Alphas and Asterisks: The Development of Statistical Significance Testing Standards in Sociology

Erin Leahey, *University of Arizona*

What is distinctive about sociology's history is the extent to which it both illustrates and challenges sociology itself.
- N. J. Demerath, III (1994)

Abstract

*In this paper, I trace the development of statistical significance testing standards in sociology by analyzing data from articles published in two prestigious sociology journals between 1935 and 2000. I focus on the role of two key elements in the diffusion literature, contagion and rationality, as well as the role of institutional factors. I find that statistical significance testing flourished in the 20th century. Contagion processes and the suitability of significance testing given a study's data characteristics encourage the diffusion of significance testing, whereas institutional factors such as department prestige and particular editorships help explain growing popularity of the .05 alpha level and use of the "three-star system" of symbolic codes (i.e., *p < = .05, **p < = .01, ***p < = .001).*

What alpha level do you use when testing for statistical significance? Many sociologists today would say .05. What symbol do you use to indicate this level of statistical significance? Many sociologists would say one asterisk. Although the choice of alpha level should technically depend on sample size, statistical power and sampling procedures, researchers routinely use the .05 percent level (signified by a single asterisk) as a benchmark despite variation in these conditions. In fact, one of the leading journals in the discipline has a policy that disallows the reporting of significance above the .05 percent level. How did statistical significance testing become normative? How did the .05 level and its symbolic code (a single asterisk) become dominant? What factors influenced the creation and propagation of these practices?

Although some scholars point to the technical superiority of significance testing as the key factor affecting its diffusion, two factors undermine this argument. First, the technical superiority of statistical significance testing relative to other methods for testing hypotheses was, and still is, hotly debated (Carver 1978; Cowger 1984; Labovitz 1972; Morrison and Henkel 1970; Raftery 1995; Schmidt 1996). Second, statistical significance testing did not become dominant in all disciplines. Researchers with goals similar to those of sociologists rely on slightly different practices; for example, public health researchers typically employ (albeit 95 percent) confidence intervals, and psychologists rely more on effect size. If statistical significance testing was technically the most appropriate procedure and technical

I would like to thank Barbara Entwisle for helping me formulate and improve the design and ideas expressed in this paper. Comments from Ken Bollen, Gail Henderson, Ted Mouw, the late Rachel Rosenfeld, Francois Nielsen, Sarah Soule, Ron Breiger, John Levi Martin, John Sonnett and Adrian Raftery were also instrumental in improving the manuscript. *Correspondence should be directed to Erin Leahey, Department of Sociology, University of Arizona, P.O. Box 210027, Tucson, Arizona, 85721-0027. E-mail: leahey@arizona.edu.*

"fit" was the sole criterion for choosing among alternative techniques, then all disciplines would have adopted the same practice. To the extent that the overall technical superiority (or what diffusion researchers call "rationality") of statistical significance testing practice can be established, prior statistical research and disciplinary comparisons suggest that technical explanations are incomplete.

However, the relevance of a related concept – *suitability* – is open to empirical investigation. Suitability refers to the appropriateness of a practice given the data at hand and is relevant to not only statistical significance testing, but also the choice of particular alpha levels. For example, because statistical significance testing is based on normal distribution theory, it should only be performed on samples obtained via probability sampling techniques. And, because sample size can affect the results of statistical significance tests, this characteristic of the data should be relevant to decisions concerning alpha levels. However, evidence that statistical significance tests and the .05 alpha level are often not suitable for specific analyses abounds (Berkson 1970; Carver 1978; Collins 1984; Cowger 1984; Labovitz 1972; Morrison and Henkel 1970; Raftery 1995; Schmidt 1996) and will be reviewed herein. Moreover, misapplication and misuse of statistical significance tests is rife and can produce results that are distorted, artificial and deceptive (Gardenier and Resnick 2001; Godlee 2000).

The suitability of statistical significance testing practices can also be compromised when researchers apply normative standards that are inappropriate for the specific research at hand. Sociologists routinely use .05 as a default alpha level even in studies with thousands of cases and very high statistical power. Regardless of sample size, statistical power and sampling procedures, sociologists commonly use a set of three alpha levels (.05, .01 and .001) with three corresponding symbols (*, ** and ***) – what I refer to as the "three-star system." Indeed, one of the top journals in the field has mandated this research practice in its editorial policy since 1991. Because these practices have become standards for publication and publication bias against null findings is common (Gardenier and Resnick 2001; Godlee 2000), researchers may be inadvertently encouraged to search for significance, to present their analytic methods or results selectively, and to neglect discussions of magnitude and other concepts that speak to the importance of findings. According to the American Statistical Association (1999), such practices infringe on statistical validity and statistical ethics. Although sociologists may be eager to incorporate statistical techniques and follow established statistical standards in order to attain scientific legitimacy, following them without attending to study-specific considerations can be costly.

What makes significance testing practice a fascinating and important case for investigation is that it appears to have diffused not *because* of its suitability in various situations, but despite of it. It may certainly be the case – and I will empirically examine – that an increase in the use of probability sampling encouraged the use of statistical significance tests, and trends in sample size were linked to the popularity of the .05 alpha level. But a large literature on everyday, practical applications of statistical significance testing (reviewed below) suggests that misuse is the norm. Moreover, in this study I find that suitability is less relevant to statistical significance testing than we might expect. The case for studying statistical significance testing is buttressed not only by the ubiquity of the practice, but also by the potentially serious methodological and ethical implications of the way it is commonly practiced.

When the suitability of a practice is expected to encourage its use (as is typically the case), perhaps adding *contagion* – individuals' proclivity to learn from and model others' practices (Strang and Macy 2001) – to the conceptual model is sufficient to explain diffusion. But when suitability is hypothesized to be less relevant (as is the case here) and

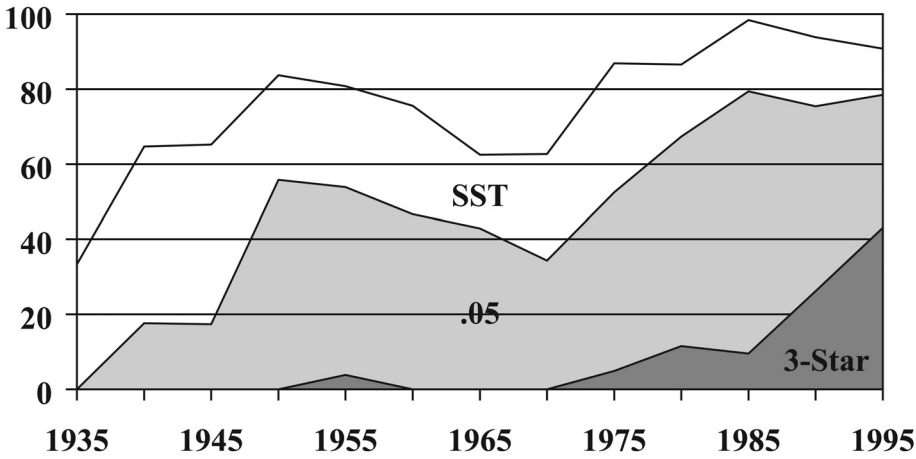
when the development of standards is not inevitable, factors in addition to forces of contagion are likely operating. While some might cite standards in other statistical practices – e.g., testing for autocorrelation (Durbin and Watson 1951) and influential cases (Bollen and Jackman 1990) – to demonstrate the inevitability of statistical standards (perhaps because of their technical superiority over alternatives), it is important to recall that statistical practice often departs from statistical theory. Moreover, there are realms of research, such as editing anomalous and inconsistent data (Leahey, Entwisle and Einaudi 2003) for which diverse practice is the norm. Thus, to understand the spread of statistical significance testing practice, I supplement the common conceptual model of diffusion to include additional social and institutional factors, such as investment in research and computing, graduate training, the status of authors and their affiliations at the time of publication, and journal editors. Incorporating these factors permits a more complete understanding of the ways in which certain statistical significance testing practices became dominant within sociology.

In this paper I heed Camic and Xie's (1994:773) call for sociologists to examine the "process by which statistical methodology [has] acquired its importance." Unlike most investigations of research practice, which explore practice in a given time and place using case study methods, I trace the development of statistical significance testing practice historically using archival research methods. The use of statistical significance testing was not advanced until the 1930s (Schmidt and Hunter 1997) when R.A. Fisher (1935) first used the .05 p-value in his book, *Design of Experiments.*[1] The use of the three-star system appears to be more recent. What influenced the adoption and preservation of these practices? To answer these questions, I collect and code data from research articles published in two top general sociology journals widely circulated between 1935 and 2000. In addition to investigating the impact of suitability and contagion, I develop and test hypotheses about the role of various institutional factors, investment in research and computing, graduate training, individual and institutional status, and journal editors.

Common vs. Formal Practice

Evidence for the existence of statistical significance testing standards abounds. Indeed, "the need for statistical methods is widely taken for granted by… social scientists." (Camic and Xie 1994: 777) Those who have analyzed quantitative data know that they will be met with suspicion if they forgo the use of statistical significance tests. In the sample of articles to be used throughout this paper, 91 percent of the articles published between 1995 and 2000 that were eligible to use statistical significance testing actually did so. Of these, 86 percent used the .05 alpha level, 67 percent used .01, and 52 percent used .001. In contrast, only 10 percent of articles used the .10 level, and 3 percent used the .02 level. The three-star system was initiated relatively recently, but has flourished; between 1995 and 2000, almost half of the articles published in this time period employed the .05, .01 and .001 alpha levels with one, two and three asterisks, respectively. This developmental trajectory of statistical significance testing practice (see Figure 1) demonstrates that statistical significance testing, the use of .05 alpha level, and the three-star system are practices have become normative in the kind of sociological research published in prestigious (i.e., primarily quantitative) disciplinary journals. (Clemens, Powell, McIlwaine and Okamoto 1995; Turner and Turner 1990.)

Figure 1. Percent of articles (calculated using five-year intervals) over time that use statistical significance testing (statistical significance test), statistical significance testing and alpha = .05 (.05), and statistical significance testing, alpha = .05, and the 3-star system (3-Star)



But to what extent are these standard practices reflective of correct practice? The rationality of a practice – how "good" it is – has been used by researchers to help explain diffusion processes (Strang and Macy 2001). It is difficult to assign a rationality score to statistical significance testing practices because research is contextual: a practice that is appropriate for one study may be inappropriate for another. Moreover, diffusion explanations that hinge on the pure technical merit of statistical significance tests are weak when we realize that many disciplines do not rely on such tests. However, it is possible to assess the extent to which a practice is suitable to the task at hand. This, which I will refer to as the *suitability* of statistical significance testing, can be ascertained for individual studies and incorporated into a diffusion model. This improves upon prior analyses of the diffusion of practice, which typically presume either the "goodness" or "badness" of a practice (Institute of Medicine 1985).

The suitability of statistical significance testing for a given study depends on how well the study (particularly its data characteristics) meets the assumptions underlying the logic of the test. In order to test hypotheses about an unobserved population parameter on the basis of an observed sample statistic, researchers assert that if the null hypothesis (typically that a parameter equals zero) is true, and the estimated standard error produces test statistics (e.g., Z, T, F, chi-square) that fall beyond the critical value corresponding to the chosen alpha level, the null hypothesis is unlikely and thus should be rejected. It is important to note that this "textbook" approach to statistical significance testing (Mohr 1990) assumes the use of simple random sampling techniques and is intended for single tests of single hypotheses. Moreover, because sample size affects the test, sample size should ideally be incorporated into decisions about alpha levels (Raftery 1995).

Current, common practice is much more informal than this textbook approach. Typically, when researchers examine the statistical significance of a regression coefficient, they note the distributional value that is exceeded by the sample result and report the corresponding level of significance. Often researchers do not, or cannot, use simple random sampling or any kind of probability sampling. They may test more than a single hypothesis without employing corrections. They may neglect to factor sample size into the choice of

alpha level.[2] This informal approach, which perhaps reflects a lack of fit or suitability between statistical significance testing logic and actual research situations, has come under particular scrutiny (Bailar 1997; Fowler 2002; Gardenier and Resnik 2001). Scholars claim that it is unsatisfactory, even misleading, for several reasons: 1) a 5 percent alpha level no longer equates to a 5 percent error risk when it is not used as decision rule but only reported as a significance level, 2) it cannot account for the model uncertainty that is evident when multiple hypotheses are tested and multiple models are specified, 3) it is easy to find statistically significant relationships when sample size is large, and 4) without simple random sampling, the rules of probability and thus classical inference do not apply.[3]

In addition to suitability, contagion is required to explain the diffusion of any practice, and significance testing practice is no exception. Contagion occurs when people mimic the practices of others and thereby propagate the practice by using it themselves and perhaps serving as a model for others; it incorporates social learning. This interdependence of outcomes (Strang and Soule 1998) is often modeled by examining inter-actor patterns of interaction and influence, which cleanly divide adopters from non-adopters. But if similarity and connectivity is assumed within a population of actors – e.g., if they are all in the same discipline aiming to publish in the same journals – contagion can be measured as the impact of intellectual currents: the prevalence of a given practice. This conception breaks with the notion of direct contagion and views potential adopters as "responsive to the distribution of present adopters in the population" (Strang and Soule 1998:284), and has been relied on extensively to understand why seemingly worthless innovations enjoy tremendous staying power (Strang and Macy 2001:153).

Although diffusion processes are typically shaped by rationality (or its situation-specific counterpart, suitability) and contagion (or prevalence in a bounded community), I argue that more than contagion is required to understand how an often-less-than-suitable practice became standard. If statistical significance testing practice were truly "by the books," then characteristics of the data that determine the appropriateness of the practice, such as sample size (which is positively related to statistical power) and sampling procedures, combined with forces of contagion, might provide a sufficiently complete model of diffusion. But to explain a more curious outcome – how an informal and occasionally unsuitable approach to statistical significance testing spread – I consider the role of additional institutional factors.

Institutional Influences

Social and institutional influences on research practice are multiple and varied, but most of them can be viewed as components of a "resource base" for sociologists' research activity and, more specifically, for their use of statistical significance testing practice. Investments in research and computing, graduate training, the professional status of authors and their institutions, as well as journals, their editors and editorial boards can all be considered institutional forces. I review literature and develop hypotheses about each of these potential influences on the use of statistical significance testing practice.

Investments in Research and Computing

Given that the "structure of sociology as an academic discipline and the production of ideas is intimately connected to the nature and level of resources that have been available to sociologists" (Turner and Turner 1990:8), support from national agencies and private foundations as well as increases in computer technology may have encouraged the diffusion of statistical significance testing. In the 1950s, "computerized commodity statistics" were not yet available (Abbott 2001:115), but the survey paradigm was dominant and financial

commitments to social science research were extensive, especially from foundations. Not only did the funding sources encourage standardization of practice, but the research institutes (e.g., ISSR and SSRC) that received research monies were typically headed by directors who steered research toward statistical analyses (Turner and Turner 1990).

Around the time that research funding for social science research began to recede, another resource – investments in computer technology – began advancing rapidly. In 1960, Samuel Stouffer, 42nd president of the ASA, commented on the role of computers and how they affected trends in statistical analysis:

> In the past decade, there seems to be a drift away from some of the conventional techniques, like correlation, in spite of some of their mathematical advantages…. There is another trend in analysis, however, which may lead in other directions. I refer to the increasing use of high-speed electronic computing machines, which are now readily available to behavioral scientists. Training in programming the new computers is now part of the experience which many of our graduate students are getting. One of these IBM monsters in an hour produces huge matrices of correlation coefficients which may have taken a clerk with a desk calculator a year to do.
>
> – Turner and Turner (1990:174)

Massive IBM computers were shared by all members of a campus community in the 1950s and early 1960s, and these gave way to individual departments' card-sorters and calculators in the mid-1960s, and statistical packages such as SPSS and SAS available for mainframe and desktop use starting in the mid-1970s. Fast computers allowed for the introduction of many independent variables, and differences significant at the .05 level became easy to find (Turner and Turner 1990:175). In essence, computers and computer programs made statistical analysis available to scholars who perhaps would have been deterred from analyses performed by hand, as this new model of research was easily performed and easily taught (Abbott 2001: 115). In sum, influence from funding agencies and foundations may have contributed to initial bursts in statistical significance testing practice. When funding for social science began to decline in the late 1970s, computer resources permitted sociologists to use statistical significance testing and find significant results easily, perhaps without contemplating the suitability of statistical significance testing for their specific dataset. It is also notable that people who adopted statistical significance testing had connections to statisticians and tended to be at large public universities (Turner and Turner 1990).

Graduate training

Because the role of graduate programs is to train future sociologists for positions that often include research, their potential influence on sociologists' research practices is strong. While pursuing their doctorate degrees, graduate students learn the process of research through formal coursework, research assistant positions, collaboration opportunities, and thesis and dissertation work guided by advisors. To the extent that graduate programs have different course offerings, faculty members, mentoring traditions (Macrina 2000), collaboration opportunities (Leahey 2004b), climate (Braxton 1991; Victor and Cullen 1988), penchants for specific types of research (Turner and Turner 1990), and preliminary exam structures (Brady, Hostetter, Milky and Pescosolido 2001), graduates' research practices might vary. Even today, the instruction and approach to statistically-oriented research that graduate students obtain depend largely on their Ph.D. program; such influence was probably greater in the earlier part of the 20th century when standards for statistical significance testing had not yet emerged.

And even for students who do not pursue quantitative research, Ph.D.-granting departments may still affect their research practices given that professional development occurs "in areas not immediately pertinent to the… dissertation." (Guston 1993) It may also be the case that higher status graduate programs have greater influence over their graduates' approach to research.

Institutional and Individual Status

Various kinds of practices endure not simply because they are good enough (perhaps "suitable") but because they do important work for actors with the most status (Casper and Clarke 1998; Timmermans and Berg 1997). Thus, in addition to prestige of one's Ph.D.-granting department, current status in the field and prestige of one's institution could influence statistical significance testing practice. Academic departments serve as a locus of professional relationships; they are the primary sites of integrative social networks within universities (Blau 1973). The power of these relatively bounded groups to shape research practice may derive from their distinct ethical work climates that provide prescriptions, proscriptions and permissions regarding moral obligations (Victor and Cullen 1988). As is true in organizational networks (DiMaggio and Powell 1983), the normative pressures inherent in professional networks of individuals may encourage consensus in research practice.

However, not all institutions are equally influential. The status of an institution may modify its influence on members' ideas and practices. While some research addresses disciplinary hierarchy (Cole 1983), I focus on hierarchy within the discipline of sociology by examining departmental prestige or status. Departmental prestige has been shown to affect individual productivity level and other opportunities (Allison and Long 1990; Long and McGinnis 1981); it may also influence research practices like statistical significance testing. Institutions with high reputations may feel more pressure to represent themselves as a cohesive unit and may have more power to make members conform.

Not only the status of institutions, but also the status of individuals that comprise them, may influence research practice and standards (Leahey 2004c). Pinch (1986) found that characteristics of authors influence an article's reception; more specifically, Galaskiewicz (1985) found that the status of individuals influences their knowledge and their evaluations of others' work. Camic and Xie (1994) demonstrate how four leading proponents of statistical research influenced the subsequent spread of statistical thinking in their respective fields. According to Platt (1996:134) "perceptions of what is customary or prestigious among colleagues" shape researchers' methodological decisions. Just as organizations "tend to model themselves after similar organizations in their field that they perceive to be more legitimate and successful" (DiMaggio and Powell 1983:152), researchers may mimic the research practices of researchers whom they deem legitimate and successful. Thus, individuals' intra-professional status (Abbott 1981) may supplement the effect of institutional status.

Journals Editorships

As scientific gatekeepers, journal editors and their boards have the potential to profoundly influence the content of work published in their journals. By determining the best research and proceeding to publish it, journals present examples of best practice to the research community (Keyfitz 1993). Because individuals place faith in published sources, especially those that are well known or prestigious (Hernon and Walters 1997), leading journals in a field may have a particularly strong influence on readers' practices (Turner and Turner 1990). This may be particularly true in sociology, a discipline with relatively low consensus (Hargens and Hagstrom (1982). Although prior research has demonstrated editors' limited capacity to monitor and influence rather invisible (i.e., unreported) practices, such as data editing (Leahey 2004a), their influence over more visible practices, such as statistical significance testing, may be enhanced.

Editors' and their boards' proclivities might become even more institutionalized – and thus have a more substantial and lasting effect – when they are translated into policy. Whereas most editorial policies concern more mundane editorial issues, such as page length and formatting style for tables, some serve to institutionalize research practice itself. For example, *Demography* requires that all authors share their data with readers. More relevant to the current study, in 1991 the *American Sociological Review* instituted an editorial policy that forbids the reporting of significance above the 5 percent level and requires one, two or three asterisks to denote significance at the .05, .01 and .001 levels, respectively. By controlling central symbols such as these, journals – their editors, editorial boards and policies – can shape both the form and content of research.

Investigations of journal editor influence on scholarly research are limited. The few studies that have been done (Coser 1975; Neavill 1975; Simon and Fyfe 1994) typically rely on anecdotal rather than systematic evidence. More recent and empirically-grounded studies have focused on the identification of organizational gatekeepers (Morrill, Buller, Buller and Larkey 1999) and their network positions (Corra and Willer 2002), but not their influence per se. Although a broad literature suggests that gatekeepers use their authority to influence research, few studies empirically test the extent of journal editors' impact on a particular research practice.

In sum, to understand the diffusion of statistical significance testing practice, I supplement the typical conceptual model in three ways. First, I substitute the focus on rationality with a more context-specific measure, suitability. Second, to tap direct, as well as indirect contagion processes, I test the impact of the prevalence on statistical significance testing practice. Third, and perhaps most importantly, I incorporate diverse institutional factors such as the funding environment, advancements in computing, graduate training, individual and institutional prestige, and journal editorships. I consider this expansion of the traditional diffusion model necessary to understand why an often less-than-suitable practice – one in which misuse and misapplication is common – spreads.

Most of my hypotheses are non-directional. High status researchers may have more knowledge of available options and be less likely to use standard practices, or they may be more likely to rationalize and accept disciplinary standards, as they probably contributed to their development. Lower status researchers may be more likely to support what is becoming apparently normative (to gain legitimacy within the discipline), or they could be more innovative. Similarly, prestigious departments may contribute more to the development of standards and may have more power to make members conform; at the same time, they could use their stature and power to encourage members to develop and use new practices. Although it may certainly be the case that particular journal editors and training programs might promote the use of statistical significance testing, the use of the .05 alpha-level, and the three-star system, I assess these affects non-directionally. I expect prevalence to have a positive affect on subsequent use of each practice, and thus use directional tests to evaluate contagion processes.

Data and Methods

To examine and explain trends in statistical significance testing, including the use of particular alpha levels and corresponding symbols, I collected historical data using archival material. I focus on the discipline of sociology[4] and collect data from scholarly journals, which contribute to paradigm propagation more than texts (Kuhn 1970) and serve as a currency of disciplinary evaluation (Clemens, Powell, McIlwaine and Okamoto 1995). Journals screen information that is permitted to circulate widely among members of the discipline; they are the gatekeepers of science (Crane 1970) and thus relevant to the establishment of research practice. I limit my

analysis to two historically and currently top sociology journals that have been at the core of the discipline since their founding the late 19th and early 20th centuries. These are general sociology journals that circulate among sociologists more widely than specialized journals. These are also top journals in the discipline that have consistently published innovative, pioneering work. For these reasons, their opportunity to influence others' work and establish trends is probably enhanced (Turner and Turner 1990). Although not representative of all sociological research, articles in leading journals tell something about "disciplinary standards and ideals." (Platt 1996:126)

I collected data from a 20 percent stratified random sample of articles published between 1935, the year in which .05 was first used by R.A Fisher, to the present. I stratified by journal and by issue in order to get a sufficient number of articles from each year. I chose not to stratify by year itself, as that could have resulted in a series of articles from the same thematic issue to the exclusion of other issues published in the same year. Stratifying by issue guarantees that oddities associated with particular issues will be fully represented. After eliminating erroneous inclusions, 1,215 articles remain.[5] I begin my analyses with the subset of 613 articles that test hypotheses using empirical, numeric data (thereby excluding theoretical and qualitative pieces), and then further limit the sample to the 496 articles that present statistical significance tests.

I recorded and coded various kinds of information about each journal article. I noted the type of data analyzed (if any) and various practices used: statistical significance testing, alpha levels, and corresponding symbols for alpha levels. These serve as the main dependent variables in the analyses and as the source for measures of prevalence. I noted data characteristics, such as the type of sampling procedures employed, if any, and sample size, because these data characteristics help determine which, if any, statistical significance testing practices are suitable for a given study. I also noted characteristics of the potential adopters (authors), including their names and their institutional affiliations, in order to create measures of professional status. I also note the journal and year of publication, as well as qualitative, textual data from the article itself that could be relevant to the practice under investigation.

Information from each sampled journal article and its respective first author is used to access additional secondary data to test hypotheses of interest. For example, the journal and year of publication allowed me to determine the appropriate journal editorship from mastheads. With the names of full authors, I was able to determine their Ph.D.-granting departments from dissertation abstracts, and with these data I created binary codes for:

a) particularly prominent institutions (the 10 most visible institutions in my sample),
b) whether the institution was public or private (given that scholars in the large public research universities had greatest access to statisticians (Turner and Turner 1990), and
c) whether the institution was located in the Northeast or Midwest (two prominent regions for research-related activity).

I documented the year of publication in order to correctly specify trends (expected to correspond to developments in research-related computer technology), to measure contagion (tapped using measures of prevalence of each practice in the prior year), and to add historic information on research funding obtained from the *Statistical Abstracts of the United States*.

With these data, I describe and explain the heterogeneity of statistical significance testing practice over time. I focus on three main practices: statistical significance testing itself, the use of the .05 alpha-level, and the use of particular symbolic codes, especially the three-star

system. The initial analysis is purely descriptive; I graphically present means and percentages to examine consistency in sociological research practice over time – or at least reports of such practices. To aid this descriptive analysis I incorporate historical information about the relevant debates and issues at hand, as well as qualitative information obtained from the sampled articles. Then, using pooled data (articles published in both journals and all years), I estimate logistic regression models using maximum likelihood techniques to explain the diffusion of statistical significance testing practices. The three dependent variables of interest are dichotomous: whether the article uses significance testing, and if so, whether the article uses a .05 alpha level and whether it uses the "three-star system" of symbolic codes.

I include explanatory variables to test hypotheses about suitability (akin to rationality), prevalence (akin to contagion), and the various institutional factors. Data characteristics can capture the *suitability* of statistical significance testing practices for a given research situation. Characteristics of the data include the type of sampling procedure employed and sample size (logged to alleviate positive skewness). Three binary indicators are included to capture sampling procedures: one for simple random sampling, one for other types of probability samples, and one for when an entire population is analyzed. Non-probability sampling is the reference category. The variable capturing prevalence is the proportion of articles in the prior year (published in either journal) that also used the research practice in question, ranging from 0 to 1. The results are robust to the specific time lag (1-year, 2-year, 5-year, etc.) used to construct the prevalence measure.

Institutional forces are operationalized in ways that exploit both the coded article-level data as well as supplemental secondary data:

Investments in Research and Computers. Because computers were developed during the latter part of the period under study, a continuous measure of investments in computer technology or the availability of computers is not available. However, historical information about the introduction of computers into academic research is used to inform the time specification used for each model. The only data on the research funding environment that is available for all years under study comes from the Statistical Abstracts of the United States and represents the percentage of federal obligations that were devoted to the National Science Foundation. Data specific to the social sciences was not available until mid-century and is thus insufficient for the present analysis.

Graduate training. Various binary variables are created to assess the effects of historically prominent departments that were likely relevant to the spread of statistical significance testing practice (e.g., Columbia and Harvard). Data on first authors' Ph.D.-granting institutions also permitted an assessment of other department characteristics, such as geographic location (e.g., perhaps the older universities in the Northeast were instrumental) and institutional type (public or private).

Institutional and Individual Status. I employ two measures to capture the status of authors and their institutions. To capture researcher status, I use the number of times the first author is represented in the sample to date (as first author or co-author)[6], which ranges from 1 to 7 (mean = 1.4), and is transformed using the square-root function to achieve normality. To capture institutional status, I include the number of times the first author's department or organization is represented in the sample to date (this includes coauthors' affiliations as well as first authors' affiliations).[7] Because this measure – which ranges from 0 to 98 – is right skewed, I use the square root in all analyses. These sample-based measures assess visibility and, to a lesser extent, reputation in the field. As Cole (1983:136) states, "our opinions of what is good science and who has done good work are based on judgments made by other people or the evaluation system in our field." Results were robust to alternative measures of status that controlled for time, such as the proportion of all articles published to date in which the institution (or researcher) was represented.

Journal editors. Using data on the journal and year in which each sampled article was published, I determine the respective journal editorship from journal mastheads and include them as binary variables in the statistical models.

Although the present study examines the spread of research practice, it is not strictly a diffusion study. Most diffusion studies limit their analyses to a constant population of potential adopters (Chaves 1996; Wejnert 2002); to replicate this I would have sampled authors instead of articles.[8] In contrast, the present study examines factors that influence the likelihood that an author employs a given significance testing practice in a given research article. Instead of assessing rates, I assess likelihood of use. I do not use the term adoption because it is possible that the authors of the articles under study have used statistical significance testing practices before; my interest is in explaining whether they use it currently. I do not examine sequences of adoption, but I do incorporate temporal heterogeneity by having the prevalence of the practice in the prior year influence current practice and include time-dependent measures of status. I do not have a control for clustering resulting from my stratified random sampling procedures, but I control the variables used to stratify the sample (time and journal) by including them in all models. The time function is year of publication, ranging from 35 (for 1935) to 100 (for 2000). Depending on the practice being modeled, I add to this linear main effect a squared term ($year^2$) and a cubic term ($year^3$), to best capture the trends depicted in Figure 1. A single binary variable is included to control for journal of publication. I control the multiple appearances of certain authors and institutions in the dataset by using robust standard error estimates.

Results

First, I consider the use of statistical significance testing itself; second, the use of the .05 alpha level; and third, the use of the three-star system. Within each of these categories, I present descriptive material as well as results from a multivariate model. The descriptive analysis includes quantitative data for the sample of articles as well as qualitative text from individual articles. For the multivariate analysis, I present results from logistic regression models in tabular form. For each pooled statistical model, I tested alternative specifications, such as a curvilinear effect of status (Phillips and Zuckerman 2001) and interaction effects between status and contagion, but these effects were not statistically significant and thus not presented in tabular form. Because many of the institutional factors of interest (e.g., journal editorships, prominent Ph.D.-granting institutions) are binary, adding all of them to one model would be inefficient. Thus I initially specify a model with all variables except the institutional characteristics, and then add each institutional factor to this model individually (results not shown). I then re-specify the model, including all of the institutional factors that were statistically significant when added individually, and present this as the final model. To further motivate and aid interpretation of these quantitative results, I integrate historical information on the relevant controversies.

In this paper, perhaps my own use of statistical significance testing will come under particularly close scrutiny. Given that I have used the most efficient type of sampling procedure (stratified random sampling), the standard errors produced through estimation will be inflated (as they assume simple random sampling), and results significant at the .08 or .10 level may be "real." At the same time, I test multiple hypotheses, which I should account for in my analysis by using a more stringent alpha level: the Bonferroni correction stipulates that the chosen alpha level (in this case .10) should be divided by the number of variables in the analysis (in this case, 10). Therefore, I focus my attention on coefficients that are significant at the .01 levels or less, though I discuss results significant at the .05 and .10 levels and also yield to current standards of asterisk usage.

Trends in Statistical Significance Testing

Of the 613 sampled articles published between 1935 and 2000 that used numerical data to test hypotheses, 81 percent did so using statistical significance tests. This proportion varied substantially over the course of the time period studied. As reflected in Figure 1, statistical significance tests were already being used in 1935, and their popularity increased rapidly (to more than 60 percent) within 15 years. The general upward trend in statistical significance testing use is the period 1965 to 1970 when its use leveled off somewhat. After 1975, statistical significance testing was widespread in quantitative research, when consistently more than 80 percent of articles eligible to use significance tests actually did so.

Despite the general increase in the use of statistical significance tests over time, concern over their usefulness and appropriateness also increased over time and climaxed in the 1960s. As early as the 1950s, sociologists lamented the "increasing acceptance" and misinterpretation of statistical significance tests (Selvin 1957). A series of articles published in the *American Sociologist* in the late 1960s presented the different sides of the debate, which are summarized in Morrison and Henkel's volume, *The Significance Testing Controversy* (1970). Researchers worried about the "fad" of statistical significance testing were convinced that assumptions about sampling were not being met. I summarize the main issues of this debate and bring qualitative evidence from the sampled articles to bear on the issues.

Concerns about the use of statistical significance testing when simple random sampling – or at least probability sampling – is not employed are evident in my sample of articles. Of the 496 articles that presented tests of statistical significance: 10 percent employed simple random sampling techniques, 30 percent used other probability sampling techniques, and 6 percent analyzed an entire population. The remaining 54 percent of articles used statistical significance tests with data collected through non-probability sampling techniques. Without knowing the probability of selection into the sample, it is impossible to correct for sampling bias, which statistical significance testing assumes is non-existent. Some authors express concern that their sampling procedures in some way invalidated the use of statistical significance tests, at least for inferential purposes. Of the 14 articles that were skeptical of statistical significance tests because of unmet sampling assumptions, half were aware of the lack of simple random sampling techniques and the possibility of non-independence of cases resulting from other types of probability samples.

There was also concern that statistical significance tests were being used even when researchers selected the entire population (universe) for study. In the present sample, 81 percent of articles that analyzed an entire population did so using statistical significance tests. The other 19 percent were too skeptical of significance tests under such conditions to use them. In one article, the authors note that "all differences are meaningful because we have the population." Others authors argued that significance tests are not necessary, meaningful or appropriate when an entire population is being investigated. A few of the articles using an entire population for analysis comment on the debate but proceed to use statistical significance tests anyway because "it is standard procedure." Although there remains controversy about whether statistical significance testing is appropriate when a population is under investigation (Berk, Western and Weiss 1995; Cowger 1984, 1985; Rubin 1985), there is an increasing sentiment in favor of it (Bollen 1995). This is also evident in my sample of articles: when the sample is restricted to articles published in 1975 or later, the percentage of articles that study an entire universe and employ statistical significance tests jumps to almost 100 percent.

Other issues in the debate, such as lack of attention to the importance of sample size

and power issues, and a neglect of substantive significance, were also evidenced in my sample of journal articles, especially around the 1960s. Two articles were concerned about the effect of sample size: "25 cases is too small a number to have any statistical significance," and more generally, "statistical tests with their exact null hypotheses tell us more about the inadequacies of our sample sizes than they do about the nature of the phenomena we are interested in." Two articles chose to focus on substantive, rather than statistical, significance: one claiming that the "magnitude of the difference is more important than [its] statistical significance." Other authors paid little attention to significance of tests because "the diversity of tests is more important" or because "post-hoc tests are open to question." Such concerns about the misuse of statistical significance testing parallel the declining use of statistical significance tests in the 1960s; but even authors who continued to use these tests voiced skepticism over their utility and purpose.

To quantitatively investigate the extent to which these concerns about the suitability of statistical significance testing, relative to the effects of contagion and institutional forces, have an impact on diffusion, I estimate a logistic regression model in which the binary dependent variable distinguishes between articles that are eligible to use statistical significance tests and use them (= 1) and those that do not (= 0). Results are presented in Table 1, Model 1. The suitability of statistical significance testing, as reflected in data characteristics, influences its use. Authors who obtained data using probability sampling techniques were more than twice as likely to use statistical significance tests as authors using non-probability samples of data. Articles that analyzed an entire population were no more or less likely to use statistical significance tests. Another data characteristic, sample size, is related to statistical significance testing, but the coefficient is negative: the larger the sample size, the less likely researchers are to use statistical significance testing. This effect does not depend on year of publication (interaction effects not shown).

In addition to the suitability of statistical significance testing given the data at hand, contagion processes and some institutional forces are also relevant to the diffusion of this practice. Even while controlling for time and thus assessing deviation from trend,[9] the extent to which significance testing was used in the prior year (its prevalence) has a strong, positive impact on a researcher's decision to use significance testing. Authors trained in the northeast, particularly at Harvard, are less likely to use statistical significance testing compared to scholars trained elsewhere. Perhaps this is attributable to Harvard's traditional theoretical focus (starting with Sorokin in the 1930s), its intellectual focus on books rather than articles, and a focus on measurement more than causal models (Turner and Turner 1990).

Moreover, authors' institutional status (the number of times their institution has appeared in the sample to date) is negatively related to statistical significance testing use. Attesting to the power that prestigious institutions have over members' practices, the univariate standard deviation for significance testing practice is smaller for more prestigious institutions than less prestigious ones. It is also important to note that in the baseline model (with all but institutional effects included), the curvilinear effects of time were statistically significant, and their direction corresponds to the "take off" of computer technology in the 1960s and 70s (results not shown). Last, two overlapping editorships discouraged the use of statistical significance testing: James Short's (at ASR 1972-74) and Arnold Anderson's (at AJS 1966-73). Short is a criminologist who considered ASR to be too narrow and consciously published a qualitative article in his first issue to try to open it up to alternative methodologies (personal communication). At AJS, Anderson was a stratification researcher who had a casual editorial style and aimed to open up the reviewer base (Abbott 1999).

Table 1: Logistic Regression Models of Statistical Significance Testing Practices: 1935-2000

	Model 1		Model 2		Model 3	
	SST		.05 Alpha Level		3-Star System	
	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.
Suitability						
sample size, logged	-.19**	.06	-.04	.06	.08	.08
simple random sampling indicator	.31	.40	.38	.36	.64	.55
other probability sampling indicator	.97**	.32	-.18	.25	.12	.36
universe/population indicator	.34	.53	-.59	.42	.91	1.09
[non-probability sampling indicator]						
Prevalence						
% articles using practice in prior year [a]	1.02*	.56	.71	.62	-1.73	1.35
Institutional Factors						
Investment in Research & Computers						
% federal obligations devoted to NSF	--	--	--	--	--	--
Graduate Training						
Public (= 1 if yes, = 0 otherwise)	--	--	--	--	--	--
Northeast (=1 if yes, = 0 otherwise)	--	--	--	--	--	--
Harvard	-.72+	.39	--	--	--	--
[Other prominent institutions]	--	--	--	--	--	--
Institutional and Individual Status						
Author's Institution	-7.61+	4.05	.02*	.008	--	--
Author	--	--	--	--	--	--

Journal Editorships						
AJS's Arnold Anderson (1966-73)	-.78+	.442	--	--	--	--
ASR's James Short (1972-74)	-1.23*	.50	--	--	--	--
ASR's Morris Zelditch (1975-77)	--	--	-1.05*	.52	--	--
ASR's Robert Faris (1952-55)	--	--	--	--	4.35**	1.60
AJS's Everett Hughes (1958-60)	--	--	--	--	5.29***	1.53
ASR's Paula England (1994-96)	--	--	--	--	1.27*	.56
[Other editorships]	--	--	--	--	--	--
Time						
year	-.08	.07	.78+	.47	.18***	.038
$year^2$	.001	.001	-.01+	.006	--	--
$year^3$	--	--	.00005+	.00003	--	--
Journal of publication						
ASR indicator	.13	.27	.40+	.22	.53	.38
Constant	3.74	2.64	-18.83+	10.54	-19.26***	3.5
Log-Likelihood	-244.59		-275.65		-126.32	
Pseudo-R^2	.13		.09		.32	
Sample Size	598		492		492	

[a] One-tailed tests

Note: I use the three-star convention: *p < = .05 **p < = .01 ***p < = .001 and note effects significant at the 10percent level (+ p < = .10). Due to the large number of institutional factors, specification of these final models was determined by the addition of each institutional factor to a baseline model, which included variables representing suitability, prevalence, time and journal. Variables whose coefficients were statistically significant when added (individually) to the baseline model were selected for inclusion in this final model. For 15 articles, I could not determine the sample size, reducing the sample size from 613 to 598 in model 1 and from 496 to 492 in models 2 and 3.

Trends in Alpha Levels

With regard to specific alpha levels, patterns in the use of the 5 percent level follow patterns of statistical significance testing use in general. Of the 496 articles that use statistical significance tests, 67 percent use the .05 alpha level. Figure 1 depicts how the use of the 5 percent level changed over time. Although no articles used the .05 alpha level in 1935, its popularity increased rapidly between 1945 and 1950. When the significance testing controversy peaked in the 1960s, the use of .05 tapered off, but regained its upward momentum in 1970. By the late 1990s, more than 80 percent of articles that use statistical significance tests used the 5 percent level of significance (not necessarily exclusively). This increasing homogeneity of practice is even more dramatic when compared to trends in the use of other alpha levels, such as .02, which failed to become standard practice in sociology. Since 1935, only 12 percent of articles have used .02, 3 percent used .03, and 2 percent used .07. As the use of .05 increased, the use of other possible and somewhat popular alpha levels declined.

In addition to increasing consensus in the choice of alpha levels, a lack of concern about such choices may also indicate acceptance of the standard that developed. In contrast to statistical significance testing overall, there seems to be less skepticism of the choice of particular alpha levels and a generally increasing acceptance of standard levels. Some skepticism about the use of particular alpha levels was voiced as early as the 1940s. In 1941, an author comments that the critical ratio of 2.0 (corresponding to two-tailed tests at the 5 percent level) was chosen "arbitrarily." An article written in 1957 noted that "we use significance levels as convenient cut-off points rather than as actual tests of inference." Another article written in the 1950s put the word "significant" in quotation marks when describing an effect found to be statistically significant at the 5 percent level. However, there are few critiques of alpha levels in later decades, possibly indicating a general acceptance of .05.

A further indication of the standardization of the .05 alpha level is when the use of it is no longer accompanied by a rationale. As early as 1941, authors began to recognize that a .05 standard was developing. Over the course of the next four decades, there were numerous references to .05 as the "typical," "ordinarily accepted," "generally considered adequate," "conventional," "standard" and "customary" significance level. Such explicit references to .05 as the standard probably fulfilled two functions: they helped legitimate and rationalize the use of .05 by the author(s), and they contributed to the idea that .05 is, in fact, standard. However, in the mid-1980s, there was a decline in explicit references to .05 as the standard. After the 1980s, the use of .05 continues almost unabated, but explicit references to .05 as the standard diminish. If implicit standards are more embedded than ones that require explicit references, then perhaps the .05 standard did not really solidify until the 1980s.

What influences whether a particular author will use the .05 alpha level for significance testing? For this practice, suitability is less relevant than it was for statistical significance testing practice overall: data characteristics, such as sampling procedures used to collect the data and the effective sample size, have no statistically significant impact on the use of the 5 percent level (See Table 1, Model 2). The prevalence of the practice in the prior year only has a strong, positive and statistically significant effect when time is incorrectly specified: with the addition of a squared and cubic term for year, prevalence loses its statistical significance. A variety of institutional characteristics were statistically significant when added to the baseline model (which included all but the institutional variables) individually; and it is these variables that are included in the final model presented here. The status of the author's institution is again significant, but here it has a positive impact on the practice of interest – use of the 5 percent alpha level. Higher status institutions (measured by their representation, to date, in this elite sample) were more likely to use the 5 percent level relative to other institutions; they also had a lower univariate standard deviation for this research practice, attesting to their power to induce conformity. Overall, ASR had a

positive influence on the use of .05, but under one particular editor – Morris Zelditch at Stanford, who edited ASR between 1975 and 1977 – the popularity of this alpha level declined. Zelditch, a social psychologist and theorist, was less concerned with statistical significance than with substantive significance. Although the ritualistic use of the .05 alpha level was a "hot topic" while he was editor, he had no intention of altering practice through his editorship and in fact pursued a rather hands-off approach to the issue (personal communication).

Trends in Symbolic Codes

The use of stricter levels of significance, including .01 and .001, appear to have followed the lead of .05. As .01 and .001 were used increasingly along with .05, a common set of symbolic codes appears to have accompanied this trend. This set of codes ("*" indicating significance at the .05 level, "**" at .01, and "***" at .001) has been used, at least partially, by 36 percent of articles that use statistical significance testing. The growth of this current standard was gradual but not monotonic, in that partial use of the three-star system preceded its complete use. The first article to use the three-star system completely (all three alpha levels and all three symbols) appeared in 1955. This system of codes, however, did not take hold immediately; it was only in the late 1970s that it was picked up again, and from 1985 forward, its use spread rapidly.

Turning to multivariate analyses permits an examination of how factors like suitability, prevalence and institutional factors influenced development of the three-star system. Results from a logistic regression model are presented in Table 1, Model 3. Like the use of the 5 percent alpha level, characteristics of the data do not factor into researchers' decisions to use the three-star system. Contagion processes (tapped with the measure of prevalence) are not influential when it comes to the spread of the three-star system. For this practice, it is the institutional factors that are most relevant. In the baseline model (in which all but the institutional factors are included), the control variables for journal of publication as well as the time variable (best represented as a linear main effect) are significant and positive. But with the addition of binary variables that capture particular combinations of time and journal – the editorships – we find that the general ASR effect can be attributed to a few particular editors. In the mid- to late-1950s, ASR editor Robert Faris and AJS editor Everett Hughes encouraged the use of the three-star system. Faris, a student of Ogburn and teacher of Leo Goodman, taught statistics for years and encouraged the use of fine distinctions – given that with the .05 level one of every 20 findings could occur just by chance (personal communication with his son, Jack Faris). Hughes' second stint as editor (1958-60), significant here, was actually a continuation of his editorship that began in 1952, during which time Peter Rossi served as editor for a year in 1958. Therefore, Hughes' positive effect on the three-star system may be attributed to Rossi. Hughes had strong ties to (statistically-oriented) Columbia, devoted much of his time to an inter-disciplinary center at Chicago (Abbott 1999), and was occasionally attacked by humanist sociologists, but his approach was still much more qualitative than Rossi's. In later years, another ASR editor, Paula England (at Arizona, 1994-96) also oversaw a rise in the use of the three-star system, but actually inherited the three-star editorial policy from the previous ASR editor, Gerry Marwell. Aside from these journal editorship variables, no other institutional factors are relevant to the spread of the three-star system of symbolic codes.

Implications

By analyzing a sample of articles published in two prestigious sociology journals since 1935, I have documented initial variation and the eventual convergence of statistical significance testing practices descriptively. I partially explained these patterns by incorporating:

1) suitability of the practice, measured by data characteristics (tapping rationality),
2) intellectual currents, measured by the prevalence of the practice in the prior year (tapping processes of contagion), and
3) various institutional factors in multivariate models that pooled articles over time.

I found that statistical significance testing flourished in the 20th century. Before 1940, only 31 percent of articles that used numeric data to test hypotheses conducted statistical significance tests and by 1995, 91 percent of such articles did so. Among articles that used statistical significance testing, I also found increasing consensus concerning the choice of alpha level(s) and the symbolic codes used to indicate particular significance levels.

This trajectory of statistical significance testing use, and the 5 percent level in particular, was not influenced heavily by trends in the suitability of the practice to the data at hand. Statistical significance testing was designed under the assumption that data are collected via simple random sampling techniques or probability samples with requisite sample weights. It is reassuring to have found that the nature of the sample is related to the use of significance tests: authors using probability samples of data were more likely to use statistical significance tests than authors with non-probability samples. The 5 percent alpha level is also suitable as long as sample sizes are not excessively large, but sample size was not related to the use of .05. Perhaps suitability is more relevant to the spread of formal practices, such as organizational policies, which are more dependent on cost and resource effectiveness (Kraatz and Zajac 1996).

Social factors, such as the prevalence of the practice and institutional forces, were more relevant to the diffusion of statistical significance testing practices. Intellectual currents in the field were contagious and positively influenced the use of statistical significance tests. Counterbalancing these trends were the effects of various institutional factors that reduced the likelihood that an author would employ statistical significance tests: receiving a Ph.D. from Harvard, current institutional status, and ASR editor James Short (1972-1974) and AJS editor Arnold Anderson (1966-1973). Regarding use of the 5 percent alpha level, high status institutions were again relevant, but this time in a positive direction: authors at institutions with greater representation in these top journals were more likely to use the .05 alpha level. One journal editor, ASR's Morris Zelditch (at Standford, 1975-77) discouraged the use of the .05 level. With respect to the use of the three-star system, the only relevant institutional factors were three journal editorships, all of which encouraged use of the now dominant symbolic codes: AJS's Everett Hughes (1958-60) and ASR's Robert Faris (at Washington, 1952-55) and Paula England (at Arizona, 1994-96).

Author status as measured in this study did not affect the use of any statistical significance testing practices. This is congruent with some specific historical instances. Despite Paul Lazarsfeld's prestigious position in the field during the post-war period, his reluctance to use statistical significance tests (Platt 1996) and his efforts to promote alternative latent class models (Raftery 2001) did not spread far beyond his department at Columbia University. Perhaps Columbia's institutional reputation, based on its variable-centered survey research paradigm that was open to statistical significance testing, transcended Lazarsfeld's. Findings of this study call for the inclusion of institutional opinion leaders in studies of diffusion, especially those tracing the spread of an informal practice.

One factor intuitively relevant to the diffusion of statistical practices is computer technology, which in this study could only be captured indirectly. The way in which time was specified in each model corresponded to accelerations in computer technology and computer availability. For example, the use of statistical significance tests and the .05 alpha-level took off in the mid-1970s, just when SPSS and SAS were released (in 1975 and 1976, respectively).

The existence of computers made statistical testing less complex and easier to apply, and the capacity and speed of computers allowed for the inclusion of many independent variables, making it easy to find at least one significant effect. Complexity is inversely related (Rogers 1962) and compatibility and portability (Abbott 1999: 212) are positively related to diffusion. Advances in computer technology may have also limited the competition by reducing the chance that more "esoteric" alternatives (Abbott 1999), such as Lazarsfeld's latent class models, would flourish. One statistical package, STATA, recently incorporated a routine to output not only coefficients, standard errors and p-values, but also asterisks that correspond to the three-star system. Apparently, ease begets use.

Understanding social forces acting upon researchers' decisions to use or disregard a particular research practice highlights both the potential benefits and drawbacks of standardization. It may help researchers to evaluate standards that have been created and propose alternatives to occasionally unsuitable standards. For example, if statistical significance testing has become habitual not because samples are increasingly collected though simple random sampling, but because researchers think that their research will be deemed unworthy if they do not use such tests, then perhaps the purpose of statistical significance testing should be critically evaluated. Recognizing the social factors influencing research practice and the more technical factors that should – but do not – influence research practice may permit a "new mode of control" over previously uncontrolled factors (Walton 1966).

Literature in the area of social movements aids an interpretation of these findings. Perhaps the spread of the informal, common approach to statistical significance testing is analogous to the spread of unsuccessful protest tactics. Statistical significance testing can be considered "unsuccessful" to the extent that it is used even when conditions for its use do not exist. Soule (1999) studied the college shantytown movement and found that despite this protest tactic, problems in helping students achieve their goals with respect to campus administrators, it diffused because it was framed as successful by external agents, such as the media, and the tactic resonated with actors' perceptions and experiences. Similar forces may be at work with statistical significance testing: the practice has been framed as successful, appropriate and scientific by journals and other gatekeepers, and it has resonated with sociologists' goals and experiences.

The results of the present study also point to the possibility of two processes that have been outlined by organizational theorists DiMaggio and Powell (1983): mimetic processes and coercive isomorphic pressures. The impact of contagion on the use of significance testing and the 5 percent level and the impact of institutional standing on the use of the 5 percent level lends support to their claim that mimicking, especially modeling the practices of those who are viewed as legitimate and prestigious, is a powerful force in the standardization process. The significance of journal of publication on use of the three-star system may be preliminary evidence of coercive isomorphism at work. Coercive isomorphism results from both formal and informal pressures or mandates – such as one journal's editorial policy on statistical significance testing – which can be felt as a force, persuasion or a simple invitation.[10]

Perhaps sociologists can capitalize on the role of social factors to reduce the influence of other factors – such as misunderstanding, lack of knowledge, psychological pressure and the ostensible lack of alternatives – that contribute to trends in statistical significance testing practice. Training that takes place within intellectual communities is likely to be heeded and transferred easily by community members. Given the significance of institutional status, researchers in highly prestigious departments may be able to influence others' research practices through their roles as researchers, teachers and reviewers. With respect to alternatives to statistical significance testing, other practices – such as the application of imagination, common sense, informed judgment and appropriate research methods given the data at hand – can be encouraged to help achieve the scope and purpose of scientific

inference (Morrison and Henkel 1970: 311). Moreover, some alternatives to traditional significance tests and traditional alpha levels have been proposed (Jones and Tukey 2000; Schneider and Darcy 1984) and Bayesian approaches are promising (Raftery 1995). Social relations will likely encourage the diffusion of knowledge about these possibilities. Some researchers, not surprisingly at Harvard University, have made efforts to initiate change in statistical significance testing research practice.[11] Such researchers, especially in their capacity as peer reviewers, may have as much power to change current standards and promote sound research practices as certain journal editors.

Notes

1. Controversy surrounding the origin of statistical significance testing and the first use of the 5 percent level remains, but is not pertinent to this paper, which forgoes an investigation of origins and focuses on explaining subsequent use. Platt (1996) notes that these are very different goals. See Camic and Xie (1994) for a historical analysis of the initial adoption of statistical methods by proponents in various social scientific fields.

2. Examples of recently published papers employing such practices are available from the author upon request.

3. Unless, of course, sampling weights are available and used appropriately when another type of probability sampling technique is used. (For an exception, see Winship and Radbill (1994) for a discussion of the inappropriateness of sampling weights in some regression analyses).

4. By controlling for discipline of the publication outlet, this design acknowledges the role of disciplines in shaping research practice. Most diffusion studies highlight either diffusion into a population or diffusion within a population (Strang and Soule 1998). I acknowledge the influence of external sources, such as other disciplines, but do not model them. If statistical significance testing practice was expanding in other fields, this trajectory might affect researchers publishing in sociology journals equally. If it affects some (e.g., scholars who are very productive researchers and knowledgeable of developments in other disciplines) more than others, then such effects may be captured by variables that capture status differentials.

5. From my sampling frame I excluded research and teaching notes, book reviews, comments and replies, bibliographies, obituaries and all articles in supplemental issues.

6. I also collected each first author's year of Ph.D. from Dissertation Abstracts Online to create a variable "professional age" (year of publication minus year of Ph.D.). Because this information is missing for 40 observations, and the results do not depend on the measure used, I use the measure described in the text.

7. Published measures of department rank are only available every few years from 1970 to the present, and even for these years, the measures are calculated differently by different foundations (e.g., The American Council on Education, and the National Research Council) and thus inappropriate for longitudinal use. Moreover, these department measures would not include values for applied settings, such as, the Census Bureau and private research corporations, which are frequently represented in the sample.

8. For practical reasons (i.e., the large number of authors represented in these journals between 1935 and 2000) this might have entailed cluster sampling techniques in which a fixed number of authors was chosen, and then all their respective articles were coded. With such a design, authors who only published once and authors who adopted the given practice before their first publication would have posed analytic problems.

9. Contagion effects become stronger when time is removed from the model.

10. Given that statistical significance practice was not mentioned in this journal's editorial policy before 1987, and that significance testing practice was relatively embedded by that time, it seems that authors' practices influence editorial policy more than the reverse. In concurrent work based on qualitative interviews with journal editors, I am trying to flush out the direction of influence.

11. For example, Morgan and Sorensen (1999: 669) use asterisks in their regression tables to accommodate these guidelines, but note that they did not include them in their original manuscript because a) they do not care for frequentist tests of point-value null hypotheses, b) some readers mistake them for substantive importance, and c) asterisks are redundant when standard errors are included.

References

Abbott, Andrew. 1999. "Department and Discipline." University of Chicago Press.

_____. 2001. "Time Matters: On Theory and Method." University of Chicago Press.

Allison, Paul D., and J. Scott Long. 1990. "Departmental Effects on Scientific Productivity." *American Sociological Review* 55:469-478.

Association, American Statistical. 1999. "Ethical Guidelines for Statistical Practice."

Bailar, John C. 1997. "Science, Statistics, and Deception." *Research Ethics: A Reader*, edited by D. E. and J. E. Stern. University Press of New England.

Berk, Richard, Bruce Western and Robert E. Weiss. 1995. "Statistical Inference for Apparent Populations." Sociological Methodology 35:421-458.

Berkson, Joseph. 1970. "Tests of Significance Considered as Evidence." Pp. 284-294. *The Significance Test Controversy*, edited by D. Morrison and Ramon Henkel. Aldine Publishing Company.

Blau, Peter. 1973. "The Organization of Academic Work." Transaction Publishers.

Bollen, Ken, and Robert Jackman. 1990. "Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases." Pp. 257-291. *Modern Methods of Data Analysis*, edited by J. F. and S. Long. Sage Publications.

Bollen, Kenneth A. 1995. "Apparent and Non-Apparent Significance Tests." *Sociological Methodology* 25:459-468.

Brady, David, Carol Hostetter, Melissa Milky and Bernice Pescosolido. 2001. "The Structure and Substance of Preparing Sociologists: The Nature of Qualifying Examinations in Graduate Education." *Teaching Sociology* 29:265-285.

Braxton, John M. 1991. "The Influence of Graduate Department Quality on the Sanctioning of Scientific Misconduct." *Journal of Higher Education* 62:87-108.

Camic, Charles, and Yu Xie. 1994. "The Statistical Turn in American Social Science: Columbia University, 1890-1915." *American Sociological Revie*w 59:773-805.

Carver, Ronald. 1978. "The Case Against Statistical Significance Testing." *Harvard Educational Review* 48:378-399.

Casper, Monica J., and Adele E. Clarke. 1998. "Making the Pap Smear into the 'Right Tool' for the Job: Cervical Cancer Screening in the USA, circa 1940-95." *Social Studies of Science* 28:255-290.

Chaves, Mark. 1996. "Ordaining Women: The Diffusion of an Organizational Innovation." *American Journal of Sociology* 101:840-873.

Clemens, Elizabeth, Walter W. Powell, Kris McIlwaine and Dina Okamoto. 1995. "Careers in Print: Books, Journals, and Scholarly Reputations." *American Journal of Sociology* 101:433-494.

Cole, Stephen. 1983. "The Hierarchy of the Sciences?" *American Journal of Sociology* 88:111-139.

Collins, Randall. 1984. "Statistics versus Words." *Sociological Theory* 2:329-362.

Corra, Mamadi, and David Willer. 2002. "The Gatekeeper." *Sociological Theory* 20:180-207.

Coser, Lewis A. 1975. "Publishers as Gatekeepers of Ideas." *The Annals of the American Academy of Political and Social Science* 421:14-22.

Cowger, Charles D. 1984. "Statistical Significance Tests: Scientific Ritualism or Scientific Method?" *Social Service Review* 58:358-371.

_____. 1985. "Reply to Allen Rubin's 'Significance Testing with Population Data." *Social Service Review* 59:521-522.

Crane, Diana. 1970. "The Gatekeepers of Science: Some Factors Affecting the Selection of Articles for Scientific Journals." Pp. 488-503 in *The Sociology of Knowledge*, edited by J. Curtis and J. Petras: Praeger Press.

DiMaggio, Paul J., and Walter W. Powell. 1983. "The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields." *American Sociological Review* 48:147-160.

Durbin, J., and G. S. Watson. 1951. "Testing for Serial Correlation in Least Squares Regression, II." *Biometrika* 38:159-178.

Fisher, R. A. 1935. *The Design of Experiments*. Oliver and Boyd.

Fowler, Floyd J. 2002. *Survey Research Methods*. Sage Publications.

Galaskiewicz, Joseph. 1985. "Professional Networks and the Institutionalization of a Single Mind Set." *American Sociological Review* 50:639-658.

Gardenier, John S., and David B. Resnick. 2001. "The Misuse of Statistics: Concepts, Tools, and a Research Agenda." Unpublished Manuscript.

Godlee, Fiona. 2000. "The Ethics of Peer Review." in *Ethical Issues in Biomedical Publication*, edited by A. H. Jones and F. McLellan. Johns Hopkins University Press.

Guston, David H. 1993. "Mentorship and the Research Training Experience." Pp. 50-65 in Responsible Science, vol. II, edited by N. A. o. *Science*. National Academy Press.

Hargens, Lowell L., and Warren O. Hagstrom. 1982. "Scientific Consensus and Academic Status Attainment Patterns." *Sociology of Education* 55:183-196.

Hernon, Peter, and Laura R. Walters. 1997. "Student and Faculty Perceptions about Misconduct: A Case Study." *Research Misconduct*, edited by E. Altman and P. Hernon. Ablex Publishing Company.

Jones, Lyle V., and John W. Tukey. 2000. "A Sensible Formulation of the Significance Test." *Psychological Methods*.

Keyfitz, Nathan. 1993. "Thirty Years of Demography and Demography." *Demography* 30:533-549.

Kraatz, Matthew S., and Edward J. Zajac. 1996. "Exploring the Limits of the New Institutionalism: The Causes and Consequences of Illegitimate Organizational." *American Sociological Review* 61:812-836.

Kuhn, Thomas S. 1970. The Structure of Scientific Revolutions. University of Chicago Press.

Labovitz, Sanford. 1972. "Statistical Usage in Sociology: Sacred Cows and Ritual." *Sociological Methods* and Research 1:13-37.

Lacy, William B. 1994. "Nineteenth Century Visions and Twentieth Century Realities." *Social Epistemology* 8:19-25.

Lavin, Danielle, and Douglas W. Maynard. 2001. "Standardization vs. Rapport: Respondent Laughter and Interviewer Reaction during Telephone Surveys." *American Sociological Review* 66:453-479.

Leahey, Erin. 2004a. "Overseeing Research Practice: The Case of Data Editing." Unpublished manuscript.

_____. 2004b. "Transmitting Tricks of the Trade: Mentors and the Development of Research Knowledge." Forthcoming, *Teaching Sociology*.

_____. 2004c. "The Role of Status in Evaluating Research Practice." *Social Science Research* 33(3): 521-537.

Leahey, Erin, Barbara Entwisle and Peter Einaudi. 2003. "Diversity in Everyday Research Practice: The Case of Data Editing." S*ociological Methods and Research* 31:4: 63-89.

Long, Scott J., and Robert McGinnis. 1981. "Organizational Context and Scientific Productivity." *American Sociological Review* 46:422-442.

Macrina, Francis L. 2000. *Scientific Integrity: An Introductory Text with Cases*. ASM Press.

Maynard, Douglas W., and Nora Cate Schaeffer. 2002. "Standardization and Its Discontents." *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, edited by D. Maynard, H. Houtkoop-Steenstra, N. C. Schaeffer, and J. v. d. Zouwen. John Wiley & Sons, Inc.

Mohr, Lawrence B. 1990. *Understanding Significance Testing*. Sage Publications.

Morgan, Stephen L., and Aage B. Sorensen. 1999. "A Test of Coleman's Social Capital Explanation of School Effects." *American Sociological Review* 64:661-681.

Morrill, Calvin, David B. Buller, Mary Klein Buller and Linda L. Larkey. 1999. "Toward an Organizational Perspective on Identifying and Managing Formal Gatekeepers." *Qualitative Sociology* 22:51-72.

Morrison, Denton, and Ramon Henkel. 1970. "The Significance Test Controversy." Aldine Publishing Company.

Neavill, Gordon B. 1975. "Role of the Publisher in the Dissemination of Knowledge." *The Annals of the American Academy of Political and Social Science* 421:23-33.

Phillips, Damon J., and Ezra W. Zuckerman. 2001. "Middle-Status Conformity: Theoretical Restatement and Empirical Demonstration in Two Markets." *American Journal of Sociology* 107:379-429.

Pinch, Trevor J. 1986. Confronting Nature: *The Sociology of Solar-Neutrino Detection*. Dordrecht: D. Reidel.

Platt, Jennifer. 1996. *A History of Sociological Research Methods in the United States*. Cambridge University Press.

Policy, Institute of Medicine Division of Health Sciences. 1985. *Assessing Medical Technologies: Report of a Study*. National Academy Press.

Raftery, Adrian. 2001. "Statistics in Sociology, 1950-2000: A Selective Review." *Sociological Methodology* 31:1-45.

Raftery, Adrian E. 1995. "Bayesian Model Selection in Social Research." *Sociological Methodology* 25:111-163.

Rogers, Everett M. 1962. *Diffusion of Innovations*. Free Press.

Rubin, Allen. 1985. "Significance Testing with Population Data." *Social Service Review* 59:518-520.

Schmidt, Frank L. 1996. "Statistical Significance: Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers." *Psychological Methods* 1:115-129.

Schmidt, Frank L., and John E. Hunter. 1997. "Eight Common but False Objections to the Discontinuation of Significance Testing in the Analysis of Research Data." Pp. 37-64 in *What if There Were No Significance Tests?* edited by L. Harlow, S. A. Mulaik and J. H. Steiger. Lawrence Earlbam Associates.

Schneider, Anne L., and Robert E. Darcy. 1984. "Policy Implications of Using Significance Tests in Evaluation Research." *Evaluation Review* 8:573-582.

Selvin, Hanan C. 1957. "A Critique of Tests of Significance in Survey Research." *American Sociological Review* 22:519-527.

Simon, Rita J., and James J. Fyfe. 1994. "Editors as Gatekeepers: Getting published in the Social Sciences." Lanham, MD: Rowman and Littlefield.

Soule, Sarah A. 1999. "The Diffusion of an Unsuccessful Innovation." *Annals of the American Academy of Political and Social Science* 566:120-131.

Strang, David, and Michael W. Macy. 2001. "In Search of Excellence: Fads, Success Stories, and Adaptive Emulation." *American Journal of Sociology* 107:147-82.

Strang, David, and Sarah Soule. 1998. "Diffusion in Organizations and Social Movements: From Hybrid Corn to Poison Pills." *Annual Review of Sociology* 24:265-290.

Timmermans, Stefan, and Marc Berg. 1997. "Standardization in Action: Achieving Local Universality through Medical Protocols." *Social Studies of Science* 27:273-305.

Turner, Stephen Park, and Jonathan H. Turner. 1990. *The Impossible Science: An Institutional Analysis of American Sociology*. Sage Publications.

Victor, Bart, and John B. Cullen. 1988. "The Organizational Bases of Ethical Work Climates." *Administrative Science Quarterly* 33:101-125.

Viterna, Jocelyn S., and Douglas W. Maynard. 2002. "How Uniform is Standardization? Variation Within and Across Survey Research Centers Regarding Protocols for Interviewing." *Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview*, edited by D. Maynard, H. Houtkoop-Steenstra, N.C. Schaeffer and J.v.d. Zouwen. John Wiley & Sons, Inc.

Walton, John. 1966. "Discipline, Method, and Community Power: A Note on the Sociology of Knowledge." *American Sociological Review* 31:684-689.

Wejnert, Barbara. 2002. "Integrating Models of Diffusion of Innovations: A Conceptual Framework." *Annual Review of Sociology* 28:297-326.

Winship, Christopher, and Larry Radbill. 1994. "Sampling Weights and Regression Analysis." *Sociological Methods and Research* 23:230-257.