THE PETER EFFECT IN EARLY EXPERIMENTAL EDUCATION RESEARCH

JOSEPH LITTLE
University of California, Santa Barbara

ABSTRACT
One of the signatures of scientific writing is its ability to present the claims of science as if they were “untouched by human hands.” In the early years of experimental education, researchers achieved this by adopting a citational practice that led to the sedimentation of their cardinal method, the analysis of variance, and their standard for statistical significance, 0.05. This essentially diverts their statistical framework from its historical conditions of production. Researchers suppressed their own agency through the use of passive voice and nominalization. With their own agency out of the way, they imbued the methods, results, and presentational devices themselves with the active agency of the situation through the use of personification. Such a depiction creates the impression that the researchers and audience stand on equal epistemic ground as interested witnesses to the autonomous activity of a third party, the method, which churns out the brute facts of science.

Well known within the Roman Catholic Church is the “commissioning of Peter,” the historic moment in which Jesus of Nazareth announced himself as the Christ, the “anointed one,” and granted the authority of the papacy to Peter [1, pp. 461-475]:

He said to them: “But you yourselves, who do you say I am?” Simon Peter answered and said: “You are the Christ, the Son of the living God!” Jesus responded and said to him: “You are blessed, Simon bar-Jonah, because flesh and blood are not the source of this revelation but my Father who is in heaven. And I say to you that you are Peter, and upon this rock I will build my church and the gates of Hades will not overpower it.” I will give you the keys of the

© 2003, Baywood Publishing Co., Inc.
kingdom of heaven and whatever you bind on earth shall have been bound in heaven; and whatever you set loose upon the earth shall have been set loose in heaven (Matthew 16:15-19).

Although verse 18, with its anachronistic reference to the “church,” stands as one of the most controversial passages in biblical textual criticism [2, 3], verse 17 offers a valuable heuristic to those of us interested in rhetorical criticism of science. For in Jesus’s response to Peter, we find an overt illustration of one of the signature features of scientific texts: the ability to present the claims of science as if they were “untouched by human hands” [4, p. 10].

In the commissioning of Peter, divine revelation stood in stark contrast to the “flesh and blood” of human agency as the means by which Peter came to know the truth. In our modern public sphere, however, the priest has been eclipsed by the scientist as the sanctioned intercessor of truth [5]. Accordingly, divine revelation has also been eclipsed, not by theory or ingenuity or the scientist per se but by the methods of science themselves. From this folk perspective, it is by way of rigorous allegiance to method that we feel freed from the specter of political ideology [contra 6; contra 7] and human bias [contra 8] and come to know truths. It is not the scientist qua human, then, but the scientist qua executor of method whom we privilege, for it is precisely in this latter capacity that the scientist seems to keep us in contact with something beyond ourselves. Put more directly, the modern culture of experimental science creates its climate of trustworthiness in part by suppressing human agency and appealing instead to a mechanical form of objectivity achieved through a steadfast commitment to method [5, 9]. This is not objectivity in its epistemologically strong or absolute sense, as in coming to know a thing “as it really is.” Rather, mechanical objectivity privileges the “impersonality of procedure” over the abilities of individual experts, thereby aiming for a scientific account independent of individual examiners [4, p. 10]. It guarantees no truth-oriented heuristic. However, through its rule-bound methodology it does promise subscribers a strong defense against the prospect of personal attack or, perhaps more relevant in our modern culture, the peril of legal liability [10].

In the early years of experimental education research, this commitment to an “objective” method materialized in the widespread adoption of R. A. Fisher’s analysis of variance (ANOVA) as the experimental method of choice [11, 12], combined with the acceptance of Fisher’s 0.05 as the professional standard for statistical significance [13, 14]. Eager to distinguish their work from the lore and missionary zeal of practitioners, educational researchers set out to inaugurate the scientific study of education through “rigorous” experimentation and “precise” statistical measurement, catalyzed in part by G. Stanley Hall’s focus on the scientific study of the child and Edward L. Thorndike’s clarion call for scientific method and measurement in education [15, 16, pp. 101-117]. Descriptive and correlational studies ensued throughout the first quarter of the 20th century. With
the publication of Fisher’s *Statistical Methods for Research Workers* [17], however, educational researchers received the “workhorse statistic” and the “sacred” level of significance that would enable them to produce reliable, respectable, and ostensibly mechanical statistical inferences from the relatively small data sets offered by educational settings [11, 13, 18, pp. 239-248]. The issue was one of legitimacy, and one of which Fisher was well aware: “Statistical methods are essential to social studies,” he argues in his 1925 text, “and it is principally by the aid of such methods that these studies may be raised to the rank of sciences” [17, p. 2]. For the claims of education to be elevated to the status of scientific knowledge, their genesis and development would have to lie beyond the realm of human judgment [4, p. 10; 9, pp. 209-213]: they would have to be delivered by the objective methods of science themselves.

Treading against this current is the realization that scientific inquiry is a human activity: underlying every experiment is an experimenter, or group of experimenters, exercising judgment and making decisions in response to a complex array of local circumstances [19, pp. 187-234; 20]. Research questions must be formulated, experiments constructed, grants secured, technicians trained, constructs measured, data discerned, results construed, and implications drawn: The entire act of making meaning must be accomplished by human effort. Talk of “rigor” and “method” ostensibly rescues the essential aspects of scientific inquiry from the fallible hands of humans, until we realize that the methods themselves are of human design, developed in response to the local circumstances of an earlier time. In the invention of the “treatment group,” for example, we find school inspector W. H. Winch fulfilling the needs of American school superintendents clamoring for a rational way to assess the effects of administrative interventions; likewise, in the development of the division between “experimenter” and “subject,” we find physician and psychologist Wilhelm Wundt responding to the pressures of a potentially troublesome organization of labor emerging at his Leipzig laboratory [16, pp. 30, 113]. Even in the development of the “psychological experiment,” we see opportunistic German physiologists responding to the disciplinary shift from organ structure to bodily function, which displaced dissection as the investigative method of choice [16, p. 25]. Ultimately, to search for the bedrock of experimental science is to begin a sort of infinite regression into the history of human experience.

By what literate means, then, did educational researchers achieve the “Peter effect” in early experimental education research? In other words, by what means did they suppress human agency and foreground their commitment to method in the accounts of experimental research published during the early years of their profession? In this article, I respond to this question by, first, tracing the ways in which educational researchers referred to Fisher’s analysis of variance (ANOVA) between 1932 and 1944 in the *Journal of Experimental Education (JXE)*, one of the few early-20th century education periodicals in which experimental studies were published. Through this “citation context analysis,” [20, 21], I show how the
changes in citational practices served to separate the ANOVA from its affiliation with Fisher, essentially effacing the memory of its human conditions of production.

This process, however, cannot be seen in the common light of “codification” [22, pp. 507-508]. For unlike the TRF(H) of Latour and Woolgar’s study [23] or that of the opiate receptor from Cozzens’s work [21], neither the ANOVA nor 0.05 carries any claim to truth. The former is a method of practice: No empirical particulars can be predicted from its theoretical structure. The latter is a convention: No amount of empirical investigation can bolster or erode its status. As such, missing from the histories of these two constructs are the signs of interplay between theoretical formulation and empirical investigation that sponsor the patterns of consolidation and incorporation to which Zuckerman and Merton refer. Instead, the results of this strand of my analysis are best viewed from the perspective of “sedimentation,” described by Berger and Luckmann [24, pp. 67-72] as a process by which highly contextualized experiences congeal into stable meanings, which are gradually stripped of context and human affiliation and transmitted abstractly through a shared sign system, such as written language, to members of an ambient social network. Once shared, these meanings are available conceptually and anonymously to the public and thus become part of the common stock of knowledge. Accordingly, my tracing of the citational practices of educational researchers constitutes a tracing of the sedimentation of the ANOVA in early experimental education research. It is a process that differs from codification in terms of its relation to theoretical and empirical activity, yet it shares with codification a common end: the “black boxing” of the symbolic form in question [23, 25].

Second, I examine the ways in which educational researchers distribute agency within the 12 ANOVA-based experimental articles published in JXE between 1932 and 1944. Through this textual analysis, I show how an overwhelming use of passive voice, combined with an occasional use of nominalization, serve as the primary means by which educational researchers suppressed their own agency within the context of experimentation. I also show how a persistent use of personification serves the important function of imbuing the methods, results, and presentation devices themselves with the active agency of the occasion. Such a depiction creates the impression that the researchers and audience stand on equal epistemic ground as interested witnesses to the autonomous activity of a third party, the method, which churns out the “brute facts” of science; these brute facts are, in turn, compiled and presented by the tables, figures, and diagrams themselves, thoroughly displacing human agency from the scene of knowledge production.

Underlying the mathematical architecture of Fisher’s ANOVA, however, is a cardinal decision point in the form of statistical significance, which has historically been operationalized within education research as 0.05 in deference to Fisher’s classic text [17]. Therefore, third, I trace the sedimentation [24] of
Fisher’s 0.05 in early experimental education research and show how changes in citational practices gradually served to separate 0.05 from its affiliation with Fisher, leading ultimately to its “black boxing” [25]. In short, as the presence of 0.05 increased in the research literature, the need for intertextual support and interpretive assistance atrophied, leaving a stable, familiar, and abstract symbol, freed from its historical conditions of production and benefiting from the ostensible objectivity that comes with quantification, to be confidently redeployed, further reified, and ultimately accepted as part of the common stock of knowledge in experimental education research.

Last, I should mention that this article represents a preliminary analysis, one that does little or nothing to protect against idiosyncratic interpretation. Future studies along this line would greatly benefit from measures of inter-rater reliability as well as from a refined set of textual categories, as we try to make progress in the messy business of studying written language.

THE SEDIMENTATION OF FISHER’S ANALYSIS OF VARIANCE

To trace the sedimentation of the ANOVA in experimental education research, I collected all ANOVA-based experimental articles published in JXE between 1932 and 1944. Rucci and Tweney [11, p. 169] have identified Reitz [26] as the only ANOVA-based JXE article published before 1940, which enabled me to narrow my manual search of the journal to the years 1940 through 1944. In doing so, I identified 12 additional articles that employ the ANOVA: Edgerton et al. [27]; Stuit and Donnelly [28]; Gabel [29]; Stuit and Lapp [30]; Sells, Loftus, and Herbert [31]; Prichard and Ojemann [32]; Long and Welch [33]; Musselman [34]; Meshke [35]; Clark [36]; Smith [37]; and Tsao [38]. Tsao [38], however, was excluded from the study, owing to my impression that, as a methodologist substantially concerned with the development and refinement of methods, he possessed a set of interests and a perspective that lay too far beyond the realm of education for his work to be considered a homogenous contribution to the collection [39, 40]. Accordingly, Reitz and the additional 11 articles identified above represent the corpus upon which I base the following citation context analysis.

In their use of “rigid statistical techniques,” argues Reitz [26, pp. 12-13], business and industry provide a promising model for the new science of education, perhaps the “biggest ‘business’ in the country,” which, by the early 1930s, had only begun to make progress in “determining their procedures.” In response to the troublesome scarcity of “rigid” educational techniques, Reitz introduces, among other methods, Fisher’s ANOVA:

The second method consists in a study of variance by the method Fisher [17] has outlined in his volume, Statistical Methods for Research Workers. . .
[This method] demonstrates whether the variance of the means is not more than might be expected through chance or not more than the variance about the means [26, p. 15].

Charles Bazerman [41, pp. 12-13] has pointed out that authors who strive for shared meaning through the use of unusual symbolic forms place substantial interpretive strains upon their readers. Therefore, it comes as no surprise that Reitz not only cites the well-known Fisher in 10 of his 11 references to the ANOVA but also provides a lesson on interpreting the results of this new method. After setting up an experimental situation to which Fisher’s method would be applied, for example, Reitz dichotomizes the range of possible outcomes and carefully explains their implications:

Should the test of significance be positive by either method, the yearly samples of student bodies may be considered as coming from different parent populations, and, therefore, as heterogeneous. If the test shows negative results, the samples may be considered as homogeneous; i.e., as coming from the same parent population [26, p. 15].

Later in the paper, Reitz follows the same form, citing Fisher and interpreting the results of the analysis:

To illustrate Fisher’s procedure, the data pertaining to Institution A are used. Fisher’s \( z \) is found to be the difference between the natural logarithms of the two standard deviations. . . . Relating \( z \) to \( SE_z \), a ratio of \(-.5897/.4097 = -1.44\) is obtained, which signifies clearly that \( z \) is insignificant in the sense that the variance about the means is much greater than that of the means. With regard to Institution A, it may again be concluded that the variation about any one mean over a period of four years is much larger than the variation of the means from year to year, i.e., Institution A has been populated over a period of four years by students of approximately the same average intelligence [26, pp. 16-17].

By 1940, citational practices involving Fisher’s ANOVA had begun to change. Intertextual references were still common but markedly on the wane, and an increasing proportion of the references were not to Fisher and his Statistical Methods for Research Workers but to second-generation textbooks, such as G. W. Snedecor’s Statistical Methods and E. F. Lindquist’s Statistical Analysis in Educational Research [42, 43]. (Fisher himself trained Snedecor in variance techniques at Iowa State College in the mid-1930s, after which Snedecor, in turn, trained Lindquist at nearby University of Iowa [11, pp. 176-179].) Of the seven ANOVA instances in Harold Edgerton et al.’s massive study [27], for example, only three involve Fisher, one involves Snedecor, and the remaining instances leave the ANOVA uncited altogether. Otto Gabel [29] cited Lindquist in conjunction with two of his six ANOVA instances, leaving the remaining four uncited as well.
As the presence of ANOVA-based articles continued to grow within educational circles in the early 1940s, citational practices as a whole rapidly atrophied. Nowhere, for example, in Sells, Loftus, and Herbert’s [31] or in Prichard and Ojemann’s article [32] is the ANOVA affiliated with any human sponsorship:

The following comparisons were made using the method of Analysis of Variance, with two criteria of classification . . . [31, p. 310].

The differences are generally not statistically significant when tested by Analysis of Variance [31, p. 319].

An application of the F-test indicated that in three of the six patterns the differences in variances are significant below the two per cent level while in three they are not [32, p. 117].

In contrast to Reitz, who affiliated the development of the method with Fisher and explained its innerworkings in great detail, educational researchers publishing in 1941 needed only to mention “the Analysis of Variance” or the “F-test” to convey a reasonably stable meaning to their readers. (The “F-test” is not entirely devoid of any trace of human origins, for the “F” in the “F-test” was introduced by Snedecor in 1934 to commemorate its inventor, R. A. Fisher [11, p. 167]. Because this fact is little known, and, I assume, little known to educational researchers in the 1940s, I treat “F-test” as a phrase without reference to human sponsorship.)

By 1943, the ANOVA had sedimented in experimental education research. Gone were the intertextual references to Fisher, or to anyone else for that matter. (Long and Welch [33] offer the last cited ANOVA instance in my corpus.) Of the 26 instances of ANOVA in Musselman’s [34] paper, for example, not a single one was cited, and the modern practice of condensing the whole of Fisher’s methods and results into a parenthetic clause appended to substantive claims had begun:

However, in the experimental group the schools were not homogeneous ($F = 5.5$, which exceeds even the one per cent tabled value) [36, p. 15].

What in 1934 was a promising new method developed by Fisher and incorporated into educational research through detailed description was now a consolidated and highly stable symbolic form, stripped of context and human affiliation and transmitted conceptually through stock phrases. In the words of Latour [25], the ANOVA had become a “black box,” a reified object whose history of production had faded away.

PASSIVITY, NOMINALIZATION, AND PERSONIFICATION OF FISHER’S ANALYSIS OF VARIANCE

To examine the distribution of agency in early experimental education research, I examined the methods-related sections of the 12 ANOVA-based experimental articles published in JXE between 1932 and 1944 [26-37]. My unit of analysis was
the verb phrase, that is, any phrase that included a verb as long as the verb was not acting as a gerund. Each unit was considered an “instance,” either of passive voice, verb nominalization, personification, human agency, or “other.” (Instances labeled as “other” fell beyond the scope of this study.) Consider the following example:

The two methods clearly indicate that the data of Institution B show the average intelligence of the student body by class and gender. The different yearly groups cannot be thought of as being drawn from the same population. Therefore, we begin to reconsider our assessment strategies in this new light.

Here, “methods . . . indicate” and “data . . . show” each count as an instance of personification because the authors have placed the agency of the verb (“indicate,” “show”) within an inanimate object (“methods,” “data”). The next sentence counts as an instance of passive voice because the agent of the action is unrepresented, eclipsed by the passive construction of the sentence. In short, the reader is left wondering, “The different yearly groups cannot be thought of . . . by whom?"

The final sentence counts as an instance of human agency in general and experimenter agency in particular: It stands as one of the rare instances in which the experimenters acknowledge their own active involvement in the context of experimentation. Based on this framework, I identified 799 instances of passive voice, 56 instances of verb nominalization, 247 instances of personification, and 221 instances of human agency within the corpus. Other instances, including instances of agency, were collected, but they fell beyond the present scope. Overall, 1,323 instances were examined for the purpose of this article.

**Passivity**

Based on the distribution of agency within the 12 articles, an overwhelming use of passive voice stands as the primary means by which educational researchers suppressed their own agency within the context of experimentation. Of the 1,323 instances examined, 799 are passive constructions whereby the experimenters essentially wrote themselves out of the experiments. In their articulation of procedures, for example, Stuit and Donnelly [28] and Long and Welch [33] offer typical examples of the passive style found throughout the corpus:

In order to make a more careful statistical study of the results, analysis of variance was employed in studying the significance of difference in means. The variance of each major group in all of the individual examinations and composite score was analyzed. By applying the “F” test, the results of which are presented in Table 111, it was found that no significant differences in means existed within any of the groups except for the Mathematics and English majors. These differences and tests of their significance are presented in Table IV [28, pp. 296-297; my emphasis].
The effect of increasing the number of antecedents was studied by regrouping the data so that the abstractness was held constant. For example, the average score for the test with two antecedents was compared with that for the test with three antecedents at the object level. The values of P for the 18 different comparisons are presented in Table XII. In nine of the comparisons a reliable difference was found between the scores, and the lower average was always associated with the sub-test having the greater number of antecedent [33, p. 260; my emphasis].

Foregrounded in these depictions are, among other things, the ANOVA, which was employed; the concept of variance, which was analyzed; results, which were presented; significant differences, which were found; the concept of effect, which was studied; the concept of abstractness, which was held constant; averages scores, which were compared; and P-values, which were presented. Yet nowhere within these accounts is the agency of either Stuit, Donnelly, Long, or Welch acknowledged. Although they prominently display their names and affiliations on the first page of their articles, the experimenters quickly assume positions “off camera” and shift the attention to the methods, results, and presentation devices themselves. The move is not out of humility but an attempt at assent through depersonalization, whether intentional or simply the consequence of following the norms of scientific writing. For only by displacing themselves from the scenes of their experiments can experimenters begin to create the perception that their results are, as Joseph Gusfield [44, p. 20] pointed out a quarter-century ago, “untainted by the obvious presence of the observer.”

That the experimenter per se stands as the intolerable actor in the scene is more noticeable in those passages that implicate humans in multiple roles. Otto Gabel, for example, treats his own agency in a way quite distinct from the way he treats the relatively innocuous agency of his experimental subjects:

Sets of material were made up in four forms, W, X, Y, and Z. Six reading selections comprised the materials for each form. Each selection dealt with one of the following topics: oil, aviation, forest conservation, coal, safety first, and reaping tools.

For each of these sets of materials the following testing procedure was used. The pupil read the six selections. After reading each selection he immediately took the tests on that selection. The tests covered two separate pages. One page contained the test questions with the definite foils; the other page contained the test questions with the indefinite foils. After answering both pages of test questions on this selection, the pupil immediately turned to the next selection and did as he had done with the first. He continued in this manner to the end of the complete set of materials. No time limit was placed upon reading the selections and taking the tests [29, p. 178; my emphasis].

Notice how Gabel systematically eschews from the fore his own agency while freely depicting his experimental subjects as active participants in various
activities. Stuit and Donnelly [28] make the same distinction between experimenter and subject agency in their account as well:

The variance of all major groups in each of the individual examinations was also analyzed. The results are reported in Table V and reveal that within every examination there are statistically significant differences in means for some of the major groups. The differences found to be significant or approaching significance are presented in Table VI. . . .

English Majors exhibit significantly higher mean scores than other groups in the following examinations: English Training, higher than all major groups except Physical Science and Mathematics; Iowa Silent Reading Test, higher than Biological Science, General Science, Political Science and Fine Arts Majors; High School Content Examination, higher than Journalism, Political Science and Fine Arts Majors; Composite score, higher than Biological Science, Journalism, Political Science, History and Fine Arts Majors.

Physical Science Majors show a significantly higher mean in the Mathematics Aptitude Examination than Political Science, History and Fine Arts Majors. General Science Majors exhibit a mean score in Mathematics Aptitude which is significantly higher than that of the Fine Arts Major [28, p. 298].

But Gabel, Stuit, and Donnelly are not alone: Of the 221 occurrences of human agency within the corpus, less than 20 percent implicate the experimenter. In the rhetoric of experimental education research, then, it is not so much human agency in general that is systematically suppressed but a particular kind of human agency: that which threatens to breach the priority of mechanical objectivity [4, 9, 45].

Nominalization

Complementing the use of passive voice within the corpus was the occasional presence of nominalization. Of the 1,323 instances examined, 56 are of a nominalized form; and of those 56, nearly half take the form of one of the following: “comparison,” “reference,” “inspection,” or “examination.” The result is a sort of aloof style of writing, which removes the experiential quality from the representation [46]. Rather than implicate her own agency in the context of interpreting scatter diagrams, for example, Smith chose the nominalized form:

Examination of the scatter diagrams indicated that this difference might be due in part to the fact that for some of the older children the ceiling of the test was reached [37, p. 99; my emphasis].

Likewise, rather than concede his involvement, Gabel chose the more distant “comparison,” which essentially imputes an activity without a responsible agent:

Obviously, then, a comparison of Form W with Form Y would compare the methods of presentation with the physical order of testing (definite-indefinite) held constant. The same would be true in a comparison of Form X with
Form $Z$ with the opposite testing order (indefinite-definite) held constant [29, p. 179; my emphasis].

Holding to this line, “examinations” indicate salient features of the experiment and “comparisons” do the comparing, all without recourse to the sponsorship of human agency.

**Personification**

In lieu of human agency, educational researchers imbued various inanimate constructs with the primary agency of the occasion through a persistent use of personification. Of the 1,323 instances examined within the corpus, 247 are cases of personification. The objects and nature of the personification vary greatly throughout the corpus: only in Edgerton et al., for example, are problems and questions depicted as autonomous objects able to “arise” of their own accord [27, pp. 262, 267], while only in Sells, Loftus, and Herbert do tables, not experimenters, “analyze” the experimental data:

Table 2 summarizes the results of Tables X to XV. This table analyzes the data of Table I to IX, summarized in Table I, arranged in a different way. Tables X to XV combined all nine Activity schools and Control schools into groups, by grade [31, p. 312].

Despite such idiosyncrasies, over half of the 247 occurrences fall into one of three categories: In general, educational researchers imbued either the methods, the results, or the presentation devices with the agency of the occasion, essentially displacing themselves from the depicted conditions of production.

Thirty-nine (15.8 percent) of the 247 occurrences personified the methods of science by placing phrases such as “method,” “ANOVA,” ”$t$-test,” “$F$-test,” “$F$-value,” and “formula” in the role of active agent. Reitz [26] and Meshke [35], for example, typify the style:

The second method demonstrates whether the variance of the means is not more than might be expected through chance or not more than the variance about the means [26, p. 15; my emphasis].

By relating $z$ to $SE_z$, .3596, a ratio of 3.46 is obtained, so that both methods clearly indicate the fluctuation of the means to be significantly greater than might be expected through mere chance [26, pp. 20-21; my emphasis].

Since the $F$-test permits fairly wide departure from normality and since inspection of the data gave no indication of such departures, it was deemed unnecessary to test the assumption of normal distribution [35, p. 5; my emphasis].

Likewise, Musselman depicts his methods as autonomous agents able not only to “bear out” hypotheses and “indicate” salient features of the experiment but also to “reveal” significant relationships between variables [34, pp. 62-63]. The
cardinal question becomes: To whom? For such accounts of scientific inquiry create the impression that the author and audience stand on equal epistemic ground as interested witnesses to the autonomous activity of the methods. Issues of authority and ethos are not at the fore. Rather, the educational researchers implicitly include themselves within the larger audience, awaiting, like everyone else, the knowledge “revealed” by the methods. That the text embodies an endless array of human choices in its symbolic representation of the event, itself the result of an endless array of human choices, all in a unified rhetorical attempt at assent is minimized within the published accounts of the experiments, achieving what Gusfield has called the “style of nonstyle” in scientific writing [44, p. 17].

Personified more frequently than the methods were the results. Of the 247 occurrences of personification, 44 (17.8 percent) were instances in which words such as “data,” “results,” “outcome,” “findings,” and “scores” received the active agency. Like the methods, the results are frequently cast as autonomous agents able to “show,” “reveal,” “indicate,” “suggest,” and “confirm” various features of epistemic interest, enabling the researcher to fade into the backdrop of the larger audience. Stuit and Donnelly [28] and Musselman [34] offer two typical passages:

The results are reported in Table V and reveal that within every examination there are statistically significant differences in means for some of the major groups [28, p. 297; my emphasis].

Data on achievement test scores for the group indicate that the superior pupils are considerably above average for achievement, and that the boys excel the girls in mean achievement test scores.

The data on mental ages indicate for the boys a slight superiority over the girls.

The data on I.Q. give the boys superiority over the girls in mean Otis Intelligence Quotients and the Henmon-Nelson Intelligence Quotients. . . .

Data on achievement ratios show for the boys definite superiority over the girls and indicate that the achievement of both boys and girls in high school subjects is considerably lower than might be expected, using their intelligence scores as measures of promise.

The data on general scholarship averages indicate that these superior pupils are doing good work on the average and that the boys are doing slightly better than the girls [34, p. 55; my emphasis].

Also embedded within such a depiction is the notion that the results stand as “brute facts,” within which immediately recognizable and unequivocal meanings reside. Sells, Loftus, and Herbert provide the most palpable example within the corpus when they claim that their findings “contain evidence of a much richer and more functional program in the Activity schools” [31, p. 319; my emphasis]. The result is a surreptitious consolidation of two distinct realms: that of quantification and that of human meaning. Literate symbols themselves contain no self-evident meaning independent of interpretation, though we often act as if they do [6, 47]. Yet it is precisely this distinction between the artifice of statistics and the reality
of the social world, whatever it is, that the rhetoric of early experimental education blurs: to move with ease between these two realms, to create the impression that the data speak for themselves—this is one of the most impressive rhetorical feats of the corpus.

Personified most frequently were the presentation devices themselves. Of the 247 occurrences of personification, 56 (22.7 percent) were instances in which such phrases as “table,” “illustration,” “diagram,” “graphic relationship,” “curve,” “graph,” and “region” received the active agency of the occasion. Edgerton et al. [27] and Musselman [34] offer two typical passages:

Table I shows considerable interrelation between different kinds of liberalism. . . . Table VII shows the correlations of Table I after the linear influence of intelligence has been removed by partial correlation [27, p. 262; my emphasis].

The graphic illustration of this relationship indicates clearly that superior pupils coming from broken families had higher achievement ratios [34, p. 64].

Notice how these two authors eschew their own agency from the fore by redirecting their audience’s attention to the tables and graphs located within the texts. This creates the impression that the audience is interacting not with Edgerton et al. or Musselman but with the presentation devices themselves, now sufficiently personified. Issues of paradigmatic allegiance [7] and interpretation [6] as well as any parallax between insider and outsider perspectives [40] recede. The effect is a sort of abdication of responsibility brought about by what Deidre McCloskey calls “unheralded assertion”:

The scientist says, It is not I the scientist who makes these assertions but reality itself . . . Scientists . . . pretend that Nature speaks directly, thereby effacing the evidence that they, the scientists, are responsible for the assertion [45, pp. 9-10].

Gusfield considers it a move toward “equivalence”:

[The experimenter] seems to say: “I will give you, the reader, all the knowledge and factual information that I have. We will reason together and achieve a consensus through fact and reason. You, as a rational person, cannot but reach the same conclusion as I” [44, pp. 21-22].

In either case, the readers are afforded the opportunity to see the “raw results” of the experiment from the mythic vantage point of the cool observer. In short, the ensemble of interaction is not that of an audience and an author but that of a universal audience, within which the author resides, and the presentation devices through which much of the work of scientific inquiry is ostensibly accomplished.
THE SEDIMENTATION OF FISHER’S 0.05

Underlying the mathematical architecture of Fisher’s ANOVA is a deeply consequential decision point in the form of statistical significance. From a philosophical point of view, it defines a threshold for ontological commitment: typically, statistically significant results are seen as empirical evidence of an underlying agent responsible for their existence, whereas results that are not statistically significant are dismissed as the product of chance. From an ethical standpoint, statistical significance embodies the balancing of two inferential risks: that of mistakenly rejecting a true null hypothesis (known as a Type I error) and that of mistakenly not rejecting a false one (known as a Type II error). From a practical perspective, statistical significance separates publishable from nonpublishable research in a world whose reward system privileges publishing above all else [20, 48, 49].

Broadly speaking, the ANOVA is a mathematical procedure used by researchers to determine whether the difference between two or more values is sufficient to eliminate the suggestion that the difference is due to chance. Although the ANOVA systematizes much of the work of experimental analysis, researchers must choose the level of significance by which to judge whether a particular difference is “statistically significant.” Such a decision cannot be made by way of science. Unlike the ANOVA, particular levels of statistical significance are not the product of theoretical or experimental development. An outsider might expect the current standard in social scientific research, 0.05, to correspond to the intersection of two special lines on an important graph; another might suspect that it owes to the assumptions of probability theory. Yet no such rationale exists for 0.05 [9, 50]. A practice reified through habit rather than owing to an immanent logic of mathematics, the standard of 0.05 emerged from the tradition of Fisher, whose 1925 classic, *Statistical Methods for Research Workers*, provided educational researchers with precedent for the norm [51].

To trace the sedimentation of Fisher’s 0.05 in experimental education research, I collected all 0.05-based experimental articles published in *JXE* between 1932 and 1944. In doing so, I identified 25 articles: Hagman [52]; Wenger [53]; Lyon [54]; Stump [55]; Edgerton et al. [27]; Gabel [29]; Suits and Lapp [30]; Sells, Loftus, and Herbert [31]; Worbois [56]; Long and Welch [33]; Stewart [57]; Appel [58]; Dawe [59]; Worbois [60]; Morse and Chittenden [61]; Meshke [35]; Smith [37]; Goldfarb [62]; Hastings [63]; Tschechtelin [64]; Willits [65]; Thelen [66]; McNiel [67]; Patterson [68]; and Tsao [38]. Tsao [38], however, was excluded from the study, owing to my impression that, as a methodologist substantially concerned with the development and refinement of methods, he possessed a set of interests and a perspective that lay too far beyond the realm of education for his work to be considered a homogenous contribution to the collection [39, 40]. Accordingly, the first 24 articles identified above represent the corpus upon which I base the following citation context analysis.
Fisher’s 0.05 first appeared in the first volume of JXE in 1932 by way of Elmer R. Hagman, a Connecticut psychologist affiliated with the Iowa Child Welfare Research Station. Here, Hagman [52] not only cites the third edition of Fisher’s *Statistical Methods for Research Workers* [69] but also provides a general definition of 0.05 to assist his readers in their interpretive task:

The next step was to find which of these factors existed coincidentally at the time of the study. For this purpose the chi-square test of independence as given by Fisher’ was used. A value of .05 for P means that there were only five chances out of 100 that the obtained difference was not a true difference [52, p. 119].

Likewise, in the following year, M. A. Wenger [53], also of the Iowa Child Welfare Research Station, cited Fisher and provided brief definitions of 0.05 in two footnotes:

In view of the small numbers in the groups to be compared, Fisher’s formula (7, p. 107) for the significance of difference in means of small samples has been employed rather than other more widely used techniques. The derived ratio of the difference to the standard error of the difference is t; P is the probability value. When P is .05 or less, the difference in the means is considered significant [53, p. 203].

Fisher’s table (7, p. 176) for different levels of significance of correlation coefficients when the samples are small has been employed throughout. Values of P not in excess of .05 indicate the correlation may be regarded as significant [53, p. 205].

In contrast to Hagman, Wenger chose a more functional definition: The salient aspect of 0.05 for Hagman was that it represents five chances in 100 that the results are not “true,” whereas for Wenger 0.05 stands on functional ground as the permissible level at which results “may be regarded as significant” [53, p. 205; my emphasis]. This rhetoric of permission, exemplified here by Wenger and also evident in the work of Lyon [54] and Gabel [29], implicitly responded to the precarious issue of justification: No one could justify the use of 0.05 on scientific grounds. Only by intertextual association to Fisher’s reputable work did 0.05 gain initial credence, a circumstance that ensured it would never appear too far from a citation to Fisher in this earliest period of modern statistical knowledge making.

By 1940, however, the close association between 0.05 and Fisher had begun to dissolve. As an increasingly familiar symbolic form, 0.05 began to appear more frequently without citation or definition. Stump, for example, cites Fisher as his source for Student’s tables of significance [55, p. 346], but nowhere within the context of 0.05 is Fisher named:

Test I (affectivity score) dealing with unpleasant objects and topics, and “disgust” words in Test I have “P” values between .05 and .02 and, therefore,
the means show significant differences between the subjects with and those without phobias [55, p. 348].

When citations did appear, they were not always to Fisher. By 1940, second-generation textbooks, such as G. W. Snedecor’s [42] *Statistical Methods* and E. F. Lindquist’s [43] *Statistical Analysis in Educational Research*, had begun to propagate Fisher’s 0.05 within education circles. Gabel [29, pp. 181-182], for example, cites Lindquist rather than Fisher in the context of 0.05:

In each of these tables, the difference in mean scores for the two methods of presentation may be regarded as significant at the 5 per cent level if the ratio of the variances for $P$ and $P \times G \times S$ exceeds 4.26, or at the 1 per cent level if this ratio exceeds 7.82, $P$ having 1 degree of freedom, and $P \times G \times S$ having 24 degrees of freedom [see Lindquist, 43, pp. 60-65].

Also gone were the definitions of 0.05, either general or functional, that guided readers in their interpretive task. In short, what was once an unusual symbolic form requiring explicit instructions for interpretive stability was beginning to sediment into a succinct and stable literate symbol: “0.05” or “the five per cent level.”

By 1942, virtually all of the intertextual references that tied 0.05 to its conditions of production, as well as to the definitions that maintained its interpretive stability, had vanished. Of the 28 occurrences of 0.05 in *JAE* between 1942 and 1944, only four (14 percent) were accompanied by citation [33, p. 255; 61, p. 275; 65, pp. 41, 42]. The far majority of the authors described their experimental differences as significant at “0.05” or at “the five per cent level” without any intertextual cue or definition:

For the control group, however, there was a difference of 5.1 points in favor of Form M, and that difference was significant at the 5 per cent level [59, p. 206].

In the comparisons of the classroom and the control groups, the differences were not large enough to reach the tabled F-values, even at the .05 level [35, p. 6].

The scores of the 3-6 to 4-6 year old children, from socio-economic level A were not significantly different from the scores of the middle group, while the differences between these same socio-economic classifications at the level 4-6 to 5-6 were significant at the five per cent level of confidence [37, p. 100].

The learning of the fourth concept in this group was significant for all three groups at the five per cent level; only five additional items in the test revealed at this level of significance for all three groups [66, p. 65].

Divorced from its highly contextualized conditions of production and transmitted as a highly stable symbolic form, 0.05 had become a black box. Accredited first through Fisher and ultimately through sedimentation, it moved beyond
the realm of critical examination to become, by 1944, part of the common stock of knowledge in experimental education research.

**CONCLUSION**

The rhetoric of experimental science creates its climate of trustworthiness in part by suppressing the “flesh and blood” of human agency (Matthew 16:17) and appealing instead to a mechanical form of objectivity achieved through a steadfast commitment to method. In the early years of experimental education research, this commitment materialized in the widespread adoption of R. A. Fisher’s analysis of variance (ANOVA) as the experimental method of choice [11, 12], combined with the acceptance of Fisher’s 0.05 as the professional standard for statistical significance.

Educational researchers suppressed the human conditions of production surrounding the ANOVA by adopting a citational practice that led to its “sedimentation,” described by Berger and Luckmann [24, pp. 67-72] as a process by which highly contextualized experiences congeal into stable meanings, which are gradually stripped of context and human affiliation and transmitted abstractly through a shared sign system, such as written language, to members of an ambient social network. In 1934, Reitz introduced the ANOVA to education researchers, citing Fisher’s *Statistical Methods for Research Workers* and assisting his audience in interpreting the results. By 1940, however, citational practices had begun to change: Intertextual references were still common, but an increasing proportion were not to Fisher but to second-generation textbooks. Within a few years, intertextual references were gone altogether, and the phrase “ANOVA” stood as a highly stable and familiar symbolic form, stripped of context and human affiliation. By 1943, it had become a “black box,” a reified object whose history of production had faded away [25].

Educational researchers suppressed their own agency within the context of experimentation in large part through an overwhelming use of passive voice combined with an occasional use of nominalization. Instances of passive voice outnumbered instances of experimenter agency 20 to 1; instances of nominalization outnumbered instances of experimenter agency almost 2 to 1. Less than 40 times within the 1,323 instances examined did educational researchers concede their own involvement within the context of experimentation. Far more likely to be acknowledged were the subjects of the experiments, or other ostensibly innocuous participants, such as teachers, administrators, or parents. In short, it is not so much human agency in general that is suppressed but a particular kind of human agency: that which threatens to breach the priority of mechanical objectivity.

With their own agency out of the way, educational researchers realized their commitment to method by imbuing the methods, results, and presentational devices themselves with the active agency of the occasion through a persistent
use of personification. Such a depiction creates the impression that the researchers and audience stand on equal epistemic ground as interested witnesses to the autonomous activity of a third party, the method, which churns out the “brute facts” of science; these brute facts are, in turn, compiled and presented by the tables, figures, and diagrams themselves.

Underlying the mathematical architecture of statistical methods, however, is a deeply consequential decision point in the concept of statistical significance, frequently operationalized as 0.05. As a practice reified through habit rather than owing to an immanent logic of mathematics, the act of choosing an appropriate level of significance falls squarely upon the shoulders of human judgment. Educational researchers suppressed the human conditions of production surrounding 0.05 by adopting, as they did with the ANOVA, a citational practice that led to its sedimentation. In short, as the presence of Fisher’s 0.05 increased within the research literature, the need for intertextual support and interpretive assistance atrophied, leaving a stable, familiar, and abstract symbol, freed from its warrantless rationale and benefitting from the ostensible objectivity that comes with quantification, to be confidently redeployed, further reified, and ultimately accepted as part of the common stock of knowledge in experimental education research.

Today, the momentum of Fisher’s 0.05 continues to shape the codification of knowledge claims throughout the social sciences. After a casual review of the sociology literature, Skipper, Guenther, and Nass remarked, “the common, arbitrary, and virtually sacred levels of .05, .01, and .001 are almost universally selected regardless of the nature and type of problem. Of these three, .05 is perhaps most sacred” [13, p. 155]. According to Cohen, psychology has fared no better: “After 4 decades of severe criticism, the ritual of null hypothesis significance testing—mechanical dichotomous decisions around a sacred .05 criterion—still persists” [70, p. 21]. In education, the “sacred” status of 0.05 is preserved not only by the sheer force of habit but by institutional pressures as well: The National Center for Education Statistics (NCES), for example, has mandated that for all statements supported by hypothesis testing, “[t]he level of significance (alpha) shall be .05” [71]. Accordingly, NCES publications, forewarn the editors of Education Statistics Quarterly, “only discuss differences that are significant at the 95 percent confidence level or higher” [72, p. 1].

So entrenched is the standard of 0.05 that professionals in related disciplines have begun to take notice: In 1977, the United States Supreme Court calculated its first standard deviation in response to a case involving ethnic discrimination in the selection of a Texas grand jury [73]. Later that year, the Court again turned to basic inferential statistics in a case involving racially motivated hiring practices in a Missouri school district, setting ablaze questions of hypothesis testing and statistical significance nationwide [74; 75, pp. 1334-1335]. By the 1980s, judges and attorneys were citing the Supreme Court’s earlier decision in Hazelwood [74] and turning to 0.05 for an objective standard by which to evaluate statistical
evidence in cases ranging from employment discrimination to food and drug regulation to environmental protection. The court of appeals in Segar v. Smith [75], for example, accepted statistical inferences of racial discrimination against the Drug Enforcement Administration that were significant at 0.05, while dismissing those that failed to meet the standard. Their justification was simple: “[S]ocial scientists usually accept a study that achieves statistical significance at the .05 level” [qtd. in 76, p. 1343].

Sedimentation, passivity, nominalization, personification, and a rhetoric of experimentation that privileges conformity over veracity, reliability over validity: Such is the anatomy of the “Peter effect” in early experimental education research. Such are the ways in which educational researchers mobilized language to convince their readers that the claims of education were the claims of science, claims that remained “untouched by human hands” [4, p. 10].

REFERENCES

**Other Articles On Communication By This Author**


Direct reprint requests to:

Joseph Little
175 Pinnacle Trail
Aurora, Ontario L4G 7G7
Canada
e-mail: joseph3141@look.ca