

Journal of Econometrics (in press)

“Waiting for Life to Arrive”: A History of the Regression-Discontinuity  
Design in Psychology, Statistics and Economics

Thomas D. Cook

Northwestern University

Thanks are due to Richard Berk, Glen Cain, Arthur Goldberger, Guido Imbens, George Knafl, Thomas Lemieux, William Shadish, Clifford Spiegelman, William Trochim and Vivian Wong for feedback on prior drafts. They are not responsible for any errors of fact or taste.

## Introduction

The regression discontinuity design (RDD) occurs when assignment to treatment depends deterministically on a quantified score on some continuous assignment variable. This score is then used as a covariate in a regression of outcome. When RDD is perfectly implemented, the selection process is fully observed and so can be modeled to produce an unbiased causal inference.

This paper is about the history of RDD. Although I am not a trained historian, I know enough to respect the primacy historians place on documenting events and trends. I also know that interpreting these events and trends has to be conditioned by independent knowledge of temporal sequence, by archived specifics that are relevant to the explanations offered, and by recourse to substantiated theories of individual and institutional behavior. Fortunately, most of this paper is about such events, trends and interpretations. But a few parts are not, and historians may well become nervous when I try to interpret events that might have happened but did not. Although some historians are developing a taste for counterfactual or virtual history (Ferguson, 1997), it is deservedly a minority taste. As an amateur historian, I will almost certainly fall into other traps professionals learn to avoid. Of those I can recognize, one is the teleological trap of inferring inevitable-seeming links between past events when, with more secure footing in the original time and place, these events might seem more contingent and many other futures possible. Another problem is that I am not an independent commentator on RDD. I have been peripherally involved in its history, albeit as a disseminator and not a theorist or practitioner. I was also marginally involved in the Northwestern University theory group that developed the design in the early 1970's after its discovery earlier (Thistlethwaite & Campbell, 1960). Doubtless I know more about RDD's history in that context than about other attempts to develop, disseminate or evaluate the design. While I have read many of the original sources reported on here, I have probably relied on secondary sources more than a real historian would. The relatively recent history of RDD has helped me, though, since many of the method's pioneers are still alive and have offered commentary on earlier drafts of this paper as it touched on their work. I have tried to incorporate their memories and sensitivities into this final version, sometimes even citing their notes to me about their work. Nonetheless, the following account is mine and not theirs; and while I respect the simplest norms of writing history, I cannot hope to dip deeply into the historian's bag of analytic tools. So, caveat lector.

This is not the first historical account of RDD. Donald Campbell, the design's originator, wrote his own version of the design's early history (Campbell, 1984), and various scholars have given snapshots of its history since then (Trochim, 1984; 2001). However, the present account is more current, detailed and interdisciplinary than its predecessors. Indeed, it is organized around academic disciplines, tracing the history of the design in Psychology and Education, then in Statistics and Biostatistics, and then in

Economics. The account speaks to many themes, including the repeated re-invention of the design across these disciplines. This was often done invoking different names for the design, the upshot being that RDD has not attained consistent “brand” status across the various behavioral, social and health sciences. Another theme speaks to the design’s differential waxing and waning by discipline, trying to describe and explain what happened. RDD was invented and initially developed in Psychology and Education, but interest in it waned there after about 1990. It has never had much visible growth in Statistics, though it was acknowledged there. And in Economics RDD had a serendipitous birth, a long period of neglect, and then a renaissance after about 1995. This special journal number is part of that revival. Since its invention in 1960, RDD has been, in Samuel Beckett’s words: “waiting for life to happen”. Will this revival breathe life into the design in Economics and, who knows, even beyond?

### Psychology and Education

No doubt exists that the first publication on RDD was an application in education by two psychologists, Thistlethwaite and Campbell (1960). No doubt also exists that Campbell was the initiator and that he continued to work on the topic while Thistlethwaite did not. What is less clear is the intuition that led Campbell to develop the design. To probe this we go to Campbell and Stanley (1963) since it provides more conceptual clarification than the earlier paper. This clarification did not involve statistical proofs, though, for Campbell was not a formal statistician. He operated from intuition and analogy embedded in deep knowledge of Fisher’s work on design structure, Brunswik’s work on representative design, and Popper’s work on epistemology, especially as regards the merits of falsification. However, Campbell regularly sought contact with statistically sophisticated colleagues, both senior and junior, though in his own oral account to me he claimed that Stanley provided little to the thinking about RDD in Campbell and Stanley.

In that work, two themes stand out in Campbell’s explanation of RDD. The first involves selection bias. Campbell and Stanley write that RDD seeks to handle selection “through representing (it) in detail, not through equation”. That is, selection depends on a measured cutoff score on a continuous assignment variable. This cutoff fully determines treatment exposure, and the regression of outcome on assignment estimates how the entire assignment variable and outcome are functionally related so as to assess whether a discontinuity occurs in the regression at the cutoff point. This strategy requires maximizing selection differences since in RDD the treated and untreated groups should not overlap at all on either side of the cutoff. Experiments handle selection quite differently—by creating treatment groups that overlap on all observables and unobservables other than for treatment, thereby “equating” them. Campbell and Stanley’s reference to RDD representing the selection process “in detail” comes close to specifying what was later identified as its most important causal feature, that selection is completely known. Campbell realized how useful a well-described selection process is, implying that selection is not a problem per se—only unknown selection is.

However, Campbell and Stanley gave more prominence to a second rationale for RDD that turns out to have a more intuitive than formal statistical warrant. They imagine a “tie breaking experiment...for a narrow range of scores at or just below the cutting point” and invite readers to see RDD as an attempt “to substitute for this true experiment by examining the regression line for a discontinuity at the cutting point”. The implication here is that RDD should be considered as though it were an experiment limited to scores immediately around the cutoff. It is there that chance plays its largest role in determining treatment assignment and true scores their smallest role, certainly when compared to what happens at more extreme points on the assignment variable. Campbell and Stanley contend, then, that RDD is like a randomized experiment at the cutoff, but nowhere else; and that knowledge of the assignment cutoff and of the function relating assignment and outcome creates a “detailed” description of the selection process.

Campbell and Stanley make little explicit mention of a third rationale for RDD. Campbell’s early work on the design coincided with his work on interrupted time-series (ITS), including the British breathalyser study in which he took a special interest (Ross, Campbell & Glass, 1970). ITS depends on an abrupt intervention occurring at a known point in time. If causally effective, this intervention should then lead to a discontinuity in the functional form relating time and outcome, either when the intervention begins or after a theoretically predicted temporal lag. The analogy with RDD is clear, for in RDD a change in intercept is also predicted to occur at a very specifically predicted intervention point on a continuum, albeit not necessarily a time continuum. Although the similarity with ITS represents a plausible background feature of Campbell’s thinking about RDD at the time, it is not emphasized in his writings on RDD. More salient are its similarity to a tie-breaking experiment and its provision of detailed knowledge about the selection process.

In discussing the statistical analysis of RDD data, Campbell and Stanley point to the importance of parallel and linear slopes on each side of the cutoff. They propose several analytic strategies for this case, without being definitive about any of them. These include a t-test at the cutoff, and covariance analysis using the assignment variable and the cutoff score as covariates. But they also note the possibility of non-linear regressions and, to deal with this, they propose data transformations or non-parametric regression. The discussion of analytic strategies is brief and tentative but shows sensitivity to the complications of non-linearity. However, the discussion assumes the computation of two separate regressions, one each side of the cutoff, a feature that Campbell would later reject under advice that, in his 1984 history, he attributes to Boruch (1973).

One of Campbell’s habits was to work closely with a graduate student on a particular topic, promoting him or her as “the world’s authority” on the topic and thus as his tutor. The first of these on RDD was Joyce Sween. Her dissertation (1971) contained two innovations. The first was conceptual, and involved arguing that RDD should be treated as an experiment rather than an observational study. The logic is that a well-implemented RDD study entails perfect knowledge of the selection process, and that perfect knowledge of the selection process is what gives the experiment its inferential power. Although the equation of groups was central in Fisher’s rationale for random

assignment and in the subsequent rationales for randomized clinical trials, Sween saw this equation as merely another instance of the more general idea that causal inference is facilitated when the selection process is fully known. Campbell (1984) reported being leery about this argument on grounds that causal inference is most justified around the cutoff and so is more local and limited in external validity than in the experiment. He also noted that, even within the narrow range of scores around the cutoff, formal random assignment is absent with RDD, making it impossible to guarantee that chance alone determines on which side of the cutoff an individual unit falls. Even so, Sween planted a seed that would later be reinvented: Is it an empty exercise to distinguish between the experiment and RDD if each involves full knowledge of the selection process?

Sween's second innovation concerned how to handle non-linearity. She proposed many ways of describing, transforming and testing the function relating assignment to outcome. She particularly focused on under-fitting and over-fitting the functional form, demonstrating in Monte Carlo analyses that fitting a lower order function than the data warranted resulted in a biased causal inference. However, over-fitting did not, though it did reduce statistical power due to the polynomial and interaction terms required. On the assumption that bias is more important than precision, Sween's dissertation resulted in Campbell's group recommending over-fitting. Her concern with mis-specified functional form also influenced the real-world examples of RDD that Campbell encouraged Northwestern graduate students to conduct. Seaver & Quarton (1976) examined how being put on the Dean's List in college based on a cutoff GPA of 3.5 affected grades in the subsequent quarter. In part, this was to explore those RDD cases where an intervention is awarded to individuals for especially high merit (or especially pressing need) and so leads to a restricted range of scores on one side of the cutoff. How should functional form be modeled in this circumstance to test hypotheses about intercept and slope differences due to treatment? Seaver and Quarton decided that responsible tests of slope differences were impossible, given the curtailed distribution on one side of the cutoff. But after visually inspecting the regressions and interpreting them as linear, they then tested intercept differences and concluded there was a treatment effect. However, Sween fitted a quadratic model to the same data and showed that the discontinuity claimed with a linear fit disappeared with a quadratic one (reported in Cook & Campbell, 1979). The danger of mis-specified functional form, already known from theory and simulations, was now demonstrated with real data.

Sween left Northwestern by 1971 but remained a peripheral part of the theory group Campbell organized around RDD in the early 1970's. It would be a mistake to see RDD as a dominant intellectual interest for Campbell at the time. As summarized in Overman (1981), he was then also involved in many other intellectual pursuits touching on methodology writ large, evaluation, social psychology and epistemology. Nonetheless, RDD was important enough to him that he did spend systematic time on it between 1970 and 1975, with his interest tapering off thereafter until the mid 1980's when it was effectively zero. But in Psychology in the 70's, the new RDD group included Robert Boruch, Charles Reichardt and eventually, William Trochim. In Mathematics, it included two mathematical statisticians on faculty, Jerome Sacks and Rose Ray, and two graduate students, Clifford Spiegelman and George Knafl. The efforts of the graduate students in

Psychology and Mathematics was supported by a National Science Foundation grant to Campbell whose largesse also provided funds for outside visitors to spend up to two summers at Northwestern. One of these visitors was William Lohr who worked on how eligibility for Medicaid, based on household income, affected the number of physician visits in the year Medicaid was passed (Lohr, 1972).

This work was important for its eventual emphasis on an embellishment to the basic RDD design. To the physician visit data from the year after Medicaid, Lohr added similar data from the year before the program. This pretest allowed him to check whether the pretest and posttest regressions differed, especially in the untreated segment of the assignment variable where a treatment effect cannot have influenced the obtained intercepts or slopes. Adding the pretest RDD function improves the causal counterfactual over the basic RDD where the counterfactual is a simple (and sometimes heroic) extrapolation of the regression slope from the untreated into the treated segment of the assignment variable. The pre-intervention RDD function allows the analyst to examine the comparability of pre- and posttest slopes in the untreated segment and also to test whether the displacement at the cutoff is different when the treatment is present versus absent. It has further advantages. It increases statistical power, enhances the communicability of results as visual plots, and if the pretest and posttest functions do not differ in the untreated segment, it permits a more credible test of slope differences due to treatment. Campbell wrote the section in Riecken et al (1974) that first presented the Lohr design, though the text there is not explicitly couched in RDD terms. Nor does it sharply catch all the advantages that pretest information offers when compared to simple RDD designs that have to rely on statistical adjustments requiring strong assumptions about functional form. With pretest data, this form is directly observed, albeit not at the same time an intervention is implemented.

Riecken et al presented another modification to the basic RDD design-- implementing the intervention at more than one point on the assignment variable. This also renders the causal hypothesis more complex, since two or more discontinuities should now result. Data corroborating this expected pattern will make many alternative interpretations less plausible. External validity will also be increased because the treatment effect is now estimated, not just at the cutoff, but at two different points on the assignment continuum. I know of no uses of this RDD variant, perhaps because it requires very stable data over a wide range on the assignment variable. Even so, like adding a pretest RDD function, implementing the intervention at multiple cutoff points captures Campbell's usual preference to deal with validity threats by design rather than model-based adjustment.

In Psychology at Northwestern, Trochim came to assume Sween's mantle as local authority and tutor to Campbell. The main emphasis in his work has been to synthesize and disseminate what is known about RDD, to work with the mathematical statistician, Spiegelman, on the fuzzy discontinuity problem, to begin the empirical study of implementing the design, and to add to the repertoire of design variants for improving the bare-bones RDD design. His 1984 book introduced several other important innovations also. For one, he implicitly challenged Campbell's assertion that RD is a design capable

of few realizations in practice. He pointed to the ways it could be used in any merit or need situation or in a first come first served situation. He also stressed that the design could be used with any kind of continuous assignment variable, whether a single item or some composite metric. He painted RDD as capable of more applications than Campbell and Stanley (1963) had implied when they wrote: “While very limited in its range of applications, the presentation (of RDD) here seems justified by the fact that those limited settings are mainly educational”.

Trochim also empirically examined the design’s use by practitioners, taking advantage of a U. S. Department of Education research program to evaluate the national Title 1 Program where eligibility depends on a specific poverty criterion. In this case, though, the Department of Education did not learn about RDD from Campbell and his theory group, but from Tallmadge and Horst (1976) and Tallmadge & Wood (1978) who may have independently invented the design. In any event, the department let school districts choose how they would evaluate Title 1, utilizing RDD as one of the three possible choices. About 2% of the districts chose RDD, about 200 in total. This low percentage implies that, for whatever reasons, school districts did not find the design attractive; but still, the absolute number of choices enabled Trochim to explore why districts had and had not used the design, developing a long list of such reasons (Trochim, 1984). He also interviewed district officials about their experiences implementing the design, again producing a list of facilitating and impeding factors. This implementation study recreates what happened in the history of both sample surveys and experiments where the necessary statistical theory had to be complemented by empirically informed propositions about better ways to implement surveys or experiments in field settings. Only when an elegant statistical theory was linked to a hodge-podge theory of implementation could surveys and field experiments become practical research tools.

Trochim also sought to extend the range of factors that could be added to the basic design over and above a pretest function and multiple cutoffs. He called attention to one section of Reichardt’s (1979) dissertation that dealt with adding RDD control functions from a non-equivalent group of units. Thus, for a state-level intervention the RDD in a treated state could be compared to that from an adjacent state. Or better yet, researchers could sample two adjacent cities on the same states’ shared border in order to achieve a more local counterfactual function for RDD analysis. The problem here that the non-equivalent intervention and comparison settings might be associated with different regression functions. Simple intercept differences pose no major problem, though; the most relevant alternative interpretations require site differences that affect intercepts or slopes especially at or around the cutoff. The inferential problems with a non-equivalent control RDD are thus more akin to those associated with non-equivalent control series in interrupted time-series work. These are less numerous and opaque than the problems that beset simpler designs with non-equivalent groups and a single pretest measure on the same scale as the outcome. It is hard to say how plausible are selection differences that are local to the cutoff area. But it seems plausible to contend that they are least likely if high power tests have failed to show slope differences in the untreated segment of the assignment variable. To my knowledge, no examples yet exist of creating the

counterfactual from RDD functions that were obtained from an untreated, non-equivalent comparison group. But this possibility is real, and with very local matches to boot.

However, another variant Trochim discussed has already been used, as we see later with Black, Galdo & Smith (2005). This is “the trickle” or “batch” variant where treatment assignment occurs in time-dependent batches. This assignment practice is often used to ensure that there are enough cases to fill all the slots in local programs that vary over time in the pattern of client entries and exits. When this happens, even the cutoff score can vary over time in order to accommodate fluctuations in client flow. Sometimes, also, relatively few units are available per site, but there can be many sites. Then, different cutoffs can ensue per site in order to handle the local variation in client flows. Such heterogeneity of cutoff values complicates analysis, requiring that data be pooled across cutoffs that vary by site and even by period within sites. Even so, two potential advantages emerge: External validity increases because the causal hypothesis is not now restricted to a single cutoff value; and at each site/period combination, the design retains its sharp discontinuity. Fuzziness arises when the site data are pooled, but at other levels of analysis sharp discontinuities still exist that should play a major role in the choice of analytic strategy.

The final design variant discussed by Trochim (1984) is use of a non-equivalent dependent variable function. An example of this would be when a compensatory language arts program is given to some students based on a particular score on, say, a prior language arts-related course. After the intervention, students are then tested in language arts as well as in some other “control” topic. Imagine, first, that this other topic is mathematics under the assumption that both language arts and math are subject to most of the same social, psychological and biological background forces that affect achievement. Accepting this assumption means that the regression for mathematics can then function as the no-cause counterfactual for language arts. Of course, it would be even better if the treatment were designed to advance some quite specific language arts skill and if, in addition to the dependent variable assessing this skill, there were also measures of intervention-irrelevant language skills. These last skills could then serve as counterfactual RDD functions; and they would be less dissimilar to the treated skill than mathematics is, so long as there is no cross-over from the intervention-relevant language arts skills to the intervention-irrelevant ones.

Trochim emphasizes RDD pretests, non-equivalent groups and non-equivalent dependent variables in order to bring more relevant data to bear for modeling functional form. With multiple cutoff points and trickle assignment, the attempt is to broaden causal inference beyond a single cutoff point. But in every instance, the aim is to use design features to compensate for limitations in the basic RDD. Boruch (1975) pushed this design preference to its extreme when he argued for combining RDD and a randomized experiment in the same study. If, in theory, RDD is most like an experiment around the cutoff, and if bias-inducing deliberate misallocations sometimes occur around the cutoff, why not persuade those persons who have the necessary authority to do a randomized experiment in the contested area around the cutoff while simultaneously conducting an RDD study across the rest of the assignment variable? This combined option has several

advantages: truly unbiased causal inference around the cutoff, an increase in statistical power, and enhanced credibility in the eyes of those suspicious of RDD.

Trochim's concern with design variants of the basic RDD structure is an important contribution, even if empirical examples of each variant are not yet available. But his major achievement was to draw attention to fuzzy discontinuities. He did not coin the term. Campbell (1969) did when he came to realize that knowledge of a cutoff can create social dynamics unique to that point--dynamics that can bias regressions because some cases around the cutoff get treatment access that violates their assigned access. At the time, Campbell was also working on the social dynamics of implementing random assignment, for example noting how social workers sometimes ensured that families with special needs received the treatment the worker thought good for them and not the treatment the coin toss had assigned them (Campbell & Boruch, 1975). Similarly, with RDD he publicized results from the Irish school-leaving exam (Madaus & Greaney, 1985) that showed how students scoring just below the passing mark were underrepresented in the population while those scoring just above it were overrepresented. The presumption is that examiners gave students close to the cutoff extra points they did not truly deserve. Campbell's original solution to the fuzzy discontinuity issue stressed observing socio-political dynamics at the cutoff and determining an area on the assignment variable where misallocation was least likely. The regression discontinuity could then be compared at the boundary points of this area rather than at the actual cutoff point. This was a poor solution, though. It offers an imperfect description of the misallocation; it throws out data in the area of uncertainty and so reduces statistical power; it requires even stronger and less transparent assumptions about functional form; and it loses the virtue of being like a random assignment experiment at the cutoff. Something better was needed, and Trochim took on the fuzzy discontinuity problem with Clifford Spiegelman, by then no longer at Northwestern. Their work is discussed in the later section on Statistics since Spiegelman played the major creative role, Trochim being the indispensable synthesizer and interpreter to those not skilled in mathematical statistics.

While a faculty member at Cornell, Trochim worked closely with a graduate student, Joseph Cappelleri, who tackled the statistical power issue with RDD. The main insight here had come earlier from the econometrician, Arthur Goldberger (1972,a), who showed two things: that the experiment was about 2.75 times more efficient than RDD under the conditions Goldberger chose to explore; and that this decrement was due to the correlation between the assignment variable and cutoff in RDD that is not present with the experiment. In a series of subsequent simulation papers, Cappelleri extended Goldberger's work by showing how this power advantage varies with where the cutoff point is located on the assignment variable, with the size of the actual effect, and with the percentage of cases that are randomly assigned in a tie-breaker study (Cappelleri, 1991; Cappelleri, Darlington & Trochim, 1994). In none of the assessed circumstances did the efficiency of RDD equal that of an experiment with the same number of cases. And since the reason for the differential is known, Cappelleri's work allowed Shadish, Cook and Campbell (2002) to develop in their Table 7.2 a list of the factors under a researcher's control that can be used to increase power in RDD work.

Together with Trochim, Cappelleri also took on the task of persuading the medical research community of the desirability and viability of combining RDD and an experiment in the same study (Trochim & Cappelleri, 1992; Cappelleri & Trochim, 1994, 1996). The context they emphasized was the ethical constraint that precludes doing experiments with those sick people most in need. They proposed giving such individuals privileged access to treatment, creating groups of the highest and lowest scorers that could be analyzed as an RDD study. However, among patients with moderate symptom levels they proposed forming a treatment and control group through random assignment, thus recreating Boruch's tie breaking experiment. This suggestion did not attain much resonance, perhaps because of its complexity or because medical researchers want to use random assignment even among volunteer patients in great physical need. Nonetheless, the effort of Cappelleri and Trochim is important in another way. Whereas Trochim had earlier used the term "regression-discontinuity" without much apparent hesitation, by the early 1990's he and Cappelleri were instead referring to it as the "cutoff-based design". Did they find "regression discontinuity" so alien that it endangered its own broader adoption, necessitating a different "brand name"?

The next real contribution from psychologists was by Aiken, West, Schwalm, Carroll & Hsuing (1998). Although simulation studies had already compared the results of experiments and RDD (Trochim & Cappelleri, 1992), Aiken et al. conducted the first ever within-study comparison of RDD and experimental results. For the RDD study, they took advantage of the fact that students entering Arizona State University with ACT or SAT scores below a given threshold were required to take a remedial writing course rather than the standard writing course. The study outcome was writing ability at the end of the standard writing class that remedial students also had to take eventually. The randomized experiment took students from a narrow but unspecified segment just below the standardized cutoff scores and randomly assigned them to the remedial or non-remedial course. Using an ANCOVA analysis that makes strong assumptions about parallel and linear regressions, the authors showed that the RDD estimates were quite close to the experimental ones. By itself, this study is not definitive. The effect is not estimated at quite the same point in the experiment and RDD study; it is not clear from the reported particulars whether the functional form was appropriately modeled; and the performance outcomes were assessed at different times in the treated and control conditions, though some data suggested this may not have been a problem. (Two other attempts to contrast RDD and experimental estimates in the same study will be reported on later, each again showing that the two design types did not result in different causal conclusions).

From the mid 1960's on, many attempts were made by Campbell and his colleagues to popularize RDD. It was discussed in Campbell and Stanley (1963) and in greater detail in books on social experimentation for which Campbell wrote sections (Riecken et al., 1974; Bennet and Lumsdaine, 1975). RDD was also extensively discussed in texts on quasi-experimentation (Cook and Campbell, 1979; Shadish, Cook and Campbell, 2002), and evaluation (Judd & Kenny, 1981; Mohr, 1985). Trochim (1984) wrote the first book devoted exclusively to the method. While the book's cover title in

caps reads Research Design for Program Evaluation, its sub-title in non-caps and only about a quarter as large in size as the title reads: the regression-discontinuity approach. It is as though program evaluation were the emphasis of the monograph and not the novelty of RDD.

Despite this publicity, RDD has not been widely adopted in Psychology or Education. Campbell and Stanley probably did not help in this when they introduced RDD as a method “very limited in its range of applications... (that are) mainly educational”. This is not a ringing endorsement, particularly when coming from the method’s own developer! Time has shown that Campbell’s early judgment was off the mark, and that the range of application is potentially large. Any continuous variable can be used for assignment; and many nations already have allocation principles that lend themselves to RDD because allocation by merit, need, first come first served, or date of birth is valued. Indeed, even more emphasis could probably be placed on them, given how normatively entrenched they are. Even cronyism lends itself to RDD, so long as those spreading their largesse are willing to progressively prioritize whom they wish to benefit! Also relevant for the practicality of RDD is that, as the Lohr (1972) example showed, it can be used with any descriptive longitudinal data set, or combination of sets, so long as they permit distinguishing an assignment variable, a cutoff and an outcome. In some ways, RDD is more flexible than experiments because RDD does not require researchers or their proxies to directly manipulate the independent variable.

Of course, some applications of RDD had dribbled into the literature in Psychology and Education; and at the institutional level, RDD was recently given special status at the Institute for Educational Sciences (IES) for when an experiment is not possible. Among the large national evaluations now underway at IES, it is being used to evaluate Reading First and Early Reading First as well as to evaluate the effects of part of the Bush administration’s No Child Left Behind legislation. But even the Reading First cases result from a misadventure, for the Congressional study authorization was inadvertently written in a way that gave school districts an easy out from participating in an experiment. Once this loop-hole was closed, later evaluations from IES have been almost all experimental. So IES’ assignment of special status to RDD hardly undermines the conclusion that its influence has been generally weak among researchers who were trained in Psychology and Education. Even the scholars who built all or some of their professional reputations around RDD became discouraged. Sween and Boruch stopped studying the issue as early as 1975; Campbell all but gave up the design after 1985; and Trochim and Cappilleri stopped about 1995. No-one from Campbell’s theory group remained actively working on the design; nor did their students.

This is ironical, for many of the theoretical, statistical and logistical issues with RDD had already been identified and solved by then; the design was discussed with approval in broadly disseminated research design and evaluation texts; successful instances of the design’s application were available for those who chose to look for them; and work had progressed on identifying factors that affect the design’s implementation in real-world education contexts in particular. By about 1990, RDD had become an established tool in Psychology and Education. But it was ossified; not a living tool that

generated novel theory and functioned as the template for numerous empirical applications whose causal yield was superior to that of other non-experimental methods. After decades of concerted but perhaps overly local and even in-bred effort by members of Campbell's extended theory group, by the mid 1990's RDD was still "waiting for life to arrive" in both Psychology and Education.

## Statistics

For the purposes of this paper I understand statistics in terms of scholars trained in that field, whether in a Department of Statistics, Mathematics or Biostatistics. Also included are those scholars trained in other fields who later came to hold their major academic appointment in a Statistics Department.

Rubin (1977) is the first published article I could find in Statistics that mentions RDD. This paper has been portrayed as another independent invention of the design together with a formal proof (Trochim, 1984; 2001; Shadish et al, 2002). But it is hardly an independent invention, since the paper cites Campbell and Stanley (1963) on RDD and also mentions Goldberger (1972, a). Nor is it a proof of RDD in the normal sense, given that Rubin's aim is broader than Campbell's. Rubin wanted to show that assignment on the basis of a covariate can lead to unbiased causal inference across the entire range of the assignment variable,  $X$ . He stipulates that all units with the same score on  $X$  must either receive the same treatment or be randomly assigned to the treatments under test. His approach uses the  $X$  points where random assignment occurs to estimate treatment and control regressions and, via model-fitting procedures that he describes, this then disciplines interpretation of the values obtained where the treatment assignment was deterministic rather than random. The tie-breaking experiment of Boruch (1975) is a special case of this model--the one when random assignment takes place symmetrically around the cutoff and deterministic assignment occurs elsewhere. But the model also captures classic RDD as another special case--where assignment occurs at only one point on  $X$  and all cases falling on any one side of this  $X$  receive the same treatment or control status. Rubin recognized this link to RDD, writing: "If the  $X$  values in the two samples do not overlap (e.g., as in the regression discontinuity design, Campbell & Stanley, 1963, pp 61-64) it is impossible to check the accuracy of (the regressions) for the full range of observed  $X$  values, and we must rely on a priori assumptions. Consequently, in order for the model-fitting efforts described above to be useful in practice, we must either have samples that overlap or strong a priori information about the functional forms" (p.11). Rubin does not detail how to get this a priori information, and it is not necessary for him to do so if the treatment and control regressions can be directly observed at the dispersed points along  $X$  where random assignment has occurred. Indeed, all the study examples that Rubin cites to illustrate his model involve random assignment at some points on  $X$ , though this is not a necessary condition for his model as his allusion to the relevance of RDD makes clear. But its presence in all the examples he does provide may be a measure of the importance Rubin attributes to directly observing functional forms and thus not having to assume them as in simple RDD. It may also reflect, though, his desire to

generate a model of assignment on the covariate that is more general than RDD because cause is estimated at multiple points on X rather than at a single point.

The next stage in Statistics is somewhat bizarre. At Northwestern, Campbell felt the need for more statistical expertise than he or his Psychology collaborators possessed. In 1973, he began the earlier mentioned active collaboration with faculty and graduate students in the Mathematics Department. Led by Sacks, the major effort they undertook was to estimate functional form. Their first relevant published paper was in many ways the most important (Sacks and Ylvisaker, 1978). As interpreted to me in a letter of April 30, 2006 by his student, George Knafl, Sacks' purpose was to use "linear combinations of the observed outcome values to estimate a special kind of nonlinear relationship, which he termed 'approximately linear'. By that he meant that the relationship was linear plus a bounded deterministic (nonrandom) error term. He first maximized the mean square error over this deterministic error and then minimized that maximum to determine the weights to assign to each observed outcome value... An observed outcome value's weight depends on how close its associated x value is to the "origin" of the current estimation problem (which need not always be 0). ...In the regression setting, ...the approximate straight line model in terms of a single predictor variable X with origin at  $x_0$  would have the following form for each x value:  $E(Y|X=x)=a+b(x-x_0)+r$  where r is the deterministic error term and is bounded by  $m(x-x_0)^2$ ...A separate linear estimate is computed for each possible predictor value  $x_0$  treating it as the associated 'origin'. Thus, the Sacks-Ylvisaker method generates a special kind of ... local nonparametric regression technique in that it is applied at each local  $x_0$  value separately to generate the regression curve as opposed to a global approach which would estimate the curve over its whole range of x values simultaneously". The paper has another important attribute brought out by its authors: "In the ideal linear model observations farthest from the origin are weighted most heavily but in the approximately linear model the observations closest to the origin have the most weight" (p. 1123).

At issue here is a set of functional forms more general than Campbell and his collaborators had developed with their emphasis on parallel and linear regressions and deliberate over-fitting. These newer methods stressed nonparametric regression in general and approximately linear relationships in particular. They also included concern for weighting observations more heavily if they are closer to a specified origin (i.e., the cutoff) and estimating effects locally at that origin/cutoff. The next step in this work was to estimate model-robust confidence intervals and bands for the approximately linear slopes discussed in Sacks & Ylvisaker (1978). A series of papers did this (Knafl, Sacks & Spiegelman, 1982; Knafl, Sacks, Spiegelman & Ylvisaker, 1984; Knafl, Sacks & Ylvisaker 1985). At the same time, the Northwestern mathematical statisticians were exploring other nonparametric regression themes. They showed, for instance, that window estimates of a nonparametric regression are "universally" consistent (Spiegelman & Sacks, 1980), just as Stone (1977) had shown for nearest neighbor estimates. Spiegelman (1980) even did some work on errors of measurement when estimating slopes, a small part of the more general problem of ascertaining functional form. Of course, nonparametric regression already had a long history in Statistics prior to Sacks and Campbell (summarized in Silverman, 1986). What the Northwestern mathematical

statisticians brought to their work was sensitivity to those parts of the nonparametric agenda that had to be developed because they were especially relevant to analyzing RDD data.

Spiegelman's main contribution to the RDD agenda was his work on the fuzzy discontinuity problem. He wrote two papers and part of his dissertation around the topic (1976; 1977; 1979). The results were summarized in Trochim and Spiegelman (1980) and extended in Trochim (1984). The essence of the approach was to construct an estimated assignment variable for each unit. Its distribution resembled, not the step function of a sharp discontinuity, but an ogive whose slope value depended on how much mis-assignment had occurred. To construct this estimated assignment variable, Spiegelman used two methods. He either calculated the percentage of mis-assignments within narrow ranges on the assignment variable or used a nearest neighbor moving average model. He also weighted the regressions. To check on these procedures he conducted many simulations that varied whether there was an effect or not, whether assignment was by the pretest plus error or was by four different types of true scores with error, and whether the error was large or small. Five analyses were then conducted for each simulation, one using the raw assignment variable, two others using the relative assignment variable either with or without weighting, and the final two using the moving average model again either with or without weighting. Trochim (1984) reports the results as follows: "Estimates from the analyses based on real assignment are biased, except when the error is random". This is as expected. "The moving average estimates of relative assignment appear to yield unbiased estimates of gain for most of the models and conditions (and) are less biased than the ones from the average assignment method" (p. 165). Since the relevant tables under the circumstances built into this simulation show a slight benefit for weighting, the weighted moving average procedure fared consistently best overall in controlling for the selection bias due to a fuzzy discontinuity. The same procedure also proved to be superior when Trochim (1984) used it to reanalyze data from a Title 1 site, again suggesting the discovery of an apparently unbiased way of dealing with a fuzzy discontinuity.

Sacks' other graduate student, Knafl, came to Northwestern slightly later. Among other things, he collaborated with his mentor to produce a computer program that incorporated all the analytic procedures the Northwestern statisticians had developed. It focused on nonparametric regression and calibration, approximately linear models, differential weighting, and local estimation at a particular value (Knafl, 1984). Campbell (1984) reports that it was successfully used with various engineering and some education examples, and Knafl confirmed the engineering uses by letter. But, as Knafl acknowledges, the program was not used by subsequent RDD researchers or incorporated into standard data-analytic packages where it would have been useful for those instances where it is difficult to assume parallel and linear regressions. It would be another decade before young economists independently developed similar methods and estimation techniques.

It is striking how small was the influence of this work. Though two papers did appear in the Annals of Statistics and another in the Journal of the American Statistical

Association (JASA), others were published in lesser outlets or were not published at all. Spiegelman reported to me how difficult it was to get the work published in premier statistics journals. Moreover, the work has not been cited in general statistical discussions of causation, and certainly not in the narrower context of debates on RDD. This last is not so surprising, however, for I could not find a single instance of the phrase “regression discontinuity” in any of the papers by statisticians associated with Northwestern, though the Thistlethwaite and Campbell publication is sometimes referenced. Even in a work as comprehensive as a dissertation is supposed to be, Spiegelman (1976) thanks Campbell and Boruch for their intellectual support, cites the Thistlethwaite and Campbell article in the reference section, but never uses the words regression discontinuity. Spiegelman reported to me that this was because statisticians had little to gain by framing their papers around a novel design from the social sciences. They believed they would be better served by framing their work around themes then active in their discipline. So they wrote about the topics above without explaining that they were partly undertaken to solve data-analytic issues in RDD. Regression discontinuity was like the love “that dared not speak its name”.

The name came out of the Statistics closet, though, in the next appearance of the design I could find. Richard Berk had been a professor of Sociology at Northwestern in the 1970’s and had interacted with the psychologists there who introduced him to RDD. He left Northwestern for UCLA where he eventually transferred his billet into the Statistics Department. Berk and Rauma (1983) published a JASA paper in which they used the design in a criminal justice application. The innovation in this paper followed from the dependent variable being categorical, requiring Berk to devise generalized least square procedures that emphasized logistic and Poisson functions rather than the linear and polynomial ones previously considered. This, and a later paper in the same journal (Berk & de Leuw, 1999), called attention to the fact that the general linear model cannot do full justice to all the ways in which the assignment and outcome variables might be related, requiring new procedures for non-continuous dependent variables. But despite dealing explicitly with RDD and appearing in JASA, these two papers do not seem to have spawned any growth in interest in RDD within Statistics. The papers presented basically uses of the design, albeit with extension to categorical outcomes. They were not textended presentations or defenses of the design.

The next important references I could find involve a mathematical statistician and a bio-statistician (Finkelstein, Levin & Robbins, 1996 a, b). RDD is renamed in these papers, now being called the “risk-based allocation design”. It is presented as an unbiased way to achieve causal inference when it is unethical to withhold treatment from those at greatest risk. It is also presented as a design with no predecessors, and the recommended analysis is based on Robbins & Zhang (1988, 1989, 1991). The editors of the American Journal of Public Health published two back-to-back articles on the topic, the first a theory paper and the second containing some examples. The latter had a very long appendix that, in his invited editorial, Mosteller (1996) likened to a third article. To publish so much on the topic, including an editorial, suggests that the editors saw the risk-based allocation design as worthy of high profile. This was probably because it seemed to solve the selection problem, and estimation procedures and examples were

already on hand. Mosteller's editorial pricked this nascent bubble by pointing out that the featured new design was RDD and that its range of application was limited compared to other attempts then underway to deal with selection. Still, these two articles constitute the first extended and explicit presentation of RDD by statisticians, albeit writing in a practice-oriented outlet rather than a mainstream statistics or biostatistics journal. Otherwise, apart from a small and earlier symposium that included papers on the use of RDD in the health sciences (Sechrest, Bunker & Perrin, 1990), little else transpired to develop or promote the design at the intersection of Statistics and Public Health.

The Finkelstein et al papers were published 38 years after Thistlethwaite and Campbell and many years after efforts had begun to feature the design in Psychology and Education. Yet Levin and Robbins thought they were inventing something new; and the journal editors presumably thought the same in giving such prominence to the work. This suggests the design had not gained much visibility in the larger community of researchers. Otherwise, why create a new name for RDD? Why else did the journal's reviewers not catch the similarity between the design being proposed as a novelty and what was already known? What would have happened had Mosteller not written his editorial? That the design was reinvented and re-labeled by Finkelstein et al. so long after its discovery is powerful testimony to RDD's low profile among the cause-probing methods then available in Statistics and Biostatistics.

Overall, RDD has not fared well with statisticians. Those who came to the topic relatively early did not stay with it. Spiegelman reported being discouraged by the great difficulty of getting his papers published and, by 1981, had dropped this line of work for more professionally rewarding topics (personal communication, March, 2006). Sacks and Knafl did the same. With the exception of Berk's JASA papers, publications explicitly devoted to RDD have not appeared in the more prestigious Statistics outlets. Even Rubin, whose work on causation has achieved very high resonance in Statistics, mentioned RDD only in parentheses in his Journal of Educational Statistics paper on assignment by a covariate. And neither he nor his trend-setting colleagues who write on causation, Holland or Rosenbaum, have ever written in detail about RDD either conceptually or data-analytically. They know about it, of course, and sometimes mention it in their general writing on observational studies. But RDD does not loom large for these well-known statisticians of causation, and ignorance is not the reason for this. Taste is more likely. It seems fair to conclude that RDD has hardly been born in Statistics as a discipline, let alone is "waiting for life to arrive".

Spiegelman thinks that the reception was poor because he and his colleagues could not provide illustrative applications that would appeal to editors. He did not feel himself well enough acquainted with the social science examples then being generated at Northwestern; and he had no other examples to present from domains that he thinks statisticians traditionally prefer, like the hard sciences or engineering. But another explanation may also be viable. One of Paul Erdos' many idiosyncracies was to assign a dollar value to mathematical problems depending on the richness of the puzzle they presented. To statisticians, does RDD seem more like an Erdos 50 cent problem than a \$5 million one? After all, RDD deals with causation, but causation is not of particular

interest to many mathematical or survey statisticians. Even for statisticians interested in causation, once it is clear that the selection process is completely known in RDD, it is not very challenging to model that process. And when one further realizes that RDD solves the selection problem only when there is a cutoff score--and not for the more numerous and more complex situations where researchers encounter selection—then the generality of RDD seems limited. It seems even more limited when one realizes that causal inference is less problematic with RDD closer to the cutoff point on the assignment variable than away from it. Also, many statisticians are suspicious of distant extrapolations, and simple RDD requires projecting a functional form into data-less segments of the assignment variable, making it seem both dangerous and naïve. To add insult to injury, RDD is a problem already solved theoretically, and hence not even a puzzle. Of course, a few problems of analysis still remain unsolved, and new ones will surely emerge. Even so, many statisticians will anticipate that solving these problems will require only minor variations in already accepted practices for modeling functional form, weighting observations, dealing with treatment crossovers, pooling RDD estimates, etc. I could find no historical record clearly laying out the case that, as a problem or tool, RDD is worth closer to 50 cents than \$5 million. So it is only a hypothesis that the factors above explain why the most visible theorists of causal methods in Statistics have not spent much time on a design they know about.

## Economics

The earliest papers on RDD in Economics were by Goldberger (1972 a, b). These unpublished papers represent two main accomplishments for RDD theory, though they were only incidental to Goldberger's main purpose. The first accomplishment was a proof of the basic design, showing formally what Campbell had only intuited. The Goldberger's papers were based on the distinction between non-equivalent groups whose difference depends on true ability in one case, and on measured ability in the other. The immediate reason for framing the paper this way was an article by Campbell and Erlebacher (1970) who, using an educational example, had shown how attempts to adjust for group pretest differences in true scores lead to biased causal conclusions because the measurement of ability contains at least error and so leaves were claiming that all controls for group pretest differences are biased, Goldberger showed formally that bias occurs when non-equivalent groups do indeed differ on a pretest true score that is incompletely observed, but that bias does not occur when selection into treatment depends only on an observed pretest score, as in RDD. The following citation captures the essence of Goldberger's proof, with the descriptions in parentheses being mine: "The explanation for this serendipitous result (no bias when selection is on an observed pretest score) is not hard to locate. Recall that  $z$  (a binary variable representing the treatment contrast at the cutoff) is completely determined by pretest score  $x$  (an obtained ability score). It cannot contain an information about  $x^*$  (true ability) that is not contained within  $x$ . Consequently, when we control on  $x$  as in the multiple regression,  $z$  has no explanatory power with respect to  $y$  (the outcome measured with error). More formally, the partial correlation of  $y$  and  $z$  controlling on  $x$  vanishes although the simple correlation of  $y$  and  $z$  is nonzero" (p. 16). Goldberger (1972, b) generalized the argument by proving that

selection on an imperfectly observed true score also leads to biased estimates of the interaction of treatment and third variables, whereas selection on the measured pretest score does not.

The second contribution Goldberger (1972, a) made was to understanding the efficiency of RDD relative to that of the randomized experiment. Under the conditions he analyzed—that included a cutoff at the midpoint of the assignment variable—its efficiency is about 2.75 times less than an experiment with the same number of units. Goldberger was also able to show why this advantage occurred, basically because the assignment variable and treatment dummy are correlated in RDD whereas they are not in the experiment. Thanks to this work, another advantage of the experiment over RDD became evident for the first time. To its greater transparency about functional form could now be added its greater power to reject the null hypothesis for a given sample size.

These papers had a great influence on Campbell and his theory group. Once they understood that Goldberger had provided a formal proof of RDD, they cited the papers relentlessly, especially the first one. However, this is ironical, for Goldberger's purpose was not to provide a proof of RDD. In a letter dated March 22, 2006, Goldberger wrote that: "The key point is that my 4/72 paper was not about the regression-discontinuity design, but rather about the notion that pre-existing differences between treatment and control groups inevitably bias estimated treatment effects. My colleague Glen Cain had persuaded me by a simple example that the Campbell-Erlebacher argument couldn't be right. In attempting to capture his argument formally, I happened to use a deterministic treatment assignment rule, hence the discontinuity. I soon realized that deterministic assignment wasn't crucial: . . . I have never thought of my paper as *proposing* a regression-discontinuity design (Goldberger's italics). Rather it was in effect distinguishing between "selection on observables" and "selection on unobservables", a distinction that became a focus of the vast econometric literature on selectivity bias."

Even so, Goldberger's papers did reinvent RDD. He did not cite Campbell's work on the topic and his serendipitous discovery was obviously also an independent one. That he was not seeking to propose a new causal method is clear from the structure of his papers. They rigorously pursue the distinction between selection on imperfectly observed true pretest scores and selection on perfectly observed yet psychometrically fallible pretest scores, showing that only the first results in bias so that group pretest differences do not constitute a universal problem for causal inference. The great irony here is that Goldberger reinvented RDD in order to critique the more general causal thinking of the very person who had earlier invented RDD! So it is no surprise that, once Campbell realized that the limiting condition Goldberger had discovered was RDD, he could accept Goldberger's criticism. After all, it fell outside the bounds of how he and Erlebacher had tried to define their problem, with its emphasis on the futility of using fallible observed pretest scores to control for group differences in true pretest scores. In RDD, selection depends on observed scores that do not need to pretend they are true scores. An even greater irony, though, is that Campbell & Erlebacher did not even mention Campbell's own earlier discovery of RDD as one condition limiting their main argument about the inadvisability of using pretests to model selection. Perfect selection adjustments are

possible when group differences are fully observed. Goldberger clearly proved that, and the proof he developed he later learned to call RDD.

By 1972, Campbell was distressed by the statistical models then in use that sought to adjust for selection processes more complex than with RDD. He began urging the increased use of social experiments, denigrating the use of non-experiments other than RDD or ITS (Campbell, 1969; Campbell & Boruch, 1975). Goldberger's Wisconsin colleague, Cain (1975), interpreted this advocacy of experiments as undercutting the value of Economics as a discipline because experiments have played a minor role in the development of that discipline's core empirical knowledge base. Non-experiments have been the method of choice, making reliance on substantive theory and statistical models to adjust for selection central, especially when interrupted time series methods could not be used. Thanks to RDD, Cain could validly point to at least one form of statistical adjustment that works in real-world settings; and he correctly generalized this to point out that causal inferences are unbiased in all other cases where the selection process is completely known and perfectly measured. But RDD's dependence on a clear cutoff means that it is irrelevant to the many other research situations where economists strive to draw causal conclusions in the face of selection processes that are more complex and difficult to measure than with RDD. So the issue became: How to generate a valid theory of selection adjustments that does not depend on random assignment and is more general than RDD?

By about 1980, RDD had fallen off the radarscope in economics, not to re-emerge for another 15 years or so. The puzzle is why Goldberger and his Wisconsin colleagues did not make more of their serendipitous discovery. After all, Goldberger did not publish his two papers cited above; he did not do any further published work specifically devoted to deterministic selection on a covariate; nor did he co-publish with Cain and Barnow on the topic other than for a small section in Barnow, Cain and Goldberger (1978). Perhaps Goldberger did not see or appreciate the practical relevance of his discovery, for in the first paragraph of his first paper he writes as one apparently disinterested in practical application: "We propose to demonstrate this point (the absence of bias when selection is on a measured pretest) in terms of a highly idealized setting, that is a formal model". But it is more likely that, as his letter above suggests, he saw RDD as a limited special case of a much broader and richer problem, understanding specification error writ large, solving the missing variables problem more generally. Is he assigning RDD an Erdos-like 50 cents as a problem of minor importance? RDD's birth in Economics was not auspicious and, yet again, the design had to "wait for life to arrive".

It had to wait all the longer because, by the early 1970's, the task of generating a general model of selection bias had captured the interest of many influential economists. It is difficult to capture in brief all the energy and creativity that went into this intellectual agenda. But crudely, scholars were taken with two main approaches to causal inference. Some were taken with instrumental variable (IV) approaches that depend on discovering instruments correlated with selection but not with errors in the outcome. However, from the mid 1980's on, more and more economists came to realize how difficult it was to achieve professional consensus about many of the causal claims

emanating from this kind of application of IV methods. It was one thing to have a highly abstract, elegant and general theory about unbiased causal inference; and quite another thing for this method to generate specific applications whose causal conclusions fellow scholars saw as beyond dispute insofar as all the central assumptions were manifestly met. Other economists interested in causation were taken with being better able to theorize about the selection process and observe it validly, as Goldberger (1972, a) and Cain (1975) had emphasized in their work on selection on observables. The problem here is that in most economic applications the selection process is more complicated than with RDD. It is usually multivariate and perhaps often non-linear; and it is almost always imperfectly observed however well it is theorized. A uniquely visible variant of selection modeling was Heckman's theoretical work. Initial enthusiasm was high that his selection models would solve the selection problem over a broad range of applications, certainly broader than what RDD could achieve with its dependence on an observed cutoff score. However, within-study comparisons from LaLonde (1985) on have shown that Heckman's selection models regularly fail to recreate the same results as experiments that share the same treatment group (Glazerman, Levy & Myers, 2003; Smith & Todd, 2005; Cook, Shadish & Wong, 2005). And while Heckman knew of RDD early, it was not something he cared to feature and promote, given his larger agenda.

What should economists do, then, if causal issues are central to their field, if substantive theory alone can only take one so far in determining causal effects, and if the practical yield of the econometric selection agenda was interpreted as disappointing? In this intellectual climate, some younger economists determined to seek out other unbiased methods for testing causal propositions, and so RDD was re-discovered along with experiments and natural experiments. By about 1995, new life was breathed into RDD in Economics and among economists in Schools of Public Policy. Some of the fruits of this revival can be found in this special edition. I will not try to recreate the very recent part of RDD in Economics, for it is more like the present than history. But I do want to mention seven things that characterize this revival for me.

The first is its hesitant birth in the form we now know it. The first papers since Goldberger to take advantage of cutoffs on an assignment variable (Angrist & Krueger, 1991; Imbens & VanderKlaauw, 1995; Angrist & Lavy, 1999; VanderKlaauw, 1999) sometimes cited Thistlethwaite and Campbell (1958), thus acknowledging a link to RDD. But the general analysis was formulated within an IV framework rather than the current linear or non-linear regression one, thus reflecting what was probably the dominant causal analytic mode of the 1990's. Only with time have these applications been fully recognized as RDD.

A second important feature is the growing number of applications in the increasing number of research domains where applied micro-economists are active (e.g., Angrist & Krueger, 1991; Imbens & VanderKlaauw, 1995; Angrist & Lavy, 1999; VanderKlaauw., 1999; Lee, 2001; Buddelmeyer & Skoufias, 2003; Jacob & Lefgren, 2004a, b; Gormley & Phillips, 2005; Black, Galdo & Smith, 2005; Ludwig & Miller, in press; Bloom, Kemple & Gamse, in preparation; Rand Corporation, in preparation). While this growth in applications has little interest per se for most econometricians, it

belies Campbell's early assertion that RDD is not of general applicability except perhaps in education. Of course, Campbell's assertion is self-evidently true in the sense that the design requires a cutoff on a continuous variable. But this is not an uncommon situation in cultures where allocation by need, by merit, or on a first-come-first-served are legitimate and where some allocation decisions are made in terms of birth dates. The normative status of such allocation principles also means that it should be possible to increase their use in many countries. As applications accrue in Economics, and as published examples come to be recognized as RDD where the authors failed to do so (e.g., Joyce, Kaestner & Colman, 2006), it becomes even more feasible to probe the hypothesis that RDD is essentially a rare hot-house flower. Inevitably, it has limiting conditions; but this does not necessarily entail that its uses must be rare or trivial when it comes to advancing substantive theory or public policy.

It is probably on the data-analytic front that most activity is taking place in RDD studies in economics. Particular emphasis has been placed both on using nonparametric regression to better model functional forms and on weighting techniques to emphasize observations closer to the cutoff. Striking is the number of sensitivity analyses done per publication to assess the robustness of results across different specifications of functional form. Also important is work on the fuzzy discontinuity problem. Random assignment can function as an IV to examine treatment-on-treated effects as well as intent-to-treat effects (Angrist, Imbens & Rubin, 1996). The cutoff in RDD can function this same way, since it is not correlated with errors in the outcome once the assignment variable is accounted for (Hahn, Todd & VanderKlaauw, 2002). If this extension of the Angrist et al thinking holds up to deeper scrutiny, researchers can use RDD to examine both intent-to-treat and treatment-on-treated causal hypotheses, thus solving the treatment mis-assignment problem that leads to fuzzy discontinuities. Other striking data-analytic advances are the emphasis being placed on the accurate measurement of behavior around the cutoff and on pooling across sites (and times) that differ in cutoff score (Black, Galdo & Smith, 2005).

On the design front, appreciation for a pretest RDD function is growing in Economics, given the roles it plays in increasing power and independently observing the functional form across all of  $X$  before the treatment is applied. The assumption here is that the pretest function is a good proxy for those parts of the functional form that have to be missing once a treatment is in place (Jacob & Lefgren, 2004,a; Bloom, Kemple & Gamse, 2006). Moreover, awareness is increasing of the possibility of implementing the intervention at different cutoff points, thus making causal inference more general than around a single cutoff value while also promoting much needed research into the pooling of RDD estimates (Black et al., 2005). Less clear is the level of current appreciation for other ways of independently assessing the likely functional form where it is not directly observable on the assignment variable. Some of these design-based alternatives are outlined in Trochim (1984) and in this paper, and others surely need to be developed. The traditional econometric emphasis on modeling need not lead to neglect of design addenda that strengthen the basic RDD structure by seeking to make functional forms more observable and to study more than one cutoff point.

The recent spate of RDD studies by economists is notable for two empirical tests of the design's validity based on comparing RDD effect estimates to those from a randomized experiment on the same topic (Cook & Wong, in press). While Aiken et al did the first such study in 1998, Buddelmeyer & Skoufias (2003) went further than them with a somewhat more sophisticated analysis. They took the first wave of Mexican Progresa data and showed that similar patterns of statistical significance were obtained when within-village program eligibles and ineligibles were compared relative to when the contrast was of eligibles in randomly selected experimental and control villages. Even so, the local average treatment effect in the RDD work was not identical with the average treatment effect in the experiment (Imbens & Angrist, 1994). An even more sophisticated test was by Black, Galdo & Smith (2005) who took pains to estimate the experimental and RDD effects at the same average treatment effect, and who used multiple sensitivity tests that varied the smoothers used as well as how close particular samples were to the cutoff point. They too showed similar patterns of statistical significance and even effect sizes, the more so closer to the cutoff given the non-linearity in their data. In some senses, it is trivial to show this empirical correspondence between the two design types, given theoretical proofs that each is unbiased. More surprising is the correspondence in statistical significance patterns, given the power difference between experiments and RDD studies and given that Aiken et al. had fewer cases in the RDD study than the experiment! Even so, the correspondence in these three design replication studies is comforting, for in each case the experiment and the RDD study were implemented in complex social settings with error and perhaps even multiple small sources of bias. So the design replications the economists have done speak to the robustness of RDD as it has been carried out recently.

Institutionalizing a design is not just an intellectual matter. Much is also sociological. In Psychology and Education, intellectual work on RDD did not spread much beyond Campbell's small theory group, and applications have been sparse. Campbell was such a polymath that RDD accounted for but a small fraction of his intellectual energy that, in the period from 1960 to 1985, was more lavishly devoted to work on evaluation theory, randomized experiments, interrupted time series, social comparison, evolutionary epistemology and meta-science. None of his students stayed with the topic or succeeded in handing it on to the next generation. RDD faded, in part because Campbell's theory group was in-bred and either failed or did not try to attract new intellectual blood. In Statistics, RDD was never really seriously taken up after the links between Campbell and the Northwestern mathematical statisticians broke up; no champions or theory groups systematically developed the idea, and its original proponents dropped it. The recent situation in Economics is quite different. For one, RDD is associated with many individuals who teach at many different universities, including some of the most prestigious ones in the USA. Individuals like Card at Berkeley, Imbens at Harvard, Smith at Michigan, Todd at the University of Pennsylvania and Vander Klauwe at North Carolina are not "marginal" economists who publish in less respected journals. Their names carry a high potential for legitimacy conferral nationally and beyond. Indeed, RDD is now being used in Italy and Mexico, and in one Italian instance it is even being used as the causal benchmark against which results from observational studies are compared (Battistin & Rettore, 2005). To add to the sense of growing

institutionalization, RDD is routinely discussed at economic conferences, and the entire number of a prestigious journal has now been devoted to the topic. These are important steps towards institutionalizing the design in Economics, though not guarantees of the success of this effort.

These are early days for predicting anything about RDD's eventual institutionalization in the Pantheon of widely used causal methods in Economics. But a theory group has at least developed; this group is more institutionally distinguished and nationally dispersed than in the other fields where RDD has been used; and the young opinion-leaders exploring the design bespeak general good taste in econometrics rather than a narrower purview limited just to RDD. This augurs well for new and rich life being breathed into RDD, especially if most of the design applications now under way turn out to be clear in their substantive findings and capable of generating new problems. Whether this upswing of interest in Economics will carry over into Psychology is unclear, though its prospects in Statistics remain grim unless it is possible to identify a rich seam of new issues of general theoretical relevance to statisticians.

### Conclusions

Several themes stand out in the half century of RDD's history. One is its repeated independent discovery. While this augurs well for the design's validity and relevance across fields, one circumstance of the reinventions has been strange. Campbell first named the design regression-discontinuity; Goldberger referred to it as deterministic selection on the covariate; Sacks and Spiegelman studiously avoided naming it; Rubin first wrote about it as part of a larger discussion of treatment assignment based on the covariate; Finkelstein et al called it the risk-allocation design; and Trochim finished up calling it the cutoff-based design. RDD comes across as the design that dares not speak its name, as though it is off-putting as a neologism. Of course, independent reinventions may spawn new labels because the existing one is not known. And reinventions that are incidental to some other purpose may not deserve a name because the discovery is, after all, incidental. Moreover, clear framing of the design's purposes and warrants can be considered more important than giving the specification so achieved a unique name. Motivations for the various design names are not clear in the historical record except for Trochim's (1990) intimation that regression-discontinuity is an unfortunate name whose meaning many people cannot immediately intuit. The dissemination of RDD has probably suffered from local labeling practices across disciplines, and so it is a relief to note that economists are unabashedly using the original Campbell terminology as they seek to revive the design that now dares to speak its name—at least in a social science like Economics where regressions and discontinuities are like mother's milk.

RDD's validity is hardly in doubt, given the proofs, the reinventions and the successful uses over 50 years. So judgments of its importance will likely depend more on the scientific consensus achieved about its generality. This last may not be as limited as first appears. Causal inference is most warranted near to the cutoff. Yet if the repertoire of sensitivity tests for functional form can be routinely extended to include a pretest, a matched but non-equivalent group or, at last resort, a non-equivalent dependent variable,

then better counterfactual estimates of functional form will be possible and reliance on unobserved extrapolations will be less. By moving to a difference in differences strategy for RDD, we may achieve reasonable estimates of whether a causal conclusion should be limited only to the cutoff area. Also, as applications pile up we may learn more about the conditions under which parallel and linear regressions are found. These imply a causal inference beyond the cutoff area, even with the most basic RDD design.

Treatment mis-allocations undermine confidence in causal claims, whatever the method used; and it is not yet clear if the problem is more serious with RDD than with experiments, say. Fuzzy discontinuities force a greater reliance both on the restricted set of cases with a sharp discontinuity and also on the validity of techniques to control for fuzziness. There is real recent cause for hope on this last score. In many contexts, the cutoff value can function as an IV and engender unbiased causal conclusions, though work is clearly needed to explore this potential solution in greater depth. The same is true of Spiegelman's continuous treatment functions. Fuzzy assignment does not seem as serious a problem today as earlier.

But RDD still only applies when assignment is by a cutoff score. How often this occurs in our society, and whether we can capitalize on norms about allocation by need, merit, birth date, and 'first come, first served', will determine much of the practical future value of RDD. I have no problem seeing it as important among other cause-testing arrows in the quiver of every non-doctrinaire social scientist. Many different causal methods are needed for the many different contexts in which quality causal inference is needed. Social scientists have been pursuing a more general pot of gold for three decades now--the discovery of methods for warranting causal inference over many different selection processes that are more complex than RDD. Solving that puzzle is indeed a problem worth Erdos' symbolic \$5 million. But must belief in this larger goal inevitably devalue RDD, explicitly or implicitly? Is not unbiased causal inference useful even when it is restricted to clearly demarcated circumstances that actually occur with some frequency in a given society? The causal silver bullet many theorists of method are seeking is indubitably preferable; and Nobel committees are right to acknowledge those who seek to crack this bigger selection problem. But arrows in a hunter's quiver can still promote physical survival even if they cannot fell prey as big as a bullet can. And let us be crystal clear. At this time, the silver bullet is a legitimate aspiration, not a current achievement like RDD is.

No advocates of RDD have seen it as superior to the randomized experiment or even equivalent to it in terms of warranting causal claims. RDD is less statistically powerful; it involves less transparent assumptions about functional form; its implementation is less well empirically understood; and methods for improving its implementation are less developed. The rationales for RDD are (1) that it can be used in a circumscribed set of circumstances where an experiment might not be feasible; (2) that it is then superior to all other known causal methods; and (3) that it can sometimes be combined with an experiment and other design features to extend causal inference along more of the assignment variable. To be bias-free in theory, as RDD is, is not necessarily to be as assumption-free or as efficient as other methods.

## References

- Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J., & Hsuing, S. (1998). Comparison of a Randomized and Two Quasi-Experimental Designs in a Single Outcome Evaluation: Efficacy of a University-Level Remedial Writing Program. *Evaluation Review*, 22(4), 207-244.
- Angrist, J.D., Imbens, G. W., & Rubin, D. B. (1996). Identification of Causal Effects using Instrumental Variables. *Journal of the American Statistical Association*, 91, 444-472.
- Angrist, J.D., & Krueger, A. (1991). Does Compensatory School Attendance affect Schooling and Earnings. *Quarterly Journal of Economics*, 106, 979-1014.
- Angrist, J. D., & Lavy, V. (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *Quarterly Journal of Economics*, 114, 533-576.
- Barnow, B. S., Cain, G. C., & Goldberger, A. S. (1980). Issues in the Analysis of Selectivity Bias. In E. W. Stormsdorfer & G. Farkas (Eds.), *Evaluation Studies Review Annual* (Vol. 5). Beverly Hills, CA: Sage Publications.
- Battistin, E., & Rettore, E. (2005). *Ineligibles and Eligible Non-Participants as a Double Comparison Group in Regression-Discontinuity Designs*. Unpublished manuscript, London, UK.
- Bennet, C. A., & Lumsdaine, A. A. (1975). *Evaluation and Experiment*. New York: Academic Press.
- Berk, R. A., & de Leuw, J. (1999). An Evaluation of California's Inmate Classification System Using a Generalized Regression Discontinuity Design. *Journal of the American Statistical Association*, 94(448), 1045-1052.
- Berk, R. A., & Rauma, D. (1983). Capitalizing on Nonrandom Assignment to Treatments: A Regression-Discontinuity Evaluation of a Crime-Control Program. *Journal of the American Statistical Association*, 78(381), 21-27.
- Black, D., Galdo, J., & Smith, J. C. (2005). *Evaluating the Regression Discontinuity Design Using Experimental Data*. Unpublished manuscript.
- Bloom, H., Kemple, J., & Gamse, B. (2005). Memo on the Evaluation Design of the Reading First National Impact Study. In I. o. E. Sciences (Ed.). Washington, DC.
- Boruch, R. (1973). *Regression-Discontinuity Designs Revisited*. Unpublished manuscript, Evanston, IL.
- Boruch, R. (1975). Coupling Randomized Experiments and Approximations to Experiments in Social Program Evaluation. *Sociological Methods and Research*, 4, 31-53.
- Buddelmeyer, H., & Skoufias, E. (2003). *An Evaluation of the Performance of Regression Discontinuity Design on PROGRESA*. Bonn, Germany: IZA.
- Cain, G. C. (1975). Regression and Selection Models to Improve Nonexperimental Comparisons. In C. A. Bennet & A. A. Lumsdaine (Eds.), *Evaluation and Experiment* (pp. 297-317). New York: Academic Press.
- Campbell, D. T. (1969). Reforms as Experiments. *American Psychologist*, 24, 409-429.

- Campbell, D. T. (1984). Forward. In W. M. K. Trochim (Ed.), *Research Design for Program Evaluation* (pp. 15-43). Beverly Hills, CA: Sage Publications.
- Campbell, D. T., & Boruch, R. (1975). *Making the Case for Randomized Assignment to Treatments by Considering the Alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects*. In C. A. Bennet & A. A. Lumsdaine (Eds.), *Evaluation and Experiment* (pp. 297-317). New York: Academic Press.
- Campbell, D. T., & Erlebacher, A. (1970). How Regression Artifacts in Quasi-Experiments can Mistakenly Make Compensatory Education Look Harmful. In J. Helmuth (Ed.), *The Disadvantaged Child* (pp. 185-210). New York: Brunner-Mazel.
- Campbell, D. T., & Stanley, J. C. (1963). Experimental and Quasi-Experimental Designs for Research on Teaching. In N. L. Gage (Ed.), *Handbook of Research on Teaching*. Chicago: Rand McNally.
- Cappelleri, J. C. (1991). *Cutoff-based designs in comparison and combination with randomized clinical trials*. Cornell University, Ithaca, NY.
- Cappelleri, J. C., Darlington, R. B., & Trochim, W. M. K. (1994). Power Analysis of Cutoff-Based Randomized Clinical Trials. *Evaluation Review*, 18, 141-152.
- Cappelleri, J. C., & Trochim, W. M. K. (1995). Ethical and Scientific Features of Cutoff-based Designs. *Medical Decision Making*, 15, 387-394.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-Experimentation: Design and Analysis for Field Settings*. Chicago, IL: Rand McNally.
- Cook, T. D., Shadish, W. R., & Wong, V. C. (2005). *Within-Study Comparisons of Experiments and Non-Experiments: Can they help decide on Evaluation Policy?* Paper presented at the Econometric Evaluation of Public Policies: Methods and Applications, Paris, France.
- Cook, T.D. & Wong, V.C. (in press). Empirical Tests of the Validation of the Regression Discontinuity Design. *Annales d'Economie et de Statistique*.
- Ferguson, N. *Virtual History: Alternatives and Counterfactuals*. London, Macmillan, 1997.
- Finkelstein, M., Levin, B., & Robbins, H. (1996a). Clinical and Prophylactic Trials with Assured New Treatment for Those at Greater Risk: I. A Design Proposal. *Journal of Public Health*, 86(5), 691-695.
- Finkelstein, M., Levin, B., & Robbins, H. (1996b). Clinical and Prophylactic Trials with Assured New Treatment for Those at Greater Risk: II. Examples. *Journal of Public Health*, 86(5), 696-705.
- Glazerman, S., Levy, D. M., & Myers, D. (2003). Nonexperimental versus Experimental Estimates of Earnings Impacts. *The Annals of the American Academy*, 589, 63-93.
- Goldberger, A. S. (1972a). *Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations*. Unpublished manuscript, Madison, WI.
- Goldberger, A. S. (1972b). *Selection Bias in Evaluating Treatment Effects: The case of interaction*. Unpublished manuscript, Madison, WI.
- Gormley, W. T., & Phillips, D. (2005). The Effects of Universal Pre-K in Oklahoma: Research Highlights and Policy Implications. *The Policy Studies Journal* 33(1), 65-81.

- Hahn, J., Todd, P., & VanderKlaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1), 201-209.
- Imbens, G.W. & VanderKlaauw, W. (1995). Evaluating the Cost of Conscription in The Netherlands. *Journal of Business and Economic Statistics*, 13(2), 72-80..
- Imbens, G.W., & Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 61(2), 467-476.
- Jacob, B. A., & Lefgren, L. (2004a). The Impact of Teacher Training on Student Achievement: Quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources*, 39(1), 50-79.
- Jacob, B. A., & Lefgren, L. (2004b). Remedial Education and Student Achievement: A regression-discontinuity analysis. *Review of Economics and Statistics*. 86(1): 226-244.
- Joyce, T., Kaestner, R. & Colman, S. (2006). Changes in abortions and births and the Texas parental notification law. *New England Journal of Medicine*, 354(10), 1031-1038.
- Judd, C. M., & Kenny, D. A. (1981). *Estimating the Effects of Social Interventions*. New York: Cambridge University Press.
- Knafl, G. (1984). MP: The Modeling Package for Model Robust Nonparametric Regression. In *American Statistical Association 1984 Proceedings of the Statistical Computing Section* (pp. 132-137). Alexandria, VA: American Statistical Association.
- Knafl, G., Sacks, J., Speigelman, C., & Ylvisaker, D. (1984). Nonparametric Calibration. *Technometrics*, 26, 233-241.
- Knafl, G., Sacks, J., Speigelman, C., & Ylvisaker, D. (1984). Calibrating for Differences. In L. J. Gleser, M. D. Perlman, S. J. Press, & A. R. Sampson (Eds.), *Contributions to Probability and Statistics: Essays in Honor of Ingram Olkin* (pp. 335-348). New York: Springer-Verlag.
- Knafl, G., Sacks, J., & Ylvisaker, D. (1982). Model Robust Confidence Intervals. *Journal for Statistical Planning and Inference*, 6, 319-334.
- Knafl, G., Sacks, J., & Ylvisaker, D. (1985). Confidence Bands for Regression Functions. *Journal of the American Statistical Association*, 80, 683-691.
- LaLonde, R. (1986). Evaluating the Econometric Evaluations of Training with Experimental Data. *The American Economic Review*, 76(4), 604-620.
- Lee, D. (2001). The Electoral Advantage of Incumbency and the Voter's Valuation of Political Experience: A Regression Discontinuity Evaluation of Close Elections. Unpublished ms., Department of Economics, University of California, Berkeley.
- Ludwig, J., & Miller, D.L. (in press). Does Head Start improve Children's Life Chances: Evidence from a Regression Discontinuity Approach. *Quarterly Journal of Economics*
- Madaus, G. F., & Greaney, V. (1985). The Irish Experience in Competency Testing: Implications for American education. *American Journal of Education* 93, 268-294.
- Marks, H.M. (1997). *The Progress of Experiment: Science and Therapeutic Reform in the United States, 1900-1990*. Cambridge: Cambridge University Press.
- Mohr, L. B. (1985). *Impact Analysis for Program Evaluation*. Chicago: Dorsey Press.

- Mosteller, F. (1996). Editorial: The Promise of Risk-Based Allocation Trials in Assessing New Treatments. *Journal of Public Health*, 86(5), 622-623.
- Overman E.S. (1981). *Methodology and Epistemology for the Social Sciences: Selected papers of Donald T. Campbell*. Chicago: University of Chicago Press.
- Rand Corporation (in preparation). Effects of No Child Left Behind: Regression Discontinuity Analyses.
- Reichardt, C. S. (1979). *The Design and Analysis of the Non-Equivalent Group Quasi-Experiment*. Northwestern University, Evanston, IL.
- Riecken, H. W., Boruch, R., Campbell, D. T., Caplan, N., Glenman, T. K., Pratt, J. W., et al. (1974). *Social Experimentation: A method for planning and evaluating social intervention*. New York: Academic Press.
- Robbins, H., & Zhang, C.-H. (1988). Estimating a Treatment Effect Under Biased Sampling. *Proceedings from the National Academy of Sciences USA*, 85, 3670-3672.
- Robbins, H., & Zhang, C.-H. (1989). Estimating the Superiority of a Drug to a Placebo When All and Only Those Patients at Risk are Treated with the Drug. *Proceedings from the National Academy of Sciences USA*, 86, 3003-3005.
- Robbins, H., & Zhang, C.-H. (1991). Estimating a Multiplicative Treatment Effect under Biased Allocation. *Biometrika*, 78, 349-354.
- Ross, H. L., Campbell, D. T., & Glass, G. V. (1970). Determining the Social Effects of a Legal Reform: The British "Breathalyser" Crackdown of 1967. *American Behavioral Scientist*, 13, 493-509.
- Rubin, D. B. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics*, 2(1), 1-26.
- Sacks, J., & Ylvisaker, D. (1978). Linear Estimates for Approximately Linear Models. *Annals of Statistics*, 6, 1122-1138.
- Seaver, W. B., & Quarton, R. J. (1976). Regression-Discontinuity Analysis of Dean's List Effects. *Journal of Educational Psychology*, 68, 459-465.
- Sechrest, L., Perrin, E., & Bunker, J. (Eds.). (1990). *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data*. Washington, DC: U.S. Department of Health and Human Services, Agency for Health Care Policy and Research.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin Company.
- Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman Hall.
- Smith, J. C., & Todd, P. (2005). Does Matching Overcome LaLonde's Critique of Nonexperimental Estimators. *Journal of Econometrics*, 125, 305-353.
- Spiegelman, C. (1976). *Two Methods of Analyzing a Nonrandomized Experiment "Adaptive" Regression and a Solution to Reiersol's Problem*. Unpublished dissertation, Northwestern University, Evanston, IL.
- Spiegelman, C. (1977). *A Technique for Analyzing a Pretest-Posttest Nonrandomized Field Experiment*. Tallahassee, FL: Florida State University.
- Spiegelman, C. (1979). On Estimating the Slope of a Straight Line when Both Variables are Subject to Error. *The Annals of Statistics*, 7(1), 201-206.

- Spiegelman, C. (2006). Personal communication. In T. D. Cook (Ed.). Evanston, IL.
- Spiegelman, C., & Sacks, J. (1980). Consistent Window Estimation in Nonparametric Regression. *The Annals of Statistics*, 8(2), 240-246.
- Stone, C. J. (1977). Consistent Nonparametric Regression. *The Annals of Statistics*, 5(4), 595-620.
- Sween, J. A. (1971). *The Experimental Regression Design: An inquiry into the feasibility of nonrandom treatment allocation*. Northwestern University, Evanston, IL.
- Tallmadge, G. K., & Horst, D. P. (1976). *A Procedural Guide for Validating Achievement Gains in Educational Projects*. Washington, DC: U.S. Department of Health, Education, and Welfare.
- Tallmadge, G. K., & Wood, C. T. (1978). *User's Guide: ESEA Title I evaluation and reporting system*. Mountain View, CA: RMC Research Corporation.
- Thistlewaite, D. L., & Campbell, D. T. (1960). Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment. *Journal of Educational Psychology*, 51, 309-317.
- Trochim, W. M. K. (1984). *Research Design for Program Evaluation*. Beverly Hills, CA: Sage Publications.
- Trochim, W. M. K. (2001). Regression Discontinuity Design. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences* (Vol. 19, pp. 12940-12945). Oxford, UK: Elsevier.
- Trochim, W. M. K. (Ed.). (1990). *Regression-Discontinuity Design in Health Evaluation*. Washington, DC: U.S. Department of Health and Human Services, Agency for Health Care Policy and Research.
- Trochim, W. M. K., & Cappelleri, J. C. (1992). Cutoff Assignment Strategies for Enhancing Randomized Clinical Trials. *Controlled Clinical Trials*, 13, 190-212.
- Trochim, W. M. K., & Spiegelman, C. (1980). *The Relative Assignment Variable Approach to Selection Bias in Pretest-Posttest Designs*: American Statistical Association.
- VanderKlaauw, W. (2002). A Regression-discontinuity Evaluation of the Effects of Financial Aid Offers on College Enrollment. *International Economic Review*, 43(4): 1249-1287.