A Historical View of Statistical Concepts in Psychology and Educational Research

Author(s): Stephen M. Stigler

Source: *American Journal of Education*, Nov., 1992, Vol. 101, No. 1 (Nov., 1992), pp. 60-70

Published by: The University of Chicago Press

Stable URL: http://www.jstor.com/stable/1085417

A Historical View of Statistical Concepts in Psychology and Educational Research

STEPHEN M. STIGLER
University of Chicago

This essay examines the historical roots of the link between statistics and psychology, exploring the manner in which the conceptual framework for the application of probability in psychological research was created by experimental design.

Statistics and psychology have long enjoyed an unusually close relationship—indeed, more than just close, for they are inextricably bound together. That tie is of an unusual nature, with historical roots in the nineteenth century, and an understanding of this peculiar historical relationship can lead to a deeper understanding of contemporary applications. I propose to examine the essence of this relationship in terms of a curious historical problem.

As is well known, statistical methods—and by "statistical methods" I mean probability-based modeling and inference—entered into psychology in 1860 in the work of Gustav Fechner (1860) in psychophysics (Boring 1961; Stigler 1986). There were some antecedents to Fechner, it is true, but for the most part the date stands: The year 1860 marks the beginning of the use of modern statistics in psychology. By the 1880s and the work of Ebbinghaus (1885) on memory, the success and acceptance of this approach was assured. Now the problem I pose is this: Those dates are at least 20 years before, and more like 30 or 40 years before, a comparable stage of development in economics or sociology (Lazarsfeld 1961; Stigler 1986). So I ask, Why this time lag? Why did psychology precede these older fields in their recognition of this new technology?

A tempting answer, and I would guess an attractive one for some psychologists, is that it is simply a matter of fact that psychologists are

American Journal of Education 101 (November 1992)

0195-6744/93/0101-0003$01.00

smarter than economists. But aside from the fact that such explanations will not do in the history of science, there is the fact that it would lead to unacceptable corollaries. For example, astronomers were using statistical methods nearly a century before psychologists—does that mean that astronomers are that much smarter than psychologists? And statistics appears in psychophysics well before it appears in education—are we to therefore put physiological psychologists above educational researchers? Of course not.

I do have an answer to this question, and I will reveal it soon, but in order to make it plausible I have to take a step back and ask, Well, what was it about astronomy other than the astronomers' native intelligence that led them to statistical methods so soon?

Statistical methods for the treatment of astronomical observations evolved over the latter half of the eighteenth century, and they were reconciled with the mathematical theory of probability early in the nineteenth century. One key figure in all of this was Pierre Simon Laplace (born 1749, died 1827); another was Carl Friedrich Gauss (born 1777, died 1855). The setting for this work was one of a refined Newtonian theory for the motions of the planets, and the types of problems for which astronomers used statistics involved the determination of the "constants" of that theory. For example, supposing that Jupiter traveled about the sun in an ellipse—what were the coefficients of the equation of that ellipse? And what would they be if we allowed for perturbations due to the gravitational effect of Saturn? And what if, to help map the stars as a background reference for the motions of planets and comets, five astronomers observed the same star with different results? How to reconcile their answers?

The key point in all of this, the anchor for the whole project, was Newtonian theory. When an astronomer resorted to statistics in the 1820s, and the tool he usually reached for was the method of least squares, there was no doubt in his mind that he was after something real, definite, objective, something with an independent reality outside of his observations, a genuinely Platonic reality inherited from the

---

STEPHEN M. STIGLER is professor in the Department of Statistics and the Committee for the Conceptual Foundations of Science at the University of Chicago. A former editor of the *Journal of the American Statistical Association,* he has written widely on the history of statistics, in addition to articles on mathematical statistics and experimental design.

then-unshakable edifice of Newtonian theory. The whole of the nineteenth century theory of errors was keyed to this point:

$$\text{observation} = \text{truth} + \text{error}.$$

Without an objective truth, this sort of a split would be impossible, for where would error end and truth begin? I might be tempted to refer to such a split as deconstruction, but as I understand the term as now used, a modern deconstructionist would be more likely to identify observation with truth and to deny the possibility of discerning error.

With an objectively defined goal, namely astronomical or Newtonian "truth," the road was free for probability-based inferences. Probability distributions for errors could be assumed or estimated, observations could be combined by maximum likelihood or least squares, and there was no ambiguity about the nature of the result. I would contrast this with the situation the economists found themselves in a century later. Suppose we wished (as William Stanley Jevons did in the 1860s) to determine the effect of the gold discoveries in California and Australia on world price levels. You could gather "before" and "after" price data on a number of different commodities important in world trade. But how can you, to use an argument put forth at that time, average pepper and pig iron? How can totally different goods subject to importantly different economic forces be combined into an index that may be considered an estimate of something as real as the position of a star or the shape of the orbit of Jupiter? Jevons *did* average such quantities, but only with copious apologies, and he did not use probability-based methods to measure the uncertainty in the result. How could he? How would he have defined the object about which he was uncertain? Perhaps Adam Smith was the Newton of economics, but there was no inverse-square law for price movements.

Educational researchers should feel a particular sympathy for Jevons's problem—for any time groups of people, or examination scores, or, in this day of meta-analysis, a collection of studies of educational interventions, any time a group of different measures are to be combined in a common analysis the question must be, What is the goal? What is it that I am estimating? There is an answer to this question, but it is not to be found in the stars, and it is not the astronomers' answer. But I am getting ahead of the story. How, I asked, could psychologists bring themselves in the 1860s to use statistical methods, when it took economists another 30–40 years? What was it about the problems the psychophysicists faced that was like the problems the astronomers faced, problems involving the measurement of sensitivity, reaction times, memory? The short answer is, nothing at all; the problems were

not intrinsically similar at all. Even the one famous "law" of early psychophysics, Fechner's law relating sensation to the logarithm of intensity, was not a Newtonian law derived from theory, but an empirical construct that fit only middlingly well, and it is one area of psychophysics where Fechner did not use probability. So what is the answer?

To understand what did happen, and how the psychophysicists managed to create a surrogate for the law of gravitation, let me look at one of the most careful and impressive investigations in nineteenth-century psychology, one that was performed in the 1880s by the American philosopher Charles S. Peirce, while on the faculty of Johns Hopkins, shortly before the president of that institution dismissed him, apparently in part because he disapproved of Peirce's handling of his marriages. I would list Peirce as one of the two greatest American scientific minds of that century (the other being physicist J. Willard Gibbs), but Peirce was a strange person, an outlier in any educational theory. He was educated principally at home, by his father, a professor of mathematics at Harvard. The younger Peirce went to Harvard—probably the tuition was lower then than now—but he graduated without distinction, seventy-ninth in a class of 91. He is said to have had a curious method of self-examination—he was ambidextrous, and he would write out questions with his left hand while simultaneously answering them with his right. He is best known as a philosopher, the father of pragmatism, a name later kidnapped by William James, but he also was a physicist, a cartographer, a mathematician, and a psychologist.

The experiments I describe took place in December 1883 and January 1884 and involved a version of the experiment that Fechner had pioneered, the application of the method of right and wrong cases to the sensation of lifted weights (Peirce and Jastrow 1885; Stigler 1978). Indeed, I could use Fechner's own work to make my point, but I have already discussed it extensively in my 1986 book, and Peirce's use affords a cleaner and more dramatic example of the idea.

For those who do not know it, I will briefly describe the method of right and wrong cases, as used here. An experimental subject is confronted with two apparently identical boxes; they differ only in that one box contains a single weight D, the other is heavier—it contains a weight equal to D and a small weight P. The subject lifts one box, then the other, and pronounces a judgment on which is heavier. The numbers of right and wrong cases are tabulated, for various D and P and with other conditions being varied (left hand vs. right hand, heavy first vs. light first, morning vs. evening, and so forth). That, basically, was Fechner's experiment.

Peirce went a step further, in an attempt to challenge one of Fechner's ideas. Fechner had proposed that, despite the fact that sensation in-

creased as stimulation increased, there was a threshold below which there was no sensation at all. He called that threshold the "just-noticeable difference." In the context of the lifted weight experiment, Fechner would maintain that, as the incremental weight P increased, the fraction of "right-cases" would increase, but there was, for each base weight D and each person, a threshold value for P—a small weight such that if P is less than that value there is no sensation, there is an even chance for a right-case. Weights P that were below the just-noticeable difference were indistinguishable from zero. Peirce did not buy this.

Now so far I have been beating around the bush, evading the question: What was it about the treatment of problems like this that opened them up to statistical treatment in the manner of the astronomers and that differed from the problems of the economists? The answer is simple: experimental design. First, there was the *possibility* of experimental design—the ability of the scientist to manipulate the conditions, in order to sharpen the hypotheses and render limited questions capable of sharp and definitive answers. This alone distanced the experimental psychologists from the economists. And second, there was the *cleverness* of the psychologists in exploiting this advantage to provide a novel surrogate for the anchor of Newtonian law. Let me return to Peirce's work, where this is clearest.

Peirce wanted to measure extremely subtle sensations, the perception of very small incremental weights. And he had a wonderful idea: a blind randomized experiment. In order to eliminate the biases attendant on factors such as which weight was lifted first, or how the weights were arranged, or whether the subject knew which was which, Peirce worked with an assistant, Joseph Jastrow, who later had a distinguished career himself in psychology. (Fechner had experimented alone, with himself as both subject and assistant.) And Peirce employed an explicit device for randomizing the order of presentation, the order of placement. He prepared a special deck of cards for this purpose, and either Peirce or Jastrow would shuffle and select a card, and prepare the weights, while the other, blind to these preparations, would be the experimental subject.

Now, the introduction of randomized experiments is usually associated with Ronald A. Fisher, in work on agricultural experimentation a half-century later, but there is no question that Peirce was clear on what he was doing and why, and his "what and why" were the same as Fisher's. As Peirce wrote, "By means of these trifling devices the important object of rapidity was secured, and any possible psychological guessing of what change the operator was likely to select was avoided. A slight disadvantage in this mode of proceeding arises from the long runs of one particular kind of change, which would occasionally be

produced by chance and would tend to confuse the mind of the subject. But it seems clear that this disadvantage was less than that which would have been occasioned by his knowing that there would be no such long runs if any means had been taken to prevent them" (Peirce and Jastrow 1885, p. 80). Writing elsewhere, in a more philosophical vein, Peirce said that the very possibility of induction depended on such randomization: "The truth is that induction is reasoning from a sample taken at random to the whole lot sampled. A sample is a *random* one, provided it is drawn by such machinery, artificial or physiological, that in the long run any one individual of the whole lot would get taken as often as any other" (Peirce 1957, p. 217).

The point is that Peirce used randomization to create an artificial baseline that was as well understood and as well-defined as any of the Platonic constants of Newtonian physics. If Fechner was correct and the just-noticeable difference existed, then Peirce's scheme would guarantee that the probability of a right-case for such a small weight would be one-half. This was the result of artifice, but it was as real as the position of any star, and it served as the basis for Peirce's subsequent probability calculations of the significance of effects, of differences.

I hasten to add that the same point holds regarding Fechner's own experiments, although not so crisply. Fechner did not randomize, but he was interested in larger differences, and from his perspective the experiments were just as good as Peirce's and, if allowance is made for the fact that Fechner was not attuned to the same level or type of subtle distinction as Peirce, would serve the same purpose as Peirce's. Fechner's control of experimental conditions, like that of Müller, Wundt, and Ebbinghaus, created an artificial baseline and a framework that made statistical investigation possible. Psychology has never been the same since.

What did Peirce find? As you might expect, he detected in his subjects a sensitivity to sensation that went far below Fechner's threshold (see table 1). The sensitivity was slight but, thanks to the design of his experiment, unmistakable. Peirce's investigation, which was one of the best examples of carefully developed and explained experimental psychology ever prepared, included other excellencies. For example, he had his subjects write down in each case an estimate of their confidence in their guess. He found that these estimates varied directly with the log odds that they actually were correct, a remarkable early appearance of the log odds as an experimentally determined measure of certainty. Summarizing the message Peirce read from his finding, that we were actually sensitive to sensations so minute that we were not consciously aware of them, Peirce wrote (with a slight trace of nineteenth-century sexism): "The general fact has highly important practical bearings,

TABLE 1

Peirce's Test of Fechner's Just-noticeable Difference

R	Errors in 50 Trials
1.1	2
1.08	4
1.06	8, 11, 7, 14, 15, 12, 6
1.05	13
1.04	15
1.03	20, 16, 20, 29, 16, 15
1.015	21, 28, 28, 20, 22

NOTE.—Values are a portion of Peirce's 1883–84 data, from Peirce and Jastrow (1885, p. 80). The values given are the counts of errors made by Peirce in 22 experiments of 50 trials each, at seven different values of $R = (D + P)/D$. The base weight D for these experiments was $D = 1$ kg. A logistic regression analysis of these data indicates a highly significant relationship between R and the log odds of a right answer (log likelihood ratio statistic of 93.7, on 1 df).

since it gives new reason for believing that we gather what is passing in one another's minds in large measure from sensations so faint that we are not fairly aware of having them, and can give no account of how we reach our conclusions about such matters. The insight of females as well as certain 'telepathic' phenomena may be explained in this way. Such faint sensations ought to be fully studied by the psychologist and assiduously cultivated by every man" (Peirce and Jastrow 1885, p. 83).

If astronomers found their object in the application of Newton's laws to the universe, and if experimental psychologists constructed theirs through the design and control of experimental conditions, what of economists, and more to the present point, of educational researchers? Here the definition of an object for inference was harder and evolved over a longer time.

Let me highlight the problem by quoting a distinction between "observations" and "statistics" that was made by Francis Edgeworth in 1885. Edgeworth, who was one of the foremost economic theorists of the time, is also a possible claimant to the title "the Father of Educational Statistics." He had in mind observations as the material of astronomers

and statistics as the material of economists—nonexperimental data from what we might now term observational studies. He wrote:

> Observations and statistics agree in being quantities grouped about a Mean; they differ, in that the Mean of observations is real, of statistics is fictitious. The mean of observations is a cause, as it were the source from which diverging errors emanate. The mean of statistics is a description, a representative quantity put for a whole group, the best representative of the group, that quantity which, if we must in practice put one quantity for many, minimizes the error unavoidably attending such practice. Thus measurements by the reduction of which we ascertain a real time, number, distance are observations. Returns of prices, exports and imports, legitimate and illegitimate marriages or births and so forth, the averages of which constitute the premises of practical reasoning, are statistics. In short observations are different copies of one original; statistics are different originals affording one "generic portrait." Different measurements of the same man are observations; but measurements of different men, grouped round l'homme moyen, are primâ facie at least statistics. [Edgeworth 1885; quoted in Sills and Merton 1991, p. 56–57]

The route that was followed to bring such statistics under the sway of what we now call statistical analysis began much earlier, with an argument by analogy. Adolphe Quetelet had looked at the distribution of human attributes—anthropometric characteristics such as stature—and seen a distribution like that of astronomers' errors: what we now call a normal distribution. For astronomers the distribution was anchored at an objective truth and represented the distribution of errors about that truth. For Quetelet, this was reversed—the curve, once noted, could be used to define the objective truth. The center of the curve would serve the same purpose as the astronomers' goal; for Quetelet it would mark the stature of *l'homme moyen.*

The acceptance and extension of this idea to more useful areas was slow. Quetelet himself wanted to apply it to moral qualities, but he lacked data. Francis Galton took up the challenge and opened new fields when he actually applied Quetelet's ideas to examination scores in 1869, using as his data scores from the admissions test for the Royal Military College at Sandhurst (Galton 1869). Galton found that what was true for height was true for these scores, and he went on to develop this idea into a framework for studying inheritance, including the inheritance of ability as reflected through examination scores.

The extension of this framework, from the simple analogies of Quetelet to the full force of modern multivariate analysis, was not

accomplished in a single step. Through the combined efforts of Galton, Edgeworth, and Karl Pearson, Quetelet's analogy becomes a technical apparatus of great power. I have, in my 1986 book, treated at some length how the simple idea that a fitted normal curve could define a population center, and thus an object for analysis, evolved. One important proponent was Ebbinghaus, who made this the key to his work on memory. The distribution would validate the average. Ebbinghaus wrote, in 1885, "I examine the distribution of the separate numbers represented in an average value. If it corresponds to the distribution found everywhere in natural science, where repeated observation of the same occurrence furnishes different separate values, I suppose—tentatively again—that the repeatedly examined psychical process in question occurred each time under conditions sufficiently similar for our purposes" (Ebbinghaus 1885, pp. 19–20). In Galton's hands this device led to the invention of correlation and eventually to modern regression analysis, in which a multivariate distribution defined a relationship among its variates through its conditional distributions. The spread of these ideas into areas such as educational research was initially slow, however, despite the fact that one of the very first to appreciate the power of the ideas was John Dewey himself.

Dewey reviewed Galton's book *Natural Inheritance* for the *Journal of the American Statistical Association* in September 1889, and he proved himself one of the most perceptive of the early critics—even quicker to appreciate the work than Karl Pearson! Dewey wrote: "It is to be hoped that statisticians working in other fields, as the industrial and monetary, will acquaint themselves with Galton's development of new methods, and see how far they can be applied in their own fields" (Dewey 1889, pp. 333–34). But I am not aware of Dewey's having pursued his own suggestion in education.

I would, however, like to call your attention to one of the earliest detailed expositions of the application of statistical methods to education, an 1888 paper by Edgeworth in the *Journal of the Royal Statistical Society* (Edgeworth 1888). Edgeworth did not treat the use of regression, but he did provide a tutorial on the use of statistics for the analysis of examination scores. He discussed the scaling of exams—how the normal distribution could be used as a scaling device, the virtues of making corrections in the mean for different examiners' propensities, whether or not it was useful to analyze results on a logarithmic scale (or to combine results by a geometric mean), and how to estimate variability (including the introduction of variance components models into this area). He illustrated these ideas with empirical work of his own; for example, he looked into the relationship between the speed with which examinations are graded and the grades. He himself graded a set of

examinations in the English language quite rapidly, set them aside for a while, then returned and studied them in great detail. He found the difference in marks small, contributing only 1 or 2 percent to the probable error of the aggregate of marks. He also looked into the loss of accuracy in having exams graded by teaching assistants, and the loss in grading only a randomly selected half of the questions put. In dealing with these questions he compared the loss in accuracy (as measured by an increase in variance) with the between-examiner component of variation.

I have tried to make the point that there was a fundamental difference between the application of statistical methods in astronomy, in experimental psychology, and in the social sciences, and that this difference had a profound effect on the spread of the methods and the pace of their adoption. Astronomy could exploit a theory exterior to the observations, a theory that defined an object for their inference. Truth was—or so they thought—well differentiated from error.

Experimental psychologists could, through experimental design, create a baseline for measurement and control the factors important for their investigation. For them the object of their inference—usually the difference between a treatment and a control group, or between two treatments—was created in the design of the experiment.

Social scientists, without experimental control over their material, had to go further. For them the statistical model itself defined the object of inference, often a set of conditional expectations given a set of covariates. The role of statistics in social science is thus fundamentally different from its role in much of physical science, in that it creates and defines the objects of study much more directly. Those objects are no less real than those of physical science. They are even often much better understood. But despite the unity of statistics—the same methods are useful in all areas—there are fundamental differences, and these have played a role in the historical development of all these fields.

Note

This article was an invited address to the annual meeting of the American Educational Research Association, Chicago, April 4, 1991. The research was supported by National Science Foundation grant DMS 89-02667.

References

Boring, Edwin G. "The Beginning and Growth of Measurement in Psychology." In *Quantification,* edited by H. Woolf. Indianapolis: Bobbs-Merrill, 1961.

Dewey, John. "Galton's Statistical Methods." *Publications of the American Statistical Association* 7 (1889): 331–34.

Ebbinghaus, Hermann. *Über das Gedächtnis.* 1885. Translated by Henry A. Ruger and Clara E. Bussenius as *Memory: A Contribution to Experimental Psychology.* New York: Teachers College, Columbia University, 1913. Reprint. New York: Dover, 1964.

Edgeworth, Francis Ysidro. "Observations and Statistics: An Essay on the Theory of Errors of Observation and the First Principles of Statistics." *Transactions of the Cambridge Philosophical Society* 14 (1885): 138–69.

Edgeworth, Francis Ysidro. "The Statistics of Examinations." *Journal of the Royal Statistical Society* 51 (1888): 346–68.

Edgeworth, Francis Ysidro. "The Element of Chance in Competitive Examinations." *Journal of the Royal Statistical Society* 53 (1890): 460–75, 644–63.

Fechner, Gustav Theodor. *Elemente der Psychophysik.* 2 vols. 1860. Vol. 1 translated by Helmut E. Adler as *Elements of Psychophysics,* edited by Davis H. Howes and Edwin G. Boring. New York: Holt, Rinehart & Winston, 1966.

Galton, Francis. *Hereditary Genius: An Inquiry into Its Laws and Consequences.* London: Macmillan, 1869.

Lazarsfeld, Paul F. "Notes on the History of Quantification in Sociology—Trends, Sources, and Problems." In *Quantification,* edited by H. Woolf. Indianapolis: Bobbs-Merrill, 1961.

Peirce, Charles S. *Essays in the Philosophy of Science.* Edited by V. Tomas. Indianapolis: Bobbs-Merrill, 1957.

Peirce, Charles S., and Joseph Jastrow. "On Small Differences of Sensation." *Memoirs of the National Academy of Sciences for 1884* 3 (1885): 75–83. Reprinted in *American Contributions to Mathematical Statistics in the Nineteenth Century,* vol. 2, edited by Stephen M. Stigler. New York: Arno, 1980.

Sills, David L., and Robert K. Merton, eds. *Social Science Quotations.* Vol. 19 of *International Encyclopedia of the Social Sciences.* New York: Macmillan, 1991.

Stigler, Stephen M. "Mathematical Statistics in the Early States." *Annals of Statistics* 6 (1978): 239–65. Reprinted in *American Contributions to Mathematical Statistics in the Nineteenth Century,* vol. 1, edited by Stephen M. Stigler. New York: Arno, 1980.

Stigler, Stephen M. *The History of Statistics: The Measurement of Uncertainty before 1900.* Cambridge, Mass.: Harvard University Press, 1986.