History and Philosophy of Psychology Bulletin Volume 15, No. 1, 2003, pp. 16-26.

Discussion with Green

Some Meta-theoretical Issues Relating to Statistical Inference

Siu L. Chow
Department of Psychology, University of Regina

Taking issue with Chow's (2002a) critique of Wilkinson and Task Force's (1999) report on statistical inference, Green (2002) raised several instructive issues, namely, (i) appealing to authority, (ii) theories for which there is no criterion of falsification, (iii) the distinction between *experiment* and *meta-experiment*, and (iv) the probability foundation of the null-hypothesis significance-test procedure (NHSTP). It is hoped that this reply can foster a better understanding of research methods in general, and of the role of NHSTP in empirical research in particular.

Contrast vs. Control Groups

Green (2002) points out that Wilkinson et al.'s (1999) Task Force did not recommend replacing the control group by the contrast group. His point is well made. The term "control group" should not be used in non-experimental or quasi-experimental studies. It is hoped that non-experimental researchers do not treat contrast groups on par with control groups for the reasons given in Chow (2002a, pp. 3334). Contrast groups cannot be used as valid comparison baselines. Nor can they reduce ambiguities as well as control groups can.

Appeal to Authority

It is sad if researchers accept Wilkinson et al.'s (1999) report simply because it was sponsored by the American Psychological Association (APA) and prepared by "a veritable 'Who's Who' of statistical expertise in the behavioral sciences" (Green, 2002, p. 42). Although the appeal to authority may be an effective rhetorical device in advocacy, it has no place in theoretical discussion. Using it in an intellectual discussion would detract from the exchange its conceptual rigor or intellectual integrity. The fact that many well-known researchers subscribe to a theoretical stance does not necessarily warrant the truth of the said theory. The relevant consideration is the validity of the evidence used to substantiate the tenability of the theory. The following statement sets in high relief the distinction between advocacy and the pursuit of knowledge.

> You've heard it, we've heard it: Global warming is the greatest threat facing humanity. One hundred Nobel laureates recently signed a statement saying so. A UN panel of scientists says so. Our government says so. Can all these people be wrong? Of course they can. Whether the earth is warming or not is a scientific question, not a political one. (Essex & McKitrick, 2002, p. 9)

Statistical Significance and Sample Size

A constant refrain of critics of NHSTP is that statistical significance is a matter of sample size (call it the "size-dependent significance thesis"). Specifically, if a sufficiently large sample is used, statistical significance is a foregone outcome. Taking this view for granted, the critics have never provided any conceptual justification or empirical evidence for the thesis. Hence, it is refreshing to find Green's (2002) explication of the reasoning behind the "size-dependent significance thesis" as follows:

> [a] Larger sample sizes give smaller standard errors.
> [b] Larger degrees of freedom that come with larger sample sizes make the critical t smaller.
> [c] For a given difference between sample means, even if it is not significant under one sample size, it would become significant when the sample size is increased because of [a] and [b].
> [d] Chow's (2002a) empirical demonstration to the contrary of [c] is something to be expected because "most people already know - as one increases sample size most aberrantly large sample differences will be diluted out" (p. 44).

The validity of Green's argument hinges on an implicit assumption. Consider the completely randomized 1-factor, 2-level design with $n_1 = n_2 = 10$. Suppose that $\overline{X}_1 = 10;\ \overline{X}_2 = 6;$ hence, $(\overline{X}_1 - \overline{X}_2) = 4.$ It seems from [c] that Green has in mind the situation in which the difference between the two sample means will still be 4 when n_1 and n_2 are both increased to 20.

It helps to recall that H_O is never used as a categorical proposition. Instead, it appears as a component of a conditional proposition twice, once as the consequent (see [S1]) and once as the antecedent (see [S2]) of a conditional proposition.

> [S1] If the data are the result of chance influences, then H_O is true.
> [S2] If H_O is true, adopt the sampling distribution with Properties A, B and C.

In other words, a fair test of the "size-dependent significance thesis" has to honor [S1] and [S2]. Consequently, everything in the experiment has to remain the same, except for the increase in sample size. At the same time, increasing the sample size means adopting a different sampling distribution. Specifically, the theoretical properties mentioned in [S2] are inevitably changed when the sample size is changed. Moreover, the change is in the direction described in [d] of Green's account. That is, points [a] [b] and [d] are inter-related in a way that renders [c] impossible.

Confounded size-increase

Chow's (2002a) simulations were carried out with the assumption that, at all the sample sizes used, all the recognized control variables present and the experimental manipulation does not have the expected substantive efficacy. To maintain these assumptions when the sample size in increased means that no confounding variable is introduced as a result of increasing the sample size. This stipulation may be difficult to satisfy in non-experimental studies.

Consider the following study conducted to determine whether or not there is a relationship between children's IQ and their preference for shoe color. Envisage the situation in which five children with red shoes and five with blue shoes are selected, and their IQ is measured. Suppose that the difference between the mean IQs of the two groups is not significant. However, when the sample size is increased to 75, red-shoe children have a higher verbal IQ score than their blue-shoe counterparts (see the "Mean Verbal IQ" row in Table 1).

Table 1: An inadvertent change in the girl-to-boy ratio when the sample size is increased from 5 (Panel 1) to 75 (Panel 2)

	Panel 1		Panel 2	
	Red Shoes	Blue Shoes	Red Shoes	Blue Shoes
Number of girls	3	2	60	15
Number of boys	2	3	15	60
Total number of	5	5	75	75
Children				
Mean Verbal IQ	101	100.5	120	103

For illustrative purposes, described in Panel 1 of Table 1 is the composition of the samples in terms of shoe-color and gender. It shows that the verbal IQ scores of 3 girls and 2 boys are measured because they wear red shoes. The mean of this group is 101. Similarly, the verbal IQ scores of 2 girls and 3 boys are measured because they wear blue shoes. Their mean is 100.5. The difference between the two means is not significant.

In Panel 2 is described the composition of the red-shoe and blue-shoe groups when the sample size is increased to 75. The red-shoe group has a higher verbal IQ score than the blue-shoe group (viz., 120 vs. 103). Taken together, Panels 1 and 2 seem to support the "size-dependent significance" view. However, the girl-to-boy ratio is 4 to I in the red shoe group, but 1 to 4 in the blue-shoe group. This is possible if more girls like red shoes and more boys like blue shoes.

This research is a non-experimental study because the variable, shoe-color, is not a manipulated variable. Instead, the variable is an assigned or subject variable (Kerlinger, 1964) in the sense that its two levels are used to select participants. This is typical of many retrospective studies in epidemiology (Mausner & Kramer, 1985). The outcomes of t-tests based on this sort of non-experimental data are ambiguous in the a way not true of the t-tests based on experimental data. The ambiguity is the result of the fact that the assigned variable may be confounded with another variable. The confounding works as follows.

Increasing the sample size also brings about a change in the gender ratio (albeit inadvertently) at both levels of the assigned variable. It turns out that the red-shoe group is also the group with a higher verbal ability by virtue of (a) its larger girl-to-boy ratio, and (b) girls are verbally more competent than boys. In view of the confounding variable, the data do not warrant accepting the "size-dependent significance" view. In short, increasing the sample size of a non-experimental study may change the composition of the groups of participants in a manner not envisaged in the original design or intent of the study. It is the resultant confounding that brings about statistical significance, not the increase in the sample size per se.

Non-refutation Theories

Not accepting the experimenter expectancy effects (EEE), Chow (1994) reported data that were contrary to EEE. To Green (2002), this state of affairs actually supported EEE (viz., Chow, 1994, found what he expected to find). Green could have also mentioned (but did not) that EEE would also be supported had Chow's data been consistent with it. In other words, EEE receives support regardless of whether the data are consistent or inconsistent with it. This state of affairs means that EEE is an example of a non-refutable theory. A non-refutable theory is not worthy of serious consideration, as may be seen from the following three statements:

> [S3] It will rain.
> [S4] It will rain tomorrow.
> [S5] There will be 1 cm of rainfall on April 1, 2004 at Location X.

[S3] can never be refuted because it does not specify when it will rain. [S4] is supported no matter what the weather is like tomorrow because there is always another tomorrow. Neither [S3] nor [S4] can be taken seriously as an implication of a theory about atmospheric changes in the future. Consequently, the theory itself is also not worthy of serious consideration if the said theory has only implications like [S3] or [S4]. [S5] is better as a testable empirical statement because the criteria for refuting it are made explicit.

EEE is not a satisfactory empirical theory because its implications fall short of the required specificity found in [S5]. The non-refutability of EEE speaks ill of EEE (call it "the irony of non-refutability").

EEE and Russell's (1940) "I am lying" Paradox

It is easy to fall for the irony of non-refutability because it is like Russell's (1940) "I am lying" paradox. When John says, "I am lying," he is saying that there is a proposition *p* such that he asserts *p*, and *p* is false. Suppose that, at 5:30 p.m., John makes Statement [S6]. Further suppose that, during the five minutes in question, John made only one statement (viz., [S6]).

> [S6]: I make a false statement between 5.28 p.m. and 5.33 p.m.

The paradox is as follows:

> If [S6] is true, the statement made by John at 5:30 p.m. must be false. However, if [S6] is false, every statement made during the crucial period must be true. As [S6] is the only statement made, [S6] must be true. The paradox is that "if *p* is true it is false, and if it is false it is true" (Russell, 1940, p. 62).

As pointed out by Russell (1940), two hierarchically arranged propositions are implicated in the example. To see his point, first let "A(*p*)" stands for "I assert *p* between 5.28 p.m. and 5:33 p.m." Then [S6] becomes [S7].

> [S7] "There is a proposition p such that A(*p*) and p is false."

It may be seen that the example implicates the two propositions, [S7] and *p*; and they belong to two levels of abstraction. Specifically, Statement p is about the world at Level *n* of abstraction. [S7] is about what is said about the world; it is at a level more abstract than that of *p* (viz., Level *n + 1*). It is true that either *p* or [S7] is capable of being true or false. However, their truth-values belong to Levels *n* and (*n + 1*), respectively. Hence, a false [S7] does not confer the value of truth to *p* (as insinuated by the paradox). In Russell's (1940) words, "The man who says 'I am telling a lie of order *n*' is telling a lie, but of order *n + 1*" (p. 63).

Meta-experiment versus Experiment

The moral of the story is that a statement like Russell's "I am lying" example owes its paradoxical nature to a failure to distinguish between the two levels of abstraction the statement implicates. The same is true with the paradoxical nature of EEE.

EEE is a theory about conducting experiment. When one tests EEE, one is conducting an experiment about conducting experiment. Hence, EEE has to be tested at a more abstract level than a theory about behavior. Hence, what is required is a meta- experiment. Rosenthal and Fode's (1963a, 1963b) did not conduct the required meta-experiment. In contrast, Chow's (1994) study was a meta-experiment. The paradox suggested by Green (2002) is the result of failing to distinguish between the two levels of abstraction between a meta-experiment and an experiment.

Experiment versus Measurement - A Technical Difference of Importance

Green (2002) finds Chow's (1994) criticism of the experimenter expectancy effects (EEE) wanting because Green considers the difference between "data collector" and "experimenter" a simple matter of technical definition. In doing so, Green is denying the important differences between a non-experimental study and the experiment. For ease of exposition, consider the measurement exercise (Panel 1 of Table 1) and the experiment (Panel 2 of Table 2).

In Panel 1 of Table 2 is described the situation in which the investigator asks two groups of senior undergraduates to collect memory span data. To the group consisted of *A, B* and *C*, the investigator insinuates the bias that the memory span should be large (the "Large span" group henceforth). The other group (viz., *M, P* and *Q*) receives the suggestion that they are expected to obtain a small memory span (the "Small span" group subsequently).

Table 2
The Distinction Between the Formal Structure of the Experiment (Panel 1) and That of the Meta-experiment (Panel 2)

Panel 1—The Formal Structure of the Experiment

Investigator									
"Large span" Expectancy					*"Small span" Expectancy*				
A		B		C	M		P		Q
S_1		S_1		S_1	S_1		S_1		S_1
…		…		…	…		…		…
S_n		S_n		S_n	S_n		S_n		S_n
$\overline{X}_A$		$\overline{X}_B$		$\overline{X}_C$	$\overline{X}_M$		$\overline{X}_P$		$\overline{X}_Q$
		$\overline{X}_L$					$\overline{X}_S$		

A, B, C, M, P and Q are data-collectors, not experimenters.

Panel 2—The Formal Structure of the Meta-experiment

Investigator									
"Large span" Expectancy					*"Small span" Expectancy*				
D			E		H			K	
S_{C1}	S_{E1}		S_{C1}	S_{E1}	S_{C1}	S_{E1}		S_{C1}	S_{E1}
…	…		…	…	…	…		…	…
S_{Cn}	S_{En}		S_{Cn}	S_{En}	S_{Cn}	S_{En}		S_{Cn}	S_{En}
$\overline{X}(E-C)_D$			$\overline{X}(E-C)_E$		$\overline{X}(E-C)_H$			$\overline{X}(E-C)_K$	
$\overline{X}(E-C)L$					$\overline{X}(E-C)S$				

A, B, M and Q are experimenters.

To subscribe to "the possibility that the [data collector] is subtly passing his or her expectancies on to the subjects" (Green, 2002, p. 43, my substitution in square brackets) is to say that the data obtained by A or M or any data collector is the result of a data collector's having received the specific instruction. However, the expectation instruction is not the only reason. For example, the data may be due to the type of words used, the presentation rate adopted, the mode of stimulus presentation employed or any other features found in the data collection procedure. As the data are collected in only one condition, there is no valid comparison baseline. Hence, there is no way to exclude any of the alternative explanations of the data in favor of Green's interpretation. What is amiss in Panel 1 of Table 2 is the absence of experimental controls. It is for this reason that Individuals A, B, C, M, P, Q are called "data collectors."

The basic structure of the experiment is depicted for each of D, E, H and K in Panel 2 of Table 2. Regardless of the expectancy condition, any one of them collects data in two conditions that are identical in all aspects but one (Boring, 1954, 1969; Chow, 2002*b*). For example, acoustically similar words may be used in the memory span task in the S_E condition, whereas acoustically dissimilar words of comparable length are used in the S_C condition. Moreover, the to-be-tested participants are assigned randomly to the S_E and S_C conditions.

The independent variable is the type of to-be-remembered words. The control variables are the word-length, the presentation rate adopted, the mode of stimulus presentation adopted, and the individual who collects the data. In other words, with the exception of the independent variable, experimental subjects are being tested in an identical manner in both the experimental and control conditions. Whatever the data are like, the control variables can be excluded as explanations (Boring, 1954, 1956) at the experimental level by virtue of the inductive principle, the method of difference (Cohen & Nagel, 1934; Chow, 2002*b*). It is for these reasons that Individuals D, E, H and K are called "experimenters."

At the meta-theoretical level, the expectation instruction is the meta-independent variable. Individuals D and E are being "tested" in the same manner as Individuals H and K are being tested, with the exception of the expectation instruction. It is for this reason that the study is a meta-experimenter to the investigator.

In other words, the difference between "data collector" and "experimenter" should not be dismissed as a mere issue in technical definition. It points to an important methodological difference, namely, whether or not the procedure has provisions for excluding alternative interpretations of the data. That controls are present in the experiment (Panel 2 of Table 2), but absent in the measurement exercise (Panel 1 of Table 2), gives the lie to the assertion that "the technical definition of experimenter has little to do with the (EEE's) reality" (Green, 2002, p. 43). Rosenthal (1976) realized as much when he said,

> But much, perhaps most, psychological research is not of this sort [the researcher collects data in one condition only, as represented by A, B, C, M, P or Q in Panel 1 of Table 2]. Most psychological research is likely to involve the assessment of the effects of two or more experimental conditions on the responses of the subjects [as represented by D, E, H or K in Panel 2 of Table 2]. If a certain type of experimenter tends to obtain slower learning from his subjects, the "results of his experiments" are affected not at all so long as his effect is constant over the different conditions of the experiment. Experimenter effects on means do not necessarily imply effects on mean differences. (Rosenthal, 1976, p. 110, explication in square brackets and emphasis in italics added).

The italicized statement shows that Rosenthal (1976) saw the crucial difference between collecting measurement data in one condition (Panel 1 of Table 2) and experimental data in two properly set up conditions (Panel 2 of Table 2). The to-be-tested statistic is the mean in the former, but the difference between two means in the latter.

The Impossibility of EEE

Being an important component of the social psychology of the psychological experiment (SPOPE), EEE implicates the possibility of the participants faking their actual performance in response to the demand characteristics of the situation (Orne, 1962, 1969; see *subject effects* and *demand characteristics* in Section 6.4 of Chow, 2002*b*). It may now be shown why Green's (2002) suggestion that "the possibility that the [data collector] is subtly passing his or her expectancies on to the subjects" (Green, 2002, p. 43) may apply to the measurement scenario in Panel 1 of Table 2, but not to the experimental situation depicted in Panel 2 of Table 2.

For EEE to be true, it is necessary to assume that the investigator succeeds in indoctrinating the individuals who have direct contact the research participants in the course of data collection. Furthermore, one has to assume further (a) that the data collectors are willing to, as well as capable of, conveying the bias to those being tested, and (b) the to-be-tested participants comply with the bias (even by faking).

If EEE (together with the rest of the SPOPE argument) were correct, the two groups of data collectors in Panel 1 would administer the memory span task in such a way that participants in their respective groups would perform in accordance to expectation. As it is possible for some one to fake poor performance (even though participants in the "Large span" group cannot fake good performance), the expectation of Panel I is not unreasonable. This would be confirmed for the investigator when $\overline{X}_L$ is larger than $\overline{X}_S$. That is, to the investigator, the arrangement in Panel 1 of table 2 is an experiment. However, the data collectors in Panel 1 of Table 2 do not conduct an experiment. They collect measurement data, not experimental data. Hence, the difference between two sample means in question is irrelevant to the truth of EEE in view of the realization that "effects on means do not necessarily imply effects on mean differences" (Rosenthal, 1976, p. 110).

Why would the participants not fake in a way that support "effects on mean differences"? The question perhaps should be why the participants and the experimenter cannot meet the EEE requirement. In discussing Panel I of Table 2, the issue is the participants' absolute performance. Faking is not impossible as discussed previously. However, there are two concerns in the case of Panel 2 of Table 2.

For EEE to be possible, each experimental subject must be capable of behaving differently in the two conditions in the direction required by the next consideration. That is, the difference between the mean differences of the two expectancy conditions must be in the direction envisaged by EEE (viz., [$\overline{X}_{(E-C)L\arg e}$ - $\overline{X}_{(E-C)Small}$]). Can any one orchestrate and monitor the experimental procedure is such a way that the difference between $\overline{X}(E - C)L$ and $\overline{X}(E - C)S$ is in the expected direction? Can all participants in the meta-experiment orchestrate such an elaborate state of affairs?

In short, underlying the "data collector" and "experimenter" distinction is the issue of whether or not there is provision in the research procedure for reducing ambiguities. It is the reason why Wilkinson et al.'s (1999) report is misleading in saying that all research methods are equally important. The experimental method is superior in its ability to exclude alternative interpretations of the data by virtue of experimental controls.

Rejecting EEE - Modus Tollens, not Null Hypothesis

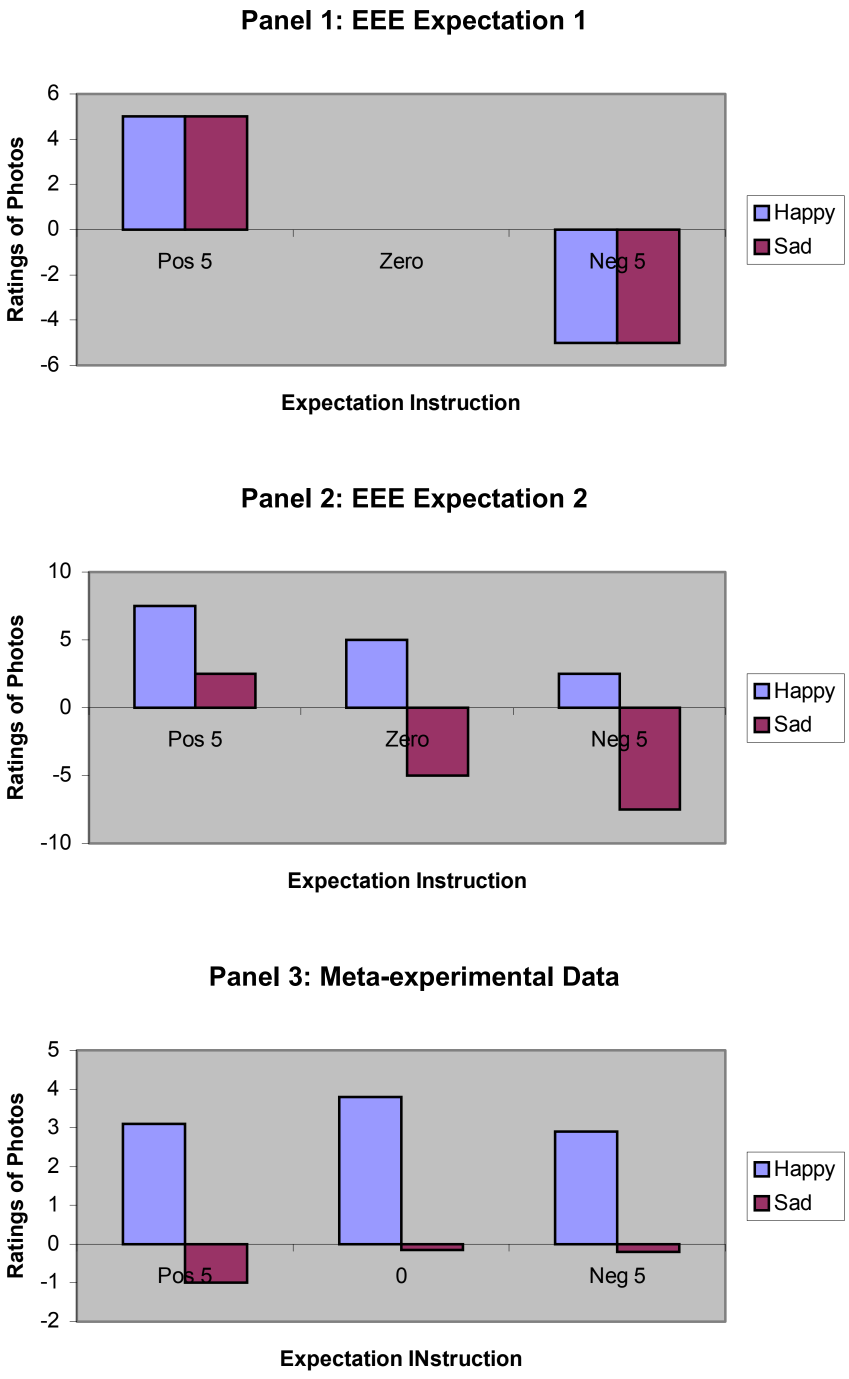
Green (2002) finds it insufficient to reject EEE on the basis of "a single failure to replicate it" (p. 42). The research community should take this good point as an invitation to test EEE (as well as the rest of the SPOPE phenomena) with properly designed meta-experiments.

Green (2002) raises an interesting question when he wonders "what conclusion we should draw when we fail to reject [the null hypothesis]" (p. 42). The answer is simply that the data are due to chance. Such a decision says nothing about the substantive hypothesis (viz., EEE in the present discussion; see Tukey, 1960, for the distinction between deciding whether or not there is statistical significance and drawing a research conclusion). However, Green's (2002) real question seems to be: How was Chow (1994) justified in rejecting EEE on the basis of a failure to reject H_O? Chow (1994) justified the rejection of EEE with modus tollens, as may be illustrated with Figure 1.

Each subject was shown a photograph on every trial. The subject's task was to rate whether or not the stimulus is a photograph of someone who had recently enjoyed a success (maximum +10) or suffered a recent failure (minimum -10). Half of the faces were "happy faces"; the other half was "sad faces." The two types of faces were randomized in the course of the experimental session. The three levels of expectancy instruction (viz., the mean rating obtained by previous researchers) were "+5", "0" and "-5," respectively.

It is not possible to derive any empirical implication from EEE without making additional assumptions. Chow (1994) considered two scenarios. It was assumed in Scenario 1 that the experimenter (e.g., D or E or H or K) and the subjects ignored completely what the stimuli were. Their behavior was determined solely by the expectancy instruction. The implication of such a scenario is shown in Panel 1 of Figure 1. The reasoning underlying the testing of Scenario 1 is Syllogism 1.

> Syllogism 1
> [Major Premise] If EEE is true, data are like what is depicted in Panel 1 of Figure 1.
> [Minor Premise] [To be supplied by the meta-experimental data.]
> [Conclusion] Retain EEE tentatively or reject EEE, depending on the data.



Panel 1: EEE Expectation 1

Panel 2: EEE Expectation 2

Panel 3: Meta-experimental Data

Figure 1. Two expectations of EEE (Panels I and 2) contrasted with meta-experimental data (Panel 3)

It was assumed in the second scenario that neither the experimenter nor the subjects could ignore completely the stimuli. However, the expectation instruction would have the following effects: [a] It minimized the effects of stimuli that were contrary to the expectation instruction. [b] It exaggerated the effects of stimuli that were consistent with the expectation instruction. The implication of the second scenario is shown in Panel 2 of Figure 1. The reasoning underlying the testing of Scenario 2 is Syllogism 2.

> Syllogism 2
> [Major Premise] If EEE is true, data are like what is depicted in Panel 2 of Figure 1.
> [Minor Premise] [To be supplied by the meta-experimental data.]
> [Conclusion] Retain EEE tentatively or reject EEE, depending on the data.

As it turned out, Chow's (1994) meta-theoretical data were like what is depicted in Panel 3 of Figure 1. The pattern is like neither the pattern in Panel 1 nor the pattern in Panel 2. In other words, the minor premise of either Syllogism 1 or Syllogism 2 denies the consequent of the syllogism in question. Hence, the antecedents of the two major premises of Syllogisms 1 and 2 is denied. It is by the force of modus tollens that Chow (1994) rejected EEE, not because he failed to reject H_O. Given the pattern in Panel 3, EEE would still be rejected even if there were a significant difference among the three instruction conditions. The non-significant result means the data may be due to chance influences (viz., the processing implied by Scenario 1 or 2 does not occur).

Statistical Significance and Substantive Importance

Green (2002) reiterates another refrain favored by critics of NHSTP. The issue is "the much more important problem is not that of small effects coming up significant, but rather of important ones (their size notwithstanding) coming up non-significant **.....**" (p. 44). A concrete example may be used to facilitate discussion.

Suppose that the two levels of the independent variable, medication, are Wonder Drug and Placebo. Ten patients each are used in the two conditions. Moreover, the putative utility of Wonder Drug is to eliminate depression (and hence prevent suicides). Obviously, the substantive consequence of administering Wonder Drug is important par excellence.. The outcome is that the mean number of suicides in the Wonder Drug conditions is nil, whereas two patients committed suicide in the placebo condition. However, the t-statistic is not significant. Green (2002) asks what one should do in such circumstances. His answer seems to be to ignore the non-significance decision, but to conclude that the two deaths in the placebo condition is not a chance event. Suggestive of this answer is the fact that the boldface ellipses in the previous paragraph stands for "because the sample sizes used were too small" (p. 44).

The purpose of doing the research is to reduce ambiguity about the efficacy of Wonder Drug. It has been shown in the previous section that the lack of statistical significance serves only to say that the experimental manipulation is not effective, and that the data-pattern is the result of chance influences. Hence, to ignore the non-significance decision is to say that Wonder Drug is actually effective even though the two deaths in the placebo condition may be an accident. The professed reason is that saving two more lives is important (so important that the researcher is willing to accept data that may be brought about by an accident). However, this reason is disingenuous for the following reasons.

To begin with, the choice of the original sample size is not be a haphazard decision in a properly designed study. At the very least, the sample sizes should be comparable to those used in other drug trials. It would be foolhardy to use a sample size of ten (or 200) when other drug trials employ a larger (or smaller) sample size. In other words, the sample size is never too small when it is a well-informed choice. Hence, in choosing a sample size, the researcher is, in fact, acknowledging that the sample size is appropriate.

The said commitment means that, if the data turn out to be non-significant, the researcher would not (and cannot) attribute the failure to the sample size. By the same token, critics of the study also cannot (and should not) simply evoke "insufficient sample size" when (a) there is no theoretical reason to do so, and (b) the sample size used is within the typical range used in related studies. The onus is on the one who suggests a change in sample size to show how the sample size is responsible for the non-significant result. As it will shown later, the appeal to statistical power does not work.

Suppose one is a disinterested, serious researcher. One would examine various aspects of the experiment, namely, the adequacy of the design, the subject-selection or assignment procedure, the dosage used, the control variables, the control procedures, the adequacy of the instruction, the competence of the data collector, the stability of the independent variable (Chow, 1985), and the like. Re-do the experiment with the same sample sizes with the improved design or procedure.

If the concern in the example is to save more lives, the researcher should test another drug rather than ignoring the evidence that does not support the drug on the mere pretext of insufficient sample size. That the researcher would persevere with Wonder Drug in the face of non-significance means that the researcher has some vested interests in Wonder Drug (rather than saving lives). For example, there are some theoretical reasons for using Wonder Drug. These issues are admittedly important and worth pursuing. However, they are not statistical problems. Obtaining statistical significance cannot solve them. Nor can they be ascertained by increasing the sample size (even if the "size-dependent significance thesis" were true).

Further Ado With Power Analysis

Why are researchers asked to select their sample size with reference to the power of the to-be-tested statistical test? The reason is "The power of a statistical test is the probability that it will yield statistically significant results" (Cohen, 1987, p. 1; italicized emphasis added). What is said in the quote is inconsistent with the fact that the statistical power is a conditional probability. Apologists of statistical power have not shown us how to resolve the inconsistence. Nor have they given an account of why researchers are asked to carry out power analysis when (a) they know that statistical power is a conditional probability, and (2) a conditional probability cannot do not what the quote says.

Figure 2 may be used to allay Green's (2002) misgivings about Chow's (2002a) critique of power analysis. From top to bottom, it depicts the transition from the population distributions (Panels A and B) to two separate sampling distribution of differences (Panel C and D) to the t-distribution in Panel E.

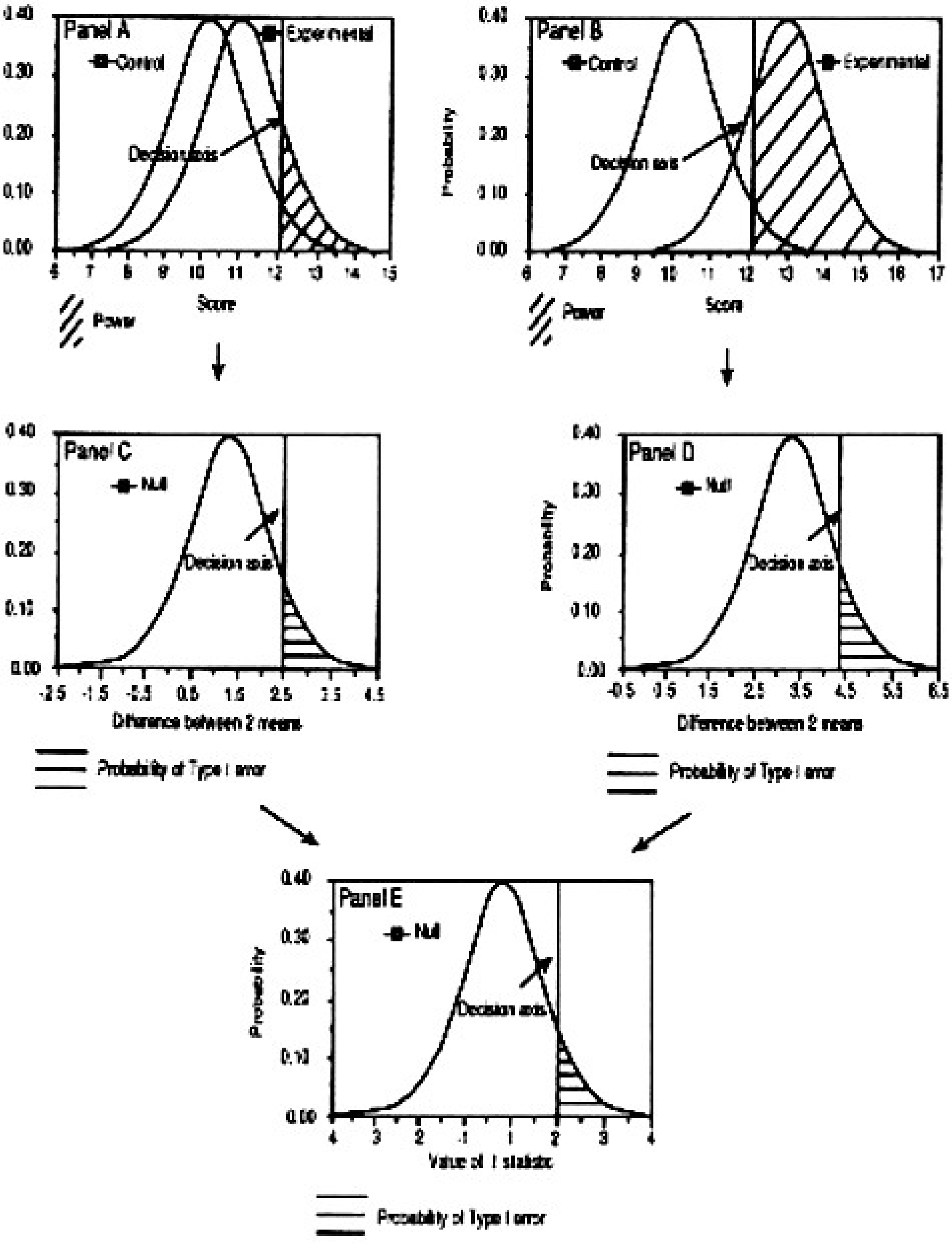


Figure 2. The graphical representation of two effect sizes (Panels A and B) and the corresponding differences between two means in raw-score units (Panels C and D), as well as in standard error units (Panel E) (adopted from Chow, 1996, Figure 6.2, p. 134).

The Population Distributions

Suppose that 20 15-year old boys are randomly selected. They are randomly assigned to the control and experimental conditions of the 1-factor, 2-level experiment (i.e., ten boys in each condition). It helps to recall Winer's (1962) insight that the researcher in the present example is dealing with samples from two underlying statistical populations (viz., 15-year old boys given no experimental treatment and 15-year old boys that are given the experimental treatment). Hence, there is a pair of distributions in Panels A and B of Figure 2.

The pair of distributions in Panel A represents a small effect size (d), whereas the pair in Panel B represents a larger effect size (where d = $\left(\overline{X}_E - \overline{X}_C\right)$ / S_C; i.e., the difference between the two sample means divided by the standard deviation of the control sample). Also shown is the fact that statistical power is larger in Panel B than the pair on in Panel A (see the area shaded with slanting lines). It may be seen from the x-axis, as well as from the denominator of the effect size, that the unit of analysis is the individual score. Despite the presence of the decision axis in either Panel A or Panel B, the decision about statistical power is not made at this level.

The Level of the Sampling Distribution

The probability foundation of inferential statistics is the sampling distribution of the test statistics. In the present example, it is the *sampling distribution of differences between means*. It is represented by Panel C or Panel D of Figure 2. Note the ranges of the two x-axes. The range goes from -2,5 through 4.5 in Panel C, but from -0.5 through 6.5 in Panel D. That is, a larger effect size (viz., the pair in Panel B) is represented by a sampling distribution displaced more to the right on the continuum of all possible mean differences between sample means.

The visual distance between the two population distributions found at the level of raw scores is not (and cannot be) represented at the level of sampling distribution. The two sampling distribution has the same dispersion means that they have the same standard error of differences. The unit of analysis at this level is a sample-pair. Again, despite the presence of the decision axis, statistical significance is not decided with the sampling distribution.

The Level of the t-distribution

The decision about statistical significance is made in terms of the t-distribution, as witnessed by the equation of the t-statistic, [El].

[El] t = $\left(\overline{X}_E - \overline{X}_C\right) - \left(u_E - u_C\right) / \ s_{(\overline{X}_E - \overline{X}_C)}$

As may be seen from [El], the t-statistic is the standardization of the sampling distribution of differences between means. The result is that, regardless of the position of the sampling distribution on the continuum of all possible sample means, the mean difference is always represented by $t = 0$. The decision about statistical significance is made with the *t*-distribution. The unit of analysis at the level is also a sample-pair.

The visual distance between the two population distributions found at the level of raw scores is not (and cannot be) represented at the level of the *t*-distribution. How may one represent statistical power at the level?

In short, it has been shown that effect size or statistical power is defined at the level of raw scores (viz., Panel A or B). The graphical representation of these two concepts require two distributions. The statistical decision is not made at the level of raw scores. Instead, it is made at the level of the standardized difference between two sample means. It is based on a lone *t*-distribution that is more abstract than the two statistical populations. At the same time, as may be recalled from Panels A and B, statistical power is not defined at the same level of abstraction as is the t-distribution. It is not clear how knowing the statistical power can inform the researcher about the probability of obtaining statistical significance.

Summary and Conclusions

Green (2002) begins with a general defense of Wilkinson et al.'s (1999) report on statistical inference in general and the use of the term "contrast group" in particular. The latter point is well made. In response to his defense of the experimenter expectancy effects (EEE), it is necessary to discuss (a) the difficulties with non-refutable theories, (b) the reason for the paradoxical nature of EEE, (c) the distinction between measurement and experiment, (d) the distinction between experiment and meta-experiment, and (e) the difference between NHSTP and modus tollens. Green's explication of the reasoning behind the "size-dependent significance thesis" is helpful to making explicit what is amiss in the thesis. Revisiting power analysis makes it possible to describe the transition from the level of raw data to the level of sampling distribution and, finally, to the level of the t-distribution. This exercise helps to show why the power-analytic argument is debatable.

References

Boring, E. G. (1954). The nature and history of experimental control. *American Journal of Psychology, 67,* 573-589.

Boring, E. G. (1969). Perspective: Artifact and control. In R. Rosenthal, & R. L. Rosnow (Eds.), *Artifact in behavioral research (pp.* 1-11). New York: Academic Press.

Chow, S.L. (1985). Iconic store and partial report. *Memory & Cognition,* 13, 256-264.

Chow, S. L. (1994). The experimenter's expectancy effect: a meta-experiment. *Zeitschrift fűr Pädagogische Psychologie* (German Journal of Educational Psychology), 8, 89-97.[http://cogprints.ecs.soton.ac.uk/archive/00000825/]

Chow, S. L. (2002a). Issues in statistical inference. *History and Philosophy of Psychology Bulletin*, *14*, 30-41.

Chow, S. L. (2002b). Experimentation in psychology--Rationale, concepts, and issues. Methods in psychological research. In Encyclopedia of Life Support Systems (EOLSS), Eolss Publishers, Oxford, UK. [http://www.eolss.net][http://cogprints.ecs.soton.ac.uk/archive/00002781/]

Cohen, J. (1987). *Statistical power analysis for the behavioral sciences* (revised edition). New York: Academic Press.

Cohen, M. R., & Nagel, E. (1934). *An introduction to logic and scientific method.* London: Routledge & Kegan Paul.

Essex, C., & McKitrick, R. (2002). *Taken by storm: The troubled science, policy and politics of global warming*. Toronto: Key Porter Books.

Green, Christopher D. (2002). Comment on Chow. History and Philosophy of Psychology Bulletin, 13, 42-46.

Kerlinger, F. N. (1964). *Foundations of behavioral research.* New York: Holt, Rinehart and Winston.

Orne, M. T. *(1962).* On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist, 17, 776-783.*

Orne, M. T. (1969). Demand characteristics and the concept of quasi-controls. In R. Rosenthal, & R. L. Rosnow (Eds.), *Artifact in behavioral research* (pp. 143-179). New York: Academic Press.

Rosenthal, R. (1976). *Experimenter effects in behavioral research* (Enlarged edition). New York: Irvington Publishers.

Rosenthal, R., & Fode, K.L. (1963a). Three experiments in experimenter bias. *Psychological Reports,* 12, 491-511.

Rosenthal, R., & Fode, K.L. (1963b). The effect of experimenter bias on the performance of the albino rat. *Behavioral Science, 8,* 183-189.

Russell, B. (1940). *An inquiry into meaning and truth*. London: George Allen and Unwin.

Tukey, J. W. (1960). Conclusions vs. decisions. *Technometrics*, *2*, 1-11.

Wilkinson, L., and Task Force on Statistical Inference, APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8), 594-604.

Winer, B. J. (1962). *Statistical principles in experimental design.* New York: McGraw-Hill.