Wayne State University
DigitalCommons@WayneState

Theoretical and Behavioral Foundations of Education Faculty Publications

Theoretical and Behavioral Foundations

11-1-2003

Deconstructing Arguments From The Case Against Hypothesis Testing

Shlomo S. Sawilowsky
Wayne State University, shlomo@wayne.edu

Journal of Modern Applied Statistical Methods
November, 2003, Vol. 2, No. 2, 467-474

1538 – 9472/03/$30.00

Deconstructing Arguments From The Case Against Hypothesis Testing

Shlomo S. Sawilowsky
Educational Evaluation and Research
Wayne State University

The main purpose of this article is to contest the propositions that (1) hypothesis tests should be abandoned in favor of confidence intervals, and (2) science has not benefited from hypothesis testing. The minor purpose is to propose (1) descriptive statistics, graphics, and effect sizes do not obviate the need for hypothesis testing, (2) significance testing (reporting p values and leaving it to the reader to determine significance) is subjective and outside the realm of the scientific method, and (3) Bayesian and qualitative methods should be used for Bayesian and qualitative research studies, respectively.

Key words: Hypothesis testing, bracketed intervals, significance testing, effect size, Bayes, qualitative

Introduction

There has been an increasing amount of journal space given to the case against hypothesis testing over the past quarter of a century. The ensuing debate has taken many directions and has been graced with many forms of argumentation (see, e.g., Sawilowsky, 2003a; Knapp & Sawilowsky, 2001). Two styles of attack against hypothesis testing are contested here.

The first is the proposition that hypothesis testing should be abandoned in favor of confidence intervals. (I prefer the term “bracketed” instead of “confidence” interval for reasons noted in Sawilowsky, 2003a.) Ancillary to this attack is the proposition that hypothesis testing is tolerable if and only if it is (a) buttressed with a report of effect sizes, (b) accompanied by graphical displays, or (c) Bayesian.

The second style of attack is that hypothesis testing should be abandoned due to philosophical arguments. An example is embodied in the question if science has benefited by hypothesis testing.

The “Confidence” Interval Attack

Neyman (1934), who discovered the bracketed interval, equated the probabilities associated with its lower and upper bound with “the ordinary concept of probability” (1934, p. 590). Initially, he seemed to equate it with the fiducial argument promulgated by Fisher (1930). The presumed lack of difference in the derivation of bracketed intervals and fiducial probabilities was the focus of the discussion subsequent to the reading of Neyman’s (1934) paper before the Royal Statistical Society. Bowley (1934) raised the question and presented his answer, “I am not at all sure that the ‘confidence’ is not a ‘confidence trick’… Does it really take us any further?... I think it does not” (p. 609). He considered bracketed intervals to be nothing more than ordinary probabilities expressed in a new form.

Neyman (1934) replied that “questions raised in the discussion on the confidence intervals would require too much space. In fact, to clear up the matter entirely, a separate publication is needed…[and] this is in preparation” (p. 623). He alluded to the nature of the response that would follow: “It has been suggested in the discussion that I used the term ‘confidence coefficient’ *instead* of the term ‘fiducial probability’. This is certainly a misunderstanding” (p. 623). Did Neyman differentiate between his proposed bracketed interval and the venerable hypothesis test?

Shlomo Sawilowsky is Professor of Education and Wayne State University Distinguished Faculty Fellow. Email: shlomo@wayne.edu.
The author gratefully acknowledges discussions on the ether with Rabbi Chaim Moshe Bergstein.

No. Neyman (1935) immediately disabused readers of the statistical literature of this notion. He stated, “The problem of estimation in its form of confidence intervals stands entirely within the bound of the theory of probability” (p. 116), as does hypothesis testing. How, then, did the claim that bracketed intervals are superior and preferred eventually arise as a weapon in the arsenal of the camp attempting to make a case against hypothesis testing?

Neyman (1941) reviewed the development of the bracketed interval, which is translated from the Polish “przedzial ufności.” He mentioned this phrase in 1930 in lectures at the University of Warsaw and the Central College (Agriculture) in Warsaw, Poland. Prior to the redaction of the theory, Pytkowksi (1932) published a practical application.

Neyman (1941) recounted that he had noticed numerical similarities obtained with his method and that of the fiducial argument. As a result, he had initially assumed the two paradigms were identical. Neyman was satisfied with considering the bracketed interval as an extension of the fiducial argument because Fisher (1930) had priority.

Eventually, Neyman (1934) became estranged from the fiducial argument. He no longer considered the two theories interchangeable. He left the reasons unstated in his opening presentation before the Society.

Fisher (1934) attended the reading as a discussant. Historical accounts of the exchange were varied. Some expressed chagrin with Fisher, who offered minimal comments on the new methodology, and instead concentrated on the relative merits of random vs purposive sampling selection. Others, in noting Bowley’s (1934) comment that the paper was difficult to understand, assumed that Fisher might have neglected to read Neyman’s paper prior to the reading and simply didn’t follow it. Still others proposed that this was Fisher’s feeble attempt at blocking his baton from being passed to Neyman, just as Karl Pearson had tried in vain two decades prior with Fisher.

These reports misrepresented Fisher’s response. Most of his comments were directed to the sampling problem because that was the primary thesis of Neyman’s (1934) paper. Moreover, a careful review of the published discussion indicates that Fisher understood the paper’s implication quite well. His response was a terse defense of the fiducial argument as the explanation of ordinary probability.

Neyman (1941) was surprised! Fiducial probability and the fiducial distribution of a parameter were “more or less, *lapsus linguae*, difficult to avoid in the early stages of a new theory” (p. 129). The fiducial argument was vague, misconceived, and vacuous in explaining ordinary probability.

The aftermath took the form of considerable and animated debate in the literature on the fiducial argument. Many mathematical statisticians, regardless of theoretical persuasion, joined in the fray by publishing their support or concern. Wald (1939), Wald and Wolfowitz (1939), and Welch (1939) sided with the bracketed interval. Fisher (1935), Starkey (1938), Sukhatme (1938), and Yates (1939) defended the fiducial argument. Pitman (1939) opined that the two theories were essentially the same, as did Bartlett (1939) to a lesser extent.

Bartlett (1936, 1939) also escalated the debate with the contention that where results diverge, the fault lies within the fiducial argument. As can be imagined, Fisher (1937, 1939a, 1939b) and Yates (1939) accepted the gauntlet. Jeffreys (1940) attempted to restore calm in claiming that the bracketed interval and the fiducial argument were both subsumed under inverse probability in the system of Bayes. This had no effect on the debate, of course, because few of the combatants were Bayesian. The controversy would only die with Fisher.

Neyman (1941) succinctly described the relationship between the two theories: “There is none” (p. 130) because “the theories of fiducial argument and of confidence intervals differ in their basic conceptions” (p. 149). He was:

> inclined to think that the literature on the theory of fiducial argument was born out of ideas similar to those underlying the theory of confidence intervals. These ideas, however, seem to have been too vague to crystallize into a mathematical theory. Instead, they resulted in misconceptions of ‘fiducial probability’ and ‘fiducial distribution of a

> parameter'… In this light, the theory of fiducial inference is simply non-existent. (p. 149)

Return to the "confidence" interval attack against hypothesis testing. Fisher's fiducial argument as the explanation of probability was challenged and defeated. However, the ordinary understanding of probability, even in its application to Fisher's F test, was never challenged, much less defeated. Those who have raised the bracketed interval attack against hypothesis testing are merely exploiting Fisher's discredited nomenclature and explanation of probability as he applied it to hypothesis testing.

Ordinary probability is synonymous in the theories of hypothesis testing and bracketed intervals. Certainly, this was Neyman's (1934) view. That is why we concluded, "There is an illogical swagger associated with criticizing hypothesis testing and subsequently advocating CIs [confidence intervals]" (Compton & Sawilowsky, 2003, p. 584).

Philosophical Attack

"Has science benefited from hypothesis testing?" The question is silly. No reputable quantitative physical, behavioral, or social scientist would overlook the breadth and depth of scholarly knowledge and its impact on society that has accrued from over a century of hypothesis testing. The definitive evidence: William Sealy Gosset created the *t* test to make better beer.

In an invited paper in this issue of *Journal of Modern Applied Statistical Methods*, Professor Dayton addresses alternative strategies to hypothesis testing. The motivating reference, Carver (1978), championed the case against hypothesis testing. Carver's (1978) attack was based on a variant of the philosophical attack: speculation and assertion. "Even if properly used in the scientific method, educational research would still be better off without statistical significance testing" (p. 398). Carver (1993) offered an "Einstein" gambit:

> An example from the history of science will help to illustrate this point. Michelson and Morley (1887) collected data relevant to the speed of light, testing the hypothesis that light travels through a medium called *luminiferous ether*. If this ether existed, then light should travel faster when moving in the same direction as the motion of the earth - similar to a boat traveling faster when going downstream compared with upstream. Michelson and Morley interpreted their published data, without tests of significance, as indicating that light traveled the same speed no matter what direction it was traveling. However, I subjected their published data to a simple analysis of variance (ANOVA) and found statistical significance associated with the direction the light was traveling (p. < .01).
>
> It is interesting to speculate how the course of history might have been changed if Michelson and Morley had been trained to use this corrupt form of the scientific method, that is, testing the null hypothesis first. They might have concluded that there was evidence of *significant* differences in the speed of light associated with its direction and that therefore there was evidence for luminiferous *ether*. If this ether existed, then light should travel faster when moving in the same ether. That conclusion would have set back Einstein's ideas many years, because his notions about relativity are based on light traveling in every direction at the same speed. Fortunately, Michelson and Morley did not corrupt the scientific method by testing the null hypothesis before they interpreted their data with respect to their research hypothesis. (p. 288)
>
> The best research articles are those that include *no* tests of statistical significance. In a single study, these tests can be replaced with estimates of effect size and of sampling error, such as standard errors and confidence intervals. Better still, by conducting multiple studies, replication of results can replace statistical significance testing. (p. 289-290)

Responses to Carver's (1993) claims appear below. In order to understand these remarks, it is necessary to preface with a description of interferometer data. Carver (1993) claimed the results were null. Indeed, the 1887 Michelson-Morley experiment is nearly unanimously touted as the most famous experiment that produced a null result. (See, e.g., Feynman, Leighton, & Sands, 1963.)

The interferometer was invented by Michelson to estimate the speed of light. It was refined by Michelson (1881) and by Michelson and Morley (1887a, 1887b) in an attempt to acquire evidence on the medium of propagation of light called ether proposed by Aristotle. The hypothesized value, equal to the Earth's orbital velocity, was approximately 30 km/s.

Michelson and Morley (1887a) did not use hypothesis tests (which had yet to be invented, not withstanding allegations regarding the dating of the sign test). Initially, they presented "the results of the observations… graphically" (p. 333). Visual inspection led to the conclusion there *was* an observed fringe shift, although it was less than what would be expected if the ether existed as hypothesized. They wrote, "It seems fair to conclude from the figure that if there is any displacement due to the relative motion of the earth and the luminiferous *ether*, this cannot be much greater than 0.01 of the distance between the fringes" (Michelson & Morley, 1887a, p. 333).

Next, they presented descriptive statistics. This led to the conclusion that "the ether is probably less than one sixth the earth's orbital velocity, and certainly less than one fourth" (p. 341). Values probably less than 5 km/s and certainly less than 7.5 km/s are not null, although different from the expected value of 30 km/s. Some results on interferometer experiments conducted from 1887 - 1935 are compiled in Table 1.

Table 1. A Sampling Of Interferometry Results.

Experimenter	Date	Velocity (k/s)
Michelson & Morley	1887	5 - ≤ 7.5
Morley & Miller	1902-4	8.7 ± 0.6
Morley & Miller	1905	7.5
Miller	4/1/1925	10.1 ± .33
Miller	8/1/1925	11.2 ± .33
Miller	9/15/1925	9.6 ± .33
Miller	9/23/1925	8.22
Miller	2/8/26	9.3 ± .33
Picard & Stahel	1926	6.9
Picard & Stahel	1927	1.45 ± .007
Illingworth	1927	< 3 - 5
Michelson, Pease, & Pearson	1929	20
Joos	1930	< 1.5
Kennedy & Thornkike	1932	24
Michelson, Pease, & Pearson	1935	20

The only null results via interferometry were obtained by Kennedy in 1926. His results were criticized by Illingsworth (1927), who found the equipment suffered from a "reduced optical system" (p. 692). Múnera (1998) noted that the Kennedy experiment was unclear regarding the local solar time of the initial orientation of the interferometer, which may have been at one of the four times per day that the expected shift tends to zero. Subsequent experiments conducted by Illingsworth (1927) with Kennedy's equipment, but with resilvered mirrors, presented nonnull results.

A variety of technical corrections were introduced to account for the non-null results. Experiments were carefully designed to rule out rival hypotheses, such as temperature, drift, sign of displacement, diurnal variation, and inter-session averaging. Nevertheless, no study produced null results.

Most interferometer experiments were conducted by Miller (1933). He took more than 200,000 readings from 1902 - 1927 based on 12,500 turns of the interferometer, including a joint effort with Morley in the early 1900s. (In comparison, Michelson and Morley made 36 turns in four days, and Piccard and Stahel made 96 turns in Belgium and 60 turns in Brussels.) Yet, Miller *never* obtained a null result.

Shankland (et al., 1955) was Miller's assistant, and subsequently was Professor of physics at Case Western Reserve University (where Morley was Professor of chemistry until 1906). After the death of his boss, he criticized Miller's work on the ether, notably with assistance from Albert Einstein. DeMeo (2000, 2001) strenuously defended Miller against Shankland's criticisms. (The reader interested in the dissident literature on ether should read DeMeo, 2000, 2001; and Múnera, 1998). Later, Shankland (1973, p. 2283) cited a letter received from Einstein dated 31 August 1954:

> I thank you very much for sending me your careful study about the Miller experiments. Those experiments, conducted with so much care, merit, of course, a very careful statistical investigation. This is more so as the existence of a not trivial positive effect would affect very deeply the fundament of theoretical physics as it is presently accepted.
>
> You have shown convincingly that the observed effect... has nothing to do with 'ether-wind', but has to do with differences of temperature.

Einstein's letter is instructive for many reasons. First, he believed the interferometer experiments on the ether "merit, of course, a very careful *statistical analysis*" [emphasis added]. Second, as late as the year of his death, Einstein *still* believed that the interferometer experiments were a threat to his special theory of relativity. Third, he had not updated his knowledge many years after the specter of temperature as a confounding variable was first raised. The *Cleveland Plain Dealer* (27 January 1926) published an exchange between Einstein and Miller, with the latter concluding,

> "The trouble with Prof. Einstein is that he knows nothing about my results," Dr. Miller said. "He has been saying for thirty years that the interferometer experiments in Cleveland showed negative results. We never said they gave negative results, and they did not in fact give negative results. He ought to give me credit for knowing that temperature differences would affect results. He wrote to me in November suggesting this. I am not so simple as to make no allowance for temperature."

In his experiments in 1923, and from 1925 - 1926 at Mt. Wilson, Miller took many steps to control for the effects of temperature. The results were consistent with earlier measurements. Similarly, Miller (cited in Joos & Miller, 1934) noted, "when Morley and Miller designed their interferometer in 1904 they were fully cognizant of this... Elaborate tests have been made... especially with artificial heating, for the development of methods which would be free from this effect [of temperature]" (p. 114). The *Cleveland Plain Dealer* (27 January 1926) added, "Speaking before scientists at the University of Berlin, Einstein said the ether drift experiments [were null in the Michelson-Morley experiment but] on Mount Wilson they showed positive results", although he attributed it to temperature and altitude.

<u>Einstein Gambit Declined</u>

There were thousands of interferomic studies conducted by dozens of physicists since 1887, and in all but one experiment the results were demonstrably non-null. The only known null result was subsequently determined to be caused by a miscalibrated instrument. When the instrument was resilvered, and the experiment replicated in the same location, the results were about 4 km/s.

Carver (1993) conducted a simple analysis of variance (ANOVA) and found statistical significance ($p < .01$). These results are tenable, assuming the null hypothesis was the observations did not differ from zero. Nevertheless, Carver's (1993) analysis suffers from a bewildering array of questions, such as:

- What data set was used? Was it from the noon readings, the afternoon readings, or a combination of readings? Was it from July 8th, 9th, 11th, or 12th of 1887; or perhaps some combination of days? Did it include all 36 turns of the interferometer, or some subset?
- What was the value of F?
- What were the degrees of freedom?
- Were the underlying assumptions of independence, homoscedasticity, and normality considered?

- Were covariates such as diurnal variation or drift considered?
- How was intersession averaging based on different calibration curves handled?
- According to Carver's (1993) advice and recommendation, why did he fail to present summary statistics or a graphic display of the results (either prior to the ANOVA or afterwards)?

Carver (1993) claimed that this significant result from the hypothesis test would have set Einstein back many years. This is unwarranted speculation. In his lecture in Berlin, Einstein rejected the 1887 Michelson-Morley results as being nonnull, despite the evidence contained within their descriptive statistics and graphs. Similarly, he would have ignored the outcome of a hypothesis test.

Einstein's theory was not based on any experimental evidence. At various times throughout his career, Einstein reminisced that it was based on the principles of Maxwell and Lorentz, and he had not relied on the Michelson-Morley experiment. Holton (1969, 1988) suggested that not only did the interferometer experiments have little or no impact, but there is evidence that Einstein was unaware of the Michelson-Morley experiment prior to developing the special theory of relativity.

Interferometer experimenters presented graphical displays, from simple scatter grams and histograms to more complex time series charts and hodograms. All pictorial representations substantiated nonzero results. Some of the latter interferometer experimenters reported standard errors. (Obviously, those who did not were remiss.) Many of the latter experimenters also reported bracketed intervals, and zero was not in them. Múnera (1998) summarized the bulk of interferometer studies with a bracketed interval, and zero was not in it. If statistical tests had been invented by 1887, it would have been easy to confirm the data were statistically significantly different from zero. Even Shankland (et al., 1955; 1973) was forced to admit this.

Carver (1993) reported an effect size (eta squared) of .005. He concluded "if Michelson and Morley had been forced … to do a test of statistical significance, they could have minimized its influence by reporting this effect size measure indicating that less that 1% of the variance in the speed of light was associated with its direction" (p. 289). The fallacy of his analysis is Michelson and Morley's (1887a, 1887b) experiment obtained results of 5 to 7.5 km/s. Regardless of what percent of variance it represents, how can anyone call a speed that exceeds the Earth's satellite orbital velocity "null" and "seek to minimize its influence"?

Of paramount importance, however, *Carver (1993) tested the wrong hypothesis*. Data inspection and graphs demonstrated interferomic data did not support the static model of luminiferous ether as a medium of propagation for light. Should a hypothesis test be desired, the correct test is whether the data were statistically significantly different – not from zero – but rather, from the hypothesized value of 30 k/s.

Carver (1993) described the process of conducting hypothesis tests prior to examining descriptive data as a corruption of the scientific method. This is a straw-person argument. Who promotes conducting hypothesis tests as a first step in the analysis of data? Who objects to examining raw data (e.g., for data entry errors, outliers), computing descriptive statistics, and inspecting graphics prior, or as a follow-up, to conducting hypothesis tests?

Carver (1993) stated the best research articles are those that contain no hypothesis tests. This regressive approach would truly set quantitative physical, behavioral, and social science back more than a century. Reasonable people have different expectations of what constitutes a rare event vs what constitutes a common event expected by chance alone. This is true with a single study, and all the more so with many replications of a study. The debate is diminished, and possibly vanishes, with the simple agreement on a threshold (i.e., nominal alpha level) prior to conducting an experiment.

Carver's (1993) reliance on reporting effect sizes as a panacea is naïve. Effect sizes are sensitive to their own underlying assumptions. In addition, the process of enclosing effect sizes in a bracketed interval relies on the same probabilities as does the obtained value of a hypothesis test. Carver (1993) also recommended the practice of reporting an effect size whether the hypothesis test "is significant or not" (p. 288). This leads to the "trouble with trivials" problem (see e.g., Sawilowsky, 2003b, 2003c).

Currently, it is a popular slogan among effect size enthusiasts to warn against "becoming stupid in another metric." Yet, Carver (1993) interpreted an eta squared of .005 as null to minimize the study outcome. *The experimental results Carver (1993) sought to minimize were speeds of over 16,750 miles per hour!*

The Next Generation of Arguments

As soon as these two lines of attack against hypothesis testing falter, three more assaults are quickly proffered. This is not the place to elaborate on them, but they are parried briefly below.

The first is to replace hypothesis testing with significance testing. P values are reported and it is left to the reader to decide if it is significant. Aside from being outside the realm of the scientific method, subjective significance testing is, in my view, a recipe for disaster (Knapp & Sawilowsky, 2001). (Note that Carver's, 1978, 1993, attack is actually against hypothesis testing, although he calls it a case against significance testing.)

The second is to abandon the frequentist approach and conduct a Bayesian analysis. I strongly promote the method of Bayes in selecting a pinch hitter in baseball because of the plethora of informative priors. However, in the absence of definitive objective priors, a condition that pervades most of physical, behavioral, and social science, Bayesian methods are not likely to be optimal.

The third is to abandon quantitative methodology altogether in favor of qualitative techniques. I discussed this option elsewhere (Sawilowsky, 1999). Qualitative methods should be used when the research hypothesis is qualitative, not because of some perceived limitation of a quantitative method in pursuing a quantitative research question.

References

Bartlett, M. S. (1936). The information available in small samples. *Proceedings of the Cambridge Philosophical Society*, *32*, 560-566.

Bartlett, M. S. (1939). Complete simultaneous fiducial distributions, *Annals of Mathematical Statistics*, *10*, 129-138.

Bowley, A. L. (1934). Discussion on Dr. Neyman's paper. *The Journal of the Royal Statistical Society*, *97*, 607-610.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378-399.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education*, *61*(4), 287-292.

Compton, S., & Sawilowsky, S. (2003). Do not discourage the use of p values. *Annals of Emergency Medicine*, *41*(4), p. 584.

DeMeo, J. (2001). Dayton Miller's ether-drift experiments: A fresh look. *Infinite Energy Magazine*, *38*, 72-82.

DeMeo, J. (2002). Dayton Miller's ether-drift experiments: A fresh look. http://www.orgonelab.org/miller.htm.

Feynman, R. P., Leighton, R. B., & Sands, M. (1963). *The Feynman lectures on physics: Mainly mechanics, radiation, and heat. Vol 1*. Reading, MA: Addison-Wesley, 15-5.

Fisher, R. A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, *26*, 528-535.

Fisher, R. A. (1934). Discussion on Dr. Neyman's paper. *The Journal of the Royal Statistical Society*, *97*, 614-619.

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics*, *6*, 391-398.

Fisher, R. A. (1937). On a point raised by M. S. Bartlett on fiducial probability. *Annals of Eugenics*, *7*, 370-375.

Fisher, R. A. (1939a). The comparison of samples with possibly unequal variances. *Annal of Eugenics*, *9*, 174-180.

Fisher, R. A. (1939b). A note on fiducial inference. *Annals of Mathematical Statistics*, *10*, 383-388.

Jeffries, H. (1940). Note on the Behrens-Fisher formula. *Annals of Eugenics*, *10*, 48-51.

Joos, G., & Miller, D. (1934, January 15). Letters to the Editor. *Physical Review*, *45*, 114.

Holton, G. (1969). Einstein, Michelson, and the 'crucial' experiment. *Isis*, *60* (1969).

Holton, G. (1988), *Thematic origins of scientific thought, Kepler to Einstein*. (Revised ed.). Cambridge, MA: Harvard University Press.

Illingworth, K. K. (1927) A repetition of the M-M experiment using Kennedy's refinement. *Physics Review*, *30*, 692-696.

Knapp, T., & Sawilowsky, S. (2001). Constructive criticisms of methodological and editorial practices. *Journal of Experimental Education*, *70*, 65-79.

Michelson, A. A. (1881). The relative motion of the earth of the luminiferous aether. *American Journal of Science*, *22*(S3), 120-129.

Michelson, A. A., & Morley, E. W. (1887a). On the relative motion of the Earth and the luminiferous ether. *American Journal of Science*, *34*(S3), 333-345.

Michelson, A. A., & Morley, E. W. (1887b). On the relative motion of the earth and the luminiferous aether. *Philosophical Magazine*, *24*(151) S5, 449-463.

Miller, D. (1933, July). The ether-drift experiment and the determination of absolute motion of the Earth. *Review of Modern Physics*, *5*(2), 203-242.

Múnera, H. A. (1998). Michelson-Morley experiments revisited: Systematic errors, consistency among different experiments, and compatibility with absolute space. *Apeiron*, *5*, 37-54.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive sampling. *The Journal of the Royal Statistical Society*, *97*, 558-625.

Neyman, J. (1935). On the problem of confidence intervals. *Annals of Mathematical Statistics*, *6*, 111-116.

Neyman, J. (1941). Fiducial argument and the theory of confidence intervals. *Biometrika*, *32*, 128-150.

Pitman, E. J. G. (1939). The estimation of the location and scale parameters of a continuous population of any given form. *Biometrika*, *30*, 391-421.

Pytkowski, W. (1932). *The dependence of the income in small farms upon their area, the outlay and the capital invested in cows*. Warsaw: Bibljoteka Pulawska.

Sawilowsky, S. (1999). Quasi-experimental design: The legacy of Campbell and Stanley. In (Bruno D. Zumbo, Ed.) *Social indicators/quality of life research methods: Methodological developments and issues, Yearbook 1999*. Norwell, MA: Kluwer.

Sawilowsky, S. (2003a). A different future for social and behavioral science research. *Journal of Modern Applied Statistical Methods*, *2*(1), 128-132.

Sawilowsky, S. (2003b). You think you've got trivials? *Journal of Modern Applied Statistical Methods*, *2*(1), 218-225.

Sawilowsky, S. (2003c). Trivials: The birth, sale, and final production of meta-analysis. *Journal of Modern Applied Statistical Methods*, *2*(1), 242-246.

Shankland, R., McCuskey, S. W., Leone, F.C., & Kuerti, G. (1955). New analysis of the interferometer observations of Dayton C. Miller. *Review of Modern Physics*, *27*(2), 167-178.

Shankland, R. (1973). Michelson's role in the development of relativity. *Applied Optics*, *12*(10), 2280-2287.

Starkey, D. M. (1938). A test of the significance of the difference between means of samples from two normal populations without assuming equal variances. *Annals of Mathematical Statistics*, *9*, 201-213.

Sukhatme, P. V. (1938). On Fisher and Behrens' test of significance of the difference in means of two normal samples. *Sankhyā*, *4*, 39-48.

Wald, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses, *Annals of Mathematical Statistics*, *10*, 299-326.

Wald, A., & Wolfowitz, J. (1939). Confidence limits for continuous distribution functions, *Annals of Mathematical Statistics*, *10*, 105-118.

Welch, B. L. (1939). On confidence limits and sufficiency, with particular reference to parameters of location, *Annals of Mathematical Statistics*, *10*, 58-69.

Yates, F. (1939). An apparent inconsistency arising from tesets of significance based on fiducial distributions of unknown parameters. *Proceedings of the Cambridge Philosophical Society*, *35*, 579-591.