History and Philosophy of Psychology Bulletin Volume 14, No. 1, 2002

# Issues in Statistical Inference

Siu L. Chow

Department of Psychology, University of Regina

Being critical of using significance tests in empirical research, the Board of Scientific Affairs (BSA) of the American Psychological Association (APA) convened a task force "to elucidate some of the controversial issues surrounding applications of statistics including significance testing and its alternatives; alternative underlying models and data transformation; and newer methods made possible by powerful computers" (BSA; quoted in the report by Wilkinson & Task Force, 1999, p. 594). Guidelines are stipulated in the report for revising the statistical sections of the APA Publication Manual.

Some assertions in the report about research methodology are reasonable. An example is the statement, "There are many forms of empirical studies in psychology, including case reports, controlled experiments, quasi-experiments, statistical simulations, surveys, observational studies, and studies of studies (meta-analyses) ... each form of research has its own strengths, weaknesses, and standard of practice" (Wilkinson & Task Force, 1999, p. 594). However, it does not follow that data collected with any two methods are equally unambiguous. At the same time, a method that yields less ambiguous data is methodological superior to one that yields more ambiguous data. That is, despite the assertions made in the report, a case can be made that "some of these [research methods] yield information that is more valuable or credible than others" (Wilkinson & Task Force, 1999, p. 594).

It is unfortunate that the report reads more like an advocacy document than an objective assessment of the role of statistics in empirical research. Moreover, non-psychologist readers of the report can be excused for having a low opinion of psychologists' research practice and methodological sophistication.

Lest psychologists' methodological competence be misunderstood because of the report, this commentary addresses the following substantive issues: (a) the acceptability of the 'convenience' sample, (b) the inadequacy of the contrast group, (c) the unwarranted belief in the experimenter's expectancy effects, (d) some conceptual difficulties with effect size and statistical power, and (e) the putative dependence of statistical significance on sample size.

## The 'Convenience' Sample, Representativeness and Independence of Observations

> If we can neither implement randomization nor approach *total control of variables* that modify effects (outcomes), then we should use the term "control group" cautiously. In most of these cases, it would be better to forgo the term and use "contrast group" instead. In any case, we should describe exactly which confounding variables have been explicitly controlled and speculate about which unmeasured ones could lead to incorrect inferences. In the absence of randomization, we should do our best to investigate sensitivity to various untestable assumptions. (Wilkinson & Task Force, 1999, p. 595, emphasis in italics added)

A non-randomly selected sample is characterized as a "convenience sample" (Wilkinson & Task Force, 1999, p.595). It is a label apparently applicable to most samples used in psychological research because most experimental subjects are college student-volunteers. However, a case can be made that using such non-random samples does not necessarily detract from the findings generality. Nor does such a practice violate the requirement that data from different subjects be statistically independent. More importantly, using non-random samples is not antithetical to experimental controls.

### *Non-random Participant-selection and Representativeness*

Suppose that, on the basis of the data collected from student-subjects, Experimenter E draws a conclusion about memory. The non-random nature of the sample would not affect the objectivity of the finding when the validity of the experiment is assessed with reference to unambiguous, theoretically informed criteria. At worst, one may question the generality of the experimental conclusion. Perhaps, this is the real point of the "Sample" section (Wilkinson & Task Force, 1999, p. 595), as witnessed by its reservations about the representativeness of the convenience sample.

Although non-random selection of research participants jeopardizes the generality of survey studies, random subject-selection may not be necessary for generality in cognitive psychology. For instance, a non-random sample in an opinion survey about an election may be selected by stationing the enumerators at the entrance of a shopping mall. The representativeness of the opinion of such a sample (of the entire electorate's opinion) is suspect because patrons of the particular shopping mall may over-represent one social group, but under-represent another social strata. This is crucial because political opinion and socio-economic status are not independent.

In contrast, consider a student-subject sample of a study of the capacity of the short-term store. As there is no reason to doubt the similarity between college students' short-term store capacity and that of the adult population at large, it is reasonable to assume that the student-subject sample is representative of all adults *in the said capacity* despite that no random selection is carried out. That is, random selection is not always required for establishing the generality of the result when there is neither a theoretical nor an empirical reason to question the representativeness of the sample in the context of the experiment.

### *Student-subjects as Theoretically Informed Samples*

The psychologist's practice of using student-subjects is further justified by the fact that psychologists employ student-subjects in a theoretically informed way. For example, in testing a theory about verbal coding, the experimenter may use only female students. The experimenter may use only right-handed students when the research concern is a theory about laterality or hemispheric specialization. Students may be screened with the appropriate psychometric tests before being included in a study about attitude. In short, depending on the theoretical requirement, psychologists adopt special subject-selection criteria even when they use student-subjects. Moreover, psychologists do select subjects from outside the student-subject pools when required (e.g., they use hyperactive boys to study theories of hyperactivity). The mode of subject-selection is always explicitly described in such an event. That is, psychologists' convenience samples do not detract from the data's generality. Furthermore, psychologists describe only those procedural features that deviate from the usual, well-understood and warranted practice.

### *Independent Observations from Non-randomly Selected Samples*

A crucial assumption underlying statistical procedures (be it significance test, confidence-interval estimate or regression analysis) is that observations are independent of one another. It can be illustrated that cognitive psychologists' use of non-randomly selected student-subjects does not violate this independence assumption. Consider the case in which, having discussed among themselves, twenty students decide to participate in the same memory experiment. This is non-random subject-selection *par excellence.*

Suppose further that subjects, whose task is to recall multiple 10-word lists in the order they are presented, are tested individually. The words and their order of appearance are randomized from trial to trial. Under such circumstances, not only would an individual subject's performance be independent of that of other subjects, the subject's performance is also independent of his or her own performance from list to list. In other words, to ensure statistical independence of observations, what needs to be randomized is the stimulus material or its mode of presentation, not individual subjects. Such a randomized procedure ensures that non-randomly selected subjects may still produce statistically independent data.

## Causal Inference-Deductive Implications of Explanatory Theories

The conclusion about any causal relationship is based on the implicative relationships among the explanatory theory, the research hypothesis, the experimental hypothesis, the statistical hypothesis, and the data (see, e.g., the three embedding conditional syllogisms discussed in Chow, 1996, 1998). The causal conclusion owes its ambiguity to deductive logic as a result of the facts that (a) hypothetical properties are attributed to the unobservable theoretical entities postulated (Feigl, 1970; MacCorquodale & Meehl, 1948), (b) it is always possible to offer multiple explanations for the same phenomenon (Popper, 1968a, 1968b), and (c) *affirming the* consequent of a conditional proposition does not affirm its antecedent (Cohen & Nagel, 1934; Meehl, 1967, 1978). In other words, the report's treatment of random subject-assignment is not helpful when it incorrectly assigns to the research design the task of making causal inference possible. Nor is the ambiguity of drawing causal conclusions a difficulty in inductive logic, as said in the report that "the causal inference problem ... one of missing data" (Wilkinson & Task Force, 1999, p.600).

## Random Subject-assignment, Control and Induction

If causal inference is independent of research design in general (and the completely randomized design in particular), what precisely is the role of the design in empirical research? The answer to this question sets in high relief the unacceptability of the report's suggestion of replacing the control group with the contrast group if the researcher is concerned with conceptual rigor or methodological validity.

### *Experimental Design and Induction*

Contrary to the induction by enumeration assumed in the report (recall the invocation of `missing data' on p. 600), underlying a valid research design is one of Mill's (1970) canons of induction (viz., Method of Difference, Joint Method of Agreement and Difference, Method of Concomitant Variation, and Method of Residues; see Cohen & Nagel, 1934, for the exclusion of Method of Agreement). The function of these inductive rules is to exclude alternative explanations, as may be seen in Table 1, which depicts the formal structure of the completely randomized one-factor, two-level experiment described in the `Independent Observations from Non-randomly Selected Samples' sub-section above.

Made explicit in Table 1 are the independent variable (viz., the similarity in sound among the ten words in the list), four control variables (viz., list length, number of lists, rate of presentation, and the length of the items used), the dependent variables (viz., the number of items recalled in the correct order), and some of an infinite number of extraneous variables. This formal arrangement of the independent, control and dependent variables satisfies the stipulation of Mill's (1973) Method of Difference. That is, psychologists rely on an inductive method that is more sophisticated than the induction by enumeration envisaged in the report.

### *Control Variables and Exclusion of Explanations*

Variables CI through C4 are control variables in the sense that they are represented by the same level at both levels of the independent variable. This feature is one type of the `constancy of condition' of experimental control (Boring, 1954, 1969). Suppose that there is a good reason to exclude chance influences as the explanation of the difference between $\overline{X}_E$ and $\overline{X}_c$ (i.e., the difference is statistically significant). This difference is found when there is no difference in any of the four control variables between the experimental and control conditions. Consequently, it can be concluded that none of the control variables is responsible for the difference between $\overline{X}_E$ and $\overline{X}_c$. This shows that experimental control in the form of using control variables serves to exclude explanations, not to affirm a causal relationship.

### *Random Subject-assignment As A Control Procedure*

Extraneous variables of the experiment are defined by exclusion, namely, any variable that is neither the independent, the control or the dependent variable is an extraneous variable. As the symbol, $C\propto$, in Table 1 indicates, there is an infinite number of extraneous variables. It follows that, in order to exclude any of them as an explanation of the data, these extraneous variables have to be controlled (in the sense of being held constant at both levels of the independent variable). Depending on the nature of the independent variable, the extraneous variables may be excluded from being confounding variables by (a) assigning subjects randomly to the experimental and

Table 1
The Method of Difference That Underlies the Completely Randomized One-factor, two-level Experimental Design

| | Independent variable | Control variables | | | | Extraneous variables | Dependent variable |
|---|---|---|---|---|---|---|---|
| | | C 1 | C2 | C3 | C4 | C5 to C∝ | |
| | Similarity in sound | List length | Number of lists | Rate of presnetation | Length of items used | | Number of items recalled in the correct order |
| E | Yes | 10 | 12 | 1 item/s | 5-letter nouns | gender, age, SES, height, ethnicity, hobbies, etc. | $\overline{X}_E$ |
| C | No | 10 | 12 | 1 item/s | 5-letter nouns | | $\overline{X}_C$ |

E = Experimental Group; C = Control Group

control conditions (the only procedure recognized in the report), (b) using the repeated-measures design, and (c) using the matched-groups (or randomized block) designs. That is, instead of rendering possible causal inference, random subject-assignment is only one of several control procedures that serve to prevent extraneous variables from being confounding variables.

### Control versus Contrast Group

That no contrast group can replace the control group may also be seen from Table 1. The control group and the experimental group are identical in terms of all the control variables. It is reasonable to assume that the two groups are comparable in terms of the extraneous variables to the extent that the completely randomized design is appropriate and that the random-assignment procedure is carried out successfully. Being different from the control group, the contrast group envisaged in the report has to be a group that differs from the experimental group in something else in addition to being different in terms of the independent variable. The additional variable involved cannot be excluded as an alternative explanation. That is, there is bound to be a confounding variable in the contrast group; otherwise it would be a control group.

The subject's gender is treated as an extraneous variable in Table 1. However, if there is a theoretical reason to expect that male and female students would perform differently on the task, gender would be controlled in one of several ways. First, gender may be used as an additional control variable (e.g., only male or female students would be used). Second, gender may be used as another independent variable, in which case the relevancy of gender may be tested by examining the interaction between acoustic similarity and gender. The third alternative is to use gender as a blocking variable, such that equal number of male and females are used in the two groups. Which male (or female) is used in the experimental or control condition is determined randomly. In other words, the choice of any variable (be it the independent, control or dependent variable) is informed by the theoretical foundation of the experiment. This gives the lie to the report's treating matching or blocking variables as 'nuisance' variables.

Giving the impossible meaning of "total control of variables" (Wilkinson & Task Force, 1999, p.545) to `control' is an example of a striking feature in the report, namely, its indifference to theoretical relevancy. It is objectionable that the confusing and misleading treatment of the control group is used in the report as the pretext to "forgo the term ["control group"] and use 'contrast group' instead" (Wilkinson & Task Force, 1999, p.595, explication in square brackets added). As had been made explicit by Boring (1954, 1969), the control group serves to exclude artifacts or alternative explanations.

The Task Force's recommendation of replacing the control group by the contrast group is an invitation to weaken the inductive principle that underlies experimental control. Such a measure invites ambiguity by allowing confounds in the research. The ensuing damage to the internal validity of the research cannot be ameliorated by explaining `the logic behind covariates included in their designs' (Wilkinson & Task Force, 1999, p.600) or by describing how the contrast group is selected (pp. 594-597). Explaining or describing a confound is not excluding it.

**Experimenter's Expectancy Effects Revisited**

> Despite the long-established findings of the effects of experimenter bias (Rosenthal, 1966), many published studies appear to ignore or discount these problems. For example, some authors or their assistants with knowledge of hypotheses or study goals screen participants (through personal interviews or telephone conversations) for inclusion in their studies. Some authors administer questionnaires. Some authors give instructions to participants. Some authors perform experimental manipulations. Some tally or code responses. Some rate videotapes. An author's self-awareness, experience, or resolve does not eliminate experimenter bias. In short, there are no valid excuses, financial or otherwise, for avoiding an opportunity to double-blind. (Wilkinson & Task Force, 1999, p. 596)

As may be seen from the quote above, the report bemoans that psychologists do not heed Rosenthal's (1976) admonition about the insidious effects of the experimenter's expectancy effects (or EEE henceforth). Psychologists are faulted for not describing how they avoid behaving in such a way that they would obtain the data they want. Given the report's faith in EEE, it helps to examine the evidential support for EEE by considering Table 2 with reference to the following comment:

> 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 to do necessarily imply effects on mean differences.* (Rosenthal,. 1976, p. 110, explication in square brackets and emphasis in italics added).

The putative evidence for EEE came from Rosenthal and Fode (1963a, 1963b), the design of both of which is shown in Panel 1 of Table 2. In their 1963a studies, students in the "+5" expectation and "-5" expectation groups were asked to collect photo-rating data under one condition. Again, students collected `rate of conditioning' data with rats in two expectation conditions in their 1963b study. Of interest is the comparison between the mean ratings of the two groups of students. A significant difference in the expected direction was reported between the two means, $\overline{X}_{+5}$ and $\overline{X}_{-5}$, in both studies.

Note that the said significant difference is an effect on means, not an effect on mean difference, in Rosenthal's (1976) terms. Moreover, Rosenthal (1976) also noted correctly that the schema depicted in Panel 1 is not the structure of psychological experiments. That is, Individuals A, B, C, M, P and Q in Panel 1 should not be characterized as `experimenters' at all because they did not conduct an experiment. While the two studies were experiments to Rosenthal and Fode (1963a, 1963b), the studies were mere measurement exercises to their students. In other words, Rosenthal and Fode's (1963a, 1963b) data cannot be used as evidential support for EEE.

What is required, as noted in the italicized emphasis above, are data collected in accordance with the meta-experiment schema depicted in Panel 2 of Table 2. While Chow (1994) was the investigator who conducted a meta-experiment (i.e., an experiment about conducting the experiment), D, E, H and K were experimenters because they collected data in two conditions which satisfied the constraints depicted in Table 1. When experimental data were collected in such a meta-experiment, Chow (1994) found no support for EEE. There was no expectancy effect on mean difference in the meta-experiment. That is, EEE owes its apparent attractiveness to the casual way in which `experiment' is used to refer to any empirical research. The experiment is a special kind of empirical research, namely, a research in which data are collected in two or more conditions that are identical (or comparable) in all aspects,

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

| Investigators (Rosenthal & Fode, 1963a, 1963b) | | | | | |
|---|---|---|---|---|---|
| *+5* | | | *-5* | | |
| **A** | **B** | **C** | **M** | **P** | **Q** |
| $S_1$ | $S_1$ | $S_1$ | $S_1$ | $S_1$ | $S_1$ |
| … | … | … | … | … | … |
| $S_n$ | $S_n$ | $S_n$ | $S_n$ | $S_n$ | $S_n$ |
| $\overline{X}_A$ | $\overline{X}_B$ | $\overline{X}_C$ | $\overline{X}_M$ | $\overline{X}_P$ | $\overline{X}_Q$ |
| | $\overline{X}_{+5}$ | | | $\overline{X}_{-5}$ | |

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

Panel 2—The Formal Structure of the Meta-experiment

| Investigator (Chow, 1994) | | | | | |
|---|---|---|---|---|---|
| +5 | | | -5 | | |
| **D** | | **E** | **H** | | **K** |
| $S_{C1}$ $S_{E1}$ | | $S_{C1}$ $S_{E1}$ | $S_{C1}$ $S_{E1}$ | | $S_{C1}$ $S_{E1}$ |
| … … | | … … | … … | | … … |
| $S_{Cn}$ $S_{En}$ | | $S_{Cn}$ $S_{En}$ | $S_{Cn}$ $S_{En}$ | | $S_{Cn}$ $S_{En}$ |
| $\overline{X}(E-C)_D$ | | $\overline{X}(E-C)_E$ | $\overline{X}(E-C)_H$ | | $\overline{X}(E-C)_K$ |

A, B, M and Q are experimenters.

except one (viz., the aspect represented by the independent variable).

### Effect Size and Meta-analysis

> We must stress again that reporting and interpreting effect sizes in the context of previously reported effects is essential to good research. It enables readers to evaluate the stability of results across samples, designs, and analyses. Reporting effect sizes also informs power analyses and meta-analyses needed in future research. (Wilkinson & Task Force, 1999, p. 599)

The Task Force's reservations about the accept-reject decision about $H_0$ and its insistence on reporting the effect size (Wilkinson & Task Force, 1999, p.599) and confidence-interval estimates (Wilkinson & Task Force, 1999, p.599) have to be considered with reference to (a) Meehl's (1967, 1978) distinction between the substantive and statistical hypotheses, (b) what the statistical hypothesis is about, and (c) Tukey's (1960) distinction between making the statistical decision about chance influences and drawing the conceptual conclusion about the substantive hypothesis. As $H_0$ is the hypothesis about chance influences on data, a dichotomous accept-reject decision is all that is required. It is not shown in the report why psychologists can ignore Meehl's or Tukey's distinction in their methodological discourse.

The main reason to require reporting the effect size is that the information is crucial to meta-analysis. This insistence would be warranted if meta-analysis were a valid way to ascertain the tenability of an explanatory theory. However, there are conceptual difficulties with meta-analytic approaches (Chow, 1987). For the present discussion, note that 'effect' as a statistical concept refers to (a) the difference between two or more levels of an independent variable or (b) the relation between two or more variables at the statistical level. Given the fact that different variables are used in the context of diverse tasks in a converging series of experiments (Garner, Hake, & Eriksen, 1956), the effects from diverse experiments are not commensurate even though the experiments are all ostensibly about the same phenomenon (see Table 5.5 in Chow, 1996, p. 111). It does not make sense to talk about the 'stability results across samples' when dealing with apples and oranges. Consequently it is not clear what warrants the assertion, "reporting and interpreting effect sizes in the context of previously reported effects is essential to good research" (Wilkinson & Task Force, 1999, p.599).

### Some Reservations about Statistical Power

The validity of the power-analytic argument is taken for granted in the report (Wilkinson & Task Force, 1999, p.596). It may be helpful to consider three issues about the power-analytic approach, namely, (a) the statistical power is a conditional probability, (b) statistical significance and statistical power belong to different levels of abstraction, (c) the determination of sample size is not a mechanical exercise.

#### Power Analysis as a Conditional Probability

Statistical power is the 1's complement of b, the probability of the Type II error. That is, statistical power is the probability of rejecting $H_0$, given that $H_0$ is false. The probability becomes meaningful only after the decision is made to reject $H_0$. As b is a conditional probability, so should be statistical power. How is it possible for such a conditional probability to be an exact probability, namely, "the probability that it will yield statistically significant results" (Cohen, 1987, p. 1; italics added)?

#### The Putative Relationship Between Statistical Power and Statistical Significance

Central to the power-analytic approach is the assumption that statistical power is a function of the desired effect size, the sample size, and the alpha level. At the same time, the effect size is commonly defined at the level of the statistical populations underlying the experimental and control conditions (e.g., Cohen's, 1987, d). It take two statistical population distributions to defined the effect size.

The decision about statistical significance, on the other hand, is made on the basis of a lone theoretical distribution in the case of the t-test (viz., the sampling distribution of the differences between two means). Moreover, the sampling distribution of difference is at a level more abstract than the distributions of the two statistical populations underlying the experimental and control conditions. Consequently, it is impossible to represent correctly both alpha and statistical power at the same level of abstraction (Chow, 1991, 1996, 1998). Should psychologists be oblivious to the `disparate levels of abstraction' difficulty noted above?

### Sample-size Determination

It is asserted in the report that using the power-analytic procedure to determine the sample size would stimulate the researcher "to take seriously prior research and theory" (Wilkinson & Task Force, 1999, p.586). This is not possible even if it were possible to leave aside the `disparate levels of abstraction' difficulty for the moment. A crucial element in determining the sample size with reference to statistical power is the 'desired effect size.' At the same time, it is a common power-analytic practice to appeal to "a range of reasonable alpha values and effect sizes" (Wilkinson & Task Force, 1999, p.597). Such a range consists typically of ten to fourteen effect sizes.

Apart from psychological laws qua functional relationships between two or more variables, theories in psychology are qualitative explanatory theories. These explanatory theories are speculative statements about hypothetical mechanisms. Power-analysts have never shown how subtle conceptual differences in the qualitative theories may be faithfully represented by their limited range of ten or so 'reasonable' effect sizes. Furthermore, concerns about the statistical significance are ultimately concerns about data stability and the exclusion of chance influences as an explanation. These issues cannot be settled mechanically in the way depicted in power-analysis. The putative relationships among effect size, statistical power and sample size brings us to the putative dependence of statistical significance on sample size.

## The Relationship Between Statistical Significance and Sample Size Examined

It is taken as a truism in the report that statistical significance depends on sample size. Yet, there has been neither empirical evidence nor analytical reason for saying that "statistical tests depend on sample size" (Wilkinson & Task Force, 1999, p.598). Consider the assertion, "as sample size increases, the tests often will reject innocuous assumptions," (Wilkinson & Task Force, 1999, p.598) with reference to Table 3. Suppose that the result of the 1-tailed, independent-sample t-test with *df = 8* is 1.58. It is not significant at the .05 level with reference to the critical value of 1.86. The *df* becomes 148 and the critical value becomes 1.65 when each of the independent samples is increased to 75. An implication of the size-dependent significance assertion may now be seen.

Table 3

An implication of the `sample size-dependent significance' thesis

| | | | |
|---|---|---|---|
| Independent-sample t | df = 8 (n, = n2 = 5) | calculated t = 1.58 | critical t = 1.86 |
| | df = 148 (n, = n2 = 75) | calculated t = ? | critical t = 1.65 |
| | df = 1498 (n, = n2 = 750) | calculated t =? | critical t = 1.65 |

In order for the `sample size-dependent significance' assertion to be true, the calculated t must become larger than 1.58 when the sample size is increased from $n_1 = n_2 = 5$ to $n_1 = n_2 = 75$. Even if there is no change in the calculated *t* when the sample size is increased to 75, the calculated *t* should become larger when the sample size is increased to $n_1 = n_2$ =750. Otherwise, increasing the sample size would not make the result significant if the *t*-ratio remains at 1.58. Six simulation trials were carried out to test the `sample size-dependent significance' thesis as follows.

### Three Simulation Trials With the Zero-null $H_0$

Two identical statistical populations were used in the zero-null case (i.e., $H_0$: $u_1 - u_2 = 0$). The two populations' size, mean and standard deviation were 1328, 4.812, and .894, respectively (see Panels 1 and 2 of Table 4). The procedure used may be described with the $n_1 = n_2 = 5$ case.

(1) A random sample of 5 was selected with replacement from each of the two statistical populations.

(2) The two sample means and their difference were calculated.

(3) The two samples were returned to their respective statistical populations.
(4) Steps (1) through (3) were repeated 5,000 times.
(5) The mean of the 5,000 differences (between two means) was determined (viz., -.007; see the last but one cell of Panel 2A of Table 4). (6) The 5,000 calculated t-values were cast into a - frequency distribution (see Panel 2A).

Steps (1) through (6) were repeated with $n_1 = n_2 = 75$, as well as with $n_1 = n_2 = 750$. As may been seen from the `Mean *t*-ratio' row, the values for the three sample sizes (viz., 5, 75 and 750) are -.007, .011 and .002, respectively. They do not differ among themselves, nor does any one of them differ from zero.

### Three Simulation Trials With the Point-null $H_0$

Does the `sample size-dependent significance' thesis hold when an effect-size is expected before data collection (e.g., $H_0$: $u_1$ - $u_2$ = half of the standard deviation of the first population)? This is the situation where the expected difference between the two conditions is larger than 0 before the experiment. Hence, three more simulations were carried out with two statistical populations whose means differ. Specifically, while $u_1 = 4.812$, $u_2 = 5.262$. This arrangement represents a medium effect size in Cohen's (1987) terms (viz.., the difference of .45 represents half of the standard deviation of the first population). Steps (1) through (6) described in the "Three Simulation Trials With the Zero-null *Ho"* section above were carried out. Each of the t-ratios was determined with ($\overline{X}_E - \overline{X}_C$ - .45) as the numerator in view of the point-null, Ho: ($u_1$ - $u_2$ = 0.45 (see Kirk, 1984; Chow, 1986, pp. 132-137). The data are shown in Panels 2D, 2E and 2F in Table 4. The mean t-ratios for sizes 5, 75 and 750 are .006, 0 and .028, respectively. They are not different.

### The Independence of Sample Size and Statistical Significance

Data from Panels 2A, 2B and 2C of Table 4 are entered into a 2-way classification scheme so as to apply the $\chi^2$ test (see Panel 1 of Table 5). The three levels of the variable Sample Size are 5, 75 and 750. The second variable is Significance-status (i.e., Yes or No) with reference to the critical value appropriate for the *df.* Each of the 5,000 t-ratios from each level of Sample Size was put in the appropriate cell of the 3 by 2 matrix (see the six boldface entries in Panel 1 of Table 5).The $\chi^2$ *(df* = 2) = 2.645 is not significant at the .05 level. Data from Panels 2D, 2E and 2F of Table 4 were treated in like manner (see Panel 2 of Table 5). The six italicized boldface entries yield a $\chi^2$ *(df = 2*) = 3.458. It is also insignificant. As there is no reason to reject chance as an explanation of the two $\chi^2$'s, the conclusion is that sample size and statistical significance are independent.

## Summary and Conclusions

It is true that "each form of research has its own strengths, weaknesses, and standard of practice" (Wilkinson & Task Force, 1999, p. 594). However, this state of affairs does not invalidate the fact that some research methods yield less ambiguous data than others. Nor does it follow that all methodological weaknesses are equally tolerable if the researcher aims at methodological validity and conceptual rigor. Having a standard of practice per se is irrelevant to the validity of the research method. To introduce the criteria of being valuable or credible in methodological discussion is misleading because "being valuable" or "being credible" is not a methodological criterion. Moreover, "being valuable" or "being credible" may be in the eye of the beholder. This state of affairs is antithetical to objectivity.

Psychologists can justify using non-randomly selected student-subjects because the representativeness of such samples is warranted on theoretical grounds. Moreover, using student-subjects does not violate the independence of observations requirement. Causal inference is made by virtue of the implicative relationships among the hypotheses at different levels of abstraction and data. Being one of several control procedures, random subject-assignment serves to exclude extraneous variables as alternative explanations of data. Psychologists can exclude many extraneous variables by using the repeated-measures or randomized-block design.

Many of the observations made about psychologists' research practice would assume a more benign complexion if theoretical relevancy and some subtle distinctions are taken into account. For example, the evidential support for the experimenter's expectancy effects has to be re-considered if the distinction between meta-experiment and experiment is made. It is necessary for power-analysts to resolve the 'disparate levels of abstraction' difficulty and to explain how a conditional probability may be used as an exact probability. Despite what is said in the report, it is hoped that non-psychologist readers have a better opinion of psychologists' methodological sophistication, conceptual rigor or intellectual integrity.

## 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.), *Artifacts in behavioral research (pp. 1-11).* New York: Academic Press.

Campbell, D. T., & Stanley, J. L. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Chow, S. L. (1987). Meta-analysis of pragmatic and theoretical research: A critique. *Journal of Psychology,* 121, 259-271.

Chow, S. L. (1991). Some reservations about statistical power. *American Psychologist,* 46, 10881089.

Chow, S. L. (1994). The experimenter's expectancy effect: A meta-experiment. *German Journal of Educational Psychology, 8,* 89-97.

Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility.* London: Sage.

Chow, S. L. (1998). A précis of "Statistical Significance: Rationale, Validity and Utility." *Behavioral and Brain Sciences,* 21, 169-194. (http://www.cogsci.soton.ac.uk/bbs/Archive/bbs.chow.html)

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist,* 45, 1304-1312.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist,* 49, 997-1003.

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

Feigl, H. (1970). The "orthodox" view of theories: Remarks in defense as well as critique. In M. Radner, & S. Winokur (Eds.), *Analyses of theories and methods of physics and psychology. Minnesota studies in the philosophy of science (Vol.* IV, pp. 3-16). Minneapolis: University of Minnesota Press.

Garner, W. R., Hake, H. W, & Eriksen, C. (1956). Operationism and the concept of perception. *Psychological Review,* 63, 149-159.

MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review,* 55, 95107.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of science,* 34, 103-115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology,* 46, 806-834.

Mill, J. S. (1973). *A system of logic: Ratiocinative and inductive.* Toronto: University of Toronto Press.

Popper, K. R. (1968a). *The logic of scientific discovery* (2°d edition, originally published in 1959). New York: Harper Row.

Popper, K. R.(1968b). *Conjectures and refutations.: The growth of scientific knowledge* (originally published in 1962). New York: Harper Row.

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.

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

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

Wilkinson & Task Force on Statistical Inference, APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist,* 54(8), 594604.

| Panel l | Score | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Total |
|---|---|---|---|---|---|---|---|---|---|---|
| | Fre uenc | 1 | 12 | 36 | 412 | 669 | 128 | 65 | 5 | 1328 |

| Panel 2 | N 1=N2=1328; el = fit = .894 | | | | | |
|---|---|---|---|---|---|---|
| | Panel 2A Panel 2B Panel 2C | | | Panel 2D Panel 2E Panel 2F | | |
| | ul =u2=4.812 | | | ul =4.812; u2=5.262 | | |
| Range of t ratio's | uE - uc =-.005<br>.564<br>n1 =n2=5 | uE - uc --.001<br>Ql~_ ~1 =.148<br>n1 =n2=75 | uE - uc = 0<br>al~_ -.046<br>n1 =n2=750 | uE - uc --.482<br>61_ =.584<br>n1 =n2=5 | uE - uc = -.45<br>al_ -.145<br>.97 AD<br>nj =n2=75 | uE - uc --.449<br>r) =.046<br>*6177_*<br>nj =n2=750 |
| | Frequency | Fre uenc | Frequency | Frequency | Frequency | Frequency |
| <_ -2.901 | 3 0 | 12 | 13 | 2 9 | 12 | 8 |
| -2.90 - -2.701 | 16 | 9 | 8 | 14 | 5 | 9 |
| -2.70 - -2.501 | 3 0 | 2 2 | 6 | 3 2 | 16 | 15 |
| -2.50 --2.301 | 35 | 26 | 23 | 24 | 22 | 20 |
| -2.30 - -2.101 | 50 | 3 5 | 28 | 64 | 31 | 3 5 |
| -2.10 --1.901 | 10 | 55 | 60 | 17 | 42 | 46 |
| -1.90 - -1.701 | 124 | 85 | 96 | 114 | 91 | 76 |
| -1.70 - -1.501 | 77 | 119 | 114 | 67 | 116 | 94 |
| -1.50 - -1.301 | 214 | 160 | 112 | 190 | 170 | 136 |
| -1.30 - -1.101 | 166 | 183 | 206 | 163 | 187 | 215 |
| -1.10 - -.901 | 264 | 254 | 282 | 216 | 214 | 253 |
| -.900 - -.701 | 240 | 271 | 245 | 227 | 288 | 245 |
| -.700 - -.501 | 366 | 362 | 365 | 408 | 333 | 298 |
| -.500--.301 | 413 | 346 | 338 | 392 | 369 | 395 |
| -.300 - -.101 | 137 | 343 | 383 | 111 | 335 | 393 |
| -.100 - .099 | 744 | 461 | 401 | 862 | 499 | 441 |
| .100-.299 | 111 | 359 | 398 | 113 | 373 | 360 |
| .300 - .499 | 373 | 353 | 368 | 395 | 354 | 358 |
| .500-.699 | 389 | 331 | 340 | 374 | 349 | 347 |
| .700 - .899 | 221 | 288 | 302 | 227 | 266 | 298 |
| .900 - 1.099 | 257 | 239 | 241 | 233 | 255 | 251 |
| 1.10 - 1.299 | 162 | 211 | 187 | 159 | 188 | 197 |
| 1.30 - 1.499 | 135 | 140 | 158 | 146 | 160 | 160 |
| 1.50 - 1.699 | 136 | 98 | 120 | 123 | 103 | 118 |
| 1.70 - 1.899 | 101 | 89 | 65 | 111 | 75 | 86 |
| 1.90 - 2.099 | 12 | 46 | 55 | 9 | 71 | 53 |
| 2.10 - 2.299 | 69 | 41 | 34 | 67 | 34 | 40 |
| 2.30 - 2.499 | 38 | 20 | 22 | 32 | 21 | 19 |
| 2.50 - 2.699 | 34 | 19 | 12 | 39 | 12 | 15 |
| 2.70 - 2.899 | 14 | 11 | 9 | 9 | 4 | 8 |
| 2.90 - 3.099 | 1 | 6 | 3 | 1 | 2 | 4 |
| >- 3.100 | 31 | 6 | 6 | 3 2 | 3 | 8 |
| Mean t-ratio | -.007 | -.011 | .002 | .006 | 0 | .028 |
| Expected t-ratio | 0 | 0 | 0 | 0 | 0 | 0 |

Table 5

The number of empirically determined t-ratios tabulated in Table 3 that exceed the critical value of the t-ratio (significant) and do not exceed the critical value (non-significant) at the .OS level when Ho is a zero-null (Panel ) and a point-null (Panel 2).

| | | *df* | $n_1$ -$n_2$ | Critical *t* | Signi-ficant | Not significant | $\chi^2$ (df = 2) |
|---|---|---|---|---|---|---|---|
| Panel 1<br>alpha = .05 (1-tailed) | | 8 | 5 | :9 -1.86 or >_ 1.86 | 462 | 4538 | |
| | | 148 | 75 | 5 -1.65 or >_ 1.65 | 510 | 4490 | 2.645 |
| | | 1498 | 750 | <_ -1.645 or > 1.645 | 490 | 4510 | |
| | | | | | | | |
| Panel 2<br>alpha = .05 (1-tailed) | | 8 | 5 | <_ -1.86 or >_ 1.86 | *449* | *4551* | _ |
| | | 148 | 75 | :5 -1.65 or >_ 1.65 | *471* | *4529* | 3.458 |
| | | 1498 | 750 | :5 -1.645 or >_ 1.645 | *503* | *4497* | |

Siu Chow is a professor of psychology at the University of Regina. He is interested in the interface between attention and memory, the rationale of experimentation, and the role of statistics, particularly significance tests, in empirical research. (email: Siu.Chow@uregina.ca)