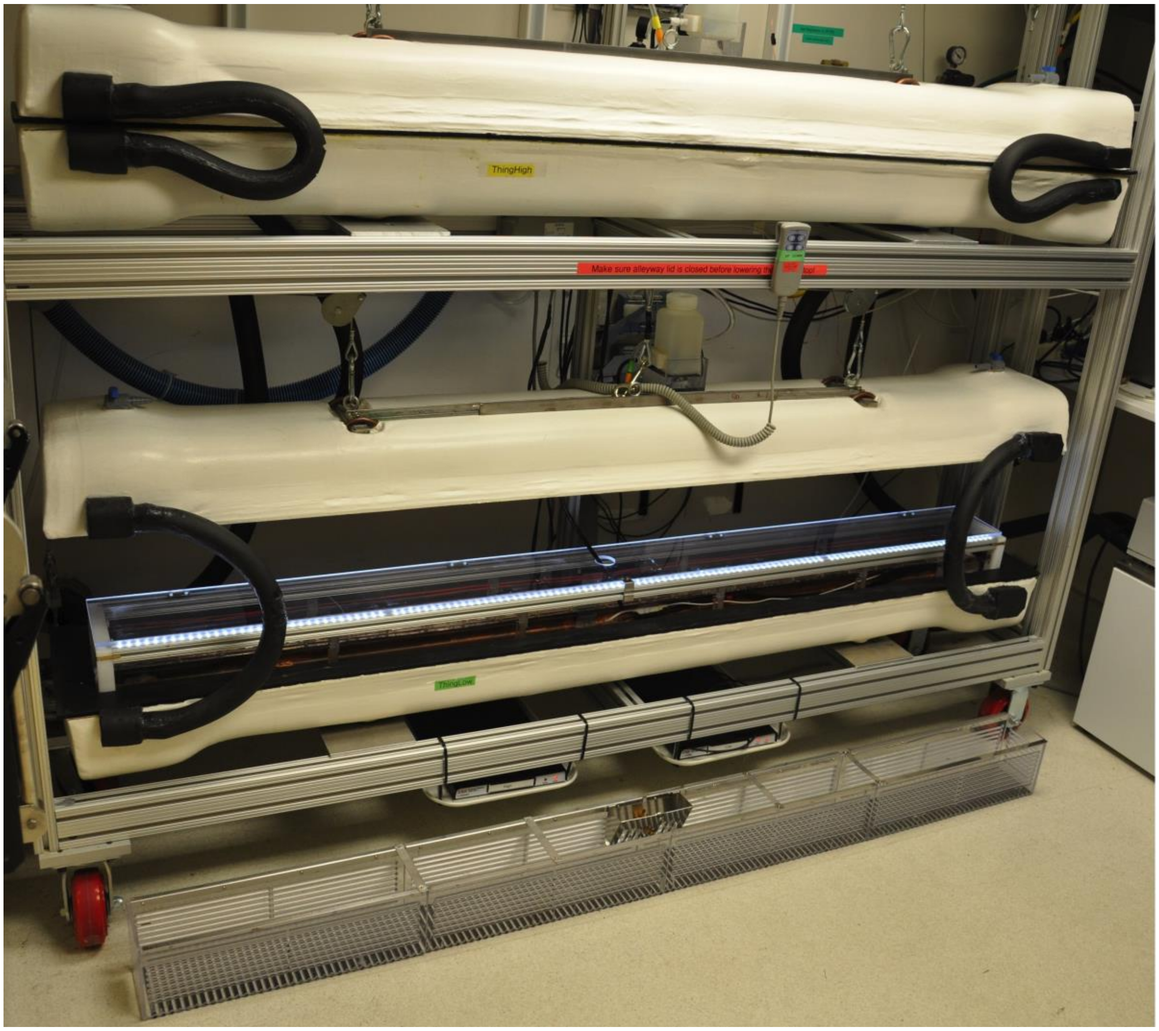
Supporting Information (Online Only)

Part I. Thermal Gradient.

Our lab's two thermal gradients (Supplemental Figure 1) are based on a previously published design (Gordon et al., 1991; Gordon, 2005) that we modified to make each gradient gas-tight. The rectangular cross-section of the gradient allows for easy installation of gas-tight fittings and provided a broad 5-cm lip between the upper and lower halves of each gradient for a 6.35-mm thick compressible closed-cell neoprene gasket (Type G-231-N, Rubatex International, Bedford, VA). Each thermal gradient is made of a 6.35-mm thick copper shell that surrounds an internal rectangular space that is 190.50 cm long, 17.14 cm high and 17.14 cm wide. A 10.16-cm wide water jacket is welded onto each end of the copper shell. Two recirculating water baths (Thermo Electron's NesLab Model RTE7, Newington, NH) pump water through the water jackets (warm water to one end and cool water to the other), thereby creating a temperature continuum along the length of the copper shell that surrounds the alleyway. The external surface of each gradient is heavily insulated and sealed with white rubberized paint. A removable acrylic alleyway (182.9 cm long, 12.06 cm high, 12.06 cm wide) is suspended from an internal support structure within the thermal gradient and removable waste trays are placed below the alleyway's rod flooring. The rat's location in the alleyway is detected using 24 infrared beams (model OPB100Z, Optek, Carrollton, TX, USA) that are placed 19 mm above the alleyway floor and positioned 76.2 mm apart along the length of the alleyway, ending 38.1 mm from either end. A thermistor with 0.1°C accuracy (model KS222J2, U.S. Sensor Corporation, Orange, CA, USA) is positioned 10 mm above each of the 24 infrared transmitters and measures the ambient temperature at that location. An antenna made from two continuous 22-gauge solid-strand insulated wires surrounds the lateral walls of the alleyway and receives the telemetric T_C signal from a sensor within the rat's peritoneal cavity. Uniform illumination during the light portion of the 24-h light/dark cycle is provided by 128 evenly-spaced white light LEDs (Osram, LINEARlight Model #OS-LM01A-W2-854, Munich, Germany) that run the length of the alleyway and project their light upward from the top edge of the internal support structure. Pelleted chow is freely available in the center of the alleyway from a food hopper hung from the alleyway wall. An Edstrom Vari-Flo drinking valve (Part #1000-8000, AgSelect, Waterford, WI) protrudes from the center of the food hopper to provide ad lib water from a source external to the gradient. The internal structure that supports the rat alleyway also supports the antenna, infrared beams, thermistors, and hinged acrylic lid to the alleyway. External to the gradient are three, 8-channel USB-Temp devices (Measurement Computing Corporation, Norton, MA, USA) that are wired to the thermistors. The electronics operating the infrared beams are computer-controlled via the 8 digital I/O lines of each USB-Temp device. Additional descriptive information about the thermal gradient is provided elsewhere (Ramsay et al., 2011).



Supplemental Figure 1. The lab's two thermal gradients are shown in this photograph. The upper gradient is fully closed while the lower gradient is in the open position. The acrylic alleyway has been removed from the lower gradient's internal support structure and placed on the floor for greater clarity. The internal support structure's illuminated LED lighting is visible and its clear acrylic lid is closed (which serves as a lid for the alleyway). A water bottle sits on the surface of the top shell of each gradient and provides water to the drinking valve located in the food dish hung from the center of the alleyway. Temperature controlled water is re-circulated around the ends of each gradient.

Gas Delivery to the Thermal Gradient. A Parker Balston Lab Gas Generator (Model 74-5041NA) used room air to create a continuous supply of purified and dehumidified compressed air, which provided gas for the control gas condition. A digital mass flow controller (Dwyer Instruments DMF-41411, range of 0-15 L/min) delivers the control gas to the thermal gradient at a flow rate of 10 L/min. The 60% N_2O gas condition is composed of 60% N_2O, 21% O_2, and 19% nitrogen (N_2) which is made by blending medical-grade oxygen (O_2), medical-grade N_2O, and control gas from the lab gas generator. Specifically, digital mass flow controllers blend 6 L/min of N_2O (Dwyer Instruments DMF-41411, range of 0-15 L/min), 2.4 L/min of control gas (Dwyer Instruments DMF-41411, range of 0-15 L/min), and 1.6 L/min of O_2 (Dwyer Instruments DMF-41409, range of 0-5 L/min) to deliver 10 L/min of 60% N_2O to the thermal gradient. [A 10 L/min blend of 79% N_2O, 21% O_2, and 0% control gas is delivered for the first 12 min of the 60% N_2O gas condition to achieve the targeted 60% N_2O gas concentration more quickly.] Either 60% N_2O or the control gas enters the gradient through an inlet port located in the top center of the upper copper shell and is dispersed by the acrylic lid covering the alleyway. Gas exits through two outlet ports centered at each end of the lower half of the copper shell. Concentrations of N_2O, O_2, and CO_2 are measured using an infrared gas analyzer (Normocapoxy, Datex Instruments Corp., Helsinki, Finland) that draws gas samples via a t-connector placed in the incurrent and excurrent gas lines connected to the copper shell.

Part II. Rationale For Not Adjusting For Multiple Comparisons.

We believe that multiple comparison adjustment (MCA) presents a host of problems that make it inappropriate for our research. The problems with MCA are discussed in detail elsewhere (Feise, 2002; Perneger, 1998; Rothman, 1990). A summary with additional references that are germane to this issue is presented below.

Original motivation emphasized automated decision-making in accordance with the needs of industrial quality control efforts, not science. In a paper with enormous influence in the adoption of MCA (Neyman & Pearson, 1933), Jerzy Neyman and Egon Pearson wrote: *"We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis. But we may look at the purpose of tests from another view-point. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong. Here, for example, would be such a "rule of behavior": to decide whether a hypothesis, H, of a given type be rejected or not, calculate a specified character, x, of the observed facts; if* $x > x0$ *reject H, if* $x \le x0$ *accept H. Such a rule tells us nothing as to whether in a particular case H is true when* $x \le x0$ *or false when* $x > x0$*. But it may often be proved that if we behave according to such a rule, then in the long run we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false."* (p. 291).

Accordingly, the Neyman-Pearson decision framework (Neyman & Pearson, 1928, 1933) was, in the words of Ronald Feise (Feise, 2002), initially adopted as a means *"…to aid decisions in repetitive industrial circumstances, not to appraise evidence in studies."* Subsequently, the concept of "statistical significance" (alpha, itself a decision metric that originated with Neyman-Pearson) took on an undeserved aura of importance among a vast army of scientists and consumers of scientific information that alpha simply does not deserve.

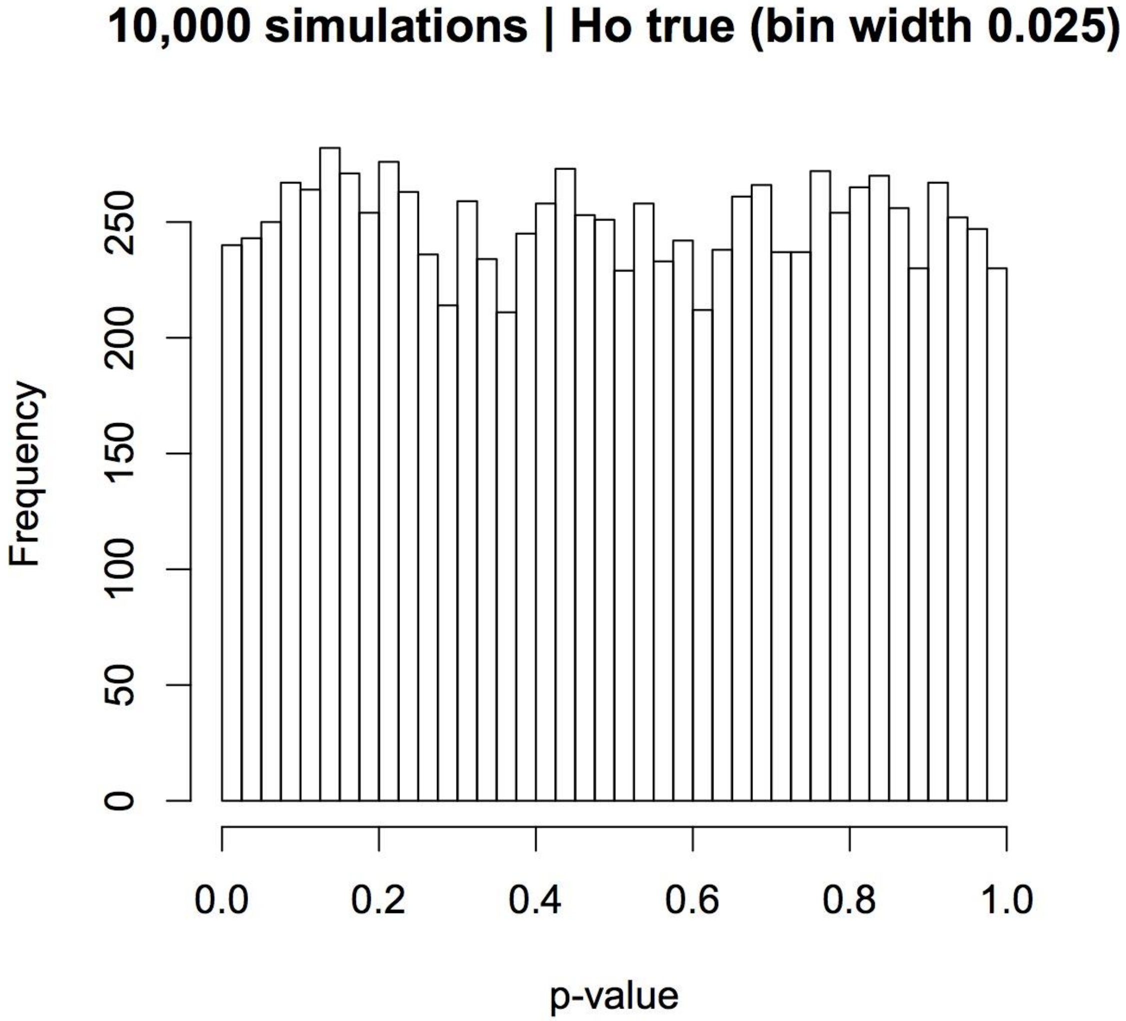
Irrational interpretation of p-values. Adjusting p-values creates the conundrum whereby the significance of each is interpreted in light of the number of dependent variables that are encompassed in the "family-wise" null hypothesis, an ambiguous concept that raises a host of vexing questions. In the words of Feise (Feise, 2002): *"Does "family" include tests that were performed, but not published? Does it include a meta-analysis upon those tests? Should future papers on the same data set be accounted for in the first publication? Should each researcher have a career-wise adjusted p-value? Should we publish an issue-wise adjusted p-value and a year-end-journal-wise adjusted p-value?"*

No consensus on methodology. Owing, in part, to the preceding conundrum, an impressive number of alternative multiple comparison adjustment methods have been developed and advocated (e.g., 18 methods are provided for the general linear model in SPSS Statistics 21), with no consensus as to which (if any) is best justified for each of an array of different study designs.

Type 2 error is a serious problem and p-values are a poor evidential metric. Increased protection against type 1 error comes at the cost of increased type 2 error, a serious problem given the contemporary context of resource scarcity and the much amplified concern regarding the overuse of animals in research. The problematic nature of under-protection against type 2 error is exacerbated by the fact that the failure to achieve statistical significance is frequently interpreted to mean that the null hypothesis is true, an utterly fallacious conclusion that confuses p-values with evidence (Goodman, 1999a, b; Johansson, 2011). As noted above, the argument for adjusting for multiple comparisons is motivated by the school of thought developed by Jerzy Neyman and Egon Pearson in 1933 (Neyman & Pearson, 1933) that emphasizes controlling long-term type 1 error rates in an automated fashion that is largely divorced from the scientific quest of judging, to the extent possible, the truth or falsity of a statistical hypothesis (let alone a scientific hypothesis), i.e., decision error control. This scheme rests, ultimately, on reducing statistical decision making to a binary choice: a statistical finding is or is not "statistically significant" based on an arbitrary p-value cutoff (usually $p<0.05$) and the arbitrary definition of a "family-wise" null hypothesis, as discussed above. Tellingly, however, Neyman and Pearson did not themselves regard p-values as an evidential metric (see above), rather they only sought to control the long run type 1 error rate using simple fixed decision rules [discussed in detail by (Johansson, 2011)]. Well (and long) understood among statisticians is that p-values are of sharply limited evidential value, as they are conditioned solely on the null hypothesis and *are uniformly distributed* if the null is true such that the probability of obtaining any given p-value is essentially constant if the null is true (see Supplemental Figure 2). A renewed effort is underway to inform working scientists as to the use and misuse of p-values. For example, a classic paper by David Lykken (Lykken, 1968) was recently recommended by distinguished University of Alabama biostatistican David Allison (http://f1000.com/prime/718373355). In this paper, Lykken wrote that *"…the finding of statistical significance is perhaps the least important attribute of a good experiment; it is never a sufficient condition for concluding that a theory has been corroborated, that a useful empirical fact has been established with reasonable confidence – or that an experimental report ought to be published.* Arguably the most engaging denouncement of p-values-as-evidence was articulated by Jacob Cohen in his wonderful 1994 paper titled, "`The Earth is Round (p<.05)`" (Cohen, 1994): *"What's wrong with [null hypothesis significance testing]? Well, among many other things, it does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does! What we want to know is "Given these data, what is the probability that Ho is true?" But …what it tells us is "Given that Ho is true, what is the probability of these (or more extreme) data? These are not the same, as has been pointed out many times over the years by [many others]."* The point is that p-values cannot be used to judge the truth or falsity of the null hypothesis, and they cannot be regarded as evidence for or against the alternative hypothesis. For detailed treatments of the problems with p-values see (Cohen, 1994; Goodman, 1999a, b; Johansson, 2011; Lykken, 1968).

If not p-values, what then? Confidence intervals, effect sizes, data patterns and replication. In his classic paper, David Lykken (Lykken, 1968) asserted that "*The value of any research can be determined, not from the statistical results, but only by skilled, subjective* *evaluation of the coherence and reasonableness of the theory, the degree of experimental control employed, the sophistication of the measuring techniques, the scientific or practical importance of the phenomena studied, and so on.*"

Jacob Cohen (Cohen, 1994) recommended the use of graphic methods (pictures!), improvements in and the standardization of measurement, an increased emphasis on estimating effect sizes using confidence intervals, *"and the informed use of available statistical methods is suggested."* [We do so by employing modern correlated-measures regression models, e.g., linear mixed model analysis]. Cohen (1994) stressed that *"psychologists must finally rely, as has been done in all the older sciences, on replication."* Astonishingly, however, it would seem that what Cohen refers to as "the older sciences" is a classification that does not as yet adequately pertain to many contemporary biomedical studies published in prestigious journals. Consider that following recent disclosures that many highly cited and very influential findings published in *Science* and *Nature* could not be replicated, NIH director Francis Collins and principal deputy director Lawrence Tabak published in *Nature* an article underscoring the importance of replication in biomedical research (and the pressing need for better statistical training), and emphasizing that journals should be more welcoming of papers that replicate noteworthy findings (Collins & Tabak, 2014). Similarly, Marcia McNutt, the Editor-in-Chief of *Science* called for increased priority on reproducibility in a 2014 editorial (McNutt, 2014).



Supplemental Figure 2. Monte Carlo simulation illustrating that p-values are uniformly distributed when the null hypothesis is true, i.e., when groups are sampled from a single population. This simulation involved two groups (n=12 each) and was performed using an analysis of covariance power simulation program written in R. The point is that the probability of obtaining any given p-value is essentially constant across p-values (in theory, the probability distribution is totally flat) such that one cannot gauge the truth or falsity of the null hypothesis when the null is true.

References for the Online Supplement

Cohen J (1994) The Earth Is Round (P-Less-Than.05). *Am Psychol* 49:997-1003.

Collins FS, Tabak LA (2014) Policy: NIH plans to enhance reproducibility. *Nature* 505:612-613.

Feise RJ (2002) Do multiple outcome measures require p-value adjustment? *BMC Med Res Methodol* 2:8.

Goodman SN (1999a) Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med* 130:995-1004.

Goodman SN (1999b) Toward evidence-based medical statistics. 2: The Bayes Factor. *Ann Intern Med* 130:1005-1013.

Gordon CJ (2005) *Temperature and toxicology: an integrative, comparative and environmental approach*. CRC Press: Boca Raton.

Gordon CJ, Lee KL, Chen TL, Killough P, Ali JS (1991) Dynamics of behavioral thermoregulation in the rat. *Am J Physiol* 261:R705-711.

Johansson T (2011) Hail the impossible: p-values, evidence, and likelihood. *Scand J Psychol* 52:113-125.

Lykken DT (1968) Statistical significance in psychological research. *Psychol Bull* 70:151-159.

McNutt M (2014) Reproducibility. *Science* 343:229.

Neyman J, Pearson ES (1928) On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrica* 20A:175-240.

Neyman J, Pearson ES (1933) On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London Series A Containing Papers if a Mathematical or Physical Character* 231:289-337.

Perneger TV (1998) What's wrong with Bonferroni adjustments. *BMJ* 316:1236-1238.

Ramsay DS, Seaman J, Kaiyala KJ (2011) Nitrous oxide causes a regulated hypothermia: rats select a cooler ambient temperature while becoming hypothermic. *Physiol Behav* 103:79-85.

Rothman KJ (1990) No adjustments are needed for multiple comparisons. *Epidemiology* 1:43-46.