

*What if Plato Took Surveys?: Thoughts about Philosophy Experiments*

**What if Plato Took Surveys? Thoughts about Philosophy Experiments**

*William Goodman*

About ten years ago, a new branch of philosophy, called ‘Experimental Philosophy,’ was introduced. Its founders sought to adapt for philosophy some experimental methods applied, for example, in psychological research. They aim to counter or supplement conventional ‘thought experiments’ used by traditional philosophers.

Traditionalists may cap an argument by saying something like this: ‘If you insist on statement A, you have to accept statement B. Yet B defies our intuitions; don’t you agree? Hence: Reject A.’ But experimental philosophers now survey real people in the world; then they counter traditional arguments by saying something like: ‘Empirically, many people do *not* share the intuition about B that you just appealed to; so your argument fails.’ Often, experimental philosophers look for and claim to find such differences in intuitions between groups with different ethnic, or other, characteristics, such as ‘Westerners’ versus ‘East Asians.’ Some experimental philosophers focus on refuting traditional intuition-based arguments, as just described; others explore more positive programs.

Methodology is the movement’s defining issue; so we examine here particularly the methods that experimental philosophers employ. This paper is neutral about specific, philosophical debates undertaken by experimental philosophers, such as in epistemology or ethics. None of this paper’s criticisms are meant to deny that experimental methods can have a legitimate role in philosophy. Instead, this paper offers a cautionary note about methodological pitfalls, relevant for those who explore the movement’s potentials.

The portrayal of experimental philosophy as a new philosophical movement is somewhat self-promoted—as based on proponents’ websites and articles in the popular press (Knobe, 2010; Stich, 2001). Yet a genuine, wider interest is signalled by increasing attention paid to the topic in conferences and special journal issues. A number of references, below, derive from such sources.

There is no one protocol or objective shared by all experimentalists. Nadelhoffer and Nahmias (2007) describe three ‘distinct kinds of projects’ for experimental philosophy, which Spicer (2010) more recently called three ‘agendas’. Most attention and controversy focus on the first two agendas; these would use the experiments to either rebut traditional-style arguments, or somehow refine them. The third agenda is less focused on traditional debates— exploring instead for underlying mechanisms for *why* people respond,

intuitively, as they do (Knobe, 2007; Hardy-Vallee & Dubreuil, 2009). Many papers have elements of more than one agenda.

For the first agendas, it makes sense to demand that philosophers’ arguments be sound (where this can be assessed) as well as valid; and this requires true premises. If a premise is really about what ‘people (universally) intuit’, and this premise is experimentally falsified, then there is an impact on the argument. But adequately interpreting an experiment’s result requires understanding the study’s design.

The experimental philosophy literature, regrettably, contains accounts of experiments with inadequate descriptions for: what hypothesis tests were used; or why those specific tests; or on how/why particular samples were selected or the findings assessed as ‘significant’; or on the validity and reliability of their survey designs and test instruments. Below we clarify these issues, the last listed of which Simon Cullen has addressed extensively in a recent paper (2010). We also look at whether quantitative experiments, in particular, are necessarily central to the potential contributions of experimental philosophy. Newer experimental philosophy papers are generally improving in methodological respects; yet some still cite, unquestioningly, earlier less careful, findings. This paper hopes to encourage, not discourage, experimentation—but with perhaps more awareness of methodological considerations.

(Note that, beyond the scope of this paper, is another school calling themselves experimentalists (see Berman, 2009). The latter is essentially about introspective methods, like Berkeley’s.)

# Examples of Methodological Issues arising in the Literature for Experimental Philosophy

Key to the program for experimental philosophy is the movement’s self- description as ‘experimental’—deliberately suggestive of objective, empirical- based, and transparently rule-following methodologies. Experimentalists do not just write essays to persuade; they also point to hard and demonstrably suitable data, and to justifiable analyses for summarizing or manipulating those data. Those claiming this type of backing for their conclusions imply a commitment to such a standard.

The experimental and reporting standards used by experimental philosophers are evolving, but some important issues still reoccur. Several key concerns are raised below, in the context of papers (presented roughly chronologically) where the issues have appeared.

We present, first, what is virtually the founding paper for experimental philosophy: *Normativity and Epistemic Intuitions*, by Weinberg, Nichols, and Stich (2001). They hypothesize that people’s epistemic intuitions about a given topic vary as a function of seemingly extraneous, individual factors, such as one’s culture, socioeconomic group, or academic background. Order of prompting, they hypothesize, also can vary the outcome. In experiments, they

exposed participants to various ‘intuition probes...similar to cases that have actually been used in the recent literature in epistemology.’

One such probe adapts Lehrer’s ‘Truetemp’ thought experiment (1990), designed to explore the ‘externalist/internalist dimensions of ... subjects’ intuitions.’ Suppose that by some scenario a person Charles endures a brain- changing event, so any time he estimates the temperature, he is always reliably correct. The next time Charles estimates a temperature (suppose he says ‘71 degrees’)—and inevitably he is right—does he *know* the temperature is 71 degrees or only believe it? Respondents’ answers to this scenario are matched to their ethnicities, to look for patterns.

At least two important methodological issues for experimental philosophy arise in this paper, which is still being cited unquestioningly (e.g., Beebe, 2010).

*Methodological Issue 1: Ensuring Representative Samples*

Respondents for the above study were recruited, in some manner, from among students at Rutgers University, in north-eastern United States. A sample is representative if it contains roughly the same mix of types of individuals as does the full population, at least regarding differences that might impact results. Otherwise, outcomes may only reflect local circumstances. Did Weinberg *et al* ensure a reasonable mix of ethnic, socioeconomic, and demographic characteristics among their presumably all student-age subjects? And can they confirm that English-speaking East Asians who can read complex English-language prompts are representative of the larger population of, presumably, the world’s mostly *non*-English-literate East Asians? (These questions matter because if the sample is not representative, then regardless of response patterns in the sample, conclusions cannot be inferred about a whole group, as a population. The problem is compounded if, as Cullen suggests, different groups do not even interpret the same survey prompts the same way.)

Moreover, the study’s authors developed and applied over the course of their research alternative versions of their prompts. Did they, then, use the same, original participants for all repeat promptings (risking a ‘carry-over effect’ from subjects’ previous ponderings and responses) or did they use different participants (risking extraneous differences between the different groups recruited)?

There are no mechanical or easy answers to such questions when designing a research protocol. But if such concerns are *not* addressed during study design, or the rationale for protocol decisions explained when presenting and interpreting results, then experimenters’ findings may not necessarily be more generalizable than those of so-called armchair philosophers.

*Methodological Issue 2: Choosing the Appropriate Test*

Almost universally, experimental philosophers use statistical tests as a means to bolster their empirical claims. At the least, the test used should be clearly identified, and its relevance explained. Only in a later version of their

paper (in Nichols *et al*, 2003) did Weinberg *et al*’s narrative explicitly name the test employed.

A common format for hypothesis testing is to posit, explicitly or implicitly, a presumption (to be either rejected or not-rejected, based on study results) that there is ‘no effect’, e.g. that ‘Westerners’ and East Asians’ intuitions about *x* are *not* different.’ This no-effect presumption is called the *null hypothesis*. If the empirical results appear inconsistent with that hypothesis (to a suitable degree), then the tester might conclude that the no-effect presumption is really untenable. However, one’s choice of test can be disputed by questioning either the technical assumptions of one’s method, or, more fundamentally, the way that the null hypothesis has been framed.

Figure 1 illustrates two alternative ways to frame the null hypothesis for the first Truetemp case. The testing method employed by Weinberg *et al* assessed—and rejected—a null hypothesis most like the version presented in Figure 1(a). The implied null hypothesis is that, for both Westerners and East Asians, the respective *ratios* of numbers of people choosing one versus the other of the two available responses are the same. I.e., their test compares the two ethnic groups’ respective response ratios for how many responded ‘Knows’ versus how many responded ‘Only Believes’. Those ratios *do* appear non-equal in the top graph, and the authors’ hypothesis test confirms this appearance. However, their actual paper does not address this matter of differing opinions *within* individual cultural groups. It speaks only of differences *between* groups.

**Figure 1.** *Framing the Null Hypothesis for Truetemp Case, Version 1*

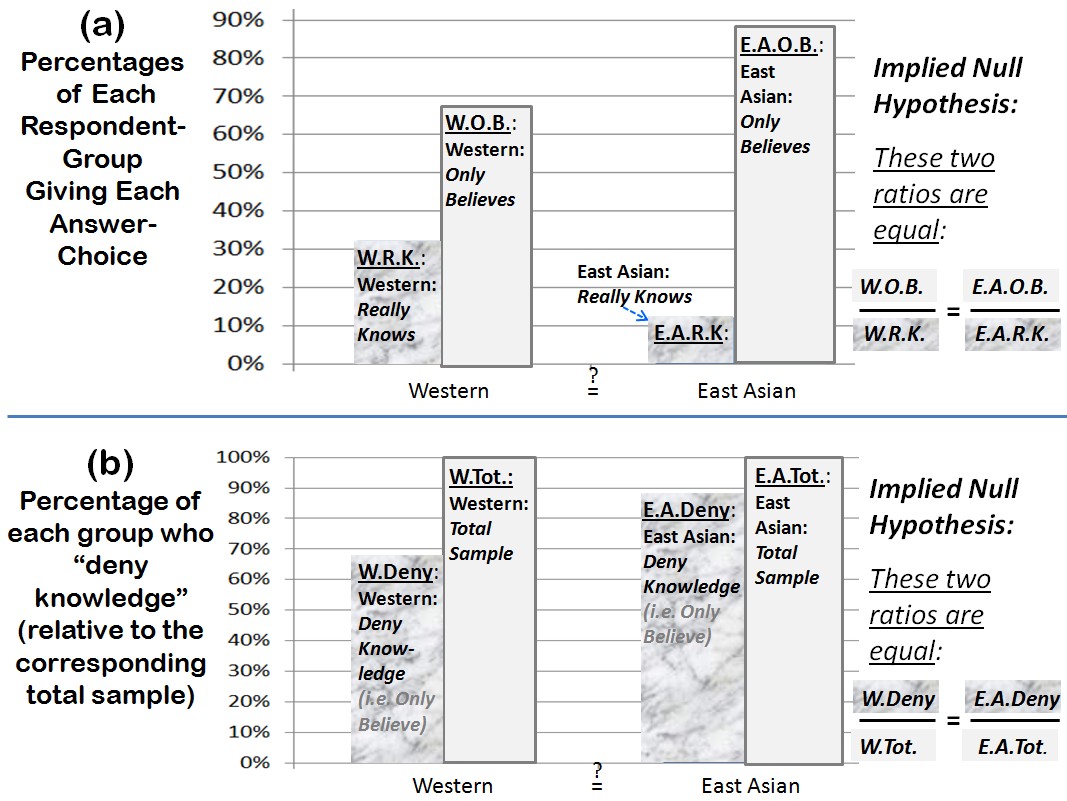


Figure 1(b) better corresponds to the actual claim that Weinberg *et al* make in their narrative. They are comparing the proportions of respondents in each

group who specifically ‘Deny knowledge,’ i.e., answer “Only believes”. A null hypothesis that better matches *that* claim is that the two groups’ numbers of knowledge deniers are consistent with (i.e., roughly proportional to) the groups’ respective numbers of participants. In this light, Figure 1(b) suggests that the data do *not* particularly show much difference between cultural groups.

In practice, the choice for ‘best’ test or null hypothesis may not be self- evident; and colleagues may disagree with one’s choices. Yet we expect philosophers to justify the methods and assumptions that they employ, and this can include *experimental* assumptions. As recently as May *et al* (2010), we find examples of failing to name the test that supposedly justifies some conclusion. Choice of test can also impact the choice of survey questions (see Issue 5).

*Methodological Issue 3: Being Mindful of Possible Confounding*

Confounding occurs when there are other, non-acknowledged variables that could impact the experimental results. Suppose you tested for the impact of people’s shoe-brand preferences on their cars’ purchase prices, but failed to consider a variable for income or wealth. Surprisingly, your test might show a ‘significant’ impact for shoe-brands; but really because price differentials for some shoes (e.g., between expensive import brands versus discount-store options) will stand in for (called being a ‘proxy’ for) the non-considered wealth factors.

An illustration where this could occur in experimental philosophy is found in Machery *et al.*’s paper (2004). Again, they contrast the intuitions of different cultural groups, by re-purposing historical thought experiments as prompts— here, with regard to proper names and reference. The authors provide appropriate details about participants (where recruited, demographics, and exclusion criteria), and give more information about the tests employed.

However, concerns arise from the authors’ response to possible objections that ‘the labels “East Asian” and “Western” are too crude […given] the enormous diversity’ within those groups. They dismiss that objection lightly, saying that ‘if we find significant results using crude cultural groupings, [...then] more nuanced classifications should yield even stronger results.’ Their assumption is that ‘cultural group’—properly bounded—is *the* key predictive variable, so if there are any misclassifications, these would be randomly scattered, with no systematic impact on the study.

But suppose the real, causal variable is ‘Was English the primary language spoken in the childhood home?’ The East/West variable would be only a proxy. Even ‘nuanced,’ this proxy would *systematically* fail to group together such cases (that *should* be grouped together) as (a) a bilingual (but primarily French speaking) French Canadian in the U.S. sample, and (b) a bilingual (but primarily Cantonese speaking) individual in the Hong Kong sample.

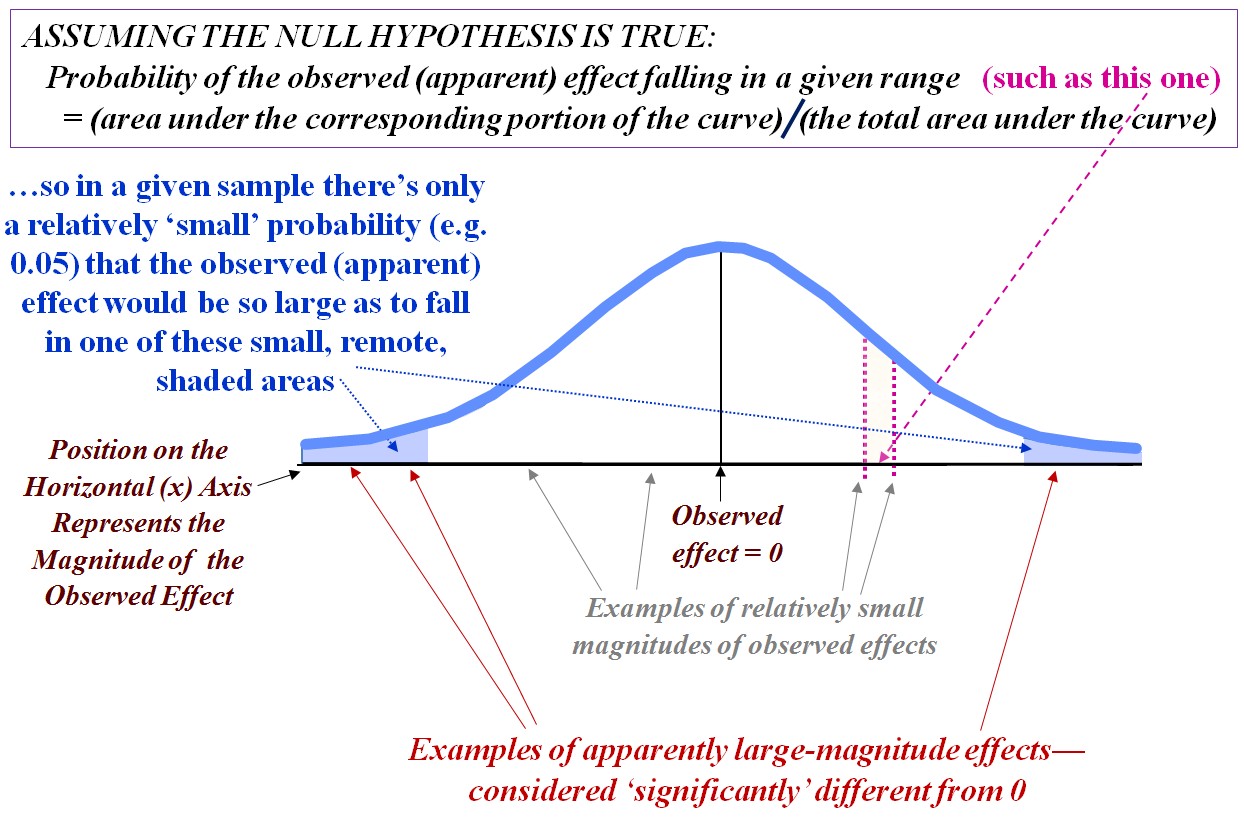
Since a confounder is something we failed to test for explicitly, the problem can often arise from an imperfect data collection instrument (see Issue 5). However, confounding factors can also arise during the *execution* of an experiment that possibly looked good on paper.

*Methodological Issue 4: Meaningfully assessing Effects*

Almost universally, experimental philosophers apply (or claim to apply) conventional statistical concepts of ‘significance’ and ‘*p*-values’. They say an effect (such as the test-measured difference between Easterners’ and Westerners’ responses) is ‘significant’ if it is so large as to make it seem implausible that both (a) there really is *no* effect in the population (i.e. the null hypothesis is true), yet (b) this apparently very large effect in our sample occurred anyway, just by a random fluke. Surely, the thinking goes, such a discrepancy suggests that assumption (a) about the population must really be incorrect.

Such conclusions are generally based on a combination of ‘*p-*values’ (discussed below) and a model like Figure 2.

**Figure 2.** *Expectations of Observed Effect Size, IF the Null Hypothesis is True*



The figure models what is *expected* as a sample result *if* the null hypothesis (that in the population, there really is no effect to be observed) is true. Interpret the curve this way: *If* the null hypothesis were really true, the probability that any random sample would yield a result (of some apparent effect) that falls in a particular range along the bottom axis, would be proportionate to the corresponding, relative area under curve, above that range of possible values.

For example, suppose there are two cultural groups for whom there really is ‘no difference’ between the intuitions of those two groups regarding some scenario (i.e., the null is true). Then what do we expect to result if, in an experiment, we sample opinions from members of those two groups? Most likely, the sampled difference of opinions between the groups (the ‘observed effect’) would be zero or relatively small—falling in the large, central area (in

white) under the curve in Figure 2. The large relative size of this white area suggests that, correspondingly, a ‘relatively small effect’ is the most likely observed outcome, again *if* the null were true.

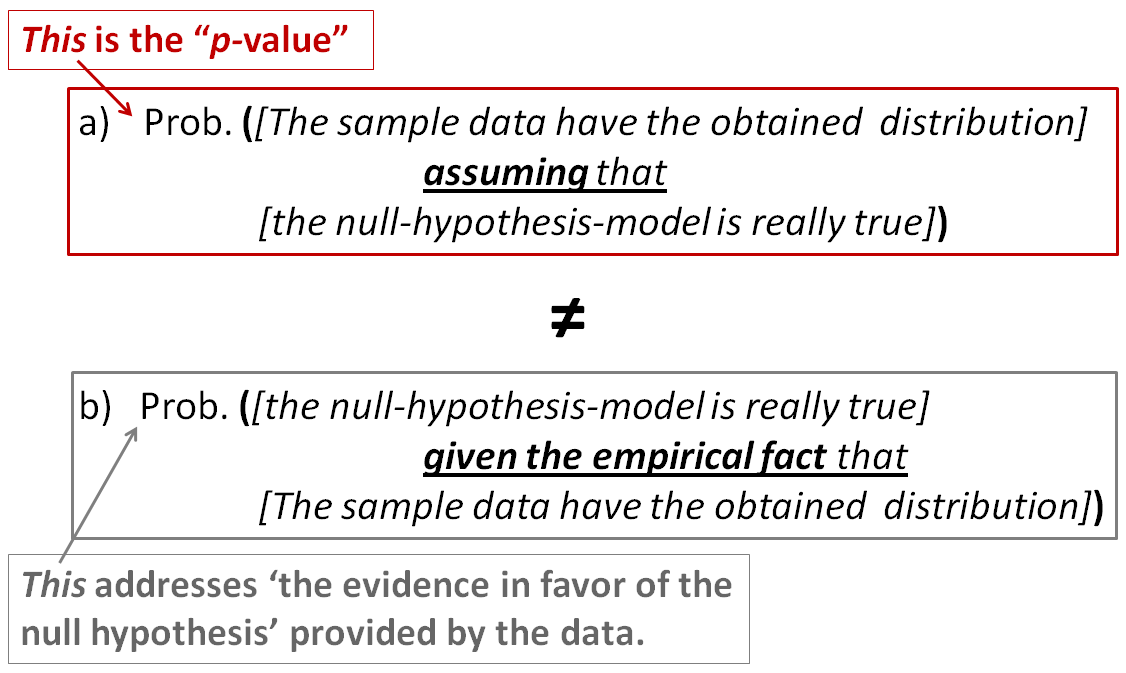
On the other hand, a given sample *could* exhibit, by random fluke, an apparently large effect—so large that in the figure, it would be depicted in one of the remote (shaded) ‘tail’ areas. These tail areas are relatively small, suggesting that *if* the null were true, results that fall there are unlikely and unexpected. If such results occur there on this occasion, we are inclined to question the veracity of the ‘no effect’ hypothesis, itself. Therefore, when an effect is large enough to fall in an ‘unexpected’ outer region of the figure, authors may report that a ‘significant’ difference (from the null expectation) has been found.

This approach is many decades old and very widespread, but is also widely misinterpreted. It is often conjoined with a concept called *p-*values, which are cited as a kind of standard for claiming significance of results. Significance is commonly claimed when the *p*-value ≤ 0.05, though some accept larger, ‘marginal’ *p*-values, such as 0.084 (Swain *et al*, 2008). A clear example where *p-*values are misinterpreted appears in Bengson *et al* (2009), where he mistakenly says ‘a *p*-value of … 0.001 [i.e., near the small tip of a tail region in Figure 2] ...means … a 99.9% chance [i.e., 1–0.001] that the [finding about the population] is genuine.’

The *p*-value is the conditional probability labelled (a) in Figure 3: It is the probability of obtaining such a sample result as we did, *presupposing that* the null hypothesis, *as happened to be postulated*, is really true. (Recall that the null hypothesis could have been formulated differently.) At best, a small *p-* value (ideally, much below the conventional 0.05 cut-off) tells us that a particular sample’s outcome is *surprising, if* the null is really true—as, for example, if my samples show a large difference between cultures’ views, although in reality there is no difference between these populations’ views. This would be a surprising result; and the assigned *p-*value is likely small.

However, (a) is *not* (as Bengson suggests) a direct measure of evidence for (or against) the null hypothesis. A truer measure of evidence is expression (b) in the figure: (b) gives the probability that the null hypothesis (e.g., of no real difference between populations) is true, *given that in fact* we obtain a sample result exactly as empirically occurred. (b) and (a) are not identical; in fact, (b) is not even, strictly speaking, a probability statement. At the time of the experiment, the null hypothesis is either factually true, or not; the null’s truth value is not subject to probability. (b) is more like a reliability check for the test method: Of many times such methods have been employed, and the test results have been similar to what we obtained, in what share of such cases has the null actually been true? Empirical studies suggest that the *p-*value method is a poor predictor of the latter proportion (Goodman, 2010).

**Figure 3.** *The p-Value versus ‘Evidence in Favor of the Null Hypothesis’*



But even if we use a *p-*value to conclude that the population exhibits a real effect, many reviewers and publishers now demand also a contextually meaningful, independent measure of *effect size* (Ellis, 2010). The apparent difference from the null may be inconsequential, in practice. Having found *some* difference between culture-groups’ perceptions, what supports a claim that this is a *non-trivial* difference? Re-citing the *p-*value does not answer this question. A fine, recent example for *not* using *p-*values just mechanically is found in Feltz and Zarpentine (2010, footnote 20): Although nominally *p* < 0.05 for a certain finding, they advise caution due to (a) confounding factors they identified, and (b) the increased risk of false positives when you do multiple testing (since each test adds more chance for error). Similarly, Schaffer and Knobe (2010) hedge their strict reliance on *p-*values—being cautious when they ‘just squeak past the conventional threshold of statistical significance.’

*Methodological Issue 5: Ensuring Validity and Reliability of Test Instrument*

For a test to be informative, it should measure what it purports to measure, and do so consistently (i.e., be valid and reliable). Cullen’s paper, cited earlier, provides a thorough treatment of ways in which experimental philosophers’ survey instruments have often not met this standard. For example, errors may arise from inattentiveness to Grice’s (1975) conversational maxims, such as to make conversational contributions (including survey prompts) as informative as required, without superfluous or ambiguous additions. Responses which experimentalists interpret as subjects’ varying their intuitions based on ‘irrelevant factors’ may really reflect the subjects’ trying gamely to meaningfully interpret all the conversational elements included in the survey prompts.

Consider this example from Weinberg *et al.*’s cancer conspiracy case (2001): Its long and complex text prompt (in one paragraph) includes at least these components: Three statements of logically relevant premises; then three non-sequiturs (about what might have happened but didn’t, and who didn’t know); followed by a conclusion. Subjects’ ambiguous responses may better reflect how they tried to parse this non-Grice-sensitive paragraph, than their specifically logical intuitions.

Being aware of such communications concerns is just a start. To ensure one’s survey questions and responses are interpreted as intended, one should try them out first by designing and running pilot studies. In the literature, Buckwalter’s paper (2010) is a rare exception in suggesting that a preliminary study was conducted. Included in Cullen’s critique (2010) are a number of post-hoc experiments that raise questions for experimental philosophers’ surveys. He shows, for example, how responses to one of Weinberg *et al*.’s vignettes would vary if responses were accepted using a 5-point scale, from least to most agreement as to whether someone ‘knew’ something, as opposed to using a forced ‘yes/no’ choice. Ideally, an experimenter would have tried this check *in advance*, on his/her own, to ensure the impact of the question- form on the answer would not remain an unknown and possibly confounding factor.

# Are ‘Experiments’ the Central Issue?

Experimental philosophy adds something new to philosophy. But is its contribution necessarily, or only, about experiments?

*Why Necessarily Experiments*

Experiments are not new to philosophers with a bent towards modeling and understanding human perceptions, at least regarding risk and decision making, or in areas like philosophy of biology. For example, what Swain *et al* (2008) identified as ‘ordering’ effects (that responses to prompts may be influenced by their presentation sequence) was foreshadowed by earlier work on ‘framing effects’ (that responses to risk cues may vary with presentation contexts) (Fagley & Miller, 1987; Tversky & Kahneman, 1981). Experiments for these other domains have generally been conducted by psychologists, economists, and others, but interested philosophers have closely followed these projects; compare Gardenfors and Nils-Eric’s compendium (1988).

But might experimental philosophers be taking the expression ‘thought *experiments’* too literally? They object to arguments that seem to hinge on statements like (for one example) ‘Almost everyone will accept [thus-and-so]’ (DeRose, 2005). If indeed an argument stands on claims about near-universal beliefs, then an empirical check of those claims is reasonable.

Yet what we call thought experiments are sometimes just exercises for ‘drawing out a contradiction in a theory, thereby refuting it,’ as Brown (1996) and recently, Ichikawa (2011), point out. Here, the better analogy may not be a

literal experiment, but instead a step in a mathematics or geometry demonstration, following what Polya (1968) describes as the ‘logic of plausible reasoning’. Compare Socrates, in Plato’s Meno (1961), where he demonstrates to a slave boy an error in the boy’s original understanding of triangles. At each step, Socrates invites the boy to reflect with him whether a certain observation or claim is reasonable; Socrates invites his discussant into a shared conceptual context, to assess the appropriateness of each step. Whether a sample of students outside this shared context of reasoning agrees with the boy’s ‘intuitions’ is not clearly relevant.

In like manner, consider the Gettier cases for epistemology, frequently cited by both traditional and experimental philosophers (Greco, 2007; Beebe, 2010). Suppose a person has a factually true belief and good (by usual standards) evidence, but, by some twist, the evidence is true only accidently. The usual case is presented with this logical structure:

* + 1. If A’s attributes with respect to ‘knowing’ what is in fact the true nationality of B’s car *satisfy the true definition for knowledge*, then A *knows* the nationality of B’s car.
    2. But (I, pondering this odd combination of facts and A’s attributes, sense that) under this unusual scenario, A does *not* ‘know’ the nationality of B’s car.
    3. Therefore (I conclude by Modus Tollens), A’s attributes etc. do not satisfy the true definition for knowledge.

This argument structure collapses for a discussant who disagrees on premise (2); just as Socrates’ triangle demonstrations might fail if encountering a discussant with *non-Euclidian* intuitions. This dependence on a particular sample evaporates if the topic is specifically Euclidian triangles.

For the Gettier case, the argument has apparent force when the person writing premises (1) and (2) *means the same thing* when she writes ‘knows’ each time. Grice (1969) suggests that meaning reflects the intention of the speaker. The speaker in (1) searches for a ‘timeless meaning’ of the word ‘knows’, but in (2) intends to somehow discredit ‘accidental knowledge.’ It is beyond the scope of this paper to determine whether ‘knows’ should mean just one thing, or whether the above syllogism captures that meaning. The point is that asking how many people would support statement (2) if presented out of context does not exhaust the philosophical interest in the original thought argument.

*Why Only (Quantitative) Experiments?*

On the other hand, if the real issue is getting empirical data, why restrict, arbitrarily, the means of getting it? Experimental philosophers have largely associated ‘experiments’ with techniques to obtain quantitative data, such as obtained from fixed-answer surveys. Some have now realized that these instruments can be too restrictive for obtaining all relevant information (Nahmias *et al*, 2004). More qualitative methods of analysis, which directly

engage participants in discussions and focus groups, are used in the social sciences. These also could be called experiments in the general sense of seeking to obtain information by an empirical discovery process. Though more systematically documenting and analyzing responses, this approach looks more like familiar dialogue methods favoured by x-Phi opponents like Kauppinen (2007).

Qualitative methods can have their own issues. For example, how can one validly compare and integrate one’s qualitative findings with more quantitative measures, where both are available (Risjord *et al*, 2001). Yet these methods offer another route, besides quantitative experiments, for accessing empirical data, where needed for sound arguments.

# Discussion/Conclusion

In summary, experimental philosophers are right to ask for empirical evidence where arguments depend on empirical claims, though not every argument or thought experiment is necessarily this type. Experimental philosophers usefully demonstrate how experiments may help to obtain empirical evidence, and show potential impacts of this on traditional philosophy.

Where quantitative experiments are applicable, this paper alerts to some methodological issues that experimenters—not just in philosophy—should address, in both the planning and reporting of their studies, to enhance their results’ credibility: (a) Are the samples sufficiently representative? (b) Are the test and null hypothesis employed appropriate for the types of claims intended?

(c) Have potential confounding factors been considered? (d) How will significance of results, and meaningfulness of effect size, be determined? (e) Have the test instrument’s validity and reliability been established?

Note that some appeals to intuition or thought experiments are inevitable in arguments. It is impossible to recursively field-test every presumed fact that serves as an assumption for testing *other* facts. At some point, you just have to make your case, for example, as to why a sample recruited from your campus is sufficiently representative. Nonetheless, it is always open to another experimenter to question and field-test any such assumptions, and experimental philosophers are commended for attempting this.

# References

Beebe, J.R. (2010). ‘The relevance of experimental epistemology to traditional epistemology.’ Paper presented at the international conference, How and why economists and philosophers do experiments, March 27-28, in Kyoto, Japan.

Bengson, J., M.A. Moffett & J.C. Wright (2009). ‘The folk on knowing how.’

*Philosophical Studies* 142(3): 387-401.

Berman, D. (2009). *A manual of experimental philosophy*. Dublin: Jeremy Pepyat Books.

Brown, J.R. (1996). ‘Thought experiments.’ In: *Stanford Encyclopedia of Philosophy*.

Available at <http://plato.stanford.edu/entries/thought-experiment/>[8 October 2010]

Buckwalter, W. (2010). ‘Knowledge isn’t closed on Saturday. A study in ordinary

language.’ *Review of Philosophy and Psychology* 1(3): 395-406.

Cullen, S. (2010). ‘Survey-driven romanticism.’ *Review of Philosophy and Psychology* 1: 275-296.

DeRose, K. (2005). ‘The ordinary language basis for contextualism, and the new invariantism.’ *The Philosophical Quarterly* 55(219): 172-198.

Ellis, P.D. (2010). *The essential guide to effect sizes*. Cambridge: Cambridge University Press.

Fagley, N.S. & P.M. Miller (1987). ‘The effects of decision framing on choice of risky

vs certain options.’ *Organizational Behavior and Human Decision Processes*

39(2): 264-277.

Feltz, A. & C. Zarpentine (2010). ‘Do you know more when it matters less?’

*Philosophical Psychology* 23(5): 683-706.

Gardenfors, P. & S. Nils-Eric (eds.) (1988). *Decision, probability, and utility. Selected readings*. Cambridge: Cambridge University Press.

Goodman, W.M. (2010). ‘The undetectable difference. An experimental look at the

‘problem’ of *p*-values.’ In: *JSM Proceedings*. Alexandria, VA: American Statistical Association.

Greco, J. (2007).‘Knowledge as credit for true belief.’ In: M. DePaul (ed.), *Intellectual*

*virtue. Perspectives from ethics and epistemology*, 111-134. Oxford: Oxford University Press.

Grice, H.P. (1969). ‘Utterer’s meaning and intention.’ *The Philosophical Review* 78 (2): 147-177.

\_. (1975). ‘Logic and conversation.’ In: P. Cole and J.L. Morgan (eds.),

*Speech acts*, 41-58. New York: Academic Press.

Hardy-Vallee, B. & B. Dubreuil (2009).‘Folk epistemology as normative social cognition.’ *Review of Philosophy and Psychology* 1(4): 483-498.

Ichikawa, J. (2011). ‘Who needs intuitions? Two experimentalist critiques.’ In: T.

Booth & D. Rowbottom (eds.), *Intuitions*, Forthcoming at Oxford University Press. Available also at <http://jonathanichikawa.net/papers/wni.pdf>[14 February 2011].

Kauppinen, A. (2007). ‘The rise and fall of experimental philosophy.’ *Philosophical Explorations* 10(2): 95-118.

Knobe, J. (2007). ‘Experimental philosophy and philosophical significance.’

*Philosophical Explorations* 10(2): 119-121.

(2010). ‘A return to tradition.’ *New York Times. Opinion Pages*.

Available at: <http://www.nytimes.com/roomfordebate/2010/08/19/x-phis-new-> take-on-old-problems/a-return-to-tradition [3 March 2011].

Lehrer, K. (1990). *Theory of knowledge.* Boulder: Westview Press.

Machery, E., R. Mallon, S. Nichols & S.P. Stich (2004). ‘Semantics, cross-cultural style.’ *Cognition* 92(3): B1-B12.

May, J., W. Sinnott-Armstrong, J.G. Hull & A. Zimmerman (2010). ‘Practical interests, relevant alternatives, and knowledge attributions. An empirical study.’

*Review of Philosophy and Psychology* 1(2): 265-273.

Nadelhoffer, T. & E. Nahmias (2007). ‘The past and future of experimental philosophy.’ *Philosophical Explorations* 10(2): 123-149.

Nahmias, E., S. Morris, T. Nadelhoffer & J. Turner (2004). ‘The phenomenology of free will.’ *Journal of Consciousness Studies* 11(7/8): 162-179.

Nichols, S., S. Stich & J.M. Weinberg (2003). ‘Metaskepticism. Meditations in ethno- epistemology.’ In: S. Luper (ed.), *The skeptics*, 227-247. Aldershot, England: Ashgate Publishing.

Plato (1961). ‘Meno.’ In: E. Hamilton & H. Cairns (eds.), *The collected dialogues of Plato*, 353-384. Princeton, N.J.: Princeton University Press.

Polya, G. (1968). *Patterns of plausible inference. Volume II of Mathematics and*

*plausible Reasoning*. Princeton, NJ: Princeton University Press.

Risjord, M., M. Moloney & S. Dunbar (2001). ‘Methodological triangulation in nursing research.’ *Philosophy of the Social Sciences* 31(1): 40-59.

Schaffer, J. & J. Knobe (2010). ‘Contrastive knowledge surveyed.’ *Nous* Pre- publication access: doi: 10.1111/j.1468-0068.2010.00795.x [12 January 2011].

Sosa, E. (2007). ‘Experimental philosophy and philosophical intuition.’ *Philosophical*

*Studies* 132(1): 99-107.

Spicer, F. (2010). ‘Cultural variation in folk epistemic intuitions.’ *Review of Philosophy and Psychology* 1(4): 515-529.

Swain, S., J. Alexander, & J.W. Weinberg (2008). ‘Instability of philosophical

intuitions. Running hot and cold on Truetemp.’ *Philosophy and Phenomenological Research* 76(1): 138-155.

Stich, S.P. (2001). ‘Plato’s method meets cognitive science.’ *Free Inquiry* 21(2): 36-

38.

Tversky, A. & D. Kahneman (1981). ‘The framing of decisions and the psychology of choice.’ *Science* 211(Issue 4481): 453-458.

Weinberg, J.M., S. Nichols & S. Stich (2001). ‘Normativity and epistemic intuitions.’

*Philosophical Topics* 29(1/2): 429-460.