**Experimental Philosophy Meets Experimental Design: 23 Questions**

**Marica Bernstein**
bernstmc@gmail.com

*This paper, with subsequent modifications, was presented at the MidSouth Philosophy Conference, February 2007.*

**Abstract**
Experimental philosophers use empirical research methods to generate quantitative data, the statistical analysis of which subsequently informs philosophical analyses. These researchers' philosophical claims thus hinge crucially on their experiments. In this paper I present the results of a pilot study aimed at addressing the question: Does the primary experimental philosophy literature satisfy the most basic demands of experimental design and data analyses? If it does not, then experimental philosophers' philosophical claims are void. If it does, then experimental philosophy is on its way to becoming a legitimate discipline and achieving its promise.

**Introduction**

A novel approach to investigating traditional philosophical questions about intuitions, moral judgments, actions, and so on is *experimental philosophy*.[1] What distinguishes experimental philosophers from their analytic counterparts (in philosophy of language, science, and the like) is their use of *empirical* research methods to generate *quantitative* data, the *statistical* analysis of which subsequently informs philosophical analyses. While some don't, I applaud this approach.

Experimental philosophers are sometimes giddy when speculating on the future of experimental philosophy, or "X-Phi": "Using survey methods for assessing folk intuitions has the potential to be delightfully liberating" (Nichols 2004). According to its proponents, the X-Phi approach can change the way philosophy is done, the experimental results can impact specific areas of research, and those results could have real-world application to e.g., legal issues. X-Phiers' enthusiastic attitude and X-phi's potential scope in part account for its current popularity, and for the "wave of the future" quality young philosophers and students sense in it. Enthusiasm and scope, however, do not guarantee solid empirical research. Without proper attention to the quality of X-Phi research, the futures of X-Phi and those attracted to it are questionable. How well do X-Phiers implement the distinguishing features of X-Phi—the empirical methods that gather quantitative, statistically analyzed data?

I assume that X-Phiers want their interpretations of experimental results to speak to broader philosophical questions. In this regard, they are aligned with experimentalists in traditional scientific fields.[2] But to convincingly conclude one thing or the other, the

[1] This contemporary brand of experimental philosophy should not be confused with naturalized philosophy, with philosophical investigations that draw on others' scientific work, or with the work of philosophers engaged in scientific research in traditional fields like animal behavior or neuroscience.

[2] There is an important distinction, however. It is one thing to want interpretations to impact some bigger question, quite another to want particular results. Even the purest experimentalists will candidly admit they design and conduct experiments in order to demonstrate some thing, but scientists are cautioned to avoid making statements about their wants regarding experimental outcomes. X-Phiers state, in print and on blog, that they designed an

experimental design and analyses of data need to conform to widely accepted standards, and enough information about the design, data collection, and the like must be presented to allow the reader to either concur or disagree with the author's interpretation and philosophical conclusion. Here I present the results from a study whose goal was to address the question, "Does the X-Phi primary literature satisfy these demands?" If it does, then experimental philosophy is on its way to becoming a legitimate discipline and achieving its promise. If it does not, then experimental philosophers' philosophical claims are void. To achieve my goal, I developed 23 questions pertaining to basic design, analysis, and reporting principles, and then evaluated a number of papers with respect to these questions. This is not an inferential statistical analysis of the X-Phi literature. This is explicitly a *descriptive* analysis—and a pilot project at best!— aimed at making the practices of this new discipline more scientifically legitimate. In my analysis of the literature, critics of X-Phi will find ammunition beyond typical philosophical objections. In my concluding comments, X-Phiers will find suggestions for doing better research. Taken together, X-Phi can take a step toward achieving its promise.

**Background**

To appreciate why I asked these 23 questions, a brief presentation of experimental design and data analysis principles is needed.[3]

Science begins with observations. Observations lead to broad general questions. If one's goal is to get more than a broad general answer—i.e., if one desires an answer that can be *asserted with a precise degree of confidence*—those questions need to be reshaped. Reshaping involves decomposing the big question into a set of more limited questions, all of which address the broad general question, but none of which by itself answers it. Thus begins experimental design. Scientific or *research questions* are well-defined: there is no ambiguity or imprecision about the subject or predicate. Research questions take one of several generic forms: "Does X affect Y?" or "Is Y the same for all X's?" or "Is there a relationship between X and Y?" What exactly are X and Y? What is meant by affect, relationship? It is the experimenter's responsibility to spell these out.

Questions framed in this way are easily restated as *research hypotheses*. A research hypothesis is not directly testable (nor can a research question be answered directly). A *statistical hypothesis* can be tested. There are two statistical hypotheses: the null hypothesis, $\mathbf{H_0}$ (e.g., with respect to Y, $X_1=X_2$), and the alternative, $\mathbf{H_A}$ ($X_1\neq X_2$). "Testable" means there is an appropriate statistical test, given the structure of the data, which will determine whether the null hypothesis will be accepted or rejected given probability conditions set out before the experiment is run. The predetermined probability condition is the *critical significance level*, $\boldsymbol{\alpha}$. It indicates how willing the experimenter is to reject the null when in fact it is true. The familiar statement, "Results were considered significant if $p<0.05$" relates to critical significance level and indicates that the researcher is willing to reject $\mathbf{H_0}$ (and accept $\mathbf{H_A}$), knowing that if the identical experiment had been run 100 times using samples collected from the same population, *due to chance alone* the statistical test of the results from five of those experiments could have dictated that $\mathbf{H_0}$ be accepted. Detailed discussion of $\boldsymbol{\alpha}$ is not possible here, but two things should be

---

experiment because they wanted "to show ___". This is a major mistake propagated to students, and this potential bias threatens the legitimacy of X-Phi.

[3] No citations are offered in this section. I consulted Zolman (1993), Zar (1999), Curran-Everett and Benos (2004), Holmes (2004), Lipton (2005), Geire *et al*.(2006), and the UCLA "Introduction to survey data analysis" web site. There are other respectable resources.

emphasized. First, **α** levels of 0.05, 0.01, 0.001 are used only by convention, but the choice of which to use is *not* arbitrary. Second—and I can't emphasize this enough— the decision of critical significance level *must be made in advance of data collection and analysis*. One cannot analyze the data, see what p-value the test spits out, and then decide that $p<0.01$ will be significant.[4]

When scientists collect data they are making measurements. Protocols for collecting measurements must be determined in advance, and be precise and unbiased. Measurements (i.e., the responses) are categorized based on scale. On ratio and interval scales the interval between consecutive units is constant, and measurements can be either continuous (45.5 grams) or discrete (number of rats weighing 45-50g).[5] In each case comparisons can be made between two numbers and a meaningful numeric value obtained (e.g., 91g = 2 X 45.5; this rat weighs twice as much as that rat). Measurements on an ordinal scale are relative to each other, but comparisons do not generate numeric values (e.g., "best" is not two times better than "good"). Measurements on a nominal scale are of attributes, e.g., hair color. No matter what the numerals are, measurements on ordinal and nominal scales are *categorical*. Truly silly mistakes are made by failing to understand scale. Coding heads as "0" and tails as "1" doesn't result in an average flip that is 0.572 tails.[6] Notice that the amount of information contained in the measurement diminishes from continuous→discrete→ordinal→nominal. The scale on which the response or dependent variable measurements ("Y" in the examples above) are collected, along with similar considerations about the independent variable, e.g., the groups ("X"), determine the structure of the data.

---

[4] [added] Some have misinterpreted my assertions here as suggesting that X-Phiers blindly follow dogmatic statistical rules. This is not the case. There are several important concepts related to critical significance level, **α**: **β** (the probability of wrongly failing to reject $\mathbf{H_0}$ when in fact $\mathbf{H_0}$ is not true of the population, i.e., the statistical test leads one to conclude there are no differences between groups when in fact there are), power (1-**β),** and two possible errors, Type I and Type II. Briefly, with respect to two groups, there is a true state of affairs in the population. Thus, the results of a statistical test on a random sample from that population can lead to one of four possibilites:

| | | Actual state of affairs | |
|---|---|---|---|
| | | Same | Different |
| Researcher's decision (from test result) | Accept $H_0$ (same) | **Correct decision** | **Incorrect decision Type II error** |
| | Reject $H_0$ (different) | **Incorrect decision Type I error** | **Correct decision** |

Two of these are wrong (due to by chance alone), and the goal is to minimize the likelihood of both types of error. **α** is the probability of a Type I error; **β** is the probability of a Type II. **α** and **β** are inversely related (which can be shown both mathematically and graphically); a lower **α** (e.g., 0.01 vs 0.05) lowers the probability of a Type I error while increasing the probability of a Type II. It's not possible to optimize one without cost to the other. Statistical power (1-**β**) is the probablility of *not* making a Type II error; by convention power=0.80 is considered adequate. In addition to **α** and **β**, power is affected by effect size (how much do two groups differ?) and by sample size.

Thus, one obvious reason to specify **α** when designing the experiment is to ensure that the sample size is large enough to result in an adequately powered experiment. This, together with decisions about **α**, increases the liklihood of seeing the hypothesized effect (if it exists) and making the correct decision (see also Appendix Experiment B for illustration).

[5] A ratio scale has a true 0 (0°K ), an interval scale does not (0°F).

[6] [Added] This point should be intuitive. Consider how an average is calculated: the sum of all responses (measurements) divided by the number of responses. Suppose a group was asked to rank some response (strongly disagree to strongly agree) on a scale of 0-4. In the (mis)calculation of an average on the data, a "4" response would count four times more than a "1", and "0" wouldn't count at all: individuals who strongly disagree would really have no voice.

For a host of reasons, data structure matters tremendously. Importantly, it determines the choice of appropriate statistical test, all of which make assumptions about data structure. Because data structure is known before data are collected, the appropriate test is dictated before data collection. Data structure and design also bear on choice of statistical software to use for analyses. All data analysis software packages have limitations, and some have acknowledged errors in computational algorithms.

Broad general questions are asked about a population of interest. The *target population* is usually large and frequently hypothetical (e.g., all mammals—past, present, future!). The *sample population* is the well-defined subset of the target population to which a researcher has access (e.g., lab rats of a particular strain). Measurements are collected on individuals randomly chosen from the sample population. These individuals comprise the *sample*. In a random sample, each member of the sample population has an equal and independent chance of being included in the sample.[7] The experimenter must be well acquainted with characteristics of the sample population to ensure a random sample of appropriate size, and to control for the effects of extraneous factors known to impact the responses.

In scientific papers, the above considerations are presented in the Introduction and especially Methods sections. And to re-emphasize, all of them—design, choice of statistical test, protocols for obtaining an adequately sized random sample—must be taken into account BEFORE a single datum is collected.

There are two remaining issues. First, the result of a statistical test, e.g., the F-statistic from an ANOVA, is a fact about the sample. The p-value that accompanies the test statistic (and takes design features such as sample size, etc. into consideration) gives the experimenter grounds for accepting or rejecting the null hypothesis—for making a statistical inference about the sample population. P-values are numeric values on a continuous scale and should be reported as such, not as $p<0.05$.[8] Second, just as the level of precision increases as one moves from general question to research hypothesis to statistical hypothesis, and from target population to sample population to sample, so too the level of uncertainly increases as one moves back up from sample to sample population to target population. These are the moves from fact about a sample to statistical inference to generalization about a target population.

## Methods

I am asking if X-Phiers' experimental designs, data collection and analyses, reports of methods and results, and legitimacy of their generalizations adhere to widely accepted scientific standards. While this could spawn testable hypotheses, I here report only descriptive statistics characterizing my sample.

### *My 23 questions*

The salient features of design and reporting dictated my 23 questions (Table 1). Most were framed such that they could be answered with a Yes/No answer, with "Yes" always objectively better. There are important questions to ask of this literature that require knowledge

[7] There are other legitimate techniques for constructing a sample from a population, but the underlying principles still apply.

[8] This is a common mistake in scientific publications. The reasons for reporting a precise p-value are: 1) it communicates more information, and 2) it allows the reader, whose interest in the result may differ from the researcher's, to assess the result within the context of his/her own research program.

and understanding of the big question and are thus beyond my area of expertise. I nevertheless included some in the list, with notation.

***Sampling, data collection and analyses***

I reviewed 14 papers from two sources (Table 2): 1) Stephen Stich's Spring 2006 Experimental Philosophy Seminar reading list[9], 12 papers; 2) "Selected papers in experimental philosophy" on Joshua Knobe's "The Experimental Philosophy Page" [10], 3 papers (1 paper appeared on both sites). All were primary papers that reported experiments and results, and were not summaries of previously reported work. All were available on-line. My sample may not have represented the target population (all X-Phi primary literature). Nevertheless, I felt justified in using this set. Given their presence on these web sites, I take it the experimental results have philosophical import. Importantly, young philosophers and students interested in learning about the X-Phi approach will turn to these sources first. These papers and their methodologies are the available models of experimental philosophy.

My review was blind to author and journal. A colleague assisted with this. After reviewing all papers, I re-analyzed the data in three because I was skeptical of the reported statistical significance. I do not report how individual papers fared. Results were tabulated in *SYSTAT 11* (Wilkinson 2004).

**Results**

***Research questions, hypotheses, and design***

All of the papers contained a statement (often a prediction) easily translated into a research question. All (implicitly) had null and alternative hypotheses. None contained a statement of critical significance level. Only six of 14 used the correct statistical test. However, in several of the papers that scored "yes" on this question (#8), a general form of the correct test was used (e.g., Person's chi-square test for independence of proportions), but given the structure of the data the general test for significance was inappropriate (e.g., small sample sizes call for Fischer's exact test). A common mistake among the eight that did not use the correct test was numerically coding categorical data, and then performing a test that assumed continuous data.

Only four designs controlled for potential confounds. Table 3 presents the remainder of results.

***Populations and sampling***

The target population in eight of 14 papers was unspecified people, i.e., the folk. For six of these the sample population was undergraduate students. In two of these six, the sample was student volunteers, but there was no indication of what motivated students to volunteer; In two, the sample was students who were fulfilling an experiment participation requirement. Two of 14 papers adequately described sampling criteria. In two the sample was truly drawn randomly from the sample population.

Total sample sizes ranged from 18-220. The number of groups per experiment ranged from 1-8. Assuming equal group size within an experiment, group sizes ranged from 9-81/group.

---

[9] Stich, S. 2006. Topics and readings on *Experimental Philosophy Seminar*. http://www.rci.rutgers.edu/~stich/Experimental_Philosophy_Seminar/experimental_philosophy_seminar_readings.htm

[10] Knobe, J. The Experimental Philosophy Page on http://www.unc.edu/~knobe/ExperimentalPhilosophy.html

However, four experiments had groups that were not sized approximately equal, and there was no mention that a correction was made for unequal group sizes.

***Data collection and analyses***

For 11 of 14 papers the person collecting the data was not blind to the research question. The X-Phiers themselves approached potential subjects, conducted the interview, or passed out questionnaires.

Eleven papers presented data that did not meet the assumptions of the statistical test used to test the null hypothesis. Only one paper took other factors into account in the analysis. Roughly two thirds (9/14) reported p-values incorrectly. Among them, many reported significant results as "$p<0.05$", "$p<0.01$", and "$p<0.001$" for multiple tests on the same data set. Table 3 provides other results.

***Conclusions and other questions***

Two papers did not attempt to generalize beyond the sample population. Of the remainder, one offered justification for doing so. Ten of 14 failed to acknowledge inevitable biases and limitations of the study. Five papers were authored by one individual with no reference to others who may have participated in data collection or analyses.

***Post-hoc re-analyses***

Using standard statistical tests for the data sets presented, I was unable to replicate purported statistically significant results in two of three papers. See Appendix, Experiments A and C, for details.

**Discussion**

My sample was small and skewed. The nature of some papers forced me to make subjective (but charitable) decisions with respect to some questions. Nevertheless, the results of my experiment point to potentially serious problems that could have tremendous impact on the future of X-Phi and its researchers. I'll let the specific results speak for themselves. In the remainder of this paper I'll speculate on why these mistakes were made, and then offer some constructive suggestions to the X-Phi community.

Every field of experimental research has an ontogeny. In early stages, statistically robust results can often be obtained by unsophisticated experiments that may have fared no better on my questions (Zolman 1993). X-phi is young. Some X-phiers may not appreciate the theoretical underpinnings and rigorous demands of experimental design. Importantly, some do. Incorrect design and analyses decisions may be being learned from earlier work. (It would be interesting to ask if the literature is getting better.) The conference audiences where work-in-progress papers are presented probably take the results at face value and move quickly to the philosophical issues. At best they may challenge that the experiment addresses the philosophical issue. Finally, the journals to which X-Phi papers are submitted may not have reviewers adept at evaluating design and statistical significance. It's clear that there are no accepted standards for how to structure papers that report X-Phi results.

To help move X-Phi forward, I offer the following suggestions.

- **Consult and learn:** There are statisticians at every university who specialize in experimental design and analysis of experimental data. As soon as a research question and an experiment that will potentially address it are conceived, consult one. This recommendation is contentious within

science itself. Purists believe that understanding design and analysis are as much a part of doing the science as knowing, say, the basics of animal husbandry, or the basic physical principles of the measurement instrument. Because X-Phiers understand the philosophical issues better than statisticians, completely turning over design and analyses doesn't make sense either. But there is middle ground—consult and learn at the same time. My point is that the time for consultation is *before* data are collected.

- **Haste makes waste**: Restatement of the above with warnings! Eventually someone is going to challenge, *in print*, the "results" of some X-Phi experiments. Addressing these challenges is going to take more time and be more painful than consulting and learning would have been. In addition, because participants in the experiments are human subjects, the experiments may fall under the purview of local IRBs (Institutional Review Boards). Once under IRB review, X-Phiers will no longer be able to stroll into a park and ask folks questions. Instead, they'll first spend hours pleading for exemption, and when that doesn't work, weeks writing IRB protocol, justifying sample sizes, gathering informed consent… .
- **Teach:** If you are leading a seminar on X-Phi, consider a guest lecture by a statistician.
- **Ask some simple questions** in your review of X-Phi literature: Mine may help. In addition, set aside the philosophy for a moment and look at the numbers. Consider two groups, 1 and 2, each with 27 individuals. Each individual is asked to choose A or B. The results are reported as:

| | Response | |
|---|---|---|
| Group | A | B |
| 1 | 74% | 26% |
| 2 | 62% | 38% |

The difference between the groups is claimed significant. But look at the numbers:

| | Response | | |
|---|---|---|---|
| Group | A | B | Total |
| 1 | 20 | 7 | 27 |
| 2 | 17 | 10 | 27 |
| Total | 37 | 17 | 54 |

If A and B were heads and tails the coin is probably not fair (37/54 compared to 17/54). But look further. Did Group 1 really flip statistically significantly more heads than Group 2 (3 out of 37 heads flips)? Could this result be due to chance?

- **Establish standards** as a community: Consider these issues; there are others. 1) The structure of the portion of the paper that describes the experiment and its results should be uniform no matter what the research question. Because design decisions must be made in advance of data collection, Methods and Results can't be mushed together. 2) Standardize the notation of the statistics themselves or at least require that the notation be defined in each paper. ("Average" and "mean" are different and use different notations; in texts I consulted, "M" always meant "median", never average or mean.) 3) If someone other than the author(s) contributed to design, collection, analyses, readers need to know. Many science journals are requiring that the contribution of all individuals who participated in experiments be spelled out. IRB protocols are even more stringent.

I asked if the X-Phi literature satisfied the demands of experimental design and statistical analyses. My analysis suggests it does not always. X-Phi isn't alone here. Design and analyses mistakes abound in well-established literatures (Zolman 1993; Curran-Everett and Benos 2004; Holmes 2004). These mistakes are not insurmountable.

**References**


Curran-Everett, D. and D. Benos. 2004. Guidelines for reporting statistics in journals published by the American Physiological Society. *Physiological Genomics* **18**: 249-251.

Giere, R. N., J. Bickle and R. F. Mauldin. 2006. *Understanding Scientific Reasoning*. Belmont, CA, Thompson Wadsworth.

Holmes, T. H. 2004. Ten categories of statistical errors: a guide for research in endocrinology and metabolism. *Am J Physiol Endocrinol Metab* **286**(4): E495-501.

Knobe, J. The Experimental Philosophy Page on http://www.unc.edu/~knobe/ExperimentalPhilosophy.html

Lipton, P. 2005. Testing hypotheses: Prediction and prejudice. *Science* **307**: 219-221.

Nichols, S. 2004. Folk concepts and intuitions: From philosophy to cognitive science. *Trends in Cognitive Science*.

Stich, S. 2006. Topics and readings on *Experimental Philosophy Seminar.* http://www.rci.rutgers.edu/~stich/Experimental_Philosophy_Seminar/experimental_philosophy_seminar_readings.htm

Wilkinson, L. (2004). SYSTAT 11. Richmond, CA, SYSTAT Software, Inc.

Zar, J. H. 1999. *Biostatistical Analysis*. Upper Saddle River, NJ, Prentise Hall.

Zolman. 1993. *Biostatistics: Experimental Design and Statistical Inference*. New York, Oxford University Press.

**TABLE 1**

**Research questions, hypotheses, and design**

0) Did the research question truly address the intended philosophical question? [Although critical, I did not address this.]
1) Was the research question specific and well defined?
2) If appropriate, was there a control level of the independent variable (a control group or treatment)?
3) Were units of measurement of the dependent variable defined or clear in context?
4) Was there more than one dependent variable?[1]
5) Were the dependent variables a true measure of the effect of the independent variable? [In some cases I was unable to evaluate this.]
6) Were confounding factors acknowledged and controlled for?
7) Was the statistical hypothesis falsifiable? [At least implicitly, were there null and alternative hypotheses?]
8) Was the appropriate statistical tests employed, given the structure of data?
9) Was $\alpha$, the critical significance level, specified before data were gathered?[2]

**Populations and sampling**

10) What was the target population?
11) What was the sample population?
12) Were sampling criteria or methods described adequately?
13) Was the sample a random sample?
14) Was sample size reported? [If so, I collected information on sample size, number of groups, group sizes.]

**Data collection and analyses**

15) Was the data collector blind to the research question (or were controls in place to ensure unbiased collection)? [If not explicitly stated, I assumed that the data were collected by one or more of the authors.]
16) Were assumptions of the statistical test met?
17) Were p-values reported accurately?
18) Where appropriate, was the correct measure of variation included?
19) With respect to responses that could be affected by sex, age, or other factors, were these factors included in the analysis?
20) Was information about data analysis software included?

**Conclusions and other questions**

21) If conclusions were generalized beyond sample population, was justification given?
22) Were inevitable biases, imprecisions, and limitations of the study, including unreliability of the responses, acknowledged?
23) Number of researchers? [Authors plus others who contributed to the experiment itself.]

---

*Twenty-three questions asked of each X-phi paper. With the exceptions of #10, 11, and 23, all could be answered Yes or No, with Yes being objectively better. [1] Goes to redundancy of effect. [2]As evidenced by inclusion of a statement, e.g., "Results were considered significant if $p<0.05$".*

**TABLE 2**


Knobe, J. 2003. Intentional action and side-effects in ordinary language. *Analysis* **63**: 190-193.

Knobe, J. 2003. Intentional action in folk psychology: An experimental investigation. *Philosophical Psychology* **16**(2).

Leslie, A., J. Knobe and A. Cohen. in press. Acting intentionally and the side-effect effect: 'Theory of mind' and moral judgment. *Psychological Science*.

Machery, E., R. Mallon, S. Nichols and S. Stitch. 2004. Semantics, cross-cultural style. *Cognition* **92**: B1-B12.

Nadelhoffer, T. Bad acts, blameworthy agents, and intentional actions: Some problems for juror impartiality.

Nichols, S. Folk intuitions and free will.

Nichols, S. 2004. The folk psychology of free will: Fits and starts. *Mind and Language* **19**(5): 473-502.

Nichols, S. and J. Knobe. Moral responsibility and determinism: The cognitive science of folk intuitions.

Nichols, S., S. Stitch and J. Weinberg (2003). Metaskepticism: Meditations in ethno-epistemology. in *The Skeptics*. S. Luper, eds. Aldershot, England, Alderrshot Publishing**:** 227-247.

Nichols, S. and J. Ulatowski. Intuitions and individual differences: The 'Knobe' effect revisited.

Rips, L., S. Blok and G. Newman. in press. Tracing the identity of objects. *Psychological Review*.

Stolz, K. and P. Griffiths. 2004. Genes: Philosophical analysis put to the test. *History and Philosophy of the Life Sciences*.

Woolfolk, R., J. Doris and J. M. Darley. 2005. Identification, simulational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition*.


---

*Primary experimental philosophy literature reviewed for this analysis; Available on-line at* http://www.rci.rutgers.edu/~stich/Experimental_Philosophy_Seminar/experimental_philosophy_seminar_readings.htm *and* http://www.unc.edu/~knobe/ExperimentalPhilosophy.html .

**TABLE 3**

| My questions[1] | Number of responses Yes | No |
|---|---|---|
| **Research questions, hypotheses, and design** | | |
| 1) Research question specific and well defined? | 14 | 0 |
| 2) Control group or treatment?[2] | 0 | 5 |
| 3) Units of measurement defined/implicit? | 14 | 0 |
| 4) More than one dependent variable? | 7 | 7 |
| 5) Dependent variables true measure of the effect?[4] | 5 | 0 |
| 6) Confounding factors controlled for? | 4 | 10 |
| 7) Statistical hypothesis falsifiable? | 14 | 0 |
| 8) Appropriate statistical tests employed? | 6 | 8 |
| 9) **α** level stated? | 0 | 14 |
| **Populations and sampling** | | |
| 12) Sampling criteria described adequately? | 2 | 12 |
| 13) Random sample? | 2 | 12 |
| 14) Sample size reported? | 13 | 1 |
| **Data collection and analyses** | | |
| 15) Data collector blind to the research question? | 3 | 11 |
| 16) Assumptions of statistical test met? | 3 | 11 |
| 17) P-values reported accurately? | 5 | 9 |
| 18) Correct measure of variation?[2] | 0 | 3 |
| 19) Other factors included in the analysis? | 1 | 13 |
| 20) Data analysis software? | 1 | 13 |
| **Conclusions and other questions** | | |
| 21) Justified generalization?[5] | 1 | 11 |
| 22) Bias, imprecision, and limitations acknowledged? | 4 | 10 |

*Results to questions that could be answered Yes or No. [1]Questions are abbreviated. See Table 1 for complete wording. [2]Inappropriate question for some papers. [3]I was unable to answer for some. [4]Two papers did not generalize to a target population.*

## Appendix

### Experiment A

**Q**: Does response differ among four vignette conditions?
**Set-up**: Two independent variables, A and B, each with two levels, 1 and 2, giving four vignette conditions: $A_1B_1$, $A_1B_2$, $A_2B_1$, $A_2B_2$. One main dependent variable, Y: response to a straightforward question using a Likert 7-point scale. [Likert-type scales assign a numeric value indicating, e.g., level of the subject's agreement with a statement.] (Seven other responses were collected for data exploration purposes).
**n= 72** (17/vignette group)
**$H_0$**: $A_1B_1 = A_1B_2 = A_2B_1 = A_2B_2$ (the response does not differ among the four groups)

**Results** of a 2-way ANOVA (analysis of variance) claimed significant: Effect of A: $F(1, 68) = 6.83$, $P < 0.02$. Effect of B: $F(1, 68) = 5.02$, $P < 0.03$. Therefore,
**Reject $H_0$ and accept $H_A$:** in this statement, $A_1B_1 = A_1B_2 = A_2B_1 = A_2B_2$, at least one "=" should be replaced with "≠" . In ordinary language the result, [as incorrectly obtained] was that if A was 1, than more people responded with answers closer to 7 on the scale than if A was 2. Similarly for conditions of B.

**Re-analysis is not possible** because given the data structure, an ANOVA is not the correct statistical test. Its assumptions are not met. The dependent variable is on an ordinal scale, not an interval or ratio scale as the test assumes. (There are other assumptions of an ANOVA but this is a moot point since the data are categorical, not continuous or discrete.)

*Likert-type scales are common in X-Phi experiments (and in other fields). All parametric statistics, such as ANOVA, assume that in the population the dependent variable (response) conforms to one of several probability distributions, e.g., a normal distribution, where the interval between units is constant, i.e., on a ratio or interval scale. In addition, parametric statistical test use location (e.g., averages) and variation (e.g., standard deviations); ordinal data have neither (although one can calculate the median for these data). Many will argue that with enough choices, the collected responses will be normally or otherwise distributed. Thus, there are those that think a 9-point scale is superior to a 7- or 5-point. This misses the point. There is no in principle reason to think that individuals perceive the difference between, e.g., "strongly disagree" and "disagree" as equal to the difference between "disagree" and "somewhat disagree" (nor any reason to think all individuals have the same perceptions about this). People may perceive the difference between 5' and 6' to be smaller than between 6' and 7', but this is a misperception that can be shown empirically to be false. There simply is no way to turn data on a categorical scale into data on an interval scale. There are nonparametric statistical tests, e.g., versions of Kruskal-Wallis test, which would have been appropriate for these data.*

**Experiment B**

**Q**: Do responses differ between two vignette conditions?
**Set-up**: Two conditions, A and B; Two responses, 1 and 2 (mutually exclusive and exhaustive)
**n=126** (63/condition)
**$H_0$**: pA = pB (proportion of responses in two conditions are equal)

Reported data, percent of responses for each condition:

| | Response | |
|---|---|---|
| Condition | 1 | 2 |
| A | 75% | 25% |
| B | 51% | 49% |

Results claimed significant: Chi-square (1, n=125) = 7.62, p<0.01, therefore
**Reject $H_0$ and accept $H_A$: pA≠pB**, the proportion of responses differs between the two conditions.

**Re-analysis, α=0.05**

Extrapolate to determine number of responses for each condition:

| | Response | |
|---|---|---|
| Condition | 1 | 2 |
| A | 47 | 16 |
| B | 32 | 31 |

Results of Chi-square test for equality of proportions:

Pearson Chi-square (1, 78) = 7.635, p=0.006, therefore
**Reject $H_0$ and accept $H_A$: pA≠pB**, the proportion of responses differs between the two conditions.

*This re-analysis illustrates two things. 1)The slight difference in chi-square values (7.62 vs. 7.635) is probably due to software differences, illustrating the need to report this information (Question #20). 2) It shows the value of design. One might ask, why n=126? Why not 120 or 130? Although detailed explanation is beyond the scope of this note, the reason is simple. N=63/group is exactly the sample size needed to achieve statistical power of 80%-- the accepted standard. Not only is this X-Phier able to assert with 99.4% confidence that the null hypothesis has not been wrongly rejected, he/she is also able to assert with 80% confidence that the alternative hypothesis is true. Slightly smaller sample sizes would have resulted in <80% confidence; larger would have been a waste of resources.*

**Experiment C**

**Q**: Do responses differ among four vignette conditions?
**Set-up**: Two main factors, A and B, each with two conditions, 1 and 2: $A_1$, $A_2$, $B_1$, $B_2$; Two responses, 1 and 2 (mutually exclusive and exhaustive)
**n=80** (20/condition)
**$H_0$**: pA = pB (proportion of responses in four conditions are equal)

Reported data, percent of responses for each condition:

| | Response | |
|---|---|---|
| Condition | 1 | 2 |
| $A_1$ | 95% | 5% |
| $A_2$ | 76% | 24% |
| $B_1$ | 79% | 21% |
| $B_2$ | 28% | 72% |

Results comparing $A_2$ with $B_2$ claimed significant: Chi-square (1, n=37) = 7.7, $p<0.01$, therefore **Reject $H_0$ and accept $H_A$: pA≠pB**, the proportion of responses differs between theses two conditions

**Re-analysis, α=0.05**

Extrapolate to determine number of responses for each condition:

| | Response | |
|---|---|---|
| Condition | 1 | 2 |
| $A_1$ | 19 | 1 |
| $A_2$ | 15 | 5 |
| $B_1$ | 16 | 4 |
| $B_2$ | 4 | 16 |

**Re-analysis not possible:** 1) All chi-square tests assume that each cell has >5 responses (or counts); 2) As data are structured, it's statistically invalid to test only a portion of the data. There are alternative ways to structure these data, but none negate the fact that half the cells are sparse. Additional problem is that in report of the statistic, above, n=37, but according to text 80 participants were assigned to four equal groups.