statistically significant, then the fraction is even larger (see the figure, right panel).

Our concern is that researchers are sometimes tempted to use the "iron law" reasoning to defend or justify surprisingly large statistically significant effects from small studies. If it really were true that effect sizes were always attenuated by measurement error, then it would be all the more impressive to have achieved significance. But to acknowledge that there may have been a substantial amount of uncontrolled variation is to acknowledge that the study contained less information than was initially thought. If researchers focus on getting statistically significant estimates of small effects, using noisy measurements and small samples, then it is likely that the additional sources of variance are already making the t test look strong. Measurement error and selection bias thus can combine to exacerbate the replication crisis.

The situation becomes more complicated in problems with multiple predictors, or with nonindependent errors. Wacholder *et al.* (*11*) discuss scenarios beyond simple two-group risk-exposure studies where misclassification can lead to exaggerated estimates. For the simpler setting, though, they conclude that while "the estimate may exceed the true value…it is more likely to fall below the true value." We agree with Wacholder *et al.* for studies in which effects and sample sizes are large. But for noisier studies, especially combined with selective filtering on statistically significant observed effects, we think that there is a greater chance that the effects are exaggerated rather than attenuated. Jurek *et al.* have also provided evidence that individual research studies can be biased away from the null (*12*).

A key point for practitioners is that surprising results from small studies should not be defended by saying that they would have been even better with improved measurement. Furthermore, the signal-to-noise ratio cannot in general be estimated merely from internal evidence. It is a common mistake to take a t-ratio as a measure of strength of evidence and conclude that just because an estimate is statistically significant, the signal-to-noise level is high. It is also a mistake to assume that the observed effect size would have been even larger if not for the burden of measurement error. Intuitions that are appropriate when measurements are precise are sometimes misapplied in noisy and more probabilistic settings.

The consequences for scientific replication are obvious. Many published effects are overstated and future studies, powered by the expectation that the effects can be replicated, might be destined to fail before they even begin. We would all run faster without a backpack on our backs. But when it comes to surprising research findings from small studies, measurement error (or other uncontrolled variation) should not be invoked automatically to suggest that effects are even larger. ■

REFERENCES AND NOTES

1. S. Messick, *Validity*, ETS Research Report Series RR-87-40 (Educational Testing Service, Princeton, NJ, 1987).
2. C. Spearman, *Am. J. Psychol.* **15**, 72 (1904).
3. K. J. Rothman, S. Greenland, T. L. Lash, *Modern Epidemiology* (Kluwer, ed. 3, 2008).
4. J. Hausman, *J. Econ. Perspect.* **15**, 57 (2001).
5. J. J. Heckman, *Boston Rev.* (1 September 2012); http://bostonreview.net/forum/promoting-social-mobility/final-response-aiding-life-cycle-james-heckman.
6. J. K. Maner, *J. Exp. Soc. Psychol.* **66**, 100 (2016).
7. S. Goldin-Meadow, *Assoc. for Psychol. Sci. Observer* (October 2016); www.psychologicalscience.org/observer/preregistration-replication-and-nonexperimental-studies.
8. J. Simmons, L. Nelson, U. Simonsohn, *Psychol. Sci.* **22**, 1359 (2011).
9. A. Gelman, E. Loken, *Am. Sci.* **102**, 460 (2014).
10. A. Gelman, J. B. Carlin, *Perspect. Psychol. Sci.* **9**, 641 (2014).
11. S. Wacholder, P. Hartge, J. H. Lubin, M. Dosemeci, *Occup. Environ. Med.* **52**, 557 (1995).
12. A. M. Jurek, S. Greenland, G. Maldonado, T. R. Church, *Int. J. Epidemiol.* **34**, 680 (2005).

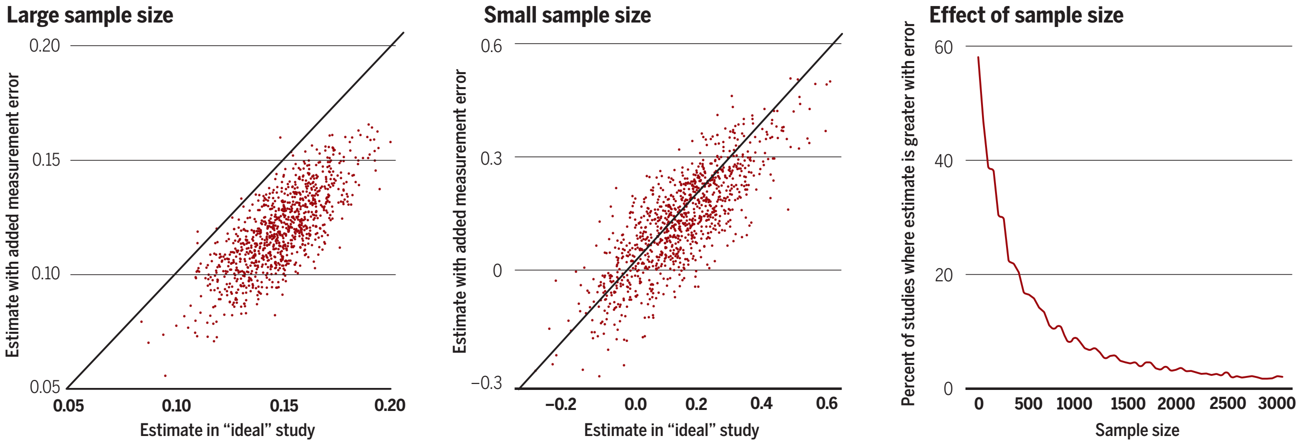
ACKNOWLEDGMENTS

We thank the Office of Naval Research for grant N00014-15-1-2541.

10.1126/science.aal3618

Distribution of statistically significant estimates in the presence of added error

To obtain the graphs, effect sizes from simulated studies were estimated in the "ideal" setting and after adding random error. For large-*N* studies, added error always reduces the effect. For small *N*, the reverse can be true. Of statistically significant effects observed after error, a majority could be greater than in the "ideal" setting when *N* is small.



GRAPHIC: G. GRULLÓN/*SCIENCE*