Supporting Information (Online Only)

Part I. Total Calorimetry, Core Temperature (Tc), and N_2O Administration Chambers

The lab has 6 independent total-calorimetry chambers that also measure Tc telemetrically and serve as gas-tight exposure chambers for N_2O. Total calorimetry simultaneously measures the rates of total heat loss (HL) and metabolic heat production (HP) that are the underlying determinants of Tc and which reflect the influence of control mechanisms involved in regulating Tc.

A thermoelectric direct calorimeter (Seebeck gradient layer calorimeter, SEC-A-0701, Thermonetics Corporation, La Jolla, CA) generates a millivolt signal that is directly proportional to the heat flowing across its gradient layer. This measure is referred to as dry heat loss (DHL), which includes HL derived from conduction, convection, and radiation. The millivolt signal from the calorimeter is sent to a custom RC time circuit (made using a 33 kilo-ohm resistor and a 33 micro-farad capacitor) to yield a 1-sec time delay that reduces the noise in the differential signal. The signal is then amplified by 100 times using a precision instrumentation amplifier (AD624, Analog Devices, Inc. Norwood, MA) prior to being sent to a Universal Interface 2 device (UI-2, Sable Systems, Las Vegas, Nevada) and then to data acquisition and instrument control software written in LabVIEW (version 6.8, National Instruments, Austin, TX). The direct calorimeter is calibrated using an electric heater that produces an appropriate range of outputs (~0.8 - 3 Watts).

Stabilizing the temperature surrounding the outer aspect of the gradient layer facilitates measuring DHL from the animal in the calorimeter because it minimizes changes in heat flow resulting from dynamic changes in the surrounding temperature of the lab. This is accomplished by sending cold tap water through an automotive radiator provided with good air exchange from an adjacent room fan. The water is then directed through a secondary heat exchanger that consists of 18 meters of coiled copper tubing (i.d. = 1.9 cm, with each coil separated by ~5 cm) immersed in a ~290-liter reservoir of room temperature water. The large mass of water in the reservoir enables dampened temperature variation throughout the day relative to the more variable temperature of the lab. The water passing through the copper tubing becomes the same temperature as the surrounding water and is then sent through insulated PVC pipe to the calorimeters. Approximately 100 ml/min of water flow is diverted to each calorimeter where it circulates through copper tubing attached in a serpentine pattern to all 6 sides of the outer surface of the gradient layer within the calorimeter's housing. The temperature of the water entering a single calorimeter and exiting all six calorimeters is measured (SS-TC thermistors, Sable Systems, Las Vegas, Nevada) before the water exits the lab via a floor drain.

The internal dimensions of the direct calorimeter are 19 x 19 x 19 cm. The calorimeter serves as the N_2O-exposure chamber and so all joints of all internal surfaces are treated with RTV silicone sealant to make the chamber gas-tight. Each calorimeter's lid and body are compressed against a closed cell rubber gasket using adjustable fasteners. Gas enters through a threaded hole in the top center of the lid and exits through a threaded

hole centered on a lateral wall. A rat is placed directly into a calorimeter on a removable floor made of 0.635-cm diameter acrylic rods with a 0.318-cm gap between each rod. A plastic waste tray is below the rod flooring and sits on the antenna platform that is on the floor of the calorimeter. The surface of the rod flooring is ~5.2 cm above the floor of the calorimeter. Thus, the effective volume for the rat in the calorimeter is 19 x 19 x 13.8 cm.

Medical grades of O_2, N_2O, and N_2 from compressed gas cylinders are mixed to provide the gas blend used for total calorimetry. Each direct calorimeter chamber receives its blended gas from three, one for each gas, independent Mass Flow Controllers (MFC, model 840L-2-0V1-SV1-E-V1-S1, Sierra Instruments, Monterrey, CA) that are controlled by a Sable Systems Mass Flow Controller 4 (MFC-4). The MFC-4 is computer-controlled via the UI-2 and the UI-2 digital I/O accessory board (Sable Systems). The designated mass flows of O_2, N_2O, and N_2 enter a custom manifold (Industrial Specialties, Englewood CO) where they are blended and the gas leaving the manifold goes through tubing containing an in-line spiral element that ensures thorough mixing of the component gasses. The blended gas enters the influent gas port centered on the ceiling of each direct calorimeter at a flow rate of 1.5 l/min standard temperature and pressure, dry (STPD). The control-gas mixture consists of 21% O_2 and 79% N_2. To rapidly establish the target concentration of 60% N_2O, the first 6.5 min of gas delivery consist of 72% N_2O, 21% O_2, and 7% N_2. After 6.5 min, the gas delivery is changed to 60% N_2O, 21% O_2, and 9% N_2 which continues until the end of the N_2O-delivery period when the gas is switched back to the control gas mixture.

Indirect calorimetry requires measuring each rat's oxygen consumption, which is determined by comparing the O_2 concentration entering the chamber (i.e., the baseline measurement) to the O_2 concentration leaving the chamber. O_2 concentration is measured using a Sable Systems O_2 analysis system (FoxBox). In order to compute evaporative HL, the humidity of the gas leaving the chamber must be determined relative to the baseline value. Water vapor pressure is measured using a sensitive Sable Systems water vapor analyzer (RH-300). A Sable System Dewpoint Generator (DG-4) is used to calibrate the RH-300. A computer-controlled Sable Systems Respirometer Multiplexer (TR-RM4) controls the gas plumbing connections that determine whether baseline or exfluent gas samples are taken for measurement. The exfluent gas is measured continuously except when baseline samples are collected once every 20 min for a duration of 1.2 min. The baseline gas sample is collected after it exits the in-line spiral blender, the RH300 and the FoxBox and has been returned to the influent gas line that enters the ceiling of the direct calorimeter. The exfluent gas sample is taken after the gas has left the direct calorimeter chamber and the digital mass flow meter (GFM-1109, Dwyer Instrument Inc., Michigan City, IN). The exfluent gas sample is sent to the RH-300, and then to the FoxBox before being returned to the exhaust gas line. Prior to being vented from the lab, the 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 exhaust gas line.

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 $x \leq x0$ accept H. Such a rule tells us nothing as to whether in a particular case H is true when $x \leq x0$ 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.

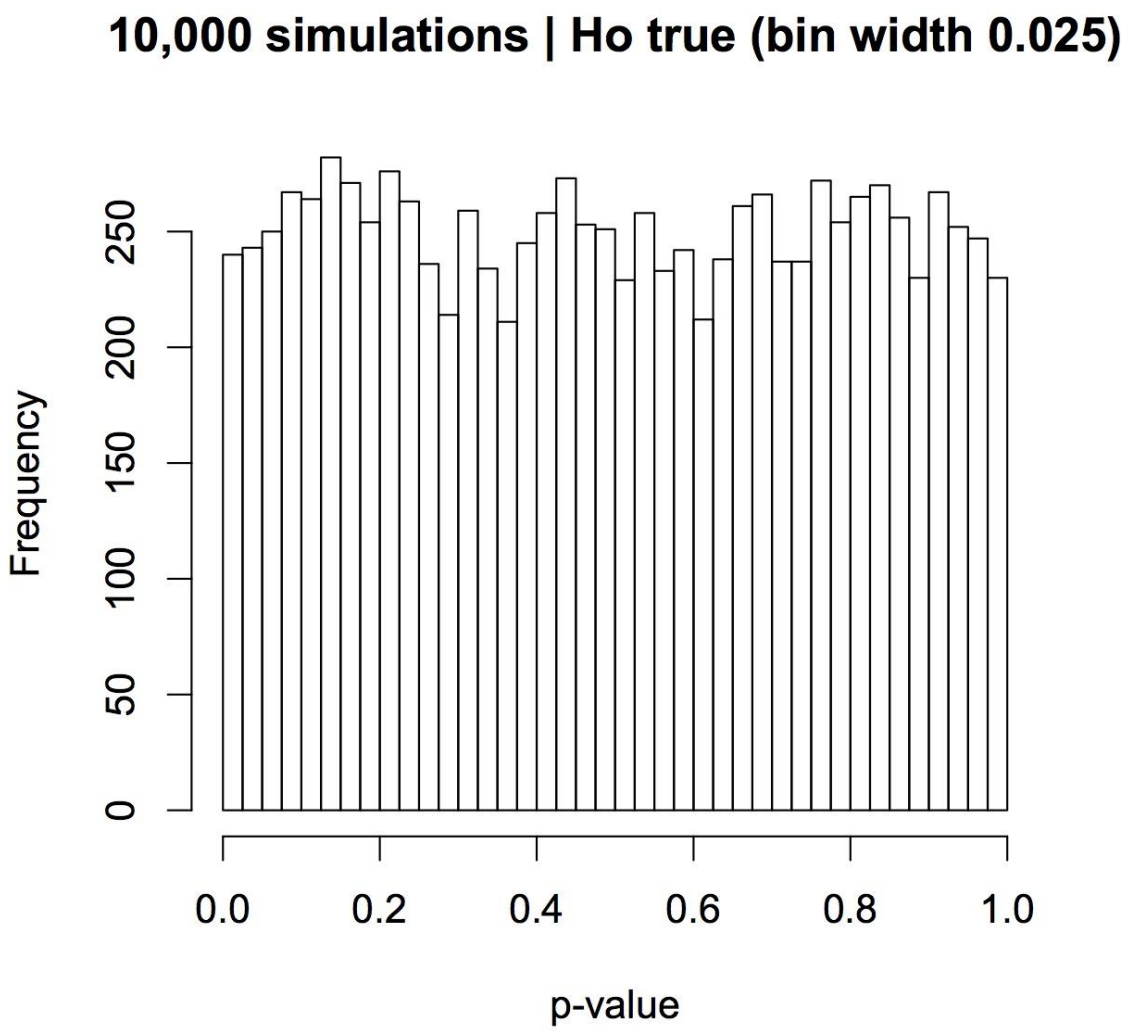
Type 2 error is a serious problem and p-values are a poor evidential metric. Increased protection against type 1 error comes at the cost of increased type 2 error, a serious problem given the contemporary context of resource scarcity and the much amplified concern regarding the overuse of animals in 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.