Psychometrika (2025), **90**, 1400–1418
doi:10.1017/psy.2025.10033

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

Evidence Factors in Fuzzy Regression Discontinuity Designs with Sequential Treatment Assignments

Youjin Lee[1] and Youmi Suk[2]

[1]Department of Biostatistics, Brown University, Providence, RI, USA; [2]Grace Dodge Hall 552, Teachers College, Columbia University, 525 West 120th Street, New York, NY 10027

Corresponding author: Youmi Suk; Email: ysuk@tc.columbia.edu

(Received 30 October 2024; revised 20 June 2025; accepted 23 June 2025; published online 8 August 2025)

Abstract

Many observational studies often involve multiple levels of treatment assignment. In particular, fuzzy regression discontinuity (RD) designs have sequential treatment assignment processes: first based on eligibility criteria, and second, on (non-)compliance rules. In such fuzzy RD designs, researchers typically use either an intent-to-treat approach or an instrumental variable-type approach, and each is subject to both overlapping and unique biases. This article proposes a new evidence factors (EFs) framework for fuzzy RD designs with sequential treatment assignments, which may be influenced by different levels of decision-makers. Each of the proposed EFs aims to test the same causal null hypothesis while potentially being subject to different types of biases. Our proposed framework utilizes the local RD randomization and randomization-based inference. We evaluate the effectiveness of our proposed framework through simulation studies and two real datasets on pre-kindergarten programs and testing accommodations.

Keywords: evidence factors (EFs); observational studies; regression discontinuity; sensitivity analysis

1. Introduction

In observational studies, researchers often conduct multiple analyses on a single dataset to answer the same or related causal questions. This strategy can help rule out non-causal explanations and enhance the robustness of causal conclusions (Cook, 1985). However, if not carefully designed, multiple analyses may not be more beneficial than a single analysis; they can provide redundant information or consistently incorrect answers if they share overlapping biases. Planning and conducting a carefully designed multifaceted approach within a single dataset have motivated a new design known as *evidence factors* (EFs), introduced by Rosenbaum (2010). Briefly, EFs are two or more independent tests of the same causal null hypothesis about the treatment effect using a single dataset, with each potentially being subject to different biases. Each EF provides *unique* information about the causal question, and combining these factors can strengthen causal conclusions. The overarching goal of this article is to propose a new framework for constructing EFs in regression discontinuity (RD) designs.

An RD design, introduced by Thistlethwaite & Campbell (1960), is one of the most credible quasi-experimental designs. RD designs are used when a subject's eligibility for treatment depends solely on a running variable (Hahn et al., 2001; Thistlethwaite & Campbell, 1960). If the value of the running variable (e.g., English proficiency) is below or equal to a cutoff, the subject is eligible for the treatment

(e.g., a testing accommodation); otherwise, the subject is ineligible. This is known as a *sharp* RD design. However, eligibility does not necessarily lead to treatment use, and noncompliance yields a *fuzzy* RD design. For example, in a fuzzy RD setting, policy makers might establish the treatment eligibility rule, while the ultimate decision to use the treatment depends on the students themselves. As such, fuzzy RD designs inherently involve sequential assignment processes, where each is driven by different decision-makers. See Section 2 for two motivating examples from fuzzy RD designs.

When data are collected from fuzzy RD designs, researchers can use various analytic approaches to estimate causal effects. For example, one can apply an intent-to-treat approach (i.e., a sharp RD method) to estimate a causal effect at or near the cutoff, or one can use eligibility status as an instrumental variable (IV) to estimate a subgroup causal effect near the cutoff (Lee & Lemieux, 2010). However, both intent-to-treat and IV-based approaches in fuzzy RD designs focus on a local effect at or near the cutoff and are potentially vulnerable to the same bias associated with the eligibility rule, such as manipulation of the running variable, i.e., subjects precisely adjusting its values (Cattaneo et al., 2024; Crespo, 2020; Lee & Lemieux, 2010). Alternatively, researchers may consider an as-treated analysis that compares subjects who actually used the treatment with those who did not, to estimate the treatment effect. While this approach ensures greater generalizability, it is susceptible to confounding bias due to self-selection. Each analytic approach relies on different sets of causal assumptions and has unique strengths and weaknesses, so researchers may benefit from applying multiple approaches within a single dataset (Suk et al., 2022; Wong et al., 2007).

However, simply conducting multiple analyses does not lead to more accurate causal conclusions. If not properly designed, multiple analyses may only be slight variations subject to the same bias. In contrast, EF analysis aims to carefully design (nearly) independent analyses and then combine them to reinforce causal conclusions on the same null hypothesis. Each EF is designed to provide a unique and independent piece of evidence on the treatment effect. Recent studies have extended the original idea of Rosenbaum (2010) to different settings (Karmakar & Small, 2023; Rosenbaum, 2023; Zhao et al., 2022). For example, Karmakar & Small (2023) and Zhao et al. (2022) propose EF analysis with IVs, and Rosenbaum (2023) introduces EFs using a second control (see Section 3.2 for more details). Although these prior studies can touch on certain features of fuzzy RD designs through IV-type analyses, none explicitly tackle unique features and challenges inherent in RD settings. Fuzzy RD designs offer a distinct setup due to their inherent locality and sequential assignment processes, neither of which has been addressed in EFs analysis.

In this article, we propose a novel framework for constructing EFs in fuzzy RD designs with sequential treatment assignments. We achieve this by restricting the sample within the window for one analysis and properly conditioning on past treatment assignments in the subsequent analysis. Our proposed framework involves three key steps: (i) constructing EFs, (ii) combining p-values from each EF, and (iii) conducting sensitivity analysis. With the proposed framework, we contribute to both RD and EFs literature as follows. First, we introduce the EFs design into fuzzy RD analysis by leveraging the inherent causal relationships among treatment variables to disentangle biases across multiple analyses. Unlike traditional fuzzy RD approaches, which rely on intent-to-treat and IV-type analyses that suffer from the overlapping biases, our EF framework provides nearly independent pieces of evidence from multiple analyses, both statistically and causally, thereby strengthening causal conclusions. Second, we formalize the causal assumptions required for constructing multiple EFs in fuzzy RD designs, a new setting for EF analysis. Specifically, we leverage sequential decision-making processes and the locality inherent in analyzing samples near the cutoff to construct valid EFs. Importantly, our design accommodates both nested and non-nested structures among treatment assignment variables and uses a single dataset from a fuzzy RD design without requiring external data. Lastly, we extend sensitivity analysis methods used in EFs designs to the fuzzy RD setting. In particular, we incorporate an assignment model and test statistic that reflect locality in RD designs, and use them in sensitivity analyses to assess the robustness of our conclusions to violations of the key assumption of the RD design.

The remainder of this article is structured as follows. Section 2 illustrates two real-world, fuzzy RD examples, exhibiting one-sided and two-sided non-compliance, respectively. Section 3 provides a brief

https://doi.org/10.1017/psy.2025.10033 Published online by Cambridge University Press

literature review on RD and EFs. Section 4 introduces our settings and formalizes our assignment models with their causal assumptions. Section 5 then describes our proposed EF approach and its properties. Section 6 provides simulation studies, and Section 7 presents two data applications. Lastly, we provide our discussion and conclusions in Section 8.

2. Two motivating examples

2.1. State pre-kindergarten (pre-K) program

Wong et al. (2007) investigate whether state pre-K programs for 4-year-old children improve academic performance in their vocabulary, math, and print awareness skills. This study employs an RD design where the running variable is a child's birth date; children with birthdays on or after a certain date are eligible for enrolling in the pre-K program, whereas those with birthdays before it are not. Additionally, there are non-compliers who do not follow the treatment assignment based on eligibility, and the presence of these non-compliers requires using a fuzzy RD design for program evaluation (Suk, 2024). Figure 1 provides an RD plot showing the relationship between the outcome of vocabulary scores and the running variable of birth date. While the majority of children comply with the eligibility status determined by the running variable, we observe non-compliers on both sides of the cutoff, i.e., two-sided non-compliance.

Rosenbaum (2023) introduces how to construct EFs with multiple control groups to test a single hypothesis, where each factor is vulnerable to different biases. To evaluate the effect of New Jersey's pre-K program, we can adapt this approach by conducting an EF analysis to test the same null hypothesis: there is no effect of using the pre-K program. However, this application raises some important questions. First, is Rosenbaum (2023)'s proposal directly applicable to a fuzzy RD design with two-sided non-compliance? Second, how can we form multiple comparisons for EFs analysis in the fuzzy RD design? Third, how many valid EFs can be constructed in such a design?

2.2. Extended time accommodation (ETA)

Testing accommodations are essential for students with disabilities or English language learners (ELLs) to accurately demonstrate their abilities during assessments. The ETA is the most frequently provided testing accommodation in many testing programs. Suk et al. (2022) and Suk & Kim (2024) examine the effects of ETA for ELLs using a fuzzy RD design. They use ELL English proficiency scores as the

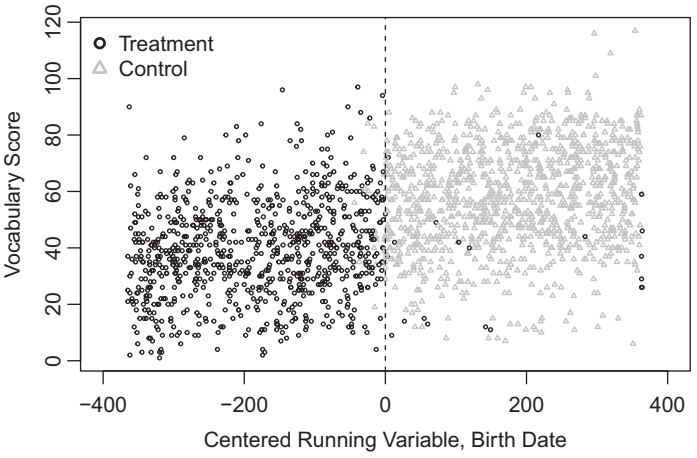

Figure 1. Visual representation of a fuzzy RD design in New Jersey's pre-K program: black points represent the treatment group (i.e., program user group), while gray points represent the control groups (i.e., non-user group).

https://doi.org/10.1017/psy.2025.10033 Published online by Cambridge University Press

Table 1. Treatment statuses and four different groups in the context of the ETA

ETA eligibility status	ETA receipt status	ETA used status
0	0	0
1	0	0
1	1	0
1	1	1

running variable, and ETA eligibility status as an IV. In ETA settings, sequential assignment processes with one-sided non-compliance occur, and they result in three treatment statuses: eligibility, receipt, and use; see Table 1 for treatment statuses and four different groups formed through the sequential assignment process.

Prior studies (Suk & Kim, 2024; Suk et al., 2022) assume that the ETA effect is only present when students make use of it; otherwise, there is no assumed ETA effect. They employ a fuzzy RD design to estimate three causal effects: the effect of being eligible for ETA at the cutoff score, the effect of receiving ETA at the cutoff among compiler students, and the effect of using ETA at the cutoff among compiler students.[1] While these different causal effects provide informative evidence on ETA's effectiveness, each effect estimate could be susceptible to different types of biases occurring at each assignment process, as well as to the overlapping biases inherent in the RD design (e.g., manipulation of the running variable). However, if carefully designed, EFs can be constructed in this ETA data to test the same null hypothesis of "no effect of using ETA," where each factor has independent pieces of information about the null hypothesis.

3. Literature review and our contribution

3.1. Local randomization framework for RD designs

Our setting is based on RD designs, and there are two frameworks available: continuity-based and local randomization. The first framework is based on the continuity assumption (Hahn et al., 2001). This framework assumes that the conditional expectations of potential outcomes, given the running variable, are continuous at the cutoff to provide valid counterfactuals in both sharp RD and fuzzy RD designs (Cattaneo et al., 2023b; Dong, 2018). This assumption often requires (local) parametric or nonparametric regressions. The second framework is the local randomization framework, which was motivated by the works of Lee (2008) and Thistlethwaite & Campbell (1960). This framework interprets RD designs heuristically as locally randomized experiments within a neighborhood of the cutoff (i.e., within the window, denoted by $\mathcal{W}$). It is based on a local unconfoundedness assumption, which assumes that the treatment assignment (e.g., eligibility) is unconfounded given the observed covariates within the window $\mathcal{W}$ (Cattaneo et al., 2015). The main distinction between continuity-based and local randomization frameworks lies in how they formalize the idea of comparability (Cattaneo et al., 2024). While comparability under the continuity-based framework arises from extrapolating the limiting distribution of local regression-based estimators at the cutoff, the local randomization framework assumes comparability within a small neighborhood around the cutoff, like in randomized experiments, and utilizes randomization-based inference without having to specify regression models.

In this work, we adopt the second framework of the local randomization based on randomization-based inference, which utilizes the randomization mechanisms of treatments while treating the observed outcomes as fixed. Randomization-based inference not only accommodates the local randomization framework but also easily accounts for other commonly used designs for causal comparisons, such

[1] In the ETA example, compliers mean students who would use (or receive) the ETA program if eligible and would not use (or receive) it if ineligible.

https://doi.org/10.1017/psy.2025.10033 Published online by Cambridge University Press

as matching and stratification (Rosenbaum, 2002). Randomization-based inference, which does not rely on large sample approximations, is particularly useful in fuzzy RD designs where at least one assignment depends on observations within a small window around the cutoff. When leveraging randomization-based inference, we assume that matched data on observed covariates come from a randomized trial; in other words, within the matched strata, the treatment status is randomly assigned to each subject. Therefore, an outcome comparison within the matched strata can provide a valid causal comparison, without additional covariate adjustment. Under the local randomization framework, such unconfoundedness must hold within one stratum that includes only subjects within a window around the cutoff. However, further covariate adjustment (e.g., matching) can be applied within this stratum to improve efficiency and power in estimating causal effects (Cattaneo et al., 2023a).

3.2. EF design with multiple analyses

An EF design can be used when multiple analyses are possible to test the same causal null hypothesis in observational studies. Instead of treating one among several analyses as primary and the others as secondary, each analysis contributes an independent piece of evidence in the EF design. Multiple analyses can arise from multiple doses (variations) of the treatment. For example, one analysis compares outcomes between the treated and control groups (e.g., smokers and non-smokers), while another analysis compares outcomes between those with intense treatment and those with weak treatment (e.g., heavy smokers and light smokers) within the treated group (Rosenbaum, 2010). To ensure these multiple analyses produce independent inferential results, Rosenbaum (2010) proposes an EFs design with multiple doses of the treatment.

Additionally, multiple control groups, who did not take the treatment for different reasons, can generate multiple analyses, where each control group is compared to the corresponding treatment group (Rosenbaum, 2023). Most recently, Rosenbaum (2023) introduces an EF design with multiple control groups. Specifically, Rosenbaum (2023) introduces two assignment variables, one of which defines secondary control, likely derived from external data sources, and deliberately nests these variables by design, even though their causal relationships are unclear. In contrast, as will be elaborated in the next section, our proposed EF design explicitly assumes causal relationships among treatment assignment variables within a single dataset. This assumption is valid because fuzzy RD designs inherently involve sequential assignment processes, influenced by different decision-makers. Our approach leverages these sequential assignment variables for EF analysis and accommodates both one-sided (nested) and two-sided (non-nested) non-compliance, each of which is illustrated in Section 2. This differs from the nested structure addressed in Rosenbaum (2023).

Furthermore, the presence of multiple IVs generates multiple analyses. These IV analyses are often correlated with each other as well as with a direct comparison between the treated and control groups. Recent research has explored EFs using such IVs (Karmakar & Small, 2023; Karmakar et al., 2020; Zhao et al., 2022). In particular, Karmakar et al. (2020) propose the reinforced design, where multiple IVs, in addition to a direct comparison (e.g., between actual users and non-users), provide EFs. Due to potential correlation among IVs, the reinforced design requires specifying the order of multiple analyses to avoid overlapping biases. Zhao et al. (2022) relax this ordering requirement by blocking or nesting procedures, each of which would only be valid under particular structures among IVs. In our context, multiple treatment assignments up to the final one can be considered IVs; however, these assignments must be causally ordered. This new requirement has not been formally addressed in the existing EF literature.

4. Settings

4.1. Notation

Subjects are grouped into strata using observed covariates to form J strata. We assume that the observed covariates of subject i in stratum j (subject ij afterward), denoted by $\mathbf{W}_{ij}$, are controlled by this stratification. In other words, $\mathbf{W}_{ij} = \mathbf{W}_{i'j}$ for all $i \neq i'$ across all strata $j = [1:J]$, where $[m,\ell]$ denotes a set

https://doi.org/10.1017/psy.2025.10033 Published online by Cambridge University Press

of an integer from m to ℓ. Let Y_{ij} denote the outcome of interest for unit ij. Consider that the first level of treatment assignment exhibits an RD design, determining eligibility for treatment. Let $A_{ij} \in \{0,1\}$ denote the eligibility for the treatment, where $A_{ij} = 1$ indicates that subject ij is eligible and not eligible if $A_{ij} = 0$. The eligibility is determined based on a running variable, denoted by X_{ij}. Given a cutoff x_c, $A_{ij} = 1$ if $X_{ij} \le x_c$ and $A_{ij} = 0$ if $X_{ij} > x_c$.[2] Let $T_{ij} \in \{0,1\}$ indicate the actual treatment use of subject ij. However, not all eligible subjects actually use the treatment, and ineligible subjects occasionally use the treatment in practice. There could be sequential treatment assignment processes driven by different levels of decision-makers. For example, in the ETA application, (1) first, students' eligibility is determined by their ELL English-proficiency scores according to government or state policy; (2) second, students receive ETA from their school administrators; and (3) lastly, students can choose to use the ETA offered.

Suppose that K denotes the number of treatment assignment processes ($K \ge 1$), resulting in a K-dimensional treatment status for each subject. When $K = 1$, we have the sharp RD with $A_{ij} = T_{ij}$. When $K \ge 2$, $Z_{ij}^{(k)}$ indicates that subject ij takes the treatment *at level* k ($k \in [1:K]$) where $Z_{ij}^{(1)} = A_{ij}$ and $Z_{ij}^{(K)} = T_{ij}$. We allow $Z_{ij}^{(k)} = 1$ even if $Z_{ij}^{(k')} = 0$ for any $k' \in [1:k-1]$, i.e., allowing two-sided non-compliance; for example, subject ij potentially uses the treatment even if they were ineligible and/or did not receive the treatment from school administrators. However, we focus on the assignment model with $Z_{ij}^{(k)}$ conditioning on $\mathbf{Z}_{ij}^{(1:k-1)} = \mathbf{1}$ for each $k \in [1:K]$, where $\mathbf{Z}_{ij}^{(\ell_1:\ell_2)} = \{Z_{ij}^{(\ell)} : \ell \in [\ell_1, \ell_2]\}$. For notational convenience, define $Z_{ij}^{(0)} = \varnothing$ and $\mathbf{Z}_{ij}^{(1:0)} = Z_{ij}^{(0)} = \varnothing$. We elaborate this conditioning in Section 4.3. We denote $\mathbf{W}$, $\mathbf{X}$, $\mathbf{Z}^{(1:k)}$, and $\mathbf{Y}$ as the collection of W_{ij}, X_{ij}, $\mathbf{Z}_{ij}^{(1:k)}$, and Y_{ij} across units, respectively.

4.2. Potential outcomes, assumptions, and hypotheses

Let $\mathbf{S}_{ij} = \{Z_{ij}^{(k)} : k = 1,2,\ldots,K\} = (A_{ij}, \mathbf{Z}_{ij}^{(2:k-1)}, T_{ij})$, which can take 2^K different values. Let $\mathcal{S}_K$ be the set of the 2^K, length-K vectors with 0 or 1 coordinates. Under a series of one-sided non-compliance, a vector of $\mathbf{S}_{ij}$ can take up to $(K+1)$ different values, where $Z_{ij}^{(k)} = 1$ implies $\mathbf{Z}_{ij}^{(1:k-1)} = \mathbf{1}$ for all $k \in [2:K]$. For example, when $K = 3$, $\mathbf{s} \in \mathcal{S}_3^* \subseteq \mathcal{S}_3$ such that $\mathcal{S}_3^* = \{(0,0,0),(1,0,0),(1,1,0),(1,1,1)\}$. Using the potential outcomes framework, let $Y_{ij}^{\mathbf{s}}$ denote the potential outcome of Y_{ij} when $\mathbf{S}_{ij} = \mathbf{s} \in \mathcal{S}_K$ with $\mathbf{s} = (s_1, s_2, \ldots, s_K)$ under the consistency assumption (Rubin, 1974). For example, $Y_{ij}^{(1,1,0)}$ is the potential outcome of subject ij if she were eligible for the treatment and received the treatment from school administrators but did not use it. The following assumption is essential for using $\mathbf{Z}_{ij}^{(1:K-1)}$ as a (conditional) IV to examine the causal effect of the actual treatment use.

Assumption 1. $Y_{ij}^{\mathbf{s}} = Y_{ij}^{s_K}$ for all $\mathbf{s} \in \mathcal{S}_K$

In other words, a whole vector of $\mathbf{S}_{ij}$ has an effect on the outcome only through the actual treatment use, T_{ij}. This means the exclusion restriction (Angrist et al., 1996). Then, our null hypothesis of interest is as follows:

$$H_{0,K} : Y_{ij}^{s_K=1} = Y_{ij}^{s_K=0} \text{ for all subjects } ij. \tag{1}$$

If our null of no treatment effect $H_{0,K}$ is true, then $Y_{ij}^{\mathbf{s}} = Y_{ij}^{\mathbf{s}'}$ for all $\mathbf{s}, \mathbf{s}' \in \mathcal{S}_K$ under Assumption 1. The null in (1) also implies that there is only one version of the potential control outcome among 2^{K-1} values in $\mathcal{S}_K$. Additionally, we require the following assumption about the (causal) relationship among K different treatment statuses, which represents the most distinct causal structure compared to the existing EF literature (Karmakar et al., 2020; Rosenbaum, 2023; Zhao et al., 2022).

Assumption 2. (a) Treatment statuses, $\mathbf{Z}_{ij}^{(1:K)}$, are causally ordered and (b) there are no unknown or unmeasured common causes among them.

[2] In some cases like our pre-K program, treatment assignment A_i can be determined in the opposite direction such that $A_i = 1$ if $X_i \ge x_c$ and $A_i = 0$ if $X_i < x_c$.

https://doi.org/10.1017/psy.2025.10033 Published online by Cambridge University Press

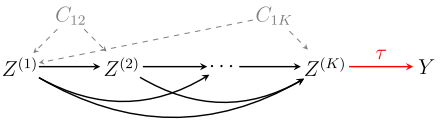

Figure 2. A directed acyclic graph (DAG) illustrating the causal relationships among treatment statuses $Z^{(k)}$ ($k \in [1:K]$), outcome Y, and unmeasured common causes $C_{\ell k}$ between $Z^{(\ell)}$ and $Z^{(k)}$ ($\ell \neq k$).
Note: Observed covariates are omitted for simplicity.

The condition (a) of Assumption 2 implies that the treatment status at level k is established *after* the treatment statuses at or earlier than level $k-1$ are determined. For example, students' decision to use the treatment should not affect their eligibility and receipt statuses. Moreover, there should be no unmeasured confounding among the treatment statuses at different levels as implied in the second condition. This assumption is more likely to hold when distinct decision-makers determine the treatment assignment process at each level, each affected by different criteria or rules. For instance, in ETA settings, assignment decisions are made separately by entities, such as government or state policy, schools, and students. Each decision-maker is likely to rely on different information (e.g., state-level cutoffs, school-level logistics, and student preferences) while also considering some shared, observable information (e.g., students' ELL English proficiency scores).

To illustrate, Figure 2 provides a causal structure among variables under Assumption 1 and condition (a) of Assumption 2. In Figure 2, the treatment statuses of $\mathbf{Z}^{(1:K)}$ are causally ordered and any status among $\mathbf{Z}^{(1:K-1)}$ does not have a direct effect on the outcome Y. Condition (b) of Assumption 2 further assumes that there are no unmeasured common causes among $\mathbf{Z}^{(1:K)}$, such as C_{12} and C_{1K}, that can confound the causal relationships among the treatment statuses.

We note that we can relax condition (b) of Assumption 2 to allow for unmeasured common causes among *some* of the K treatment statuses. In Section S2 of the Supplementary Material, we provide a method for selecting a valid set of treatment statuses by excluding one of the two statuses that are presumed to share unmeasured common causes.

4.3. Treatment assignment models

In this section, we illustrate the treatment assignment models at each level of a sequential assignment process, which can seamlessly incorporate Rosenbaum's sensitivity analysis (Karmakar et al., 2020; Rosenbaum, 1987; Zhao et al., 2022). Let $\mathcal{F}$ denote the collection of the potential outcomes and observed and unobserved covariates. Then, we posit the following assignment model at level $k = 1$:

$$Pr(A_{ij} = 1 \mid \mathcal{F}, \mathcal{W}) = \frac{\exp\{\kappa_1(\mathbf{w}_{ij}) + \gamma_1 u_{ij,1}\}}{1 + \exp\{\kappa_1(\mathbf{w}_{ij}) + \gamma_1 u_{ij,1}\}}, \tag{2}$$

where κ_1 is an arbitrary function of the observed covariates, and $u_{ij,1}$ denotes the unmeasured covariate. When $\gamma_1 = 0$, this model entails the key assumption of the local randomization framework in an RD design, where eligibility is randomly assigned among subjects given the observed covariates within the window $\mathcal{W}$. In this case, $Pr(A_{ij} = 1 \mid \mathcal{F}, \mathcal{W})$ only depends on the observed covariates' values, and thus, subjects within stratum j have the same probability of $A_{ij} = 1$. Therefore, $Pr(A_{ij} = 1 \mid \mathcal{F}, \mathcal{W}) = Pr(A_{i'j} = 1 \mid \mathcal{F}, \mathcal{W})$ for all $i \neq i'$.

At subsequent levels of $k \in [2:K]$, we consider the following assignment model among those who were assigned to treatment at previous levels $\ell \in [1:k-1]$:

$$Pr(Z_{ij}^{(k)} = 1 \mid \mathcal{F}, \mathbf{Z}_{ij}^{(1:k-1)} = \mathbf{1}) = \frac{\exp\{\kappa_k(\mathbf{w}_{ij}) + \gamma_k u_{ij,k}\}}{1 + \exp\{\kappa_k(\mathbf{w}_{ij}) + \gamma_k u_{ij,k}\}}. \tag{3}$$

Similarly, κ_k is an arbitrary function of the observed covariates, and $u_{ij,k}$ is the unmeasured covariate for each $k \in [2:K]$. Note that as stated in Assumption 2, unmeasured covariates at each assignment level

https://doi.org/10.1017/psy.2025.10033 Published online by Cambridge University Press

are unique to that level and do not influence other levels. The model (3) implies that given that subject ij was assigned to treatment at previous levels $\ell \in [1:k-1]$, whether subject ij is assigned to treatment or not at level k depends on $(\mathbf{w}_{ij}, u_{ij,k})$; if $\gamma_k = 0$, then the assignment does not depend on any unobserved covariates.

We do not posit any model for treatment status $Z_{ij}^{(k)}$ among subjects with $Z_{ij}^{(\ell)} = 0$ for *any* $\ell \in [1:k-1]$ and $k \in [2:K]$. In practice, it is expected that the assignment mechanisms among those subjects (e.g., assignment of receipt status among ineligible students) would be very different from those who were assigned to treatment at previous assignment processes. We consider the former case as "exceptions" or "nuisances," while the latter assignment models are assumed to be known given $\mathcal{F}$. In Section S3 of the Supplementary Material, we discuss alternative treatment assignment models for $k \in [2:K]$ that further condition on the window $\mathcal{W}$ or the running variable X_{ij}.

5. EFs in fuzzy RD designs

5.1. Constructing multiple comparisons

We propose constructing up to K number of test statistics (EFs) using $\mathbf{Z}_{ij}^{(1:K)}$. We consider model-free (one-sided) test statistics, such as the stratified Wilcoxon's signed rank statistic (Wilcoxon, 1992). We denote the test statistic at level k as $t_{\mathcal{D}_k}(\mathbf{Z}^{(k)}, \mathbf{Y})$, where $\mathcal{D}_k$ represents the subset of units used in the test, which possibly depends on $\mathbf{Z}^{(1:k-1)}$, $\mathbf{W}$, and $\mathbf{X}$. The statistic $t_{\mathcal{D}_k}(\mathbf{Z}^{(k)}, \mathbf{Y})$ uses $Z_{ij}^{(k)}$ as the treatment assignment variable and $\mathbf{Y}$ as fixed outcomes. Under Assumption 1 and the null hypothesis in (1), we have $\mathbf{Y} = \mathbf{Y}^{\mathbf{s}}$ for all $\mathbf{s} \in \mathcal{S}_K$. Therefore, each statistic $t_{\mathcal{D}_k}(\mathbf{Z}^{(k)}, \mathbf{Y})$ for $k \in [1:K]$ tests whether the potential outcomes under a particular treatment vector in $\mathcal{S}_K$ differ and can also provide a valid test for the null in (1) albeit with different power rates. For each test at level k, we refer to the group with $Z_{ij}^{(k)} = 1$ as the treatment comparison group, and the other group with $Z_{ij}^{(k)} = 0$ as the control comparison group.

Specifically, we first propose testing the null (1) with the test statistic $t_{\mathcal{D}_1}(\mathbf{Z}^{(1)}, \mathbf{Y})$, using $Z_{ij}^{(1)}$ (or equivalently, A_{ij}) as a treatment assignment variable. We perform the RD analysis under the local randomization framework, resulting in the comparison only within the window $\mathcal{W}$, i.e., $\mathcal{D}_1 = \{ij : X_{ij} \in \mathcal{W}\}$. This analysis would be valid to test for the null $H_{0,K}$ under Assumption 1 and would not be affected by any of $\{u_{ij,k} : k \in [2:K]\}$ under Assumption 2. The subsequent EFs at level k ($k \in [2:K]$) are constructed using only the subset of subjects with $\mathcal{D}_k = \{ij : \mathbf{Z}_{ij}^{(1:k-1)} = \mathbf{1}\}$. By conditioning on $\mathbf{Z}_{ij}^{(1:k-1)} = \mathbf{1}$, any bias occurring at previous levels would not affect the validity of the test at level k.

Consider the simplest case with $K = 2$ under two-sided non-compliance, as in our pre-K program example. Table 2 illustrates the treatment and control groups used for EF analysis within stratum j, stratified by the indicator for the window inclusion (i.e., whether $X_{ij} \in \mathcal{W}$) and their treatment statuses $\mathbf{S}_{ij} = (A_{ij}, T_{ij})$. Two levels of treatment assignments result in two control groups: the first control group (Control 1) consists of subjects assigned control at $k = 1$, while the second control group (Control 2) consists of subjects assigned treatment at $k = 1$ but control at $k = 2$. The treated group contains the actual users of the treatment. In this case, our proposed framework allows us to construct two EFs. The first EF analysis (EF 1 in Table 2) uses only subjects within the window and compares ineligible subjects with $A_{ij} = 0$ to eligible subjects with $A_{ij} = 1$. Based on a value of A_{ij}, this analysis pretends that the subjects in the second control group are part of the treatment comparison group. It compares the first control group to the other two groups within a window $\mathcal{W}$ (i.e., those with $X_{ij} \in \mathcal{W}$). This analysis corresponds to the intent-to-treat analysis in an RD design, which only uses the eligibility assignment, but ignores the actual treatment use. On the other hand, the second EF analysis (EF 2 in Table 2) only compares the second control group and the treated group while excluding the first control group (e.g., any ineligible subjects) from the analysis but includes those outside of the window.

For further illustration, consider the case with $K = 3$ under one-sided non-compliance as in the ETA study. In the ETA setting, we have three different control groups: the first control group consists of subjects with $\mathbf{Z}_{ij}^{(1:3)} = \mathbf{0}$; the second control group consists of subjects with $Z_{ij}^{(1)} = 1$ but with $\mathbf{Z}_{ij}^{(2:3)} = \mathbf{0}$;

https://doi.org/10.1017/psy.2025.10033 Published online by Cambridge University Press

Table 2. Treatment statuses and the treatment and control comparison groups used for each EF analysis, resulting from a two-level treatment assignment process ($K = 2$) with two-sided non-compliance

$X_{ij} \in \mathcal{W}$	A_{ij} (Treatment eligible)	T_{ij} (Treatment used)	Group	EF 1	EF 2
1	0	0	Control 1	C	·
1	0	1	Control 1	C	·
1	1	0	Control 2	T	C
1	1	1	Treatment	T	T
0	0	0	Control 1	·	·
0	0	1	Control 1	·	·
0	1	0	Control 2	·	C
0	1	1	Treatment	·	T

Note: At each level, subjects with status 0 represent the control comparison group, while those with status 1 represent the treatment comparison group. T: Treatment comparison group; C: Control comparison group; ·: Data excluded.

Table 3. Treatment statuses and the treatment and control comparison groups used for each EF analysis, resulting from a three-level treatment assignment process ($K = 3$) with one-sided non-compliance

$X_{ij} \in \mathcal{W}$	A_{ij} (Treatment eligible)	$Z_{ij}^{(2)}$ (Treatment received)	T_{ij} (Treatment used)	Group	EF 1	EF 2	EF 3
1	0	0	0	Control 1	C	·	·
1	1	0	0	Control 2	T	C	·
1	1	1	0	Control 3	T	T	C
1	1	1	1	Treatment	T	T	T
0	0	0	0	Control 1	·	·	·
0	1	0	0	Control 2	·	C	·
0	1	1	0	Control 3	·	T	C
0	1	1	1	Treatment	·	T	T

Note: At each level, subjects with status 0 represent the control comparison group, while those with status 1 represent the treatment comparison group. T: Treatment comparison group; C: Control comparison group; ·: Data excluded.

and the third control group includes those with $\mathbf{Z}_{ij}^{(1:2)} = \mathbf{1}$ but with $Z_{ij}^{(3)} = 0$ (see Table 3). With our proposed framework, we can construct three EFs. The first EF includes all of these three control groups using $Z_{ij}^{(1)}$ as the treatment assignment variable within a window $\mathcal{W}$; the second EF excludes the first control group and uses $Z_{ij}^{(2)}$ as the assignment variable; and lastly, the third EF only compares the last control group to the treated group. In both Tables 2 and 3, we do not use variation in the treatment status $Z_{ij}^{(k)}$ among those with $Z_{ij}^{(\ell)} = 0$ for any $\ell \in [1 : k-1]$.

5.2. Randomization-based inference and multiple p-values

We apply randomization-based inference for testing the treatment effect with multiple comparisons at each level $k \in [1 : K]$. With the test statistic $t_{\mathcal{D}_k}(\mathbf{Z}^{(k)}, \mathbf{Y})$ for $k \in [1 : K]$, we assume that all outcomes $\mathbf{Y}$ are fixed and randomization of treatment assignment $Z_{ij}^{(k)}$ is the only source of randomness (Rosenbaum, 2002). The assignment models provided in (2) and (3) enable us to leverage randomization mechanisms (within strata) in a finite population, avoiding any specific modeling or reliance on asymptotic conditions. Without an unmeasured covariate in each treatment assignment model (i.e., $\gamma_k = 0$ for $k \in [1 : K]$),

https://doi.org/10.1017/psy.2025.10033 Published online by Cambridge University Press

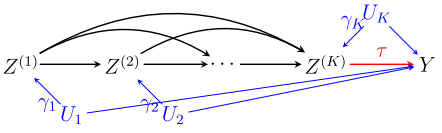

Figure 3. A DAG illustrating the hypothetical causal relationships among variables with unmeasured covariates, U_k's, denoted in blue. *Note*: An edge between U_k and $Z^{(k)}$ is present if and only if $\gamma_k \neq 0$ ($k \in [1:K]$). Observed covariates $\mathbf{W}$ and running variable X are omitted for simplicity.

the treatment status $Z_{ij}^{(k)}$ is randomly assigned to each subject within each stratum, which is constructed from the observed covariates within the window $\mathcal{W}$ if $k = 1$, or conditional on $\mathbf{Z}_{ij}^{(1:k-1)} = \mathbf{1}$ if $k \in [2:K]$.

As a result of performing randomization-based inference at each level, we obtain one p-value for each comparison. Let P_k denote a p-value from the EF analysis at level k ($k \in [1:K]$). Suppose that there is no unmeasured covariate in each level k's assignment model (i.e., when $\gamma_k = 0$). Then, under Assumption 1, we have $Pr(P_k \leq p_k) \leq p_k$ for all $p_k \in [0,1]$ when the null $H_{0,K}$ is true (i.e., each P_k is a valid p-value for testing $H_{0,K}$). Theorem 1 further demonstrates the properties of the joint distribution among $\{P_k : k \in [1:K]\}$ under the null $H_{0,K}$, where we use a nested conditioning structure in our test statistics $t_{\mathcal{D}_k}(\mathbf{Z}^{(k)}, \mathbf{Y})$, for both one-sided and two-sided non-compliance among treatment assignment variables. See a detailed proof in Section S1 of the Supplementary Material.

Theorem 1. *Under Assumption 1, suppose that the assignment models (2) and (3) hold with $\gamma_k = 0$ for all $k \in [1:K]$. Then p-values from the above design are stochastically larger than the uniform under the null. In other words,*

$$Pr(P_1 \leq p_1, \ldots, P_K \leq p_K) \leq \prod_{k=1}^{K} p_k. \tag{4}$$

However, it is possible that some comparisons/analyses out of K assignments are invalid, i.e., biased. Let a subset $\mathcal{I} \subset [1:K]$ denote the analyses that are biased, while $\mathcal{V} = [1:K] \setminus \mathcal{I}$ represents the set of indices that produce valid comparisons. For example, suppose that $2 \in \mathcal{I}$ with $K = 3$, where the second comparison is invalid due to an unmeasured covariate (i.e., $\gamma_2 \neq 0$ in model (3)). Then, the following results imply that this unmeasured covariate would not affect the properties of p-value in (4) among valid comparisons, e.g., P_1 and P_3.

Theorem 2. *Under Assumptions 1 and 2,*

$$Pr(P_k \leq p_k,\ \forall k \in \mathcal{V}) \leq \prod_{k \in \mathcal{V}} p_k. \tag{5}$$

Theorem 2 demonstrates that the bias in the treatment assignments from other analyses would not affect the (near) independence properties among valid EFs, in addition to their individual validity; see a proof in Section S1 of the Supplementary Material. This independence is attributed to the fact that a value of $\{\gamma_\ell : \ell \in \mathcal{I}\}$ would not affect the validity of $\{P_k : k \in \mathcal{V}\}$ (see Figure 3).

In Figure 3, a variable U_k denotes an unmeasured covariate that can confound the relationship between $Z^{(k)}$ and Y ($k \in [1:K]$) when $\gamma_k \neq 0$ ($k \in [1:K]$). Then we have $\mathcal{V} = \{k \in [1:K] : \gamma_k = 0\}$ and $\mathcal{I} = \{k \in [1:K] : \gamma_k \neq 0\}$. Suppose that $1 \in \mathcal{I}$. Then bias due to U_1 could represent bias in the local randomization in RD designs. This bias would not necessarily invalidate analyses at any level $k \in [2:K]$, as their analyses all condition on $Z^{(1)} = 1$ (e.g., conditioning on eligible subjects). Similar arguments can be made for the impact of $U_{k'}$ on analyses at level $k \in [k'+1:K]$ for each $k' \in [1:K-1]$. Furthermore, the presence of U_k ($k \geq 2$) results in invalid analysis with $Z^{(k)}$ testing the null, but bias does not necessarily imply bias in $Z^{(\ell)}$ with $\ell \in [1:k-1]$ under the condition (b) of Assumption 2 that there is no unmeasured common cause between $Z^{(k)}$ and $Z^{(\ell)}$ for $k \neq \ell$. For example, when $K \in \mathcal{I}$ and $2 \in \mathcal{V}$, bias due to U_K would

https://doi.org/10.1017/psy.2025.10033 Published online by Cambridge University Press

not affect the validity of P_2. This is because the condition (b) of Assumption 2 excludes the presence of unmeasured confounding between $Z^{(2)}$ and Y through U_K.

Based on Theorem 2, we can ensure that at least $v := |\mathcal{V}|$ number of nearly independent p-values are obtained. Having nearly independent p-values enables us to easily isolate evidence from any set of p-values. In such a case, we can utilize v largest p-values among K to form a single valid p-value without specifying which v analyses provide valid EFs. Define $P_{(r)}$ as the r^{th} order statistic (i.e., r^{th} smallest value) of $\mathbf{P} = \{P_k : k \in [1:K]\}$ and let $f(\mathbf{P}; v)$ be the combining function of $\mathbf{P}$, with $v = |\mathcal{V}|$. Then with $\{P_{(K-k+1)} : k \in [1:v]\}$ given v, the results from EFs can be combined into one valid p-value, using relatively simple methods of combining independent p-values, such as Fisher's method or truncated product method (Zaykin et al., 2002), as follows:

$$\text{(Fisher's method)} \qquad f(\mathbf{P}; v) = -2\sum_{k=1}^{v} \ln(P_{(K-k+1)})$$

$$\text{(Truncated product method)} \qquad f(\mathbf{P}; v) = \prod_{k=1}^{v} \ln(P_{(K-k+1)})^{I(P_{(K-k+1)} \le \varkappa)}, \qquad 0 < \varkappa \le 1,$$

where $I(\cdot)$ is an indicator function. The combined p-value is obtained using the known null distribution of each statistic $f(\mathbf{P}; v)$ when at least v p-values in $\mathbf{P}$ are (nearly) independent. This combined p-value from multiple EFs, compared to using individual p-values, can provide stronger evidence of the treatment effect when each factor suggests concurrent agreement about the null hypothesis.

We remark a potential concern regarding a lack of power, despite the validity of the test statistics $t_{\mathcal{D}_k}(\mathbf{Z}^{(k)}, \mathbf{Y})$ for $k \in [1:K]$ in testing the target null of $H_{0,K}$. This is because the treatment comparison group at each level k can include a large portion of subjects who do not actually use the treatment, potentially diluting the effect of the treatment at k level on the outcome. To avoid such dilution, researchers can use robust test statistics (see Section 4 of Rosenbaum (2023) for more details).

5.3. Sensitivity analysis

Researchers can use their conjecture on $|\mathcal{V}|$, say q, as a sensitivity parameter and examine the changes in their causal conclusions as a value of q varies ($q \in [1:K]$). Researchers may reflect their prior knowledge about treatment assignment variables in their choice of q, for example, the maximum number of treatment assignments that would satisfy Assumption 2. Selecting q can easily lead to conservative results, however, as it discards $(K-q)$ smallest p-values, which may contain p-values from valid EFs. Instead of discarding several small p-values, one can conduct sensitivity analysis to a specific unmeasured covariate in each assignment model by varying a value of parameter γ_k's in (2) and (3). A non-zero value of γ_k in (2) or (3) (or equivalently, strictly larger than one of $\Gamma_k := \exp(\gamma_k)$) implies a biased assignment at level k within strata formed by observed covariates ($k \in [1:K]$).

Given $u_{ij,k}$ and Γ_k, assuming that the observed difference between the two comparison groups could be caused by this $u_{ij,k}$ by an amount of Γ_k, one could calculate a corresponding p-value for testing the sharp null (1). This resulting p-value is often more conservative compared to the case with $\Gamma_k = 1$ where all the observed differences are attributable to the existing causal effect. Since $u_{ij,k}$'s are unknown, we can use the maximum p-value over $u_{ij,k}$ values in [0,1] when the sensitivity parameter is at most Γ_k (≥ 1). We denote this adjusted p-value as $\overline{P}_{k,\Gamma_k}$ ($k \in [1:K]$). Then each of $\overline{P}_{k,\Gamma_k}$ is a valid p-value and they also maintain the nearly independence properties under the null. This is because non-zero values of $\gamma_{k'}(k' \ne k, k' \in [1:K])$ would not necessarily affect the validity of $\overline{P}_{k,\Gamma_k}$; in other words, we have $Pr(\overline{P}_{k,\Gamma_k} \le p_k) \le p_k$ for any $p_k \in [0,1]$ under the null regardless of the values of $\gamma_{k'}$.

Similar to the argument regarding q, as a value of Γ_k increases, we are more likely to obtain conservative inferences. This is because we consider the observed difference between two comparison groups to be partially attributable to existing imbalances between these groups on an unmeasured covariate. In such a sensitivity analysis, researchers may decide the presumed maximum level of Γ_k for each assignment k (e.g., eligibility assignment would be at most biased by $\Gamma_1 = 1.5$, while the assignment

https://doi.org/10.1017/psy.2025.10033 Published online by Cambridge University Press

of treatment use would be at most biased by $\Gamma_2 = 2.0$ for those who are eligible), and then examine whether the results from multiple EFs are still against (or supporting) the null hypothesis.

6. Simulation study

We numerically evaluate the validity and the performance of our proposed approach through simulation studies that mimic the real data settings on pre-K programs and ETAs. We focus on examining (i) the non-overlapping bias between the proposed EFs and (ii) the validity of the combined p-value in the presence of invalid analyses. To examine (i), we intentionally introduce multiple sources of bias, each of which can invalidate at least one EF, and investigate whether one source can bias multiple factors. To examine (ii), we set q ($q \in [1:K]$) as the conjectured minimum number of valid analyses among K. Our goal is to demonstrate that the combined p-value given q using Fisher's method, denoted as P_q^c, can control the type-I error under the null hypothesis when q is correctly specified (i.e., when $q \leq v$) and that it increases power as the true treatment effect increases.

We further demonstrate the utility of conditioning on treatment statuses at earlier levels to separate biases among EFs. Our proposed EF analysis is compared to analyses that do not condition on the treatment statuses at earlier levels (i.e., not conditioning on $\mathbf{Z}_{ij}^{(1:k-1)} = \mathbf{1}$) while other procedures (e.g., forming strata by matching on observed covariates) remain the same. Each of (P_1^*, P_2^*, P_3^*) denotes the p-value for an unconditioned comparison for each level. For example, P_3^* is the p-value for an as-treated analysis after controlling for observed covariates through stratification. In the absence of the previous treatment statuses (i.e., at level $k = 1$), P_1^* is equivalent to P_1.

In our simulation studies, we consider two designs: one with $K = 3$ under one-sided non-compliance (Design 1) and the other with $K = 2$ under two-sided non-compliance (Design 2). In each design, we correctly specify q, i.e., the number of valid analyses. We also conduct additional simulations with a misspecified q to assess its impact on the performance of our approach. See Section S4 of the Supplementary Material for details of the simulation implementation. We present the results for Design 1 (with correctly specified q) below, and include additional results with a misspecified q in Section S5 of the Supplementary Material. We also include the results for Design 2 in Section S6 of the Supplementary Material. R code for our simulation study is available at: https://github.com/youjin1207/EFinFuzzyRD.

6.1. Simulation setup

We generate simulation data for $n = 1000$ subjects. For each subject, we first generate the baseline covariate and the running variable as follows: $W_i \overset{i.i.d.}{\sim} Unif(-1,1)$ and $X_i = 0.5W_i + X_i^*$, where $X_i^* \overset{i.i.d.}{\sim} Unif(-0.5,1)$. Then, we generate each subject's eligibility as $A_i = I(X_i \leq 0)$; that is, each subject i is eligible for treatment if and only if $X_i \leq 0$. In addition to the observed covariate, we generate an unmeasured covariate $U_{i,1} \overset{ind}{\sim} Bern(0.5I(X_i \leq 0) + 0.3)$, which may violate the exclusion restriction of the eligibility status (A_i) on the outcome (Y_i). Two unmeasured covariates $U_{i,k} \overset{i.i.d.}{\sim} Unif(-1,1)$ are also generated, each of which can introduce confounding in the treatment assignment at level k ($k = 2,3$). Next, we generate the subsequent treatment statuses, $(Z_i^{(2)}, T_i)$, and the potential outcomes with $\varepsilon_{z,i} \overset{i.i.d.}{\sim} N(0,0.01)$, $\varepsilon_{t,i} \overset{i.i.d.}{\sim} N(0,0.01)$, and $\varepsilon_{y,i} \overset{i.i.d.}{\sim} N(0,1)$ as follows:

$$
\begin{aligned}
Z_i^{(2)} &= I(Z_i^* > 0.5), \text{ where } Z_i^* = (5 + 6A_i - 0.5X_i + W_i + 3U_{i,2})/20 + \varepsilon_{z,i} \\
T_i &= I(T_i^* > 0.5), \text{ where } T_i^* = (6 + 4Z_i^{(2)} - 0.5X_i + W_i + W_i^2 + U_{i,3})/20 + \varepsilon_{t,i} \\
Y_i^0 &= 8 + X_i + 0.5X_i^2 + 0.5W_i + 0.5W_i^2 + \gamma_1 U_{i,1} + \gamma_2 U_{i,2} + \gamma_3 U_{i,3} + \varepsilon_{y,i} \\
Y_i^1 &= Y_i^0 + \tau.
\end{aligned}
$$

Figure 4 illustrates the causal structure of the observed and unmeasured variables used in our simulation design with $K = 3$. Our null hypothesis of interest is $H_{0,K} : \tau = 0$. We conduct one-sided tests, testing

https://doi.org/10.1017/psy.2025.10033 Published online by Cambridge University Press

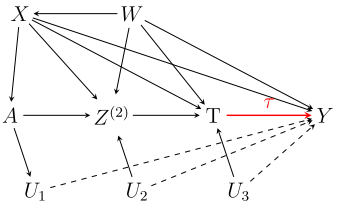

Figure 4. A causal structure of the simulated data.
Note: The dashed lines indicate the presence of a causal relationship between U_k and Y when the corresponding γ_k value is non-zero ($k \in [1:3]$).

the null of no causal effect against the alternative of a greater effect of the treatment. We introduce unmeasured bias through $U_{i,k}$'s ($k = 1,2,3$). For example, the covariate $U_{i,1}$ mediates the effect of A_i on Y_i, potentially violating the exclusion restriction by generating a causal pathway between A_i and Y_i not through T_i when $\gamma_1 \neq 0$. This bias would not necessarily be avoided by the window selection around the cutoff (i.e., around zero). On the other hand, the covariates $U_{i,2}$ and $U_{i,3}$ are associated with $Z_i^{(2)}$ and T_i, respectively, potentially confounding their relationships with Y_i when $\gamma_2 \neq 0$ and $\gamma_3 \neq 0$, respectively.[3]

In this simulation design, we consider seven different cases depending on the value of $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)$. In case (1), $\boldsymbol{\gamma} = (0,0,0)$, implying that there is no effect of the three unmeasured covariates on the outcome. In cases (2)–(4), one of the three treatment assignment processes is biased, whereas two of the three are biased in cases (5)–(7). For each case, we report the rejection rates of the three p-values from each EF, denoted as (P_1, P_2, P_3), and their combined p-value using Fisher's method, P_q^c, given a pre-specified q. Similarly, we present the results of our comparison analysis, denoted as (P_1^*, P_2^*, P_3^*), and their combined p-value, $P_q^{c^*}$. In case (1), we set $q = 3$; in cases (2)–(4), $q = 2$; and in cases (5)–(7), $q = 1$.

6.2. Simulation results

Figure 5 summarizes the results for four out of seven cases: (1) $\boldsymbol{\gamma} = (0.0, 0.0, 0.0)$, (2) $\boldsymbol{\gamma} = (1.0, 0.0, 0.0)$, (3) $\boldsymbol{\gamma} = (0.0, 0.5, 0.0)$, and (5) $\boldsymbol{\gamma} = (1.0, 0.5, 0.0)$. When all the three treatment assignment processes are unbiased, i.e., in case (1), both the proposed EF analyses and the unconditioned comparisons produce valid p-values. We observe that the testing power of the proposed EFs at $k = 2$ and 3 is higher than that of the unconditioned comparisons under our data-generating process. Specifically, in case (1), the unconditioned comparisons of P_2^* and P_3^* show lower rejection rates than P_2 and P_3. This might sound counterintuitive, as unconditioned comparisons typically include a larger number of subjects by not conditioning on previous treatment statuses. Our further investigation shows that unconditioned comparisons can produce either conservative or anti-conservative p-values depending on simulation scenarios (see Section S5 of the Supplementary Material for additional simulations).

When at least one level of the treatment assignment is biased, the bias from that level does not affect EF analyses at other levels. However, in the unconditioned analyses, bias occurring at level k ($k \in [1:K-1]$) may spill over to subsequent levels ℓ ($\ell \in [k+1:K]$). This expectation is supported by the results in cases (2), (3), and (5). Inflated type-I errors are observed both in the analysis at level k where $\gamma_k \neq 0$ and in subsequent unconditioned comparisons at level ℓ, where $\ell \in [k+1:K]$. Consequently, the combined p-value becomes invalid under unconditioned analyses (indicated by dashed black lines) as fewer than q unconditioned analyses can be valid. On the other hand, our proposed EFs with a correctly specified number of q, i.e., 2 in cases (2) and (3) and 1 in case (5), show that the biased treatment assignment at level k does not affect the EF analysis at level k' ($k \neq k'$). For example, in case (2), while P_1 is invalid, P_2 and P_3 are valid p-values. Also, in case (3), while P_2 is invalid, P_1 and P_3 are valid. As a

[3] Either the violation of the exclusion restriction or the presence of unmeasured confounders can invalidate an analysis using an IV, $Z_i^{(k)}$ ($k \in [1:K-1]$). In practice, we cannot distinguish if an IV analysis is biased due to the violation of the exclusion restriction or to unmeasured confounders.

https://doi.org/10.1017/psy.2025.10033 Published online by Cambridge University Press

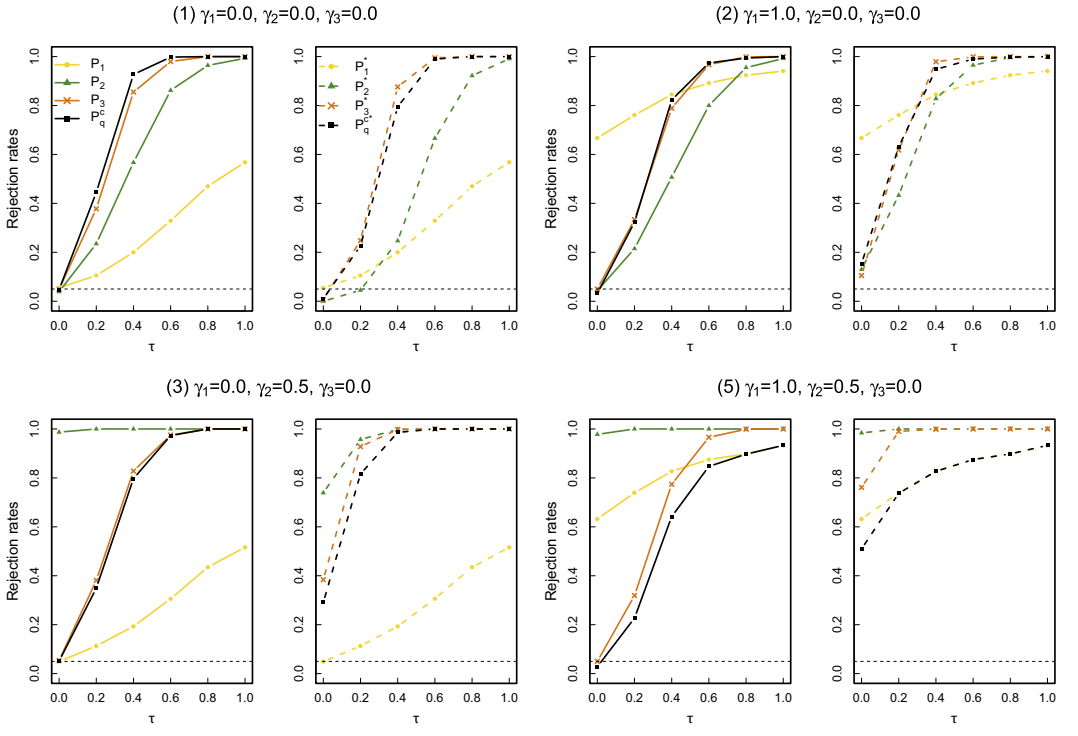

Figure 5. Comparison of rejection rates at the $\alpha = 0.05$ level between the proposed EFs and the unconditioned comparisons, based on 1000 replicates with a sample size of 1000 across four selected cases: (1) $\gamma = (0.0, 0.0, 0.0)$, (2) $\gamma = (1.0, 0.0, 0.0)$, (3) $\gamma = (0.0, 0.5, 0.0)$, and (5) $\gamma = (1.0, 0.5, 0.0)$.
Note: A value of τ denotes the effect of T_i on Y_i; P_k is a p-value from each proposed EF for $k \in [1:3]$; P_q^c is a combined p-value of (P_1, P_2, P_3); P_k^* is a p-value from each comparison at level k without conditioning on $\mathbf{Z}^{(1:k-1)} = \mathbf{1}$ $(k \in [1:3])$; $P_q^{c^*}$ is a combined p-value of (P_1^*, P_2^*, P_3^*); $q = 3$ in case (1), $q = 2$ in cases (2) and (3), and $q = 1$ in case (5).

result of such non-overlapping bias across EFs, the combined p-values of the EF analyses are still valid in these cases, even without knowing exactly which assignment is biased.

One caveat we note is that the way we generate the assignment variables may lead to misspecification of the assignment models described in Section 4.3, even though they do not necessarily affect the causal assumptions outlined in Section 4.2. Our results demonstrate robustness to such misspecification, although this robustness may not be guaranteed in other settings.

7. Empirical examples

In this section, we demonstrate real-world applications of our proposed EF analysis for fuzzy RD designs. The first data analysis evaluates a pre-K program using two different treatment statuses in the context of two-sided non-compliance. The second data analysis evaluates an ETA program using three treatment statuses under a series of one-sided non-compliance.

7.1. The effect of New Jersey's Pre-K program

We use New Jersey's data from Wong et al. (2007) to evaluate the effect of New Jersey's Abbott Preschool Program—one of three state-funded pre-K initiatives in New Jersey—on 4-year-old children's vocabulary skills. In our data analysis, we consider two different treatment statuses: eligibility and use. Eligibility status A_{ij} (i.e., $Z_{ij}^{(1)}$) represents whether a child is eligible for the pre-K program or not, which

https://doi.org/10.1017/psy.2025.10033 Published online by Cambridge University Press

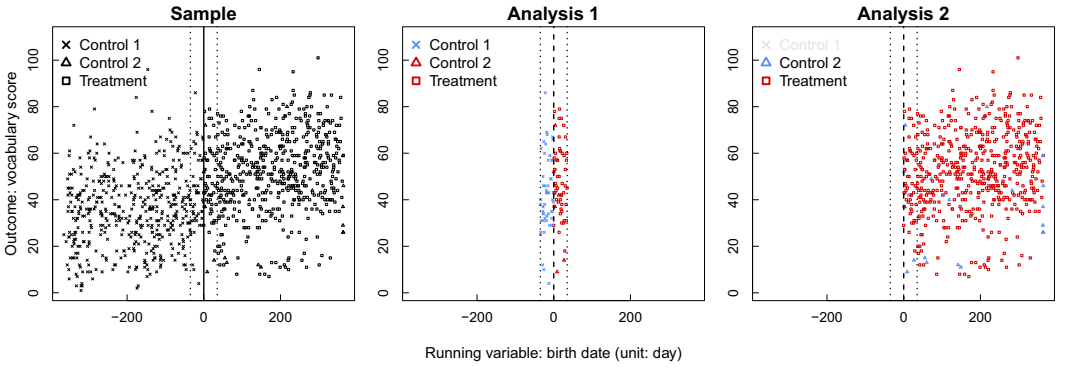

Figure 6. Distributions of three different treatment groups in two EFs analysis in New Jersey's Pre-K program: for each analysis, red symbols represent the treatment comparison group, while the skyblue symbols represent the control comparison group.

is determined by the running variable of a child's birth date, X_{ij} (unit: day); children with birthdays on or after a specific date are eligible for the pre-K program, whereas those with birthdays before it are not. This eligibility status differs from a child's actual use of or participation in the program, and this non-compliance occurs in a two-side format. Use status T_{ij} (i.e., $Z_{ij}^{(2)}$) represents whether he/she completed the pre-K program in spring 2004. Our outcome of interest is children's vocabulary scores measured in early fall of the 2004–2005 school year. Measured covariates $\mathbf{W}_{ij}$ include gender, race/ethnicity, free lunch status, and test language type (English or Spanish). Potential unmeasured covariates (U_{ij}) include prior abilities and distance to facilities.

In this application, we construct two EFs to test the same null hypothesis ($H_{0,K}$) that there is no effect from participating in the pre-K program. The first EF compares eligible students ($A_{ij} = 1$) and ineligible students ($A_{ij} = 0$) in a small neighborhood of the cutoff. This factor is susceptible to bias associated with the eligibility rule under the RD design; for example, ineligible children in New Jersey with high prior abilities might relocate to be eligible for and participate in pre-K programs in other states, thereby "manipulating" the running variable based on unmeasured prior abilities. Moreover, the second EF compares students who participate in the pre-K program ($T_{ij} = 1$) and those who do not ($T_{ij} = 0$) among the eligible ($A_{ij} = 1$). This factor is vulnerable to bias associated with children's actual participation. For example, if the program is located too far away or lacks transportation options, some parents might be unable to assist their children in getting there.

For our data analysis, we first stratify individual children based on the four measured covariates and then conduct a stratified Wilcox rank sum test for each EF. For the first EF, we further constraint our sample near the cutoff by using window selection under the local randomization RD framework. The largest possible window around the cutoff, set at zero, in our data is (−35, 35), which ensures that the p-value of the covariate balance test exceeds 0.15. For the first EF, we also use the residual outcome, obtained by regressing the outcome Y on X, to remove the impact of X on Y within the window; see Figure 6 for the distributions of three different treatment groups in two EF analyses. Lastly, we combine one-sided p-values from multiple EFs using Fisher's method and conduct sensitivity analyses by varying Γ_k for each factor.

Table 4 presents the results of EF analysis for New Jersey's pre-K program. In the absence of unmeasured confounding ($\Gamma_k = 1.0$), the first EF does not reject the null, whereas the second EF rejects the null. The combined p-value, calculated using Fisher's method, is 0.045. Therefore, our EF analysis leads us to reject the null. This suggests a positive effect from participating in the pre-K program. However, sensitivity analyses reveal that adjusting the p-values at $\Gamma_1 = 1.1$ or at $\Gamma_2 = 1.1$ results in combined p-values above 0.05. These findings indicate that our causal conclusion are sensitive to even small degrees of unmeasured confounding at $\Gamma_k = 1.1$ ($k = 1,2$).

https://doi.org/10.1017/psy.2025.10033 Published online by Cambridge University Press

Table 4. Results of EFs analysis for New Jersey's pre-K program

	Evidence factor 1	Evidence factor 2	Fisher's
J_k (|control| vs. |treated|)	8 (51 vs. 57)	9 (21 vs. 555)	
p-value ($\Gamma_k = 1.0$ for $k = 1,2$)	0.240	0.032	0.045
p-value ($\Gamma_1 = 1.1$ & $\Gamma_2 = 1.0$)	0.314	0.032	0.056
p-value ($\Gamma_1 = 1.0$ & $\Gamma_2 = 1.1$)	0.240	0.048	0.063
p-value ($\Gamma_1 = 1.1$ & $\Gamma_2 = 1.1$)	0.314	0.048	0.078

Note: J_k represents the number of strata in EF k. Γ_k represents the sensitivity parameter in EF k. Fisher's method is used to combine p-values from two EFs. |control| and |treated| represent the sample size for the control comparison group and treatment comparison group, respectively.

7.2. The effect of ETA

We analyze the Grade 8 Mathematics Process Data from the 2017 National Assessment of Educational Progress (NAEP) to study the effect of using ETA on students' math scores. In this ETA example, we consider three different treatment statuses: eligibility, receipt, and use. Eligibility status A_{ij} (i.e., $Z_{ij}^{(1)}$) is binary and determined by an ordinal running variable. The running variable is ELL English proficiency with six ordinal levels: *No Proficiency, ELL Beginning, ELL Intermediate, ELL Advanced, Formerly ELL,* and *Never ELL*. The eligibility status differs from the receipt status $Z_{ij}^{(2)}$ and further differs from the use status T_{ij} (i.e., $Z_{ij}^{(3)}$). We use students' performance on a 15-item math test block as the outcome Y_{ij}. Measured covariates $\mathbf{W}_{ij}$ include gender, race/ethnicity, primary language, free lunch status, duration of U.S. school education (less than 1 year or more), and perseverance. Potential unmeasured covariates include students' years of residence in the US, test anxiety, and academic motivation. In our analysis sample, we exclude students with *Never ELL*, and non-compliance occurs sequentially in a one-sided format.

In this example, we construct three EFs to test the null hypothesis ($H_{0,K}$) that there is no effect of using ETA. The first EF compares eligible students ($A_{ij} = 1$) and ineligible students ($A_{ij} = 0$) in a small neighborhood of the cutoff. This factor is susceptible to bias associated with the eligibility rule under the RD design, as in the pre-K example. The second EF compares students who receive ETA from school administrators ($Z_{ij}^{(2)} = 1$) and those who do not ($Z_{ij}^{(2)} = 0$) among the eligible ($A_{ij} = 1$), and this factor is susceptible to bias associated with school administrator's decision on ETA receipt. For example, school staff may consider students' test anxiety levels to determine which students receive ETA in practice. The third EF compares students who actually use ETA ($T_{ij} = 1$) versus non-users ($T_{ij} = 0$) among recipients ($Z_{ij} = 1$). This factor is susceptible to bias associated with students' individual characteristics. For example, students who have higher motivation may actually make use of ETA.

Similar to the pre-K example, we stratify individual students based on the six measured covariates and conduct a stratified Wilcoxon rank sum test for each EF. For the first EF, however, we estimate propensity scores to create a surrogate continuous running variable from the oridinal one, following a recent approach proposed by Li et al. (2021). We select a window of (−0.01, 0.01) around the median of the propensity scores among students with the cutoff category (i.e., *ELL Advanced*). We also use the residual outcome for the first EF. Then, we calculate one-sided p-values for each EF and combine them using Fisher's method. In this application, we use four different treatment groups, similar to those in Figure S2 in Supplementary Appendix S4 from our simulation study, except that we employ the surrogate continuous running variable.

Table 5 provides the results of EFs analysis for ETA. In the absence of unmeasured confounding, all the three EFs do not reject the null, and the combined p-value using Fisher's method is 0.135. As a result, our EF analysis does not reject the null. This concurrence reinforces our conclusion that there is no effect of using ETA on math scores in the NAEP assessment. Since none of the EFs is significant, we do not conduct sensitivity analyses by adjusting Γ_k for each factor.

https://doi.org/10.1017/psy.2025.10033 Published online by Cambridge University Press

Table 5. Results of EF analysis for the ETA

	EF 1	EF 2	EF 3	Fisher's
J_k (|control| vs. |treated|)	6 (40 vs. 80)	10 (510 vs. 120)	10 (90 vs. 30)	
p-value	0.231	0.556	0.059	0.135

Note: J_k represents the number of strata in EF *k*. Fisher's method is used to combine *p*-values from two EFs. |control| and |treated| represent the sample size for the control comparison group and treatment comparison group, respectively. Numbers for |control| and |treated| are rounded to nearest tens. Details may not sum to a total due to rounding.*Source*: U.S. Department of Education, National Center on Educational Statistics (NCES), 2017 NAEP Grade 8 Mathematics Process Data, Student Features Data File Partial Form, and Response Data File.

8. Discussion and conclusions

In this article, we propose a new EF approach in fuzzy RD designs, which involve sequential treatment assignments and the locality inherent in analyzing samples near the cutoff. This approach is particularly useful when multiple decision-makers are involved in determining treatment statuses sequentially, each producing different types of biases. We establish the causal assumptions required for constructing valid EFs in fuzzy RD designs and demonstrate the effectiveness of our proposed EFs approach through simulation studies by comparing it with unconditional analyses (e.g., as-treated analysis). We also illustrate its effectiveness using two real-world data examples that involve fuzzy RD designs with one-sided and two-sided non-compliance, respectively.

Even though we introduce the setting where the first level of treatment assignment involves RD designs, our proposed approach has broader applicability. More specifically, our approach is applicable to any setting with sequential assignment processes, where one of the comparisons involves an RD analysis. In our ETA example, suppose that ELL English proficiency levels (i.e., running variable) are not available for some students, making it difficult to determine their eligibility status. In this case, the availability for the English proficiency test can be considered as the first treatment status, while the second treatment status is eligibility based on the running variable, which exhibits an RD design. We can still apply our proposed framework in this scenario to construct multiple EFs, as long as the analysis with the RD design properly adjusts the analysis sample to satisfy the local randomization assumption.

There are several limitations in our proposed approach. First, we found that empirical results can be impacted by the way the observations are stratified, which is crucial for randomization-based inference. Therefore, the window selection in the local randomization framework, as well as the choice of matching methods, can impact the empirical results. However, strict criteria for constructing strata, such as using a narrow caliper or imposing a hard threshold for balance tests, may considerably reduce statistical power. Second, we focus on multiple levels of treatment assignment that are potentially affected by individual-level characteristics, e.g., due to individual-level eligibility, constraints, or preferences. However, often such assignment can also be affected by macro-level factors beyond individuals. Future research will formally discuss different sources of biases from multilevel or clustered data in EF analysis, even when treatment is assigned to individual subjects. Third, choosing a presumably valid subset of treatment statuses that satisfy both conditions of Assumption 2 could be challenging in practice. Future research would examine ways to relax this assumption or develop robust methods that are less sensitive to its violation. Lastly, in our simulation and data application studies, we use residuals obtained by regressing the outcome on the running variable (rather than the true errors), which induces correlation between the residuals. As a result, this residual randomization approach remains asymptotically valid under certain conditions (Freedman & Lane, 1983; Sales & Hansen, 2019), even though, in our simulation studies, the rejection rates under the null achieve their nominal levels in finite samples. Despite these limitations, our proposed approach complements the existing RD literature by introducing a novel EF design with sequential treatment assignments embedded in fuzzy RD designs. We hope that our approach serves as a useful tool for reinforcing our causal conclusions in observational studies with fuzzy RD designs.

https://doi.org/10.1017/psy.2025.10033 Published online by Cambridge University Press

Supplementary material. R code for our simulation study and real data applications, along with example datasets, is available at: https://github.com/youjin1207/EFinFuzzyRD.

Data availability statement. The datasets used in this article are not publicly available. One dataset is secondary data obtained from Wong et al. (2007), and the other is the restricted-use version of the Grade 8 Mathematics Process Data from the 2017 National Assessment of Educational Progress (NAEP). A restricted-use data license can be obtained through the IES Electronic Application System (https://nces.ed.gov/statprog/instruct.asp).

Acknowledgements. The authors would like to thank Vivian Wong for granting permission to use her data from Wong et al. (2007) for the purpose of demonstrating our proposed evidence factors approach in this work.

Funding statement. Open Access funding was provided by Columbia University under an agreement between Columbia University and Cambridge University Press.

Competing interests. The authors declare that there are no financial or other conflicts of interest in relation to the work.

References

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, *91*(434), 444–455. https://doi.org/10.3386/t0136

Cattaneo, M. D., Frandsen, B. R., & Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the U.S. senate. *Journal of Causal Inference*, *3*(1), 1–24. https://doi.org/10.1515/jci-2013-0010

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2024). *A practical introduction to regression discontinuity designs: Extensions*. Cambridge University Press. https://doi.org/10.1017/9781009441896

Cattaneo, M. D., Keele, L., & Titiunik, R. (2023a). *Covariate adjustment in regression discontinuity designs* (pp.153–168). https://doi.org/10.1201/9781003102670-8

Cattaneo, M. D., Keele, L., & Titiunik, R. (2023b). A guide to regression discontinuity designs in medical applications. *Statistics in Medicine*, *42*(24), 4484–4513. https://doi.org/10.1002/sim.9861

Cook, T. D. (1985). Post-positivist critical multiplism. In: R. L. Shotland & M. M. Mark (Eds.), *Social science and social policy* (pp. 21–62). Newbury Park, CA: Publications.

Crespo, C. (2020). Beyond manipulation: Administrative sorting in regression discontinuity designs. *Journal of Causal Inference*, *8*(1), 164–181. https://doi.org/10.1515/jci-2019-0009

Dong, Y. (2018). Alternative assumptions to identify late in fuzzy regression discontinuity designs. *Oxford Bulletin of Economics and Statistics*, *80*(5), 1020–1027. https://doi.org/10.1111/obes.12249

Freedman, D., & Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, *1*(4), 292. https://doi.org/10.2307/1391660

Hahn, J., Todd, P., & Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*(1), 201–209. https://doi.org/10.1111/1468-0262.00183

Karmakar, B., & Small, D. S. (2023). Constructing independent evidence from regression and instrumental variables with an application to the effect of violent conflict on altruism and risk preference. *Biostatistics & Epidemiology*, *7*(1), e2109910. https://doi.org/10.1080/24709360.2022.2109910

Karmakar, B., Small, D. S., & Rosenbaum, P. R. (2020). Reinforced designs: Multiple instruments plus control groups as evidence factors in an observational study of the effectiveness of catholic schools. *Journal of the American Statistical Association*, *116*(533), 82–92. https://doi.org/10.1080/01621459.2020.1745811

Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. house elections. *Journal of Econometrics*, *142*(2), 675–697. https://doi.org/10.1016/j.jeconom.2007.05.004

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355. https://doi.org/10.1257/jel.48.2.281

Li, F., Mercatanti, A., Mäkinen, T., & Silvestrini, A. (2021). A regression discontinuity design for ordinal running variables: Evaluating central bank purchases of corporate bonds. *The Annals of Applied Statistics*, *15*(1), 304–322. https://doi.org/10.1214/20-AOAS1396

Rosenbaum, P. R. (2010). Evidence factors in observational studies. *Biometrika*, *97*(2), 333–345. https://doi.org/10.1093/biomet/asq019

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, *74*(1), 13. https://doi.org/10.2307/2336017

Rosenbaum, P. R. (2002). *Observational studies*. Springer. https://doi.org/10.1007/978-1-4757-3692-2

Rosenbaum, P. R. (2023). A second evidence factor for a second control group. *Biometrics*, *79*(4), 3968–3980. https://doi.org/10.1111/biom.13921

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Sales, A. C., & Hansen, B. B. (2019). Limitless regression discontinuity. *Journal of Educational and Behavioral Statistics*, *45*(2), 143–174. https://doi.org/10.3102/1076998619884904

Suk, Y. (2024). Regression discontinuity designs in education: A practitioners guide. *Asia Pacific Education Review*, *25*, 629–645. https://doi.org/10.1007/s12564-024-09956-3

Suk, Y., & Kim, Y. (2024). Fuzzy regression discontinuity designs with multiple control groups under one-sided noncompliance: Evaluating extended time accommodations. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/10769986241268902

Suk, Y., Steiner, P. M., Kim, J.-S., & Kang, H. (2022). Regression discontinuity designs with an ordinal running variable: Evaluating the effects of extended time accommodations for English-language learners. *Journal of Educational and Behavioral Statistics*, *47*(4), 459–484. https://doi.org/10.3102/10769986221090275

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51*(6), 309–317. https://doi.org/10.1037/h0044319

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In: S. Kotz & N. L. Johnson, (Eds.), *Breakthroughs in statistics*. Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-1-4612-4380-9_16

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2007). An effectiveness-based evaluation of five state pre-kindergarten programs. *Journal of Policy Analysis and Management*, *27*(1), 122–154. https://doi.org/10.1002/pam.20310

Zaykin, D., Zhivotovsky, L. A., Westfall, P., & Weir, B. (2002). Truncated product method for combining pvalues. *Genetic Epidemiology*, *22*(2), 170–185. https://doi.org/10.1002/gepi.0042

Zhao, A., Lee, Y., Small, D. S., & Karmakar, B. (2022). Evidence factors from multiple, possibly invalid, instrumental variables. *The Annals of Statistics*, *50*(3), 1266–1296. https://doi.org/10.1214/21-aos2148

Cite this article: Lee, Y. and Suk, Y. (2025). Evidence Factors in Fuzzy Regression Discontinuity Designs with Sequential Treatment Assignments. *Psychometrika*, 1400–1418. https://doi.org/10.1017/psy.2025.10033