# IDENTIFICATION OF MISCONCEPTIONS IN LEARNING STATISTICAL POWER WITH DYNAMIC GRAPHICS AS A REMEDIAL TOOL

**Yu, Chong Ho and John T. Behrens, Arizona State University**
**John T. Behrens, 325 Payne Hall, ASU, Tempe AZ 85287-0611**

**Key words: Power, Simulations**

Despite Neyman and Pearson's work stressing the role of statistical power in experiments, and the increasing accessibility of power analysis to applied researchers, these concepts continue to be misunderstood and misapplied in the psychology literature (Sedlmeier & Gigerenzer, 1989), the education literature (Brewer, 1972), and the medical epidemiology literature (Freiman, Chalmers, Smith & Kuebler, 1986; Cohen, 1990).

While several reasons for the difficulty of power analysis may exist, we have been impressed by the number of misconceptions students have regarding power analysis. The objective of this study was to catalog undergraduate and graduate students' misconceptions in this area, and to examine the efficacy of a computer simulation to remediate these misconceptions. Identification of misconceptions is important because misconceptions often propagate errors through the belief system, misconceptions are more difficult to change than uninformed opinion, and misconceptions can be used to guide instructional goals. Because our knowledge of this area is limited, we sought to spread a wide net in our data collection efforts. First, we prepared a short-answer test based on our previous teaching experiences and pilot work to help in the identification of misconceptions. In addition, we collected more process-oriented data by taping subjects as they used the computer simulation.

## Problems of Power Analysis

Statistical testing is often presented as a mismatch of Fisherian and Neyman-Pearson approaches (Huberty, 1987; Gigerenzere, 1987), which leave the students to reconcile the unmentioned differences. Fisherian legacy, which offers a clear-cut solution of yes/no, tends to dominate behavioral sciences (Galarza-Hernandez, 1993). Sedlmeier and Gigernzer (1989) observed that Fisher's theory was historically first and got institutionally entrenched. Consequently, it is difficult for researchers to make the switch to power analysis. In Fisherian statistical testing, the null hypothesis is zero effect. Without specifying an effect size, the only thing it can concluded after achieving statistical significance is that "the effect is not nil." Following this strict Fisherian tradition, researchers would find no rooms for power analysis because power depends on the unknown alternate distribution (Lehmann, 1993).

Countering the above problem, Haase, Ellis, and Ladany (1989) introduced the "good enough principle"--rather than testing whether the mean difference is nil, testing whether the difference is at a certain level. Nonetheless, effect size determination is a subjective process and there is no generally accepted convention for its determination. This ambiguity of effect size choice may be one of the factors to the lack of use of power analysis.

## Rationale of Computer Simulations

Mastery of the concept of statistical power requires not only understanding of elementary terms such as $\alpha$, $\beta$, and power, but also their inter-relationships. A student may know increasing sample size increases power, but not know why. This may result, in part, from the difficulty of explaining or drawing all the relevant aspects of the sampling distributions and the difficulty of presenting the influence of each aspect of the set up on the other relevant aspects. We concluded that instruction and assessment must go beyond treatment of individual elements of factual knowledge and focus on the development of a coherent mental model of the sampling distributions.

Toward this end, we developed a dynamic graphic that depicts two sampling distributions and their relevant parts. In addition, the graphic is modifiable in terms of effect size, sample size, and $\alpha$-level.

Graphical presentations have been shown to foster learning and appropriate mental model development Bauer & Johnson-Laird, 1993) and to be an effective means of teaching abstract statistical concepts (Beniger & Robyn, 1978; Fienberg, 1979). In addition, the benefits of interactive computer learning environment have been well-documented in literature pertaining to computer-based instruction (Lynett, 1985; Riskin, 1990). Furuther, it is well known that active participation and interactivity is an important aspect of any instructional setting (Schaffer & Hannafin, 1986)

## Method

Material

The program "Power-Sim" was written in XLISP-STAT (Tierney, 1991) by Dr. John Behrens and Yu, Chong Ho, operating on the Macintosh platform. This software module is composed of the following units: a).

a pull-down menu, which provide definitions of concepts; b). a figure displaying the null and alternate distributions; c). a dialog box indicating the power; d). slider bars for controlling effect size and sample size; a button panel for choosing a discrete α-level, and a button panel for turning coloring on in different parts of the graphic. Any change in a control panel lead to immediate updating of the graphic (See Figure 1).

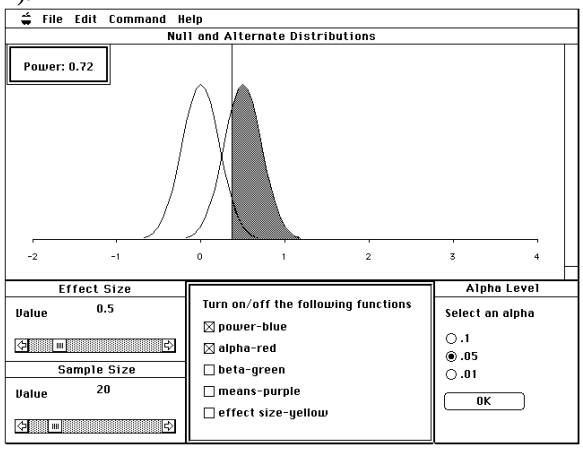


**Figure 1. Display of dynamic graphic for illustrating the interrelationships of the components of power analysis.**

Effect size can be conceptualized in many ways. In this software we adopted the definition given by Cohen (1988): the mean difference between groups in standard deviation units. To simplify the simulation, a z-test was depicted with population standard deviations of both populations set to one.

Initial test

In this study, essay-type test, concurrent think aloud protocol, participant observation, and post-treatment interview were used. The test includes questions of power and the essential components for understanding power, including null distribution, alternate distribution, sampling distributions, β, α, and effect size. The content of each question is listed in the discussion section.

The initial data collection consisted of administering the test to 31 undergraduate students at a large midwestern university. The students were enrolled in a statistics course offered by the department of psychology and had already read, and been lectured to about power analysis. The data collection was introduced as educational research that was conducted by individuals from outside the department. Subjects signed informed consent forms and were informed of their rights as participants. Responses were evaluated by a panel of three graduate assistants. and a coding scheme was developed to categorize the responses.

Concurrent Think Aloud Protocol

In a second round of data collection, four doctoral students in departments of psychology or educational psychology at the same university were asked to think aloud while working with the "Power-Sim" program. At the same time the proctor observed how they operated the program and what strategies they applied. Afterwards, they were interviewed to express their attitudes toward learning power analysis and using the software module.

The goal of having the participants think aloud is to reveal what information is kept in short-term memory during the problem solving process. Short-term memory corresponds the "voice" one is aware of during self-talk. Because most cognitive theories hold short-term memory is the working space for problem solving and holding information from long-term memory, think aloud protocols have been suggested as an essential data collection device for understanding cognitive processes (e.g. Anderson, 1987; Ericsson, 1993). Because the presence of video-tape equipment is highly reactive, session were recorded by a tape recorder and augmented by field notes from the proctor. Participants completed a brief questionnaire concerning their reaction to the program and recommendations for improvement.

Participant Observation

Because many of the participants were not familiar with the computer keyboard use, mouse use, or, in some cases, what to expect from a computer program, the proctor conducted sessions with no more than four participants in the room at one time, and closely monitored the students. When students appeared to face an insurmountable impasse the proctor moved into a Socratic teaching mode While this does not provide a clean assessment of the use of the computer simulation, it was consistent with the expected use by students who will see it used in class and have access to help when needed.

## Results and Discussion

Findings of the Test

From the initial test it was found that very few students understood the concepts of sampling distributions, effect size, hypothesis testing and the relationships among power, β, and α level. Their error patterns are summarized as the following:

1. Misunderstandings of sampling distribution. When asked to explain what a sampling distribution is, six of the 31 respondents (19.4 %) gave a correct answer with nine skipping the question entirely. Misconceptions centered on a) confusing a sampling distribution with a distribution of a sample, b) confusing it with the distribution of population raw scores and c) failing to generalize beyond the sampling

distribution of the mean.

When asked to explain what an effect size is, only one participant was able to produce a correct definition. Here there seemed to be only a few misconceptions because most participants indicated no conception at all. When misconceptions did occur they reflected vague knowledge that the effect size was related to power. Several subjects confused the variability of the sampling distribution with the effect size.

The basic laws governing the sampling distribution of the means was also absent. When asked to explain how the sampling distribution changes with increase sample size, one-third of the respondents equated the sampling and population raw score distributions. Related to this were the responses by a number of subjects that large sample effects (such as normality in the sampling distribution of the means) magically occurred at, and above, but not before, samples of size 30. This belief was probably distorted from the notion that N=30 is the cut-off between a t-distribution and a normal sampling distribution (Gordon, 1989). We found that this belief is detrimental to the learners. As a matter of fact, besides the sample size and the number of samples, the shape of the population is also a factor in determining the appearance of the sampling distribution. A normal population will result in a normal sampling distribution even if the sample size is smaller than 30.

2. Confused tentative hypotheses with preset conclusions. Question 4 and 5 are "Explain what a null distribution is" and "Explain what an alternate distribution is." Many subjects (56.7%, 17 subjects in Question 4; 60%, 18 subjects in Question 5) tended to over-generalize the role of null and alternate hypotheses. i.e. they regarded the null hypothesis as "the hypothesis that you always don't believe," and the alternative hypothesis as "the hypothesis that you always believe."

The problem of the above beliefs can be explained by the following example. In a medical study examining the safety of a new drug, the hypotheses are as below:

$H_0$: The new drug is unsafe.

$H_1$: The new drug is safe.

In this case the researcher does not necessarily believe in the alternate hypothesis because a Type II error (claiming the drug is safe but indeed it is unsafe) may result in killing patients! There was a real life example in Europe: the tranquilizer thalidmide damaged fetus but unfortunately at first it was concluded as a safe drug (Miller, 1978).

3. Failure to understand the attributes of and relationships among null and alternate distributions, and effect size. In Question 6 ("How does change in effect size affect the location of the null distribution?"), many testers (13 persons, 41.9%) thought that a change in effect size would move the location of the null distribution, while indeed the null is a theoretical distribution fixing in one position. Figure 2 is a scanned image of a graphic response by one participant.

When another question asked "How does the mean of the null distribution differ when the effect size is small compared to when the effect size is big?", eight learners (25.8%) reported a change in effect size would lead to a change of the mean of null distribution.

Because pilot work had led to the observation that some students confuse central tendency with dispersion, we asked "How do the population standard deviation differ when the effect size is small compared to when the effect size is big?" Nine subjects (29%) thought that when effect size differs, the population standard deviation differs correspondingly.

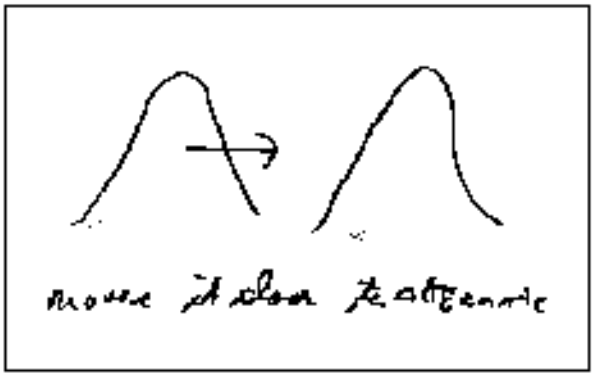

**Figure 2 Example of the misconception of null distribution. The text is "move it closer to alternate".**

4. Most students had vague conceptualizations of power, β and α. When asked to explain what β is," four subjects (12.9%) regarded the β as a distribution. Three subjects (9.7%) could point out that the β is associated with the alternate distribution, but confused β with power. In addition, many subjects simply wrote down the term "Type II error," (41.9%). Three subjects (9.7%) demonstrated that they only memorized the term without understanding because they wrote down wrong explanations to Type II error. For example, one explanation is that "$H_O$ is true but you have to rejected."

In two subsequent questions we presented a graph of overlapped distributions of null and alternate with a small effect size, and a graph of non-overlapped distributions with a large effect size, and asked the testers to explain the characteristics of power, β, α level and effect size. In the case of no observable overlap in sampling distributions, seven participants (22.6%) mis-identified β as power and reversed power to β.

5. Some subjects failed to generalize past

examples. Closely related to the problem of vague conceptualizations, students faced with non-prototypical scenarios showed difficulty applying general rules. For instance, when faced with the graphics of the non-overlapping sampling distributions, approximately 20 percent of the students drew lines in the alternate distribution and marked areas indicating β. This seemed to be in an effort to make the graphics conform the prototypical graphics they had seen in class. This suggests that they had been attending primarily to the surface features of the graphics and not understanding the mechanics which lead to the construction of the graphs. Figures 3 is an example of such responses.

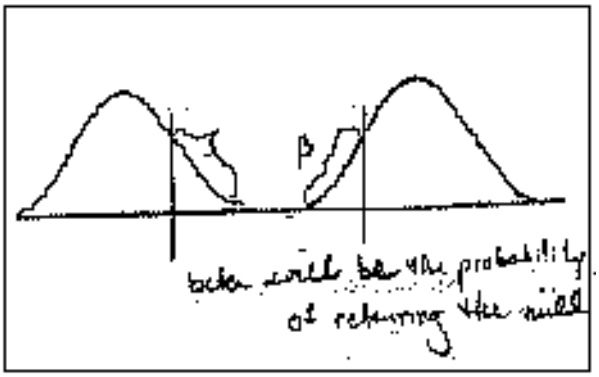


**Figure 3 Example of failure to generalize past examples (drawing an α level to make β while the two distributions are not overlapped)**

6. Knowledge of parts of concepts but not their procedural rules. When asked "Starting with an α level of .05, how should be α level be modified to increase power when all other aspects of the statistical test are held constant," 18 subjects (58.1%) did not know in which direction the α level should be changed in order to increase power. Another six participants (19.4%) even pointed to the opposite direction.

The α level and the region of rejection seem to confuse the learners. When asked "How is α level related to the region of rejection?", six students (19.4%) knew that the α level is the cut-off for the determination of rejection or retention, but they could not clearly differentiate α level and the region of rejection. Six other students (19.4%) did not understand the criterion of rejection at all. For instance, one student answered that "you look up α to find the number that cut the tail off. So if answer higher than or lower than alpha then reject."

Findings of Think-Aloud Protocols

Based on the think aloud protocol, observation and in-depth interview, the following errors of learning power analysis were spotted:

1. Failure to understand β and power have an inverse relationship. When the simulation showed a high power (.99), many users were puzzled by the "missing" β. Some students misidentifiied α as β in attempt of searching for the "missing" β. This seems related to the difficulties reported above in confusing areas under the curve with probabilities that are a function of distances in sampling space. When the areas were not obvious, students could not infer the underlying probability.

2. Failure to categorize power, effect size, α and sample size into proper dependent and independent variables. For example, some users expected that changing in the α level would change the effect size. Another user misunderstood it in the opposite way--he thought that changing the effect size could affect the α level. This mistake is partly caused by mis-interpreting the changes in the graphics as discussed below

3. Oversimplification of power analysis and failure to realize the multi-dimensionality of the issues. One subject concluded that the sample size, rather than the effect size, has more impact on power, when she set the sample size to 15 and effect size to 1 and got the power as .99. Another subject concluded that a small sample size could still achieve a high power, and thus we need not "waste money, effort, time, and energy to go after too many people." These erroneous conclusions appeared to be based on the participants focusing on a single aspect of the multi-dimensional set up of sample-size, effect-size, alpha level and power. It seems that without proper feedback, students could easily develop a number of misconceptions from the use of the computer simulation, rather than being corrected by it.

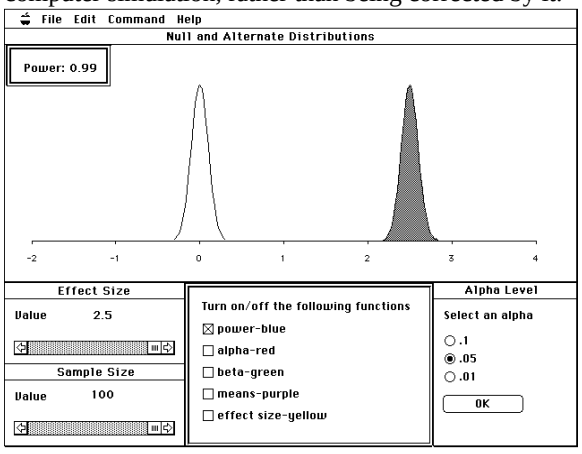


**Figure 4a**

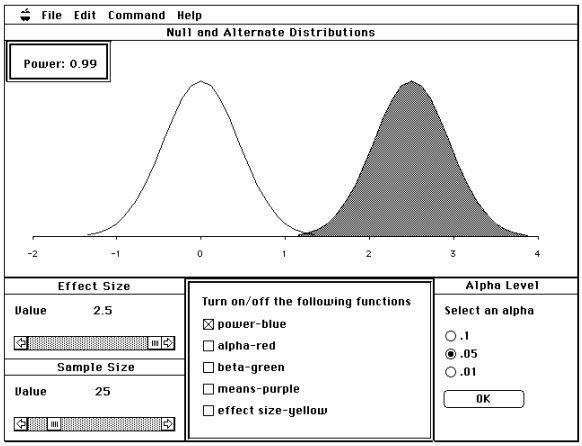


**Figure 4b: Figures 4a and 4b show equivalent effect size with small and large variability.**

4. Failure to identify the attributes of null and alternate distributions. First, some students failed to associate power with the alternative distribution. Second, the fixed scale for the ordinate (pdf) gave some students the impression that sampling distributions with smaller variability were farther apart. In this case it seems the subjects were interpreting areas such as β in terms of absolute size, rather than in terms of relative size compared to the rest of the distribution. Screens under this condition are presented in Figure 4a and 4b.

Observations Concerning the Use of the Computers

When tapes of the student-computer interaction were examined, it was clear that a number of assumptions we had made about the prerequisite skills of the students were not met.

First, students seemed to be overwhelmed by the multivariate nature of the control panel and did not appear to have appropriate problem-solving strategies for isolating and testing the effect of individual variables on the system. Rather, the most common approach was a sort of random search through the controls, usually guided by the idea of incremental movement of controls that often made the patterns indiscernible.

In this case it is not clear whether the students do not have the understanding necessary to analyze a multivariate system like the one presented here or whether they simply did not recognize the instructional task as one that required problem solving. Some other issues are addressed in the following:

1. Subjects tended to prioritize the importance of variables by their location on the screen. All subjects started from the effect size panel, then the sample size panel. This is because the effect size panel is located at the upper left, which is perceived as the most important one by English speakers. The control panel of α level, at the far right was ignored by most users. They did not alter the α level until the tutor told them to do so. Some subjects even thought that the α level is not important at all. It is recommended to group three panels together at the left so that users can manipulate them altogether.

2. Learners did not use hidden features even though these features provide fundamental information for getting started. The pull down menu contains "help," which provides the definitions of power, β, α, sample size, and effect size. The experimenter told the subjects about this feature at the beginning, and implied that it is important to start with prepositional knowledge such as definitions before moving to relationships among propositions. However, only one subject read the "help" carefully. This subject is very proficient in statistics. Her purpose of reading the text was to compare her definitions with our definitions in order to establish a common ground to start with. One participant said that she did not use the pull down menu because she expected the tutor to guide her all the way through. Other subjects simply said they never thought of it. Unlike dialog boxes and control panels, pull down menus are hidden. "Out of sight, out of mind" may be an explanation of the idleness of pull down menu.

3. Most users expected "real time" simulations bring results instantly. When they dragged an object or clicked a button and nothing happened yet, they were so impatience as to drag and click the mouse again and again. Unexpected consequences followed, of course. The users slowed down after the tutor told them that the machine has only a slow 68200 processor. Thus, computing speed is crucial in computer-based instruction. In supplement to high-end hardware, more efficient source code is also highly desirable.

## Conclusion and Recommendations

Analysis of the errors reveals that many students had not acquired the prerequisite understanding of sampling distributions to properly interpret graphic or algebraic representation of power-related concepts. Failure to attain these subskills appear to force the students to rely on surface features of explanations since any more substantial mental models would not be coherent. Accordingly, we are beginning to construct more general simulations aimed at the concepts surrounding sampling distributions. We hope to be able to provide sufficient experience in that domain to serve as preparation for the power simulations.

A second implication of the surface feature learning is that it seems to be tolerated by the educational system we observed insofar as the test items used in those classes could be answered without a deep understanding of the concepts. This suggests

curriculums should be reassessed to ensure deeper conceptual issues are addressed and assessed.

With regard to interacting with the simulation we found that it could be used as a successful remedial tool when the learner was given strong direction and feedback. There are many nuances both to interface use and the interrelationships among the variables affecting power analysis that can easily confuse the user. As we describe above, without close supervision it is possible for a student to develop superstitious behavior and make unwarranted conclusions that further, rather then repair, their misconceptions. We are beginning to develop both work-book based, and computer-based exercises that lead students through the concepts and provide feedback as they develop their mental model.

The computer simulation has shown us that the mental model a student must build is rather complex and easily misdirected. The simulation may have its greatest value in the fact that it can portray numerous scenarios which the student may never have considered (e.g. what if the effect size is 0, or there is no obvious overlap in the distributions). When the myriad possibilities are combined with guided exploration, it seems the simulation can have great benefits for the students. The program of this paper is available via the Arizona State University WWW server (http://seamonkey.ed.asu.edu/~alex/multimedia/power/power.html).

**Acknowledgments**

We would like to thank Professor Sam DiGangi at the Arizona State University for his valuable comments on the draft of this paper.

**References**

Anderson, J. R. (1987). Methodologies for studying human knowledge. Behavioral and Brain Sciences, 10, 467-505.

Bauer, M. L., & Johnsonson-Laird, P. N. (1993). How diagrams can improve reasoning. Psychological Science, 4, 372-378.

Beniger, J. R., & Robyn, D. L. (1978). Quantitative graphics in statistics. American Statistician, 32, 1-11.

Brewer, J. K. (1972). On the power of statistical tests in the "American Educational Research Journal." American Educational Research Journal, 9, 391-401.

Cohen, J. (1990). Things I have learned (So Far). American Psychologist, 45, 1304-1312.

Ericsson, K. A. (1993). Protocol analysis: Verbal reports as data. Cambridge, MA: MIT Press.

Feinberg, S. E. (1979). Graphical methods in statistics. American Statistician, 33, 165-178.

Freiman, J. A., Chalmers, T. C., Smith H. & Kuebler, R. R. (1986). The importance of beta, the type II error, and sample size in the design and interpretation of the randomized controlled trial: Survey of '71 "negative" trials. In J. C. Bailar III & F. Mosteller (Eds.) Medical uses of statistics (p.357-374). Boston, MA: NEJM Books.

Galarza-Hernandez, A. (1993, November). What is probability of rejecting null hypothesis?: Statistical power in research. paper presented at the Annual Meeting of the Mid-South Educational Research Association, New Orleans, LA.

Gigerenzer, G. (1987). Cognition as intuitive statistics. Hillsdale, NJ: L. Erlbaum Associates.

Gordon, F. (1987). Computer Graphics Simulation of the Central Limit Theorem. Mathematics and Computer Education, 21, 48-55.

Hasse, R., Ellis, M., & Ladany, N. (1989). Multiple criteria for evaluating the magnitude of experimental effects. Journal of Consulting Psychology, 36, 511-516.

Huberty, C. J. (1987). On statistical testing. Educational Researcher, 16, 4-9.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? Journal of the American Statistical Association, 88, 1242-1249.

Lynett, P. (1985). The current and potential uses of computer assisted interactive videodisc in the Education of Social Workers. Computers in Human Services, 1, 75-85.

Miller, K. K. (1978). The importance of statistical power in educational research. Occasional paper 13. Bloomington, Indiana: Phi Delta Kappa (ERIC Document Reproduction Service No. ED 152 838)

Riskin, S. R. (1990). Teaching through interactive multi-media programming. A New philosophy of the social sciences and a new epistemology of creativity. California. (ERIC Document Reproduction Service No. ED 327 133)

Sedlmeier, P. & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105, 309-316.

Tierney, L. (1991). LISP-STAT: An object-oriented environment for statistical computing and dynamic graphics. New York: John, Wiley & Sons.