Submitted to the Annals of Statistics

EVIDENCE FACTORS FROM MULTIPLE, POSSIBLY INVALID, INSTRUMENTAL VARIABLES

BY ANQI ZHAO*,[1], YOUJIN LEE*,[2], DYLAN S. SMALL[3] AND BIKRAM KARMAKAR[4]

[1]*Department of Statistics and Data Science, National University of Singapore, staza@nus.edu.sg*

[2]*Department of Biostatistics, Brown University, youjin_lee@brown.edu*

[3]*Department of Statistics, University of Pennsylvania, dsmall@wharton.upenn.edu*

[4]*Department of Statistics, Univeristy of Florida, bkarmakar@ufl.edu*

Valid instrumental variables enable treatment effect inference even when selection into treatment is biased by unobserved confounders. When multiple candidate instruments are available, but some of them are possibly invalid, the previously proposed reinforced design enables one or more nearly independent valid analyses that depend on very different assumptions. That is, we can perform evidence factor analysis. However, the validity of the reinforced design depends crucially on the order in which multiple instrumental variable analyses are conducted. Motivated by the orthogonality of balanced factorial designs, we propose a balanced block design to offset the possible violation of the exclusion restriction by balancing the instruments against each other in the design, and demonstrate its utility for constructing approximate evidence factors under multiple analysis strategies free of the order imposition. We also propose a novel stratification method using multiple, nested candidate instruments, in which case the balanced block design is not applicable. We apply our proposed methods to evaluate (a) the effect of education on future earnings using instrumental variables arising from the disruption of education during World War II via the balanced block design, and (b) the causal effect of malaria on stunting among children in Western Kenya using three nested instruments.

1. Introduction.

1.1. *Bias, instrumental variable, and replication.* In an observational study, a comparison between the treated and control groups might be biased because of unmeasured confounders, that is, because of pretreatment difference between the groups in an unmeasured variable that is associate with the outcome. Often this bias is systematic and recurs in replicated studies (Rosenbaum, 2001). In such a case, an instrumental variable, a "random nudge" to accept a treatment, provides a unique opportunity to separate a meaningful causal effect from bias (Angrist, Imbens and Rubin, 1996; Rosenbaum, 2010a). An instrumental variable that is associated with a treatment variable enables us to compare the two treatment arms by mimicking an experimental design. Consider studying the causal effect of educational attainment on earnings which we use as our running example throughout this paper. Education levels are not randomly assigned to individuals, and education and earnings are both likely to be affected by many factors, e.g., self-motivation and exposure to financial resources, that are difficult to measure accurately. Card (1993) proposed to use proximity to a college as an instrument since it is highly related to educational attainment and may be considered random.

*The authors contributed equally.

MSC2020 subject classifications: Primary 62G10; secondary 62K10.

Keywords and phrases: Bias in observational studies, causal inference, exclusion restriction, nonparametric tests, replication, sensitivity analysis.

However, a valid usage of an instrumental variable is subject to a few conditions. There must be no direct effect of the instrument on the outcome, i.e., it must satisfy the exclusion restriction assumption, and no unmeasured confounders between the instrument and the outcome. However, in practice, these assumptions are often violated even after controlling for observed covariates and are untestable without solid scientific knowledge. In Card (1993), for example, living near a college might have a direct effect on earnings because it can lead to different employment opportunities and/or might be associated with self-motivation or parents' ambition for their children that are also associated with earnings, which in the end manifests in a spurious effect between education and earnings.

Depending on the context, other instruments can be used for the same question. For example, to examine the same question of the causal effect of education on earnings, Angrist and Krueger (1991) used quarter of birth interacted with year of birth as an instrument, and Harmon and Walker (1995) used changes in the minimum school leaving age as an instrument. Card (1999) reviewed other instrumental variable choices, too. To see that quarter of birth could affect education, consider a person who leaves school at the minimum school leaving age. If the cutoff for entering the 1st grade is being born in a certain year (as it used to be in the US), then he or she will have less education if born in the first quarter than the fourth quarter. Even though each of these proposed instruments may violate the exclusion restriction and no unmeasured confounding assumption, there is no clear reason to think that biases due to invalid instruments recur in the same direction by multiple studies. For such a setting, we may obtain randomness in bias from multiple instrumental variables each of which is subject to different sources of bias that do not completely overlap with each other. For this reason, if results from the multiple independent studies based on different instrumental variables concur, their common conclusion will be more reliable than those from multiple studies that directly compare treated and control subjects. Compared to the former studies, the latter studies are more likely to be exposed to the same or similar bias which consistently draws the conclusion away from the truth.

Is there any way to obtain randomness in bias and independent analyses from using multiple candidate instruments based on a single study? If multiple instruments tend to induce bias in similar directions, e.g., subjects who live near colleges are likely to be more self-motivated and have birth dates earlier in the year in Card (1993), then there would be less gain in using multiple instruments as one bias may affect multiple results in the same direction. However, if each instrumental variable analysis could produce an orthogonal piece of evidence, we can obtain randomness in bias from multiple instruments even without replicating studies. To achieve this, we adopt the framework of evidence factor analysis (Rosenbaum, 2010b) — multiple, nearly independent analyses of the same data set that are subject to different potential biases — to multiple instrumental variables that would introduce unsystematic uncertainties in bias from each instrument. In this framework, a conclusion from several pieces of evidence tends to replicate when it is correct, and tends to fail to replicate when it is incorrect or the studies do not have sufficient power (Rosenbaum, 2001).

This paper proposes the *balanced block design* and *mutual stratification* as two new evidence factor analysis methods to implement multiple causal comparisons with instrumental variables when invalid instruments may be present. In our proposed methods, each analysis from multiple instruments is valid to study the causal effect when the putative instrument is valid, and even if it is invalid without us knowing so, its bias will not affect the validity of other analyses. We introduce three different non-parametric randomization based tests for each candidate instrument: *marginal tests*, *conditional tests*, and *reinforced tests*. We show that under the balanced block design, (a) for an instrument that is conditionally valid given the other instruments, any one of these three types of tests is asymptotically valid under mild regularity conditions; (b) using any of the three tests, the valid p-values calculated are

asymptotically jointly stochastically larger than the uniform distribution on the unit cube, which implies their near independence; and (c) the results of (a) and (b) hold in sensitivity analysis to unmeasured confounders with the upper bounds of the p-values (in place of the p-values) when the instruments are conditionally valid only after allowing for some amount of unmeasured confounding. When the candidate instruments are nested, the balanced block design is not applicable. We show that for nested instruments, (a)–(c) continue to hold with the proposed conditional tests under mutual stratification.

We present two empirical studies using our methods. The balanced block design method is applied to investigate the effect of education on future earnings using two instrumental variables arising from the disruption of education during the war. The mutual stratification method is applied to examine the effect of malaria on stunting using three nested instruments from a cluster randomized trial involving children in Western Kenya. By applying the proposed methods for each study, we can learn several nearly independent pieces of evidence from a single data set even in the presence of invalid instruments.

1.2. *Literature review.* *Evidence factor analysis* performs multiple and nearly independent analyses, also known as *factors*, each depending on assumptions that do not completely overlap (Rosenbaum, 2010b; Rosenbaum et al., 2017). When two or more nearly independent analyses for a causal hypothesis provide supportive evidence, the evidence for a causal effect is strengthened. This is because neither bias nor statistical error that might invalidate one piece of the evidence can invalidate supportive evidence from the other factors. Further, an evidence factor analysis allows sensitivity analyses to unmeasured confounders. Thus, evidence for a causal effect is strengthened if multiple evidence factors provide support for a causal effect, and such evidence is further robust to a moderate degree of unmeasured confounding in different directions. Evidence factor analysis has been developed and used in prospective studies (Zhang et al., 2011; Zubizarreta et al., 2012), in case-control studies (Karmakar, Doubeni and Small, 2020), to deconfound the effect of exposure biomarkers in a study of the effect of an external exposure (Karmakar, Small and Rosenbaum, 2020), and with multiple treatments (Rosenbaum et al., 2017; Nattino et al., 2021). Karmakar, French and Small (2019) and Karmakar and Small (2020) studied design sensitivity of an evidence factor analysis.

One source of multiple analyses is multiple instruments. Karmakar, Small and Rosenbaum (2021) proposed the *reinforced design* to use multiple instrumental variables in an evidence factor analysis where one comparison can be valid even in the presence of invalid instruments. When there are K candidate instruments, this method develops $K + 1$ evidence factors from those instruments plus the direct comparison of treatment and matched controls. To apply this method, one needs to delicately build independent analyses using multiple instrumental variables that are possibly correlated with each other. Previously, this was done by specifying an order of analyses within the instrumental variables set and imposing a set of partial exclusion restrictions for each instrument where all previous instruments are fixed by conditioning. However, this method may collapse when at least one instrumental variable violates the partial exclusion restriction assumptions and when we have no information on which instrumental variable does. This is further elaborated in Section 2.2. Our proposed methods in this paper disentangle the order imposition and thus have wider applications to studies with multiple candidate instruments.

Prior work on the violation of the exclusion restriction often constrains the number of invalid instruments, instead of the order of analyses. The methods provided by Han (2008), Bowden et al. (2016), Kang et al. (2016), Guo et al. (2018) and Windmeijer et al. (2019) require that at least 50% among the candidate instruments are valid. Furthermore, several robust methods have been proposed in Mendelian randomization by adapting the approaches

used in a meta-analysis, viewing combining evidence from multiple instruments as combining results from multiple studies (Burgess, Butterworth and Thompson, 2013; Bowden, Davey Smith and Burgess, 2015; Burgess, Dudbridge and Thompson, 2016). Under the summarized genome-wide association studies (GWAS) setting, different versions of the inverse variance weighted methods have been developed, often assuming no correlation between genetic instruments (Han, 2008; Greco M et al., 2015). Kolesár et al. (2015) and Bowden, Davey Smith and Burgess (2015) proposed the causal effects estimation method provided that the instruments' effects on the treatment are orthogonal to their effects on the outcome. All of the methods above utilize multiple instruments, some of which are possibly invalid, to produce one valid causal inference under some additional assumptions. In contrast, in our proposed methods, the nearly independent factors can be combined to provide a robust causal conclusion when they corroborate each other, or they can be used to caution against a causal conclusion when there are large discrepancies among them. Since the factors are by construction susceptible to separate sources of biases, the proposed methods provide useful detail regarding the strength of the evidence in the presence of invalid instruments.

Literature on sensitivity analysis for invalid instruments is limited. Small and Rosenbaum (2008) considered sensitivity to a nonrandom assignment of an instrument. Small (2007) proposed multivariate sensitivity parameters from a regression model to evaluate the effect of possible violation of the exclusion restriction. In summarized GWAS data, sensitivity analyses to each instrument's assumption can be performed using visualization approaches such as a funnel plot or a scatter plot. See Burgess et al. (2017) for more sensitivity analysis tools in a Mendelian randomization analysis. Wang et al. (2018) developed the sensitivity analysis method using a sensitivity parameter that quantifies the associations of the candidate instruments with unmeasured confounders and with the outcome. Recently, Spieker et al. (2020) proposed a generalized version of the exclusion restriction, of which strength is controlled by a suggested sensitivity parameter. Most of the sensitivity analyses above heavily depend on the treatment and/or outcome models. On the other hand, Kang et al. (2021) viewed the maximum number of invalid instruments as a sensitivity parameter, resulting in more conservative inference with a higher maximum number. Our work will extend the sensitivity analysis tools in Small and Rosenbaum (2008) and Kang et al. (2021).

1.3. *Modification of dependence structure via the balanced block design.* Violation of the exclusion restriction or no unmeasured confounding assumptions subjects randomization tests of the candidate instruments to liberal type-I error rates. In the presence of multiple candidate instruments, some instruments may directly violate the exclusion restriction whereas others are conditionally-valid after conditioning on the directly-invalid ones. The test for a treatment effect using a conditionally-valid instrument may be invalidated due to the lack of proper conditioning.

One way to relax the requirement on proper conditioning is to form balanced blocks in which there are an equal number of units under each combination of the candidate instruments. Analogous to the balanced factorial design in the context of randomized experiments, the blocking intuitively balances the distributions of the directly-invalid instruments within each level of the conditionally-valid ones, and thereby ensures the validity of the resulting tests even in the absence of proper conditioning. The explicit proof of the utility of the balanced blocking is not trivial, however. This is because the balanced blocking is imposed ex-post facto, instead of in a pre-planned design. We establish the ability of the proposed balanced block design to restore the test validity with multiple, possibly invalid, instruments in Section 3.

Previously, for a treatment-control analysis without unmeasured confounders, Rosenbaum, Ross and Silber (2007) noted the appeal of balancing the empirical distribution of categorical variables in matching. They proposed the fine balancing method as a useful addition to

traditional approaches like propensity score and close individual matches when exact matching is impractical. This can be seen as an attempt to mimic the randomized experiments as the gold standard for causal inference, which, by stochastically balancing the distributions of both observed and unobserved covariates between treatment groups on average, ensures unbiased estimation of the average treatment effects. In contrast, the proposed balanced blocking method balances the distribution of candidate instruments instead of covariates.

1.4. *A single ordinal instrument to multiple nested instruments.* Quite often in applications, instruments are hierarchically nested either by nature or by design. But balanced blocking is not feasible with nested instruments as at least one level of each instrument is fixed given the value of its nesting instrumental variable. A few applications of nested instruments follow. Consider the Mendelian randomization where one candidate instrument often indicates the presence of a minor allele of single-nucleotide polymorphism (SNP) or the number of the minor alleles (e.g., G), e.g., 0, 1, and 2 for having AA, AG, and GG, respectively. When having two minor alleles has a larger association with the exposure variable than having a single minor allele does, we may benefit from dividing the instrument into two, nested instruments. Here the first instrument indicates having at least one minor allele of SNP or not and the second instrument indicates having two minors alleles or not. Similarly, in measuring the causal effect of education, the number of (older) siblings has been commonly used as an instrument for attending higher education (e.g., Sander (1995) and Tan (2006)). In such a case, we can translate one ordinal instrument into K instrumental variables, each of which denotes having at least k number of siblings $(k = 1, 2, \ldots, K)$. Furthermore, the instrumental variables sometimes denote the distance to available services (e.g., Hadley et al. (2003); Lorch et al. (2012); Voors et al. (2012), and Zeng et al. (2019)) or the intensity of a single variable (e.g., Walker et al. (2020) and Zeng, Li and Ding (2020)). Then each of these continuous or ordinal instruments can be converted into multiple nested instruments. Such conversion to the coarser set of nested instruments might be more reasonable than using a single instrument when the validity of the instrument may depend on the level or intensity of its value. For example, Voors et al. (2012) examined the causal effect of exposure to violence on economic behaviors using distance to Bujumbura (where much of the war occurred) as an instrument. Here it might be conceivable that being beyond (or within) a certain distance to Bujumbura is associated with unmeasured confounders that affect the outcomes (economic behaviors). In this case, the continuous instrument that denotes the distance to Bujumbura might be invalid, whereas the binary instrument that would be only active at a certain range of the distance is valid.

Given these numerous situations where nested instruments are possible, and where our proposal of a balanced blocking is not possible, a different method is required. We propose a stratification method that conditions on all the rest of candidate instruments. We demonstrate that the resulting p-values from each instrumental variable analysis after conditioning on proper variables are likely to produce negatively correlated conclusions, providing non-redundant findings akin to evidence factors.

1.5. *Outline of this paper.* Section 2 defines the notation and reviews the key concepts in the evidence factor literature. Section 3 introduces the balanced block design as a way to construct approximate evidence factors in the presence of possibly invalid instruments. In Section 4, we illustrate the mutual stratification method for multiple, nested instruments. Section 5 illustrates the finite sample performance of the balanced block design and mutual stratification. We apply the proposed methods to two real data examples in Section 6. Section 7 concludes the paper with some practical recommendations. All the relevant code can be found at `https://github.com/youjin1207/EvidenceIV`.

2. Notation and key definitions.

2.1. *Notation.* Consider a study population of N units in I strata, $i = 1, \ldots, I$, with n_i individuals ij in stratum i ($j = 1, \ldots, n_i$; $N = \sum_{i=1}^{I} n_i$). There are K binary candidate instruments, $Z_{ij,1}, \ldots, Z_{ij,K}$, in addition to a binary treatment variable D_{ij}. Let $\mathbb{I}(\cdot)$ denote the indicator function. In the educational attainment and earnings example, one could take $K = 3$ with (i) $Z_{ij,1} = \mathbb{I}\{\text{grew up in an area with a college}\}$, (ii) $Z_{ij,2} = \mathbb{I}\{\text{born in the first quarter}\}$, (iii) $Z_{ij,3} = \mathbb{I}\{\text{minimum school leaving age} \leq 15\}$, and $D_{ij} = \mathbb{I}\{\text{attending high school}\}$ for individual ij. Individual ij has an observed covariate vector $\mathbf{x}_{ij}$ and K unobserved covariates $\{u_{ij,k} : k = 1, \ldots, K\}$ corresponding to the K candidate instruments, respectively. The observed covariate vector is controlled by stratification with $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$ for $1 \leq j, j' \leq n_i$ whereas the unobserved covariates may be different for two units in the same stratum, i.e., $u_{ij,k} \neq u_{ij',k}$ for some i, k and $j \neq j'$. First consider two potential outcomes of individual ij: r_{Tij} when $D_{ij} = 1$ and r_{Cij} when $D_{ij} = 0$. We aim to test the following Fisher's sharp null hypothesis of no treatment effects for all units:

$$H_0 : r_{Tij} = r_{Cij}, \qquad \forall ij. \tag{1}$$

Let $[m] = \{1, \ldots, m\}$ be the set of 1 to m for positive integer m; we will hence abbreviate $i = 1, \ldots, I$, $j = 1, \ldots, n_i$, and $k = 1, \ldots, K$ as $i \in [I]$, $j \in [n_i]$, and $k \in [K]$, respectively, when no confusion would arise. Write $\mathbf{Z}_k = (Z_{11,k}, \ldots, Z_{In_I,k})^{\mathrm{T}}$ for the N-dimensional assignment vector for instrument k. Let $\mathcal{K}$ be a subset of $[K]$, and let $\widetilde{\mathbf{Z}}_{\mathcal{K}} = (\mathbf{Z}_k)_{k \in \mathcal{K}}$ be the $N \times |\mathcal{K}|$ matrix concatenating the assignment vectors of the instruments in $\mathcal{K}$. Abbreviate $\widetilde{\mathbf{Z}}_{[K]}$ as $\widetilde{\mathbf{Z}}$ for $\mathcal{K} = [K]$.

Let $\mathbf{A}_{ij} = (Z_{ij,1}, \ldots, Z_{ij,K})$ be the assignment vector of the K instruments for unit ij. Let $\mathcal{A} = \{0, 1\}^K$ be the set of the 2^K possible values $\mathbf{A}_{ij}$ can take. For $\mathcal{K} \subseteq [K]$, let $\mathbf{A}_{ij,\mathcal{K}} = (Z_{ij,k})_{k \in \mathcal{K}}$ be the subset of $\mathbf{A}_{ij}$ containing the instruments within $\mathcal{K}$; thus, $\mathbf{A}_{ij,[K]} = \mathbf{A}_{ij}$. Let $\mathbf{A}_{ij,-\mathcal{K}} = (Z_{ij,k})_{k \notin \mathcal{K}}$ be the complement of $\mathbf{A}_{ij,\mathcal{K}}$ with regard to $\mathbf{A}_{ij}$.

Let R_{ij} be the observed outcome for unit ij, vectorized as $\mathbf{R} = (R_{11}, \ldots, R_{In_I})^{\mathrm{T}}$. Let $r_{ij,(\mathbf{a},d)}$ be the potential outcome of individual ij if $(\mathbf{A}_{ij}, D_{ij}) = (\mathbf{a}, d) \in \{0, 1\}^{K+1}$. Let $\boldsymbol{r}_{ij}$ be the collection of $r_{ij,(\mathbf{a},d)}$ for $(\mathbf{a}, d) \in \{0, 1\}^{K+1}$. Write $\mathcal{F} = \{(\boldsymbol{r}_{ij}, \mathbf{x}_{ij}, u_{ij,k}) : i \in [I];\ j \in [n_i];\ k \in [K]\}$. Distinct individuals are assumed to have independent values of $(\mathbf{A}_{ij}, D_{ij})$ and $(\mathbf{A}_{i'j'}, D_{i'j'})$ conditioning on $\mathcal{F}$. The observed outcome satisfies $R_{ij} = r_{ij,(\mathbf{A}_{ij}, D_{ij})}$ for unit ij.

The above notation imposes no assumptions on the effects of the candidate instruments on the outcome, and results in 2^{K+1} potential outcomes for each individual ij. Different types of the exclusion restriction assumption impose different restrictions on how the instruments may affect the outcome, and constrain the number of potential outcomes in various ways to facilitate inference for the causal effects. We formalize this in the following section.

2.2. *Variants of the partial exclusion restriction and approximate evidence factors.* The classic exclusion restriction assumption asserts that a valid set of instruments can change the outcome only by changing the value of the treatment, and in no other way (Angrist, Imbens and Rubin, 1996). It is the strongest form of the exclusion restriction, resulting in only two potential outcomes for each individual ij, that is, $r_{ij,(\mathbf{a},1)} = r_{Tij}$ and $r_{ij,(\mathbf{a},0)} = r_{Cij}$ for all $\mathbf{a} \in \{0, 1\}^K$. In the educational attainment and earnings example, this means neither the proximity to school, the birth quarter, or the minimum school leaving age has any direct effect on earnings once we condition on the educational attainment. There is uncertainty whether this exclusion restriction holds for applications in practice.

Karmakar, Small and Rosenbaum (2021) proposed the *partial exclusion restriction* that is less stringent in the way it constrains the set of potential outcomes.

DEFINITION 2.1. Let $\mathcal{K} \subseteq [K]$, and let k be the smallest element of $\mathcal{K}$. The *partial exclusion* holds for $\mathcal{K}$ if, with $\mathbf{A}_{ij,[k-1]} = (Z_{ij,1}, \ldots, Z_{ij,k-1})$ fixed by conditioning, each individual ij has two potential outcomes depending upon the value of D_{ij}, namely r_{Tij} if $D_{ij} = 1$ or r_{Cij} if $D_{ij} = 0$.

Under this restriction, they proposed the *reinforced design* that conducts K analyses on the K candidate instruments, in which the step k analysis performs a stratified randomization inference on $\mathbf{Z}_k$ conditioning on the $\mathbf{A}_{ij,[k-1]}$'s. When the partial exclusion restriction holds for $\mathcal{K} \subseteq [K]$, the reinforced design delivers $|\mathcal{K}|$ valid evidence factors for $k \in \mathcal{K}$. In the educational attainment and earnings example, the step 1 analysis compares the earnings for individuals who live closer to a college to the others by an unconstrained randomization test, the step 2 analysis compares earnings of individuals who were born in the first quarter of the year to the others after stratifying on the geographical proximity to a college, and the step 3 analysis compares earnings of individuals who went to a school with a school leaving age of at least 16 to the others after stratifying on geographical proximity to a college and birth-year cohort. The partial exclusion restriction will fail to hold for $\mathcal{K}$ when $1 \in \mathcal{K}$ if, for example, areas with a college also attract businesses which leads to different career opportunities for the residents. The increase in the minimum school-leaving age to 16 in the United Kingdom enforced from September 1, 1972 led to over-crowding in schools and labor market shortage, thus possibly affecting future job opportunities and earnings (Halsey et al., 1980). Thus, the partial exclusion restriction may also fail to hold when $3 \in \mathcal{K}$. Additionally, this change in minimum school-leaving age does not affect individuals born in the first quarter of 1972; hence $Z_{ij,2}$ and $Z_{ij,3}$ are correlated. See further discussion of these instruments in the references given in the introduction.

The definition of the partial exclusion restriction implies an inherent order among the K candidate instruments under consideration. If a suboptimal order is used, the reinforced design may produce biased tests. Consider $K = 2$ for concreteness. The set $\mathcal{K}$ can take three possible values, $\mathcal{K} = \{1\}, \{2\}, \{1, 2\}$, corresponding to scenarios (i) $Z_{ij,1}$ satisfies the exclusion restriction, (ii) $Z_{ij,2}$ satisfies the exclusion restriction after conditioning on $Z_{ij,1}$, and (iii) both instruments satisfy the exclusion restriction, respectively. The reinforced design in the order of $Z_{ij,1}$ to $Z_{ij,2}$ runs an unconditional test of $Z_{ij,1}$ and a conditional test of $Z_{ij,2}$ conditioning on $Z_{ij,1}$, returning $|\mathcal{K}|$ valid evidence factors under either of these three scenarios. The above list of possible $\mathcal{K}$'s, however, does not cover the fourth scenario in which (iv) $Z_{ij,1}$ satisfies the exclusion restriction only after conditioning on $Z_{ij,2}$, whereas $Z_{ij,2}$ violates the exclusion restriction both unconditionally and after conditioning on $Z_{ij,1}$. In this fourth scenario, the exclusion restriction is violated in both steps of the above reinforced design in the order of $Z_{ij,1}$ to $Z_{ij,2}$, subjecting the analysis to two biased tests.

Since the optimal ordering of the instruments assumed by the partial exclusion restriction is unlikely to be known in practice, Definition 2.2 gives a variant of the original partial exclusion with no order implication.

DEFINITION 2.2. Let $\mathcal{K} \subseteq [K]$. The *unordered partial exclusion* holds for $\mathcal{K}$ if, with $\mathbf{A}_{ij,-\mathcal{K}} = (Z_{ij,k})_{k \notin \mathcal{K}}$ fixed by conditioning, each individual ij has two potential outcomes depending upon the value of D_{ij}, namely r_{Tij} if $D_{ij} = 1$ or r_{Cij} if $D_{ij} = 0$.

The unordered partial exclusion ensures that the potential outcomes depend on only the treatment and the instruments outside $\mathcal{K}$, namely, $r_{ij,(\mathbf{a},d)} = r_{ij,(\mathbf{a}_{-\mathcal{K}},d)}$ for $\mathbf{a} = (z_1, \ldots, z_K)^{\mathrm{T}}$ and $\mathbf{a}_{-\mathcal{K}} = (z_k)_{k \notin \mathcal{K}}$. The instruments in $\mathcal{K}$, as a result, satisfy the exclusion restriction after conditioning on $\mathbf{A}_{ij,-\mathcal{K}}$. The unordered partial exclusion restriction allows for possible violation of no direct effects and/or no unmeasured confounding assumptions in other instruments

that are conditioned on. For example, in our running example, the unordered partial exclusion restriction with $\mathcal{K} = \{2\}$ allows for the possibility of ambitions driving families living near good colleges and different labor markets that pupils may face, but stipulates that the birth quarter has no direct effect on the outcome after we condition on the proximity to college and the minimum school leaving age.

Note that, in practice, sometimes we cannot distinguish whether an invalid instrument $k \notin \mathcal{K}$ violates the no direct effects assumption or the no unmeasured confounders assumption. For instance, higher ambitions might be interpreted as a confounder so that conditioning on ambitions renders the first instrument valid, but without measurements of ambitions, the instrument is invalid because of the failure of the no unmeasured confounders assumption. Or, one might argue differently that living near good colleges comes along with ambitions, resulting in a direct effect on earnings when ambitions affect earnings. Since we cannot observe unmeasured confounders (e.g., ambitions) nor can we manipulate the instrument while fixing the treatment, we do not know the sources of invalidity of conditioning instrument(s) (Angrist, Imbens and Rubin, 1996; Heng, Small and Rosenbaum, 2020).

Absent knowledge of $\mathcal{K}$ in practice, the reinforced design does not provide appropriate evidence factors under this unordered exclusion restriction. We propose two novel strategies to produce valid evidence factors irrespective of the order of the analyses.

Evidence factors are formally defined as being *nearly* independent using the relationships among p-values from each evidence. Definitions 2.3 and 2.4 review the definitions of approximate evidence factors refined from Rosenbaum (2011).

DEFINITION 2.3. A vector of p-values $(P_1, \ldots, P_\nu) \in [0,1]^\nu$ is *stochastically larger than the uniform* when for all coordinate-wise non-decreasing bounded continuous function $g : [0,1]^\nu \to \mathbb{R}$, we have, under H_0, $E\{g(P_1, \ldots, P_\nu)\} \geq E\{g(U_1, \ldots, U_\nu)\}$, where $U_1, \ldots, U_\nu$ are i.i.d. uniform[0,1] random variables.

DEFINITION 2.4. Multiple analyses are *approximate evidence factors* when (i) bias that invalidates one analysis does not necessarily bias other analyses; and (ii) the p-values from the analyses are stochastically larger than the uniform under the null.

Under the unordered exclusion restriction for $\mathcal{K}$, we will show that the proposed balanced block design and mutual stratification method yield $|\mathcal{K}|$ approximate evidence factors based on instruments in $\mathcal{K}$.

2.3. *Combination of multiple evidence factors.* When multiple independent p-values are available, various methods exist to combine them to form a single p-value for decision making. Typically, the combined statistic, e.g., Fisher's method (Fisher, 1926) and Simes' method (Simes, 1986), is a monotone function of the component p-values. Becker (1994) gave a comprehensive survey of these methods. Rosenbaum (2011, Lemma 1) showed that these combination methods also yield valid combined p-values when the components are stochastically larger than the uniform.

When several of the instruments are possibly invalid, we may not know $\mathcal{K}$. Assume it is guessed that a subset $\mathcal{K}$ (with size $\nu = |\mathcal{K}|$) of K p-values ($\{P_1, \ldots, P_K\}$) are stochastically larger than the uniform distribution of the ν-dimensional unit cube. Lemma 1 affords the basis to define a p-value by combining the K p-values in this situation. Let $P_{(k)}$ denote the kth order statistic of $\{P_1, \ldots, P_K\}$.

LEMMA 1. *Assume $P_1, \ldots, P_K$ are K p-values among which a subset of size ν are stochastically larger than the uniform distribution of the ν-dimensional unit cube. For any f*

that is coordinate-wise non-decreasing with $\Pr\{f(U_1,\ldots,U_\nu)\le\alpha\}\le\alpha$ *for all* $0\le\alpha\le 1$*, where* $U_1,\ldots,U_\nu$ *are i.i.d. uniform[0,1] random variables, we have* $f(P_{(K)},\ldots,P_{(K-\nu+1)})$ *is a valid p-value.*

PROOF OF LEMMA 1. Let $\mathcal{K}$ be the index set of the ν p-values stochastically larger than the uniform distribution of the ν-dimensional unit cube. Without loss of essential generality, assume $\mathcal{K}=[\nu]$ with $P_1\ge P_2\ge\cdots\ge P_\nu$. The result follows from Definition 2.3 by taking $g(p_1,\ldots,p_\nu)=\mathbb{I}\{f(p_1,\ldots,p_\nu)>\alpha\}$. Since $\Pr\{f(P_{(K)},\ldots,P_{(K-\nu+1)})\le\alpha\}\le\Pr\{f(P_1,\ldots,P_\nu)\le\alpha\}=1-E\{g(P_1,\ldots,P_\nu)\}\le 1-E\{g(U_1,\ldots,U_\nu)\}=\Pr\{f(U_1,\ldots,U_\nu)\le\alpha\}\le\alpha$. □

The above combination method is similar to the partial conjunction method of Benjamini and Heller (2008). We propose to use the truncated product method of Zaykin et al. (2002) to combine independent p-values by calculating the product of those p-values smaller than some truncation point, $\varkappa$ $(0<\varkappa\le 1)$. Hsu, Small and Rosenbaum (2013) demonstrated that the truncated product has higher power than Fisher's method in sensitivity analyses of observational studies. The implementation of the (truncated) product of p-values is available through the `sensitivitymv` package in R (Rosenbaum, 2015).

3. Approximate evidence factors via the balanced block design.

3.1. *Bias in the absence of proper conditioning.* Assume the following model for the instrument assignment:

$$\Pr(Z_{ij,k}=1\mid\mathcal{F},\mathbf{A}_{ij,[k-1]})=\frac{\exp\{\kappa_k(\mathbf{x}_{ij},\mathbf{A}_{ij,[k-1]})+\gamma_k u_{ij,k}\}}{1+\exp\{\kappa_k(\mathbf{x}_{ij},\mathbf{A}_{ij,[k-1]})+\gamma_k u_{ij,k}\}} \tag{2}$$

for $k\in[K]$, where $\gamma_k\ge 0$. Let $n_{i,\mathbf{a}}$ be the number of units with instrument combination $\mathbf{A}_{ij}=\mathbf{a}\in\mathcal{A}$ in stratum i, vectorized in lexicographical order of $(i,\mathbf{a})$ as $\mathbf{n}=(\mathbf{n}_i)_{i=1}^I$ with $\mathbf{n}_i=(n_{i,\mathbf{a}})_{\mathbf{a}\in\mathcal{A}}$. Let Ω be the set of possible values of $\widetilde{\mathbf{Z}}=(\mathbf{Z}_1,\ldots,\mathbf{Z}_K)$ subject to the joint counts constraint $\sum_{j=1}^{n_i}\mathbb{I}(\mathbf{A}_{ij}=\mathbf{a})=n_{i,\mathbf{a}}$ for all $i\in[I]$. Model (2) implies

$$\Pr\{\widetilde{\mathbf{Z}}=(\mathbf{z}_1,\ldots,\mathbf{z}_K)\mid\mathcal{F},\mathbf{n}\}=\frac{\exp(\sum_k\gamma_k\mathbf{u}_k^{\mathrm{T}}\mathbf{z}_k)}{\sum_{(\mathbf{b}_1,\ldots,\mathbf{b}_K)\in\Omega}\exp(\sum_k\gamma_k\mathbf{u}_k^{\mathrm{T}}\mathbf{b}_k)}\quad\text{for }(\mathbf{z}_1,\ldots,\mathbf{z}_K)\in\Omega,$$

where $\mathbf{u}_k=(u_{11,k},\ldots,u_{In_I,k})^{\mathrm{T}}$. The assignment is thus as-if randomized if $\gamma_k=0$ for all k. Non-zero values of γ_k parameterize the influence of unmeasured confounders on instrument k. In sensitivity analyses we will let γ_k take any value less than some non-negative bound to assess the robustness of the findings to the violation of as-if randomization.

In addition, model (2) assumes the effect of the unmeasured confounders $u_{ij,k}$'s to be separate from the covariates and in a logistic model. Recently, some generalizations of this sensitivity analysis model have been considered that relax these restrictions (Hasegawa and Small, 2017; Fogarty and Hasegawa, 2019; Heng and Small, 2021). These generalizations often require additional sensitivity parameters, and can lead to harder computation of the p-values from randomization inference. We do not consider them here.

Consider $K=2$ and the unordered partial exclusion restriction with $\mathcal{K}=\{1\}$ as an illustrating example. We have $Z_{ij,2}$ directly violates the exclusion restriction, whereas $Z_{ij,1}$ on its own does not. The potential outcomes satisfy $r_{ij,(\mathbf{a},d)}=r_{ij,(z_2,d)}$ for $\mathbf{a}=(z_1,z_2)\in\mathcal{A}=\{(00),(01),(10),(11)\}$ such that $R_{ij}=r_{ij,(Z_{ij,2})}$ depends only on the value of $Z_{ij,2}$ under

the null hypothesis. Write $\mathbf{R} = \mathbf{R}(\mathbf{Z}_2)$ to highlight the dependence of $\mathbf{R}$ on $\mathbf{Z}_2$ even in the absence of the direct effects of the treatment.

Let $n_{i,(z_1 z_2)}$ be the number of units with instrument combination $(Z_{ij,1}, Z_{ij,2}) = (z_1, z_2)$ in stratum i for $z_1, z_2 = 0, 1$, and let $n_{i,k(z_k)}$ be the marginal count of units with $Z_{ij,k} = z_k$. Let $\Omega_{(1,2)} = \{(\mathbf{z}_1, \mathbf{z}_2) : \sum_{j=1}^{n_i} \mathbb{I}(\mathbf{A}_{ij} = \mathbf{a}) = n_{i,\mathbf{a}},\ i \in [I],\ \mathbf{a} \in \mathcal{A}\}$ be the set of possible values of $(\mathbf{Z}_1, \mathbf{Z}_2)$ subject to the joint counts constraint, and let Ω_1 be the set of possible values of $\mathbf{Z}_1$ subject to the marginal stratum constraint $\sum_{j=1}^{n_i} Z_{ij,1} = n_{i,1(1)}$.

For $t_1 = t_1(\mathbf{Z}_1, \mathbf{R})$ as an arbitrary test statistic for studying $\mathbf{Z}_1$, the marginal test of $\mathbf{Z}_1$ randomly permutes $\mathbf{Z}_1$ within Ω_1 and induces a uniform distribution over $\mathcal{T}^\pi = \{t_1(\mathbf{z}_1, \mathbf{R}(\mathbf{Z}_2)) : \mathbf{z}_1 \in \Omega_1\}$ conditioning on $\mathbf{R}(\mathbf{Z}_2)$. This defines the *randomization distribution* of t_1 under the marginal test. The true sampling distribution of $t_1(\mathbf{Z}_1, \mathbf{R})$ under the null hypothesis, on the other hand, is a distribution over $\mathcal{T} = \{t_1(\mathbf{z}_1, \mathbf{R}(\mathbf{z}_2)) : (\mathbf{z}_1, \mathbf{z}_2) \in \Omega_{(1,2)}\}$ with the probability mass function determined by the joint distribution of $(\mathbf{Z}_1, \mathbf{Z}_2)$. The violation of the exclusion restriction for $Z_{ij,2}$ causes $\mathbf{R}(\mathbf{z}_2)$ to vary over $(\mathbf{z}_1, \mathbf{z}_2) \in \Omega_{(1,2)}$ such that $\mathcal{T}^\pi$ and $\mathcal{T}$ differ even under H_0. This discrepancy in the supports of the sampling distributions causes the randomization distribution to deviate from the sampling distribution under the null hypothesis, and thereby subjects the marginal test of $\mathbf{Z}_1$ to possibly liberal type-I error rates even if $\gamma_1 = \gamma_2 = 0$ in (2) (Wu and Ding, 2020).

The same intuition extends to the general K-instrument study and illustrates the source of possible biases for the step $k \in \mathcal{K}$ analysis in the absence of proper conditioning on $\mathbf{A}_{ij,-\mathcal{K}}$. We define below three conditioning strategies to facilitate the discussion.

DEFINITION 3.1. Consider three types of randomization tests for studying the effects of variations in instrument $k \in [K]$. Refer to a test as *marginal* if it does not fix the values of any of the other $K-1$ instruments when studying the effects of variations in $Z_{ij,k}$. Refer to a test as *reinforced* if it fixes the values of $\mathbf{A}_{ij,[k-1]}$ by conditioning but not those of $\mathbf{A}_{ij,\{k+1,\ldots,K\}}$ when studying the effects of variations in $Z_{ij,k}$. Refer to a test as *conditional* if it fixes the values of all the other $K-1$ instruments when studying the effects of variations in $Z_{ij,k}$.

The *reinforced test* characterizes the step k analysis in the reinforced design proposed by Karmakar, Small and Rosenbaum (2021). Consider the educational attainment and earnings example with $K = 3$ instruments for concreteness. For the step $k = 2$ analysis studying the effects of being born in the first quarter, $Z_{ij,2}$, the marginal test performs an unconstrained randomization test on $Z_{ij,2}$, fixing neither geographical proximity, $Z_{ij,1}$, nor minimum school-leaving age, $Z_{ij,3}$. The reinforced test performs a stratified randomization test on $Z_{ij,2}$ by permuting $\mathbf{Z}_2$ within each stratum of $Z_{ij,1}$, namely $\{ij : Z_{ij,1} = 1\}$ and $\{ij : Z_{ij,1} = 0\}$, but puts no constraints on the value of $Z_{ij,3}$. The conditional test performs a stratified randomization test on $Z_{ij,2}$ by permuting $\mathbf{Z}_2$ within each stratum of $(Z_{ij,1}, Z_{ij,3})$, namely $\{ij : (Z_{ij,1}, Z_{ij,3}) = (z_1, z_3)\}$ for $z_1, z_3 \in \{0, 1\}$, fixing the other $K - 1 = 2$ factors while studying the effects of instrument 2.

The reinforced design thus consists of K reinforced tests of instruments 1 to K, respectively, and generates K candidate evidence factors under the partial exclusion assumption. Assume the unordered partial exclusion with $\mathcal{K} \neq [K]$ instead. Both the marginal and reinforced tests of instrument $k \in \mathcal{K}$ are subject to possible biases by not conditioning on the full set of $\mathbf{A}_{ij,-\mathcal{K}}$ even if $\gamma_k = 0$ for all k. Let $P_{k,\text{FRT}}$ denote the p-value produced by the randomization test of instrument k. Of interest is if certain study designs can help retain the test validity, in the sense of $\Pr(P_{k,\text{FRT}} \leq p_k) \leq p_k$ for all $p_k \in (0, 1)$, for all $k \in \mathcal{K}$.

A possible remedy is to form blocks in which there are an equal number of units under each instrument combination $\mathbf{a} \in \mathcal{A}$. This intuitively balances the distributions of $\mathbf{A}_{ij,-\mathcal{K}}$ over different levels of $Z_{ij,k}$ for $k \in \mathcal{K}$, and thereby offsets the bias due to the influence of

$\mathbf{A}_{ij,-\mathcal{K}}$ over R_{ij}. If $Z_{ij,k}$ were randomized we would expect the distribution of $\mathbf{A}_{ij,-\mathcal{K}}$ to be balanced stochastically. The blocking strategy, on the other hand, enforces the balance by design. Refer to such a strategy as a *balanced block design* with precise definition given in Definition 3.2 in Section 3.2. We show in Section 3.2 its utility in restoring the validity of marginal and reinforced tests for instrument $k \in \mathcal{K}$ under the unordered partial exclusion restriction, and demonstrate in Sections 3.3 and 3.4 its utility to create approximate evidence factors under all three types of tests.

3.2. *Valid marginal and reinforced tests under the balanced block design.* We demonstrate in this section the utility of the balanced block design in restoring the validity of marginal and reinforced analyses in the absence of proper conditioning. To this end, we start by introducing some additional notation and regularity conditions to facilitate the discussion.

Let $\mathbf{A} = \{\mathbf{A}_{ij} : i \in [I],\ j \in [n_i]\}$ be the collection of $\mathbf{A}_{ij}$. For an arbitrary finite population $(\mathcal{F}, \mathbf{A})$ of N units nested in I strata, let $p_i = n_i/N$ be the relative size of stratum i. Recall $n_{i,\mathbf{a}}$ as the number of units with instrument combination $\mathbf{A}_{ij} = \mathbf{a} = (z_1, \ldots, z_K) \in \mathcal{A} = \{0,1\}^K$ in stratum i. Let $p_{i,\mathbf{a}} = n_{i,\mathbf{a}}/n_i$ be the proportion of units with combination $\mathbf{a} \in \mathcal{A}$ in stratum i. For $k \in [K]$, let $n_{i,k(z_k)}$ be the marginal count of units with $Z_{ij,k} = z_k$, and let $p_{i,k(z_k)} = n_{i,k(z_k)}/n_i$ be the marginal proportion of level z_k of instrument k.

Under H_0, the unordered partial exclusion restriction ensures that $r_{ij,(\mathbf{a},d)} = r_{ij,\mathbf{a}_{-\mathcal{K}}}$ for $\mathbf{a} = (z_1, \ldots, z_K)$ and $\mathbf{a}_{-\mathcal{K}} = (z_k)_{k \notin \mathcal{K}}$. That is, with the treatment having no effect on the potential outcomes of unit ij, the potential outcomes under instrument combinations $\mathbf{a} = (\mathbf{a}_{\mathcal{K}}, \mathbf{a}_{-\mathcal{K}}) \in \mathcal{A}$ and $\mathbf{a}' = (\mathbf{a}'_{\mathcal{K}}, \mathbf{a}'_{-\mathcal{K}}) \in \mathcal{A}$ are identical as long as $\mathbf{a}_{-\mathcal{K}} = \mathbf{a}'_{-\mathcal{K}}$. The observed outcome equals $R_{ij} = r_{ij,(\mathbf{A}_{ij,-\mathcal{K}})}$. Let $\bar{r}_{i,\mathbf{a}_{-\mathcal{K}}} = n_i^{-1} \sum_{j=1}^{n_i} r_{ij,\mathbf{a}_{-\mathcal{K}}}$, $S^2_{i,\mathbf{a}_{-\mathcal{K}}} = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (r_{ij,\mathbf{a}_{-\mathcal{K}}} - \bar{r}_{i,\mathbf{a}_{-\mathcal{K}}})^2$, and $S_{i,(\mathbf{a}_{-\mathcal{K}},\mathbf{a}'_{-\mathcal{K}})} = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (r_{ij,\mathbf{a}_{-\mathcal{K}}} - \bar{r}_{i,\mathbf{a}_{-\mathcal{K}}})(r_{ij,\mathbf{a}'_{-\mathcal{K}}} - \bar{r}_{i,\mathbf{a}'_{-\mathcal{K}}})$ with $S_{i,(\mathbf{a}_{-\mathcal{K}},\mathbf{a}_{-\mathcal{K}})} = S^2_{i,\mathbf{a}_{-\mathcal{K}}}$ be the finite-population means, variances, and covariances of $\{r_{ij,\mathbf{a}_{-\mathcal{K}}} : j \in [n_i],\ \mathbf{a}_{-\mathcal{K}} \in \{0,1\}^{K-|\mathcal{K}|}\}$ in stratum i, respectively, for $i \in [I]$.

CONDITION 1. As $n_i \to \infty$ for $i \in [I]$, for all $\mathbf{a}, \mathbf{a}' \in \mathcal{A}$, (i) p_i and $p_{i,\mathbf{a}}$ have limits in $(0,1)$, (ii) $\bar{r}_{i,\mathbf{a}_{-\mathcal{K}}}$, $S^2_{i,\mathbf{a}_{-\mathcal{K}}}$, and $S_{i,(\mathbf{a}_{-\mathcal{K}},\mathbf{a}'_{-\mathcal{K}})}$ have finite limits, and (iii) there exists a $c_0 < \infty$ independent of n_i such that $n_i^{-1} \sum_{j=1}^{n_i} r^4_{ij,\mathbf{a}_{-\mathcal{K}}} \le c_0$.

We first consider the balanced block design with two instruments to illustrate the basic ideas, and then extend the results to the general K-instrument setting. Assume $K = 2$ candidate instruments, $Z_{ij,1}$ and $Z_{ij,2}$, with $\mathcal{K} = \{1\}$. The notation above simplifies to $\mathbf{a} = (\mathbf{a}_{\mathcal{K}}, \mathbf{a}_{-\mathcal{K}}) = (z_1, z_2)$, $(n_{i,\mathbf{a}}, p_{i,\mathbf{a}}) = (n_{i,(z_1 z_2)}, p_{i,(z_1 z_2)})$, and $r_{ij,(\mathbf{a},d)} = r_{ij,\mathbf{a}_{-\mathcal{K}}} = r_{ij,(z_2)}$ under H_0 with mean $\bar{r}_{i,\mathbf{a}_{-\mathcal{K}}} = \bar{r}_{i,(z_2)}$, variance $S^2_{i,\mathbf{a}_{-\mathcal{K}}} = S^2_{i,(z_2)}$, and covariance $S_{i,(\mathbf{a}_{-\mathcal{K}},\mathbf{a}'_{-\mathcal{K}})} = S_{i,(z_2,z'_2)} = S_{i,(0,1)} = S_{i,(1,0)}$. A *two-instrument balanced block design* stipulates

$$n_{i,(11)}/n_{i,(10)} = n_{i,(01)}/n_{i,(00)} \qquad \text{for } i \in [I] \tag{3}$$

in addition to $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$ for $1 \le j, j' \le n_i$, and thereby ensures identical distribution of $Z_{ij,2}$ across different levels of $Z_{ij,1}$ within each stratum and vice versa. A simple example is $n_{i,(11)} = n_{i,(10)} = n_{i,(01)} = n_{i,(00)} = 1$, where we form blocks of size $n_i = 4$ with one observation under each possible combination of $(Z_{ij,1}, Z_{ij,2}) \in \{0,1\}^2$. The proportion of units in each block with $Z_{ij,2} = 1$ equals $1/2$ for each level of $Z_{ij,1} \in \{0,1\}$ and vice versa.

Consider the difference-in-means statistic

$$\hat{\tau}_1 = \hat{\tau}_1(\mathbf{Z}_1, \mathbf{R}) = \sum_{i=1}^{I} w_i \hat{\tau}_{i,1}$$

for the marginal test of $\mathbf{Z}_1$, where $\hat{\tau}_{i,1} = \hat{R}_{i,1(1)} - \hat{R}_{i,1(0)}$ is the difference in means in stratum i with $\hat{R}_{i,1(z_1)} = n_{i,1(z_1)}^{-1} \sum_{j:Z_{ij,1}=z_1} R_{ij}$, and $w_1, \ldots, w_I$ are some prespecified weights. Let $\hat{\tau}_1^\pi$ represent a random variable from the randomization distribution of $\hat{\tau}_1$ induced by the permutation of $\mathbf{Z}_1$ within Ω_1. Assume the finite-population asymptotic framework that embeds $\mathcal{F}$ and $\mathbf{A}$ into a sequence of finite populations and assignments for $N = 1, 2, \ldots, \infty$. Wu and Ding (2020, Proposition 4) ensured that the marginal randomization test based on $\hat{\tau}_1$ controls type-I error rates asymptotically under H_0 if, under H_0, the asymptotic distribution of $|\hat{\tau}_1^\pi|$ stochastically dominates that of $|\hat{\tau}_1|$ for almost all sequences of $\mathbf{A}$. Building on this implication, Proposition 1 uses the permutation central limit theorem to show the utility of the balanced block design in restoring the asymptotic validity of the marginal test of instrument 1.

Let $e_{i,z_1} = p_{i,(z_1 1)}/p_{i,1(z_1)}$ be the conditional proportion of $Z_{ij,2} = 1$ within the subset of units with $Z_{ij,1} = z_1$. The balanced block design ensures $e_{i,1} = e_{i,0}$ for all $i \in [I]$. Recall $r_{ij,(\mathbf{a},d)} = r_{ij,(z_2)}$ under H_0. Let $\tau_{ij} = r_{ij,(1)} - r_{ij,(0)}$ be the corresponding difference in potential outcomes when the level of $Z_{ij,2}$ changes from 0 to 1. Write a.s. to indicate a statement holds for almost all sequences of $\mathbf{A}$.

PROPOSITION 1. Assume Condition 1 for $K = 2$, the unordered partial exclusion with $\mathcal{K}$ where $1 \in \mathcal{K}$, and the assignment model (2) with $\gamma_1 = \gamma_2 = 0$ for a sequence of finite populations $(\mathcal{F}, \mathbf{A})$ not necessarily balanced.

(a) If H_0 is true, then $\sqrt{N}(\hat{\tau}_1 - \mu) \rightsquigarrow \mathcal{N}(0, v)$, and $\sqrt{N}\hat{\tau}_1^\pi \rightsquigarrow \mathcal{N}(0, v')$ a.s. under the marginal test, where $\mu = \sum_{i=1}^I w_i(e_{i,1} - e_{i,0})\bar{\tau}_i$ and

$$v = \sum_{i=1}^I w_i^2 p_i^{-1} \left\{ \left(\frac{e_{i,0}}{p_{i,1(0)}} + \frac{e_{i,1}}{p_{i,1(1)}} \right) S_{i,(1)}^2 + \left(\frac{1-e_{i,0}}{p_{i,1(0)}} + \frac{1-e_{i,1}}{p_{i,1(1)}} \right) S_{i,(0)}^2 - S_{i,\tau}^2 (e_{i,1} - e_{i,0})^2 \right\},$$

$$v' = \sum_{i=1}^I w_i^2 p_i^{-1} p_{i,1(0)}^{-1} p_{i,1(1)}^{-1} \left(p_{i,2(1)} S_{i,(1)}^2 + p_{i,2(0)} S_{i,(0)}^2 + p_{i,2(1)} p_{i,2(0)} \bar{\tau}_i^2 \right)$$

with $\bar{\tau}_i = n_i^{-1} \sum_{j=1}^{n_i} \tau_{ij} = \bar{r}_{i,(1)} - \bar{r}_{i,(0)}$ and $S_{i,\tau}^2 = (n_i - 1)^{-1} \sum_{j=1}^{n_i} (\tau_{ij} - \bar{\tau}_i)^2$.

(b) Further assume the balanced block design with (3). Then $\mu = 0$ and $v \le v'$ such that $|\hat{\tau}_1| \le_{\text{st}} |\hat{\tau}_1^\pi|$ a.s. as n_i's go to infinity under H_0.

Let $P_{1,\text{m-FRT}}$ be the p-value from a two-sided marginal test of $\mathbf{Z}_1$. Under the balanced block design, Proposition 1 ensures $\Pr(P_{1,\text{m-FRT}} \le p_1) \le p_1$ for all $p_1 \in (0,1)$ a.s. as n_i's go to infinity under H_0, and thereby restores the asymptotic validity of the marginal test on instrument $1 \in \mathcal{K}$. The result extends immediately to the reinforced test, which coincides with the marginal test for instrument 1. We generalize below the result to the marginal and reinforced tests under general K-instrument studies.

DEFINITION 3.2. Assume K binary candidate instruments, $Z_{ij,k}$ for $k \in [K]$. A *K-instrument balanced block design* creates I strata, $i \in [I]$, of n_i units, ij $(j \in [n_i])$, with

$$n_{i,\mathbf{a}} = n_i \prod_{k=1}^K p_{i,k(z_k)} \tag{4}$$

in addition to $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$ for all $1 \le j, j' \le n_i$. Refer to a K-instrument balanced block design as a 2^K *balanced block design* if $n_{i,\mathbf{a}} = 2^{-K} n_i$ for all $\mathbf{a} \in \mathcal{A}$ with $p_{i,k(1)} = p_{i,k(0)} = 1/2$ for all $i \in [I]$ and $k \in [K]$.

Refer to a stratum in a balanced block design as a *balanced block* interchangeably when no confusion would arise. Balanced block design ensures $p_{i,(z_1,\ldots,z_K)} = \prod_{k=1}^{K} p_{i,k(z_k)}$, imposing independence between instruments by design. In our running example with $K = 3$ instruments, a simple 2^K balanced block design forms blocks of size $n_i = 8$ with one unit under each instrument combination $(z_1, z_2, z_3) \in \{0,1\}^3$ and $p_{i,k(1)} = 0.5$ for $k = 1, 2, 3$. Alternatively, we could let $p_{i,2(1)} = 0.25$ to match the population prevalence, assuming approximately $1/4$ of the population were born in the first quarter. The resulting balanced block features $n_{i,(z_1,1,z_3)} = 3^{-1} n_{i,(z_1,0,z_3)}$ for all $(z_1, z_3) \in \{0,1\}^2$.

Depending on if we condition on the first $k-1$ instruments in the analysis of instrument k, the strata we use for the randomization test equal either the original I balanced blocks for the marginal test or the $I \times 2^{k-1}$ strata by further conditioning on $\mathbf{A}_{ij,[k-1]}$ for the reinforced test. Index the strata for the randomization test by $s = [S]$, with $S = I$ under marginal test or $S = I \times 2^{k-1}$ under reinforced test. The difference-in-means statistic for the study of instrument k equals

$$\hat{\tau}_k = \hat{\tau}_k(\mathbf{Z}_k, \mathbf{R}) = \sum_{s=1}^{S} w_s \hat{\tau}_{s,k}, \tag{5}$$

where $\hat{\tau}_{s,k} = \hat{R}_{s,k(1)} - \hat{R}_{s,k(0)}$ with $\hat{R}_{s,k(z_k)} = n_{s,k(z_k)}^{-1} \sum_{j:Z_{sj,k}=z_k} R_{ij}$, and the w_s's are some prespecified weights. The unordered partial exclusion restriction implies that $\mathbf{R}$ is a function of the assignment vectors $\mathbf{Z}_{k'}$ for $k' \notin \mathcal{K}$ even under H_0. The $\hat{\tau}_k$ as such is a function of $\mathbf{Z}_k \cup (\mathbf{Z}_{k'})_{k' \notin \mathcal{K}}$, and is stochastic because of the stochasticity of the assignment $\widetilde{\mathbf{Z}}$.

Theorem 3.3 extends Proposition 1 to the general K-instrument study and ensures $|\hat{\tau}_k| \leq_{\text{st}} |\hat{\tau}_k^{\pi}|$ a.s. under H_0 and the balanced block design with either marginal or reinforced test for all $k \in \mathcal{K}$. To simplify the presentation, we state the results in terms of the marginal test on the I original balanced blocks; those for the reinforced test follow by replacing i with s and $\mathcal{K}$ with $\mathcal{K} \cap \{k+1, \ldots, K\}$ for the analysis of instrument k.

THEOREM 3.3. *Assume Condition 1, the unordered partial exclusion with $\mathcal{K}$, the assignment model (2) with $\gamma_k = 0$ for all k, and the balanced block design (4). For all $k \in \mathcal{K}$, if H_0 is true, then $\sqrt{N}\hat{\tau}_k \rightsquigarrow \mathcal{N}(0, v_k)$, and $\sqrt{N}\hat{\tau}_k^{\pi} \rightsquigarrow \mathcal{N}(0, v_k')$ a.s. with $v_k \leq v_k'$ under either marginal or reinforced test as n_i's go to infinity.*

Let $P_{k,\text{m-FRT}}$ and $P_{k,\text{r-FRT}}$ be the p-values in the step k analysis produced by the marginal and reinforced tests, respectively. Theorem 3.3 ensures that they both control the type-I error rates asymptotically under the balanced block design; we relegate the explicit forms of v_k and v_k' to the supplementary material.

PROPOSITION 2. Assume Condition 1, the unordered partial exclusion with $\mathcal{K}$, the assignment model (2) with $\gamma_k = 0$ for all k, and the balanced block design (4). For all $k \in \mathcal{K}$, if H_0 is true, then $\Pr(P_{k,\text{m-FRT}} \leq p_k) \leq p_k$ and $\Pr(P_{k,\text{r-FRT}} \leq p_k) \leq p_k$ for all $p_k \in (0,1)$ a.s. as n_i's go to infinity.

The stochastic dominance properties (cf. Definition 2.2) under the reinforced design, as the collection of K reinforced tests, then follow from Karmakar, Small and Rosenbaum (2021).

THEOREM 3.4. *Assume Condition 1, the unordered partial exclusion with $\mathcal{K}$, the assignment model (2) with $\gamma_k = 0$ for all k, and the balanced block design (4). If H_0 is true, then the joint distribution of the p-values from the $|\mathcal{K}|$ reinforced tests in $\mathcal{K}$ with $\hat{\tau}_k$'s as in (5), $\{P_{k,\text{r-FRT}} : k \in \mathcal{K}\}$, is stochastically larger than the uniform distribution of the $|\mathcal{K}|$-dimensional unit cube as n_i's go to infinity.*

Theorem 3.4 illustrates the utility of the balanced block design in producing $|\mathcal{K}|$ approximate evidence factors from reinforced tests even in the absence of proper conditioning, i.e., ordered partial exclusion restriction. We establish results similar to Theorem 3.4 for marginal and conditional tests in Section 3.3 and Section 3.4, respectively.

3.3. *Asymptotically independent p-values with K marginal tests.* Consider the m-statistic for the marginal comparison of the group with $Z_{ij,k} = 1$ versus the group with $Z_{ij,k} = 0$:

$$t_k(\mathbf{Z}_k, \mathbf{R}) = \sum_{i=1}^{I} \sum_{j:Z_{ij,k}=1} \sum_{j':Z_{ij',k}=0} \psi_k\{(R_{ij} - R_{ij'})/\sigma\}, \tag{6}$$

where $\psi_k(\cdot)$ is an odd function with $\psi_k(x) \geq 0$ for $x \geq 0$ and $\sigma > 0$ is a constant. The $\hat{\tau}_k$ in (5) is a special case of t_k with $\psi_k(x) = x$. Like $\hat{\tau}_k$, $t_k(\mathbf{Z}_k, \mathbf{R})$ is stochastic under the randomization inference framework due to its dependence on $\mathbf{Z}_k$ and $\mathbf{R}$, which is a function of $(\mathbf{Z}_{k'})_{k' \notin \mathcal{K}}$ even under H_0. Theorem 3.5 below states the asymptotic independence of $\{t_k(\mathbf{Z}_k, \mathbf{R}) : k \in \mathcal{K}\}$ when $\gamma_k = 0$ for all $k \in \mathcal{K}$. This affords the basis for the resulting p-values to qualify as approximate evidence factors.

Recall that $\widetilde{\mathbf{Z}}_i = (\mathbf{Z}_{i,1}, \ldots, \mathbf{Z}_{i,K})$ is the $n_i \times K$ sub-matrix of $\widetilde{\mathbf{Z}}$ with columns $\mathbf{Z}_{i,k} = (Z_{i1,k}, \ldots, Z_{in_i,k})^{\mathrm{T}}$. Let $t_{i,k} = \sum_{j:Z_{ij,k}=1} \sum_{j':Z_{ij',k}=0} \psi_k\{(R_{ij} - R_{ij'})/\sigma\}$ be the component of t_k from stratum i, which is a function of $\widetilde{\mathbf{Z}}_i$. As the $K = 2$ case, let Ω_i be the set containing all possible values of $\widetilde{\mathbf{Z}}_i$ subject to the joint counts constraint $\sum_{j=1}^{n_i} \mathbb{I}(\mathbf{A}_{ij} = \mathbf{a}) = n_{i,\mathbf{a}}$ for all $\mathbf{a} \in \mathcal{A}$, and let $t_{i,k}(\tilde{\mathbf{z}}_i)$ be the potential value of $t_{i,k}$ when $\widetilde{\mathbf{Z}}_i = \tilde{\mathbf{z}}_i \in \Omega_i$. Let $v_{i(kk')} = |\Omega_i|^{-1} \sum_{\tilde{\mathbf{z}}_i \in \Omega_i} t_{i,k}(\tilde{\mathbf{z}}_i) t_{i,k'}(\tilde{\mathbf{z}}_i)$.

CONDITION 2. As $I \to \infty$, $\max_{i=1,\ldots,I} v_{i(kk')} / \sum_{i'=1}^{I} v_{i'(kk')} = o(1)$ for all $k, k' \in \mathcal{K}$ and $I^{-1} \sum_{i'=1}^{I} v_{i'(kk')}$ has a finite limit.

Intuitively, we can always form smaller balanced blocks from larger ones of possibly different ratios. A simple example is to divide a balanced block of sizes $\{n_{i,\mathbf{a}} : \mathbf{a} \in \mathcal{A}\}$ into $n_i' = \min_{\mathbf{a} \in \mathcal{A}} n_{i,\mathbf{a}}$ smaller blocks each of which has one unit under each instrument combination. The condition of $n_i \to \infty$ for $i \in [I]$ thus implies $I \to \infty$ for a related balanced block design as long as $p_{i,\mathbf{a}}$ has a limit in $(0, 1)$ for all $i \in [I]$ and $\mathbf{a} \in \mathcal{A}$. This illustrates the connection between Conditions 1 and 2.

THEOREM 3.5. *Assume K marginal tests under the 2^K balanced block design, the unordered partial exclusion with $\mathcal{K}$, the assignment model (2), and Condition 2. If H_0 is true, then (i) for $k \in \mathcal{K}$ and $k' \neq k$, $(t_k, t_{k'})$ are asymptotically independent provided $\gamma_k = 0$; and (ii) $\{t_k : k \in \mathcal{K}\}$ are asymptotically jointly independent provided $\gamma_k = 0$ for all $k \in \mathcal{K}$. Further assume $\mathcal{K} = [K]$. Then the 2^K requirement can be relaxed to any general K-instrument balanced block design (4).*

Let P_k denote the true p-value based on the sampling distribution of t_k under H_0. Theorem 3.5 ensures the asymptotic independence of $\{P_k : k \in \mathcal{K}\}$ provided $\gamma_k = 0$ for all $k \in \mathcal{K}$. This, together with the fact that $P_{k,\text{m-FRT}} \geq P_k$ for $k \in \mathcal{K}$ under the balanced block design from Proposition 2, affords the basis for constructing asymptotically independent evidence factors based on the K marginal tests.

COROLLARY 1. Assume the unordered partial exclusion with $\mathcal{K}$ and $\gamma_k = 0$ for all k in the assignment model (2). If H_0 is true, then the p-values from the $|\mathcal{K}|$ marginal tests in $\mathcal{K}$, $\{P_{k,\text{m-FRT}} : k \in \mathcal{K}\}$, are stochastically larger than the uniform distribution of the $|\mathcal{K}|$-dimensional unit cube asymptotically provided either (i) $n_i \to \infty$ for $i \in [I]$ and we test with $\hat{\tau}_k$'s as in (5) under Condition 1 and the 2^K balanced block design; or (ii) $I \to \infty$, $\mathcal{K} = [K]$, and we test with t_k's as in (6) under Condition 2 and the general K-instrument balanced block design (4).

In the absence of as-if randomization, we cannot be sure if a comparison is valid. The asymptotic independence between t_k and t'_k from Theorem 3.5(i) ensures that the violation of random assignment by instrument $k' \neq k$ does not affect the asymptotic validity of $P_{k,\text{m-FRT}}$ as long as $\gamma_k = 0$. We formalize the intuition in Proposition 3 below. Let

$$\Pr(Z_{ij,k} = 1 \mid \mathcal{F}) = \frac{\exp\{\kappa'_k(\mathbf{x}_{ij}) + \gamma'_k u'_{ij,k}\}}{1 + \exp\{\kappa'_k(\mathbf{x}_{ij}) + \gamma'_k u'_{ij,k}\}} \qquad \text{for } k \in [K]$$

be the marginal assignment model for $Z_{ij,k}$ after integrating out the other instruments from (2) with $0 \leq u'_{ij,k} \leq 1$ and $\gamma'_k = \log(\Gamma_k)$. Let $\overline{P}_{k,\Gamma_k}$ be the upper bound on the p-value for the marginal test of instrument $k \in \mathcal{K}$.

PROPOSITION 3. Assume the 2^K balanced block design, unordered partial exclusion with $\mathcal{K}$, and the assignment model (2). If H_0 is true, then for $k, k' \in \mathcal{K}$, $(\overline{P}_{k,\Gamma_k}, \overline{P}_{k',\Gamma_{k'}})$ from the marginal tests using $(t_k, t_{k'})$ are asymptotically stochastically larger than the uniform distribution of the two-dimensional unit cube provided either $\gamma_k = 0$ or $\gamma_{k'} = 0$ as $I \to \infty$ under Condition 2. Further assume $\mathcal{K} = \{1, \ldots, K\}$. The 2^K requirement can be relaxed to any general K-instrument balanced block design (4).

Theorem 3.5, Corollary 1, and Proposition 3 together illustrate the utility of the balanced block design for constructing approximate evidence factors from marginal tests when the instruments are possibly invalidated by the violation of either the strict exclusion restriction or as-if randomization or both. We move on to addressing its properties under conditional tests.

3.4. *Asymptotically independent p-values with K conditional tests.* The conditional test of the kth candidate instrument fixes the value of $\mathbf{Z}_{-k} = (\mathbf{Z}_{k'})_{k' \neq k}$ when making inference using $\mathbf{Z}_k$. The unordered partial exclusion restriction with $\mathcal{K}$ ensures that the comparison $k \in \mathcal{K}$ is valid under conditional test as long as $\gamma_k = 0$. Of interest is whether we can combine the $|\mathcal{K}|$ comparisons for $k \in \mathcal{K}$ as $|\mathcal{K}|$ independent p-values.

Consider two ways of conducting conditional randomization tests under the balanced block design. The first, "restricted" strategy uses the same m-statistic $t_k(\mathbf{Z}_k, \mathbf{R})$ in (6) as under the marginal tests yet permutes $\mathbf{Z}_k$ within the subset of Ω_k compatible with $\mathbf{Z}_{-k}$ under the balanced block constraint (4). To illustrate, consider $K = 2$ and a 2^2 balanced block with observed assignments $\{(0,0),(1,0),(0,1),(1,1)\}$. The observed values of $\mathbf{Z}_k$'s equal $\mathbf{Z}_1 = (0,1,0,1)^\intercal$ and $\mathbf{Z}_2 = (0,0,1,1)^\intercal$, respectively. Then, in this first strategy, a test using $\mathbf{Z}_1$ permutes $(Z_{i1,1}, Z_{i2,1}, Z_{i3,1}, Z_{i4,1})$ in $\{(0,1,0,1),(1,0,0,1),(0,1,1,0),(1,0,1,0)\}$, which is the subset of $\Omega_1 = \{(z_{i1,1}, z_{i2,1}, z_{i3,1}, z_{i4,1}) \in \{0,1\}^4 : z_{i1,1} + \cdots + z_{i4,1} = 2\}$ compatible with $\mathbf{Z}_2$ under the joint counts constraint. The second, "stratified" strategy further divides each balanced block into 2^{K-1} strata by the values of $\mathbf{A}_{ij,-k} = (Z_{ij,k'})_{k' \neq k}$, and conducts a stratified randomization test based on the resulting $S = I \times 2^{K-1}$ strata. Index

by sj the jth unit in stratum s under the second strategy. The m-statistic for comparison k equals

$$t_k^*(\widetilde{\mathbf{Z}}, \mathbf{R}) = \sum_{s=1}^{S} \sum_{j:Z_{sj,k}=1} \sum_{j':Z_{sj',k}=0} \psi_k\{(R_{sj} - R_{sj'})/\sigma\} \tag{7}$$

for $\psi_k(\cdot)$ and σ as in (6). Theorem 3.6 below states the asymptotic independence of $\{t_k(\mathbf{Z}_k, \mathbf{R}) : k \in \mathcal{K}\}$ and that of $\{t_k^*(\widetilde{\mathbf{Z}}, \mathbf{R}) : k \in \mathcal{K}\}$ under the two conditional test strategies respectively when the assignments of the instruments in $\mathcal{K}$ are as-if randomized.

Let $\widetilde{\mathbf{Z}}_s = (\mathbf{Z}_{s,1}, \ldots, \mathbf{Z}_{s,K})$ be the sub-matrix of $\widetilde{\mathbf{Z}}$ corresponding to stratum s. Let $t_{s,k}^* = \sum_{j:Z_{sj,k}=1} \sum_{j':Z_{sj',k}=0} \psi_k\{(R_{sj} - R_{sj'})/\sigma\}$ be the component of t_k^* from stratum s, which is a function of $\widetilde{\mathbf{Z}}_s$. Let Ω_s^* be the set containing all possible values of $\widetilde{\mathbf{Z}}_s$ under the balanced block constraint (4), and let $t_{s,k}^*(\tilde{\mathbf{z}}_s)$ be the potential value of $t_{s,k}^*$ when $\widetilde{\mathbf{Z}}_s = \tilde{\mathbf{z}}_s \in \Omega_s^*$. Let $v_{s(kk')}^* = |\Omega_s^*|^{-1} \sum_{\tilde{\mathbf{z}}_s \in \Omega_s^*} t_{s,k}^*(\tilde{\mathbf{z}}_s) t_{s,k'}^*(\tilde{\mathbf{z}}_s)$.

CONDITION 3. As $S \to \infty$, $\max_{s=1,\ldots,S} v_{s(kk')}^* / \sum_{s'=1}^{S} v_{s'(kk')}^* = o(1)$ for all $1 \le k, k' \le K$ and $S^{-1} \sum_{s=1}^{S} v_{s(kk')}^*$ has a finite limit.

THEOREM 3.6. *Assume the unordered partial exclusion with $\mathcal{K}$ and $\gamma_k = 0$ for all $k \in \mathcal{K}$ in the assignment model (2). If H_0 is true, then (i) $(t_k)_{k\in\mathcal{K}}$ as in (6) from the $|\mathcal{K}|$ restricted conditional tests are asymptotically jointly independent Gaussian as $I \to \infty$ under Condition 2 and the general K-instrument balanced block design (4); and (ii) $(t_k^*)_{k\in\mathcal{K}}$ as in (7) from the $|\mathcal{K}|$ stratified conditional tests are asymptotically jointly independent Gaussian as $S \to \infty$ under Condition 3 and the 2^K balanced block design.*

COROLLARY 2. Assume the unordered partial exclusion with $\mathcal{K}$ and $\gamma_k = 0$ for all $k \in \mathcal{K}$ in the assignment model (2). If H_0 is true, then the p-values from the $|\mathcal{K}|$ conditional tests in $\mathcal{K}$, $\{P_{k,\text{c-FRT}} : k \in \mathcal{K}\}$, are asymptotically stochastically larger than the uniform distribution of the $|\mathcal{K}|$-dimensional unit cube provided (i) we conduct the restricted conditional tests with t_k's as in (6) under Condition 2 and the general K-instrument balanced block design (4) as $I \to \infty$; or (ii) we conduct the stratified conditional tests with t_k^*'s as in (7) under Condition 3 and the 2^K balanced block design as $S \to \infty$.

REMARK. The 2^K balanced design requirement in Theorem 3.6(ii) and Corollary 2(ii) for the stratified conditional tests can be relaxed to accommodate $p_{i,k(1)} \neq 1/2$ for one $k \in \mathcal{K}$ in each stratum i. In the case of two-instrument studies, this allows us to form balanced blocks with $(n_{00} : n_{01} : n_{10} : n_{11}) = (1 : \rho : 1 : \rho)$ or $(\rho : 1 : \rho : 1)$ where $p_{i,1(1)} = 1/2$ or $p_{i,2(1)} = 1/2$. This is of practical interest when the distribution of instrument combinations differs markedly from $(1 : 1 : 1 : 1)$ in the population prior to blocking. The flexibility in ρ allows more units to be grouped into the balanced blocks and thereby improves the power.

3.5. *Some practical remarks.* This concludes our discussion on the utility of the balanced block design in ensuring evidence factors from either marginal, reinforced, or conditional tests. In practice, we recommend using the balanced block design when the validity of multiple candidate instruments is in doubt and choosing the marginal ratios, namely the $p_{i,k(z_k)}$'s, to maximize the number of units that can be included into the balanced blocks. Then, if it is guessed that at least ν of the candidate instruments are valid, we can combine the ν largest p-values to form one single p-value. We show these steps in our application in Section 6.1.

We have considered marginal, reinforced, or conditional inference for the sharp null hypothesis H_0. As is known from the seminal work of Angrist, Imbens and Rubin (1996), an instrumental variables analysis, under a monotonicity assumption, provides inference only for the local treatment effect on the corresponding compliers. Thus, the tests under the marginal, reinforced, or conditional inference provide evidence for the local treatment effects on the corresponding subpopulations of the compliers which are defined by the instrument and the conditioning event. For example, in our running example of educational attainment and earnings, the marginal test of $Z_{ij,1}$ concerns the effect for the subpopulation that would change their schooling decision if they lived closer to a college than not, irrespective of the other instruments. The conditional test of $Z_{ij,1}$, on the other hand, concerns the conditional effect for the subpopulation that, with the other instruments fixed, would change their schooling decision if they lived closer to a college than not. If there is evidence to reject the sharp null for any subpopulation, we reject the sharp null for the whole population.

Thus, two guidelines may be followed to choose among a marginal, a reinforced, and a conditional inference (Deaton, 2010; Imbens, 2010). First, if it is expected that the local treatment effects will be large in magnitude for the tests under one of these inferential methods, then this method may be preferred, as it would suggest that the power will be large. Second, if the subpopulations of the compliers are expected to be sufficiently diverse for the tests under an inferential method, then such method may be preferred, as it would increase generalizability of the inference. Sufficient subject matter knowledge is needed to make a potentially best choice following these guidelines.

From a technical point of view, the marginal and reinforced tests require a 2^K balanced block design when not all K instruments are valid, whereas the restricted conditional test requires only a general K-instrument balanced block design. The latter is thus preferred when the $n_{i,\boldsymbol{a}}$'s differ greatly in the original study population, allowing more units to be formed into the balanced blocks, thereby increasing power.

We presented our results using the test statistics $t_k(\mathbf{Z}_k, \mathbf{R})$'s and $t_k^*(\widetilde{\mathbf{Z}}, \mathbf{R})$'s based on differences between the observed outcomes $\mathbf{R}$. These results also hold when we additionally use covariance adjustment. In a covariance-adjusted inference for H_0, one uses the residuals $R_{ij} - g(\mathbf{w}_{ij})$'s in place of the R_{ij}'s in the test statistics for some fitted model $g(\mathbf{w}_{ij})$ of the outcomes on pre-treatment covariates $\mathbf{w}_{ij}$'s that are possibly different from the $\mathbf{x}_{ij}$'s. Many authors have suggested using covariance adjustment as it is typically more efficient; see e.g., Rubin (1979); Gail, Tan and Piantadosi (1988); Tukey (1993); and Rosenbaum (2002).

A key limitation of the balanced block design is that it requires observations from all possible combinations of the instruments of interest. We move on to Section 4 to address this concern and present an alternative strategy for instruments that are nested.

4. Approximate evidence factors with nested instruments. In this section, we introduce an evidence factor analysis with nested instruments where we are not able to form balanced blocks. Consider multiple tests from K binary instruments for individual j from stratum i: $\mathbf{A}_{ij} = (Z_{ij,[K]})$ for $i \in [I]$ and $j \in [n_i]$. Unless otherwise mentioned, suppose that K instruments are positively nested in the order of $Z_{ij,1}, Z_{ij,2}, \ldots, Z_{ij,K}$; that is, $Z_{ij,k'} = 1$ only if $Z_{ij,k} = 1$ for $k \in [k'-1]$. Define $Z_{ij,0} \equiv 1$ and $Z_{ij,K+1} \equiv 0$. In the educational attainment example, $Z_{ij,k}$ can denote living within d_k-distance from a college with $d_1 > d_2 > \ldots > d_K$.

With $\mathcal{F} := \{(\boldsymbol{r}_{ij}, \mathbf{x}_{ij}, u_{ij,k}) : i \in [I],\ j \in [n_i],\ k \in [K]\}$, we have the following instrument assignment of $Z_{ij,k}$ with κ_k being an arbitrary function for $i \in [I],\ j \in [n_i],\ k \in [K]$:

$$
\begin{aligned}
\Pr(Z_{ij,k} = 1 | \mathcal{F}, \mathbf{A}_{ij,-k}) &= \Pr(Z_{ij,k} = 1 | \mathcal{F}, Z_{ij,k-1}, Z_{ij,k+1}) \\
&= \mathbb{I}(Z_{ij,k-1} = 1, Z_{ij,k+1} = 0) \frac{\exp\{\kappa_k(\mathbf{x}_{ij}) + \gamma_k u_{ij,k}\}}{1 + \exp\{\kappa_k(\mathbf{x}_{ij}) + \gamma_k u_{ij,k}\}} \\
&\quad + \mathbb{I}(Z_{ij,k-1} = 1, Z_{ij,k+1} = 1),
\end{aligned} \tag{8}
$$

where $u_{ij,k}$'s with $0 \leq u_{ij,k} \leq 1$ are unmeasured covariates. The first line of equation (8) is due to the nested structure in $\mathbf{A}_{ij}$, i.e., $Z_{ij,k}$ is conditionally independent of $Z_{ij,k'}$ for all $k' \neq k-1, k+1$.

We consider the unordered partial exclusion restriction for $\mathcal{K} = [K]$ in Definition 2.2. Under this restriction, we can rule out the influence from the invalidity of $Z_{ij,k}$ merely due to its correlation with other instruments in $\mathbf{A}_{ij,-k}$, i.e., not due to its direct effect on the outcome. To exclude such violation that would be dismissed after conditioning on the variables in $\mathbf{A}_{ij,-k}$, we further stratify the unit of inference by the other $K-1$ instruments. We call this stratification *mutual stratification* within $\mathbf{A}_{ij}$, which was also used for the second type of conditional test in Section 3.4. In this way, as long as $Z_{ij,k}$ satisfies the unordered partial exclusion within $\mathbf{A}_{ij}$, the hypothesis H_{0k} of no effect of $Z_{ij,k}$ on R_{ij} conditioning on $(\mathbf{X}_{ij}, \mathbf{A}_{ij,-k})$ is the same as the null in (1).

4.1. *Negatively correlated p-values from the mutual stratification.* As a result of the mutual stratification that conditions on $\mathbf{A}_{ij,-k}$ when doing inference with $Z_{ij,k}$, we have K p-values, one from each of the K mutually stratified instrumental variable analysis for instrument $k \in [K]$. As a test statistic from one mutually stratified analysis is typically associated with those from other analyses due to the correlations in $\mathbf{A}_{ij}$, the tests are not necessarily independent of each other. Despite possible dependencies between the test statistics, we will show that under certain conditions, the mutual stratification method still can produce approximate evidence factors (cf. Definition 2.4) that preserve the validity of the multiple comparisons.

Let us first investigate the relationships of a pair of p-values from two nested instruments, $Z_{ij,k}$ and $Z_{ij,k'}$ where $Z_{ij,k'} = 1$ only if $Z_{ij,k} = 1$. We consider a one-sided nonparametric test to test the sharp null (1). For simplicity, suppose that there is no stratum constructed by the observed covariates $\mathbf{x}_{ij}$ so that $i = 1$ for all j's and $N = n_i$. Consider the following general form of two test statistics, which can encompass Huber's m-statistics, stratified Wilcoxon rank-sum statistics, and Hodges-Lehmann aligned rank statistics (Rosenbaum, 2011). Let

$$T_k = \sum_{j=1}^{N} (1 - Z_{ij,k'}) Z_{ij,k} q_{ij,k}, \tag{9}$$

$$T_{k'} = \sum_{j=1}^{N} Z_{ij,k} Z_{ij,k'} q_{ij,k'},$$

where $q_{ij,k}$ and $q_{ij,k'}$ are some functions of the observed outcomes with the exact form of the functions determined by the test statistics. Due to the nested structure, we only compare the outcomes under $Z_{ij,k}$ variable when $Z_{ij,k'} = 0$ in T_k and compare the outcomes under $Z_{ij,k'}$ when $Z_{ij,k} = 1$ in $T_{k'}$. Let $v_{ij,\boldsymbol{\lambda}}^{(k,k')} = \text{var}\{\lambda_1(1 - Z_{ij,k'}) Z_{ij,k} q_{ij,k} + \lambda_2 Z_{ij,k} Z_{ij,k'} q_{ij,k'}\}$ with any non-zero vectors $\boldsymbol{\lambda} = (\lambda_1, \lambda_2) \in \mathbb{R}^2$. Suppose that the unordered partial exclusion restriction holds for $\mathcal{K} \subseteq [K]$.

THEOREM 4.1. *Assume the following two conditions hold for $k \in \mathcal{K}$ and $k' = \min_{l \in \mathcal{K}}\{l > k\}$: (i) $\sum_{j=1}^{n_i} q_{ij,k} q_{ij,k'} \to c \geq 0$ for a constant c and (ii) $\max_{j \in [n_i]} v_{ij,\boldsymbol{\lambda}}^{(k,k')} / \sum_{j'}^{n_i} v_{ij',\boldsymbol{\lambda}}^{(k,k')} \to 0$ as $n_i \to \infty$ for each stratum i. Then under the assignment model (8) with $\gamma_k = 0$ for all $k \in \mathcal{K}$, p-values from the aforementioned test statistics from the mutual stratification are stochastically larger than the uniform among valid instruments, i.e.,*

$$\Pr(P_k \leq p_k, \forall k \in \mathcal{K}) \leq \prod_{k \in \mathcal{K}} p_k,$$

for any $0 \leq p_k \leq 1$. *Moreover, when the unordered exclusion restriction is satisfied for a set* $\mathcal{K}$ *with* $|\mathcal{K}| \geq \nu$, $f\left(P_{(K)}, \ldots, P_{(K-\nu+1)}\right)$ *is a valid* p*-value when* $P_{(k)}$ *denotes the* k*th order statistic of* $(P_1, \ldots, P_K)$ *and* f *is as in Lemma 1.*

The one-sided tests generally satisfy condition (i) in the above theorem, not only with positive $\{q_{ij,k}, q_{ij,k'}\}$ but also with the standardized $q_{ij,k}$ and $q_{ij,k'}$ having mean zeros (e.g., m-statistics). This is because, roughly speaking, $q_{ij,k}$ and $q_{ij,k'}$ both consider subjects ij with $(Z_{ij,k}, Z_{ij,k'}) = (1, 0)$. For example, suppose that $q_{ij,k}$ and $q_{ij,k'}$ denote the rank of R_{ij} among individuals with $Z_{ij,k'} = 0$ and individuals with $Z_{ij,k} = 1$, respectively. Then $q_{ij,k}$ and $q_{ij,k'}$ of those subjects would be both small (or large) depending on the value of observed outcomes R_{ij}'s, thus they are positively correlated. On the other hand, condition (ii) excludes the case where $q_{ij,k}$ or $q_{ij,k'}$ is constant over all $j \geq M$ for some large M so that the denominator in (ii) does not increase as n_i increases.

COROLLARY 3. The result of Theorem 4.1 holds in the presence of the stratification by observed covariates $\mathbf{x}_{ij}$ when the number of the strata, I, induced by $\mathbf{x}_{ij}$ is bounded and conditions (i) and (ii) hold for all $i \in [I]$.

In practice, we may consider (i) dividing a single instrument with multiple values into multiple, nested instruments, (ii) deriving a p-value from each mutually stratified instrumental variable analysis, and (iii) performing an evidence factor analysis under different levels of sensitivity parameters. If at least ν of the candidate instruments are presumed to be valid, then we can combine the ν largest p-values to form one single p-value. In our forthcoming simulations study, we will explore the finite sample performance of two different analyses with the same data: one using a single ordinal instrument and the other using multiple nested instruments.

REMARK. Even though it seems like the nested structure of instruments determines the order of the analyses, the mutual stratification still exhibits different implications from the reinforced design. Under the mutual stratification, invalid descendent instruments that are nested within a precedent instrument cannot invalidate the analysis with the precedent instrument. On the other hand, under the reinforced design, the validity of the precedent instrument may collapse when there is an unmeasured associational or a causal path through the descendant instruments to the outcome variable.

4.2. *Sensitivity analysis with nested instruments.* One of the advantages of the mutual stratification is that we can investigate the sensitivity to bias due to the unmeasured factors directly associated with each instrument that would be still present after conditioning on other instruments.

The parameter $\Gamma_k := \exp(\gamma_k)$ in (8) quantifies the influence of unmeasured covariates $u_{ij,k}$ that is present conditioned on $\mathbf{x}_{ij}$ when $Z_{ij,k-1} = 1$ and $Z_{ij,k+1} = 0$; that is, bias due to $u_{ij,k}$ comes from the violation of the unordered partial exclusion of $Z_{ij,k}$ for $\{k-1, k, k+1\}$, thus for $[K]$. When each comparison from the mutually stratified analysis is biased by *at most* $\Gamma_k \geq 1$, denote the maximum of p-values given $\mathcal{F}$ and $\{\mathbf{A}_{ij,-k} : i \in [I],\ j \in [n_i]\}$ over all possible $\{u_{ij,k} : i \in [I], j \in [n_i]\}$ as $\overline{P}_{k,\mathbf{A}_{-k},\Gamma_k}$ or, for simplicity, $\overline{P}_k$. The following proposition shows that the upper bounds $\overline{P}_{k,\mathbf{A}_{-k},\Gamma_k}$'s are stochastically larger than the uniform distribution under the same conditions as in Theorem 4.1.

PROPOSITION 4. Consider K nested instruments and assignment model (8). For $k \in [K]$, let $\overline{P}_{k,\mathbf{A}_{-k},\Gamma_k}$ be the maximum p-value from the mutually stratified instrumental variable

analysis with $Z_{ij,k}$ at $\Gamma_k \geq 1$. Assume the same regularity conditions as in Theorem 4.1. Then under the null (1), when the unordered exclusion restriction is satisfied for a set $\mathcal{K}$ with $|\mathcal{K}| \geq \nu$, $f\left(\overline{P}_{(K)}, \ldots, \overline{P}_{(K-\nu+1)}\right)$ is a valid p-value when $\overline{P}_{(k)}$ denotes the kth order statistic of $(\overline{P}_1, \ldots, \overline{P}_K)$ and f is as in Lemma 1.

Given a fixed ν, we may vary $\{\Gamma_k : k \in [K]\}$ to introduce the amount of bias due to unmeasured confounders in $Z_{ij,k}$ $(k \in [K])$ that would be present even after mutual stratification and adjustments for the measured confounders. On the other hand, the number of (in)valid instruments among K candidates can be viewed as a sensitivity parameter as well (Kang et al., 2021). In practice, we do not know the true number of valid instruments. Instead, we can allow one to vary the minimum number of valid instruments ν and observe how sensitive the results are. The decreased value of ν from c to $c+1$ essentially dismisses the contribution of one p-value (often small p-value) to the combined p-value. Therefore, underestimating ν compared to the true number of valid instruments will result in conservative inference, but as long as we have at least ν valid instruments the combined p-value is conservative for type-I error.

5. Simulation studies.

5.1. *Numerical examples under the balanced block design.* We illustrate in this part the validity of the balanced block design under the null hypothesis even in the absence of proper conditioning. See Section S3 in the supplementary material for additional simulation studies on the power of sensitivity analysis and design sensitivity.

We consider a study with two candidate instruments, $Z_{ij,1}$ and $Z_{ij,2}$. The goal is to test the null hypothesis of no treatment effect for any unit: $H_0 : r_{ij,(z_1,z_2,1)} = r_{ij,(z_1,z_2,0)}$, $\forall ij$. Assume that instrument 2 directly violates the exclusion restriction whereas instrument 1 is valid after conditioning on instrument 2. This ensures the unordered partial exclusion restriction holds with $\mathcal{K} = \{1\}$. The marginal test of instrument 1 is invalid without further adjustment when the instruments are correlated as illustrated in Section 3.1. We show below the utility of the balanced block design to restore its validity.

Consider eight blocking strategies, with $(n_{i,(00)}, n_{i(01)}, n_{i(10)}, n_{i(11)})$ equaling (a) $(1,1,1,1)$, (b) $(1,2,1,2)$, (c) $(1,4,1,4)$, (d) $(1,2,2,4)$, (e) $(2,2,2,2)$, (f) $(4,4,4,4)$, (g) $(1,1,1,2)$, and (h) $(1,1,1,3)$, respectively. Strategy (h) forms units into blocks of size $1+1+1+3=5$ in which there is one unit with each of $(z_1, z_2) \in \{(0,0),(0,1),(1,0)\}$ and three units with $(z_1, z_2) = (1,1)$. It then conducts block randomization test by permuting the assignment vector of interest within each block; likewise for strategies (a)–(g). A blocking strategy is balanced if $\omega = n_{i,(11)}/(n_{i,(01)} + n_{i,(11)}) - n_{i,(10)}/(n_{i,(00)} + n_{i,(10)})$ equals 0. Strategies (a)–(f) thus correspond to block randomization tests under balanced blocks, whereas strategies (g)–(h) correspond to block randomization tests under unbalanced blocks. In addition, consider strategy (i) that conducts the unblocked randomization test on a population with instrument assignment model:

$$\begin{aligned} &\Pr(Z_{ij,1} = 1) = 0.33, \qquad \Pr(Z_{ij,2} = 1) = 0.4, \text{ and} \\ &\Pr(Z_{ij,1} = 1 \mid Z_{ij,2} = 1) - \Pr(Z_{ij,1} = 1 \mid Z_{ij,2} = 0) = 0.14. \end{aligned} \tag{10}$$

The resulting instrument combination frequencies are roughly $(n_{i,(00)}, n_{i,(01)}, n_{i,(10)}, n_{i,(11)}) = n_i \times (0.45, 0.22, 0.15, 0.18)$. Generate the potential outcomes for a total of $N = 3,600$ units under each strategy from

$$r_{ij,(z_1,z_2,d)} = \lambda_2 z_2 + \epsilon, \qquad \text{where } \epsilon \sim \mathcal{N}(0, \sigma^2). \tag{11}$$

This data generating model ensures $\mathcal{K} = \{1\}$ and the treatment d, missing from the right-hand side of (11), has no effect on the potential outcomes, such that H_0 holds.

Table 1 shows the type-I error rates of the marginal tests of $Z_{ij,1}$ under the eight blocking strategies at 5% significance level over $10,000$ independent replications. The result accords with the theory and illustrates the ability of the balanced block designs to deliver valid p-values even in the absence of proper conditioning. In particular, all six strategies with the balanced blocking, namely strategies (a)–(f), preserve the correct type-I error rates even without conditioning on $Z_{ij,2}$, whereas the three strategies without the balanced blocking, namely strategies (g)–(i), do not. Provided the blocks are balanced, the exact sizes and ratios, namely whether it is $(1,1,1,1)$ or $(1,2,2,4)$ or $(4,4,4,4)$, result in no material differences in the type-I error rates. Moreover, the valid tests under the six balanced block designs become increasingly more conservative as σ decreases. Intuitively, this is because the observed outcomes $R_{ij} = R_{ij}(Z_{ij,2})$'s reveal more information about the likely values of the $Z_{ij,2}$'s as σ diminishes, with units with larger R_{ij}'s more likely to have $Z_{ij,2} = 1$. As the $Z_{ij,1}$'s are balanced across different values of $Z_{ij,2}$, the marginal tests of $Z_{ij,1}$ tend to produce less extreme test statistics values and thus more conservative type-I error rates.

We next put the balanced block design in perspective, and compare its validity with the two-stage least squares regression and the original reinforced design proposed by Karmakar, Small and Rosenbaum (2021). Consider a study population of $N = 2,500$ units nested in $I = 50$ initially unbalanced strata. We generate the finite population of $(Y_{ij}, Z_{ij,1}, Z_{ij,2}, D_{ij})$ by (10)–(11) and $\Pr(D_{ij} = 1 \mid Z_{ij,1}, Z_{ij,2}) = \max\{\min(\xi_{ij}, 1), 0\}$ with $\xi_{ij} = 0.3Z_{ij,1} + 0.25Z_{ij,2} + \eta_{ij}$, where $\eta_{ij} \sim \mathcal{N}(0, 0.06)$. The null hypothesis is true and should not be rejected at a rate higher than the nominal rate of 5%. The two-stage least squares regression performs a joint analysis using $(Z_{ij,1}, Z_{ij,2})$ as instruments and reports one p-value as the final conclusion. The balanced block and reinforced designs run one randomization test for each of $Z_{ij,1}$, $Z_{ij,2}$, and D_{ij}, and produce three evidence factors to be weighed against each other. Table 2 shows the type-I error rates for testing H_0 under three combinations of (λ_2, σ). The two-stage least squares regression yields biased joint analysis. The reinforced design yields unbiased tests of D_{ij}, yet is biased for both $Z_{ij,1}$ and $Z_{ij,2}$. In contrast, the balanced block design restores the validity of the marginal tests based on $Z_{ij,1}$ in addition to those based on D_{ij}, yielding the right "do not reject" decision by majority vote which is lacking from the other two methods.

TABLE 1

Type-I error rates at 5% significance level for the marginal tests of $Z_{ij,1}$ with van Elteren statistics under different blocking strategies over $10,000$ repetitions. The potential outcomes are generated from model (11) with $\lambda_2 = 0.1$ for $N = 3,600$ units. The balanced block design features $\omega = 0$ with treatment group sizes determined by the ratio of the $n_{i,(z_1 z_2)}$'s. The treatment proportions for the unblocked analysis in column (i) is generated using assignment model (10). Results with Hodges-Lehmann statistics are similar and thus omitted.

	$\omega = 0$						$\omega \neq 0$		
	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
σ	(1,1,1,1)	(1,2,1,2)	(1,4,1,4)	(1,2,2,4)	(2,2,2,2)	(4,4,4,4)	(1,1,1,2)	(1,1,1,3)	Assignment model (10)
1	0.053	0.050	0.048	0.049	0.052	0.049	0.114	0.157	0.106
0.5	0.053	0.050	0.047	0.049	0.051	0.049	0.219	0.346	0.192
0.1	0.039	0.036	0.036	0.035	0.036	0.035	0.357	0.012	0.67

5.2. *Numerical experiments with nested instruments.* In this section, we examine (i) multiple causal comparisons using the mutual stratification with nested instruments and (ii) a single comparison using the Kruskal-Wallis test (Breslow, 1970). Set $I = 50$ strata with $n_i = 50$

TABLE 2

Type-I error rates at 5% significance level under the two-stage least squares regression, reinforced design, and balanced block design using marginal tests over 1,000 repetitions of $N = 2,500$ units nested in $I = 50$ strata generated from models (10)–(11).

		Two-stage	Reinforced design			Balanced block design		
λ_2	σ	least squares	Z_1	Z_2	D	Z_1	Z_2	D
0.5	0.5	1	1	1	0.047	0.030	1	0.043
0.1	0.5	0.946	0.265	0.996	0.047	0.048	0.968	0.043
0.1	1	0.436	0.124	0.771	0.047	0.047	0.558	0.043

units in each stratum $i = [I]$. Generate two nested instruments with $\Pr(Z_{ij,1} = 1) = 0.8 := p_1$ and $\Pr(Z_{ij,2} = 1 \mid Z_{ij,1} = 0) = 0$, and $\Pr(Z_{ij,2} = 1) = \delta p_1$. We control the dependency between the two nested instruments by varying δ in $\{0.3, 0.5, 0.7\}$. Let

$$\begin{aligned} \xi_{ij} &= \phi_1 Z_{ij,1} + \phi_2 Z_{ij,2} + \eta_{ij} \\ \Pr(D_{ij} = 1 \mid Z_{ij,1}, Z_{ij,2}) &= \max\{\min(\xi_{ij}, 1), 0\} \\ R_{ij} &= \lambda_1 Z_{ij,1} + \lambda_2 Z_{ij,2} + \beta^* D_{ij} + \epsilon_{ij}, \end{aligned} \tag{12}$$

where $(\epsilon_{ij}, \eta_{ij})$ are bivariate Gaussian with zero means, fixed variances of 0.25 and 0.06, respectively, and covariance of 0.3. Set $\phi_k = 0.5$ for $k = 1, 2$. After (i) the mutual stratification, we combine two p-values using Fisher's method (Fisher, 1926) without truncation. Consider an ordinal instrument $Z^*_{ij} = 1\mathbb{I}(Z_{ij,1} = 1) + 1\mathbb{I}(Z_{ij,2} = 1)$ and use this variable Z^*_{ij} for (ii) the Kruskal-Wallis test (or one-way ANOVA on ranks) to test whether the medians of all groups defined by Z^*_{ij} (e.g., $Z^*_{ij} = 0, 1, 2$) are equal. Figure 1 shows the rejection rates when two instruments are both valid (upper panel; $\lambda_2 = 0.0$) and one of them ($k = 2$) is invalid (lower panel; $\lambda_2 = 0.1$). Across all panels in Figure 1, λ_1 is set to zero.

The lower panels of Figure 1 present the results when the second instrument is invalid. Assuming we know that at most one of the two instruments is invalid, we use $\nu = 1$ in Lemma 1. We observe that the use of a single variable Z^*_{ij} becomes invalid to use when either $Z_{ij,1}$ or $Z_{ij,2}$ is invalid, failing to control a type-I error under the null ($\beta^* = 0.0$). On other hand, as long as ν is correctly specified, we can control a type-I error with two instruments where one of them is invalid.

Next, we generate five hierarchially nested instruments with increased stratum size $n_i = 200$ for $i = [I]$ with $I = 50$. Set $\phi_k = 0.1$ for $k = 1, \ldots, 5$. With $(\epsilon_{ij}, \eta_{ij})$ generated from the same bivariate Gaussian distribution,

$$\begin{aligned} &\Pr(Z_{ij,1} = 1) = 0.8;\ \Pr(Z_{ij,k} = 1) = \Pr(Z_{ij,k-1} = 1)\delta,\ k = 2, 3, 4, 5. \\ &\xi_{ij} = \phi_1 Z_{ij,1} + \phi_2 Z_{ij,2} + \phi_3 Z_{ij3} + \phi_4 Z_{ij4} + \phi_5 Z_{ij5} + \eta_{ij} \\ &\Pr(D_{ij} = 1 \mid Z_{ij,1}, Z_{ij,2}, Z_{ij3}, Z_{ij4}, Z_{ij5}) = \max\{\min(\xi_{ij}, 1), 0\} \\ &Y_{ij} = \lambda_1 Z_{ij,1} + \lambda_2 Z_{ij,2} + \lambda_3 Z_{ij3} + \lambda_4 Z_{ij4} + \lambda_5 Z_{ij5} + \beta^* D_{ij} + \epsilon_{ij}. \end{aligned} \tag{13}$$

Figure 2 presents the rejection rates when five instruments are all valid (upper panel; set $\nu = 5$) or when one of the five instruments is invalid (lower pannel; set $\nu = 4$). In the latter case, the Kruskal-Wallis test with a single instrument (Z^*_{ij} defined similarly) fails to control the type-I error; while the combined p-value using the mutual stratification is valid with at least ν valid p-values and often more conservative for type-I error. The power from the Kruskal-Wallis test is higher than the power using the mutual stratification when $\boldsymbol{\lambda} = (0, 0, 0, 0, 0)$. This is because instruments often become stronger when combined (Davies et al., 2015).

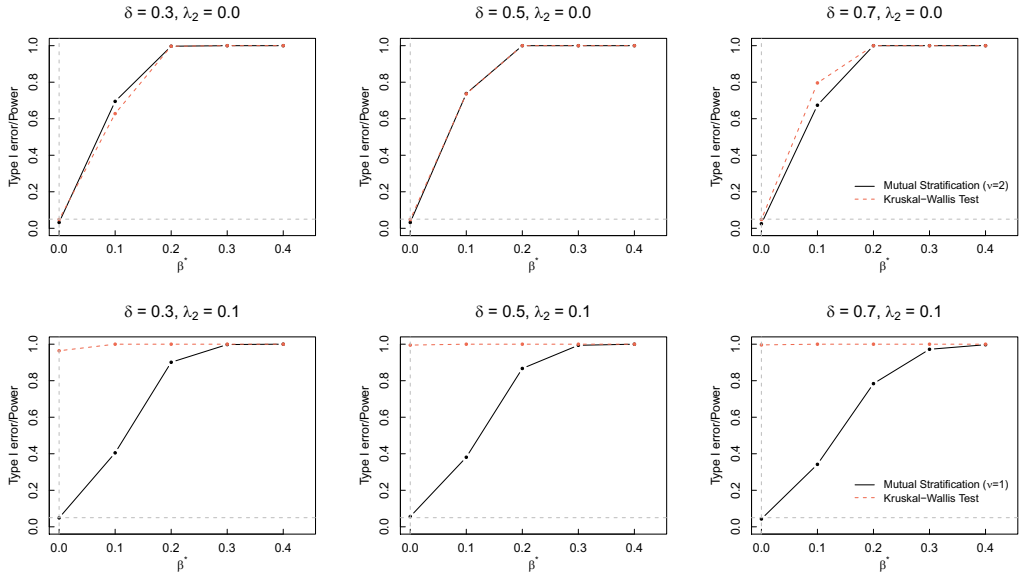

FIG 1. *Rejection rates under model* (12)*. When* $\lambda_2 = 0.0$*, both methods are valid and the Kruskal-Wallis test has a higher power than the combined p-value. On the other hand, when* $\lambda_2 = 0.1$*, the Kruskal-Wallis test is not valid anymore but the proposed method gives valid inference as long as the minimum number of valid instruments* ν *is correctly specified.*

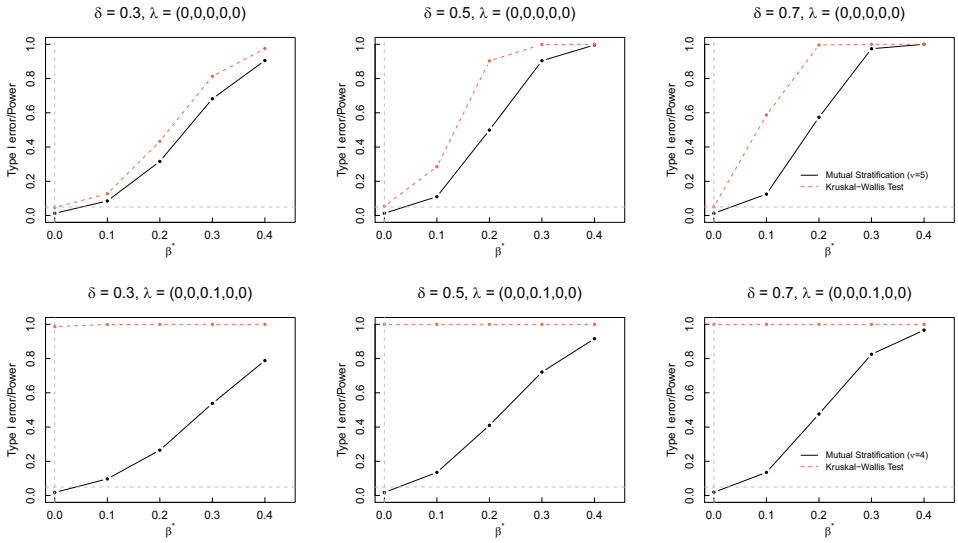

FIG 2. *Rejection rates under model* (13)*. When* $\boldsymbol{\lambda} = (0, 0, 0, 0, 0)$*, both methods are valid and the Kruskal-Wallis test has a higher power than the combined p-value using mutual stratification. On the other hand, when* $\boldsymbol{\lambda} = (0, 0, 0.1, 0, 0)$*, the Kruskal-Wallis test is not valid anymore but the proposed method gives valid inference as long as the minimum number of valid instruments* ν *is correctly specified.*

6. Data applications.

6.1. *The long-run education cost of World War II.* Social science studies sometimes involve multiple candidate instruments with unclear dependence structure and uncertain valid-

ity. We illustrate in this section the utility of the balanced block design in providing additional guarantees.

The disruption of education suffered by school-age children remains an important aspect of the long-run cost of a war and affords a promising instrumental variable for studying the effect of education on earnings. Ichino and Winter-Ebmer (2004) studied the effect of education attainment on the hourly wage in 1984–86 for individuals that were 10 years old during World War II via a two-stage least squares analysis, and documented sizable earnings loss among Germans born between 1930–39 compared to their cohorts born between 1925–29 and 1940–49. The original analysis of the German population included the indicators of 1930–39 cohort and father without high-school degree as candidate instruments, and the indicators of gender, living in big city, and father being a blue-collar worker as covariates.

Using data from the Socio-Economic Panel (SOEP) (Socio-Economic Panel (SOEP), 2018; Wagner, Frick and Schupp, 2007), we examine the hourly wage in 1986 for $2,792$ Germans born between 1925 and 1949. We develop three evidence factors from previously used strategies for identifying the effect of education on earnings, namely (i) the indicator of born between 1930–39 as an instrument, denoted by Z_1, (ii) the indicator of living in city until age 15 as an instrument, denoted by Z_2, and (iii) the indicator of receiving 10 or more years of education as the treatment, denoted by D, to address this question, each depending on very different assumptions for its validity. Table 3 depicts the structure of the three factors. See Ichino and Winter-Ebmer (2004) for discussion on age 10 being a crucial age in the German educational system for pupils to decide their future education pathways. The cohort born between 1930–39 reached age 10 during or immediately after the war, and thus suffered more serious disruptions to education at this crucial stage compared with other cohorts. See also Ichino and Winter-Ebmer (2004) for discussion on the civilian population in big cities being hit by the war more severely than in small villages or in the countryside, hence exposed to more serious disruptions to education. Following Ichino and Winter-Ebmer (2004), we use the indicators of born between 1930–39 and living in city until age 15 as two plausible instruments.

Following Ichino and Winter-Ebmer (2004), we adjust for the effects of age with a cubic polynomial in age for each gender. The residuals become the covariance-adjusted outcomes for the randomization tests under the balanced block design. The first analysis uses an indicator of whether a person was born in the 30s as a candidate instrument. The second analysis uses an indicator of whether a person lived in city until age 15 as a candidate instrument. The third analysis directly compares individuals getting 10 or more years of education with those otherwise, viewing both city/country and 30s/non-30s as covariates. The partial F-statistic from the ordinary least squares regression of the treatment D on Z_1 and Z_2 is 74.41 jointly, suggesting both candidate instruments have reasonable strengths. Following Ichino and Winter-Ebmer (2004), we adjust for gender and whether father had a high school degree by stratifying the individuals into four initial strata, and then form balanced blocks within each stratum based on the approximate proportions of (Z_1, Z_2) within that stratum. We exclude individuals for which any of the covariates or instruments are missing.

Table 4 shows the results at biases of $\Gamma = 1, 1.1, 1.2, 1.25$ under the balanced block design with marginal tests and reinforced design, respectively. The inferences under the reinforced design use all data points, whereas those under the balanced block design use only those that can be formed into the balanced blocks. See Karmakar, Small and Rosenbaum (2021) for interpretation of the parameter Γ in terms of the impact of unobserved covariates on treatment assignment and amplification. Table 4 summarizes the raw p-value upper bounds from the tests of Z_1, Z_2, and treatment D, respectively, and the combined p-value upper bounds assuming at least two of them are valid using the truncated product method with $\varkappa = 0.2$. The analyses under the balanced block design are in general more conservative than those under the reinforced design; both suggest that the full concurrence of all three evidence factors depends critically on the validity of the candidate instruments.

TABLE 3

Counts and percents of receiving 10 or more years of education for the four strata created by the two candidate instruments in the World War II study.

Year of birth	Place grew up	Education	Count	%
Born in 1930–39	City	$\geq$ 10 years	488	79
		<10 years	132	21
	Country	$\geq$ 10 years	330	63
		<10 years	196	37
Otherwise	City	$\geq$ 10 years	851	88
		<10 years	112	12
	Country	$\geq$ 10 years	495	72
		<10 years	188	28
Total			2,792	

TABLE 4

p-value upper bounds under the three analysis strategies for the World War II study. "BB" stands for the balanced block design with marginal tests for Z_k $(k = 1, 2)$ and stratified test for D. "RD" stands for the original reinforced design by Karmakar, Small and Rosenbaum (2021) in the order of Z_1 to Z_2 to D. "RD reversed" stands for the reinforced design in the order of Z_2 to Z_1 to D. The combined p-values (column "C") are computed using the truncated product method with $\varkappa = 0.2$ assuming at least two are valid.

	BB				RD				RD reversed			
Γ	Z_1	Z_2	D	C	Z_1	Z_2	D	C	Z_1	Z_2	D	C
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.1	0.05	0.04	0.00	0.01	0.02	0.00	0.00	0.00	0.02	0.00	0.00	0.00
1.2	0.26	0.23	0.00	1.00	0.13	0.04	0.00	0.02	0.13	0.05	0.00	0.03
1.25	0.44	0.40	0.00	1.00	0.25	0.11	0.00	0.22	0.26	0.12	0.00	0.23

6.2. *The effect of malaria on stunting.* In this section we provide an application of evidence factor analysis using our proposed mutual stratification with nested instruments.

Understanding the effect of malaria on stunting among children is important for building models for public health priorities like the Lives Saved Tool (Jackson and Black, 2017). To rule out the effect of unmeasured confounding between malaria and stunting among children, Ateba et al. (2021) used the randomized insecticide-treated bednets as an instrument based on a cluster randomized trial conducted in Western Kenya. This is based on a previous finding that bednet usage has a negative association with the onset of malaria. On the other hand, the time at which each child was 'assigned the bednet' was different among children. The bednet trial started in 1997. Therefore, a child born after 1997 could be exposed to the bednet since birth while a child born before 1997 could be only partially exposed to the bednet. Year born may be associated with unmeasured confounders (e.g., improvement in sanitation) which may make the instrument invalid.

Following the study in Ateba et al. (2021), we used the longitudinal observational study of the Asembo Bay Cohort study that includes a cluster randomized trials of bednets. Figure 3 shows that the proportion of lifetime exposed to the bednet has a negative correlation with symptomatic malaria per age, with 0 indicating the control and 1 indicating the intervention before birth among $N = 20,521$ children aged 0 to 5 years old. There is evidence that bednet use provides some protection against other infectious diseases besides malaria, such as cutaneous leishmaniasis, that are transmitted by insects (Wilson et al., 2014). These diseases may also affect stunting, leading to the possible invalidity of the bednet intervention as an instrument.

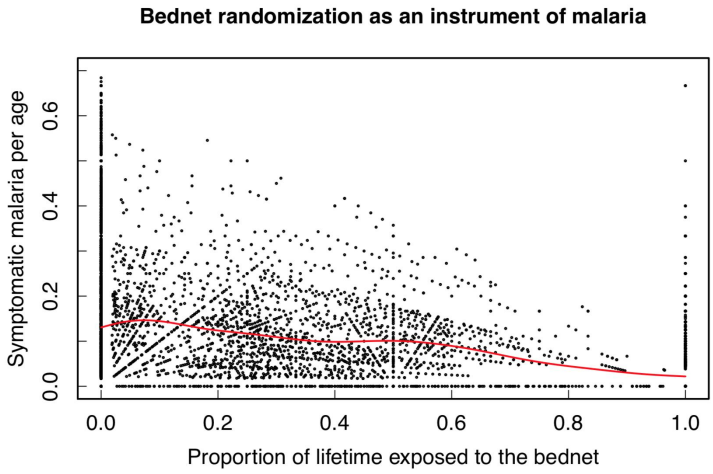

FIG 3. *Relationship between proportion of lifetime exposed to the bednet (instrument) and symptomatic malaria per age (exposure). The red line is a fitted cubic spline of the data, showing the negative correlation between the instrument and the exposure.*

TABLE 5

Results of the effect of malaria on stunting study. A t-statistic (p-value) of the regression coefficient representing the association between each instrument $Z_{ij,k}$ and the exposure variable D_{ij} conditioned on $(\mathbf{X}_{ij}, \mathbf{A}_{ij,-k})$. The second row presents the number of strata (S_k) used for each mutual stratification ($k = 1, 2, 3$) and the number of the control and the treated individuals to be compared to. The last four rows show the resulting p-value upper bounds at different sensitivity parameter Γ_k.

	IV1 ($k = 1$)	IV2 ($k = 2$)	IV3 ($k = 3$)
t-statistic (p-value) of each IV	-9.59 (< 0.001)	-7.73 (< 0.001)	-9.04 (< 0.001)
S_k (control vs. treated)	42 (745 vs. 3988)	39 (1262 vs. 740)	37 (1773 vs. 1158)
p-value ($\Gamma_k = 1.0$)	0.501	0.309	0.001
p-value ($\Gamma_k = 1.1$)	0.835	0.647	0.015
p-value ($\Gamma_k = 1.2$)	0.968	0.880	0.073
p-value ($\Gamma_k = 1.3$)	0.996	0.972	0.214

In our analysis, instead of using a binary instrumental variable that indicates having received the bednet intervention or not, we render it into three instruments depending on whether and when the bednet intervention was assigned to the child. The first instrument is an indicator of any exposure to the bednet intervention; the second instrument is an indicator of no less than 20% of exposure to the bednet intervention; the third instrument is indicating no less than 50% of exposure to the bednet intervention.

We first stratify the children in this study whose outcome (stunting) is measured in or after year 1995 by the following baseline covariates: age with four categories by quartiles, the sickle cell trait, and the child's village to account for the study design of a cluster randomized trial. We denote the exposure variable for subject ij as D_{ij} indicating the child's clinical malaria incidence rate (the number of symptomatic malaria episodes divided by the child's age) and the outcome variable as R_{ij} indicating child ij's height adjusted for age calculated as the Z-score using Epi-Info version 2000 (Ateba et al., 2021). We finally performed evidence factor analysis using the mutual stratification with $\mathbf{A}_{ij,[3]} = (Z_{ij,1}, Z_{ij,2}, Z_{ij,3})$.

Table 5 presents the results of mutual stratification analysis with each instrument. The t-statistics in the first row illustrate that all three instruments are negatively associated with the

exposure variable conditioned on the other instruments and the observed covariates. In the absence of unmeasured confounding, i.e., $\Gamma_k = 1.0$, we could reject the null only with the third instrument, with which we still reject the null at $\Gamma_k = 1.1$.

On the other hand, when all of the three instruments are valid, we can use other methods of combining the p-values. For example, when $\Gamma_k = 1.0$, using Fisher's p-value combination and the truncated p-value combination with $\varkappa = 0.2$, the combined p-values are 0.007 and 0.018, respectively. At $\Gamma_k = 1.1$, the combined p-values are 0.141 and 0.108, respectively, using the two methods. When the minimum number of valid instruments is set to $\nu = 2$, allowing at most one invalid instrument, the combined p-value is 0.44 under no unmeasured bias using Fisher's p-value combination method.

7. Discussion, limitation, and recommendations. This paper proposed two novel methods to conduct valid inference with possibly invalid instruments. Under the more liberally defined exclusion restriction condition, we allowed a subset of the candidate instruments to be conditionally valid after conditioning on those that directly violate the exclusion restriction. Along the lines of Karmakar, Small and Rosenbaum (2021), the proposed methods deliver multiple and nearly independent inferential results from multiple instruments, but remove the stringent requirement on proper ordering of that paper, as opposed to much of the existing literature that produces a single result when considering multiple candidate instruments. This is done by an evidence factor analysis, which further enables us to perform a separate sensitivity analysis for each instrument.

We showed through theory and numerical results that the balanced block design restores the validity of individual marginal and reinforced tests in the absence of proper conditioning and ensures asymptotically nearly independent p-values under all three types of inferences, namely the marginal, reinforced, and conditional tests. We thus recommend using the balanced block design for constructing approximate evidence factors when the validity of any of the multiple instruments available is in doubt. The choice between marginal, reinforced, or conditional tests, on the other hand, may be made based on subject matter knowledge to boost the power and generalizability of the tests.

A key limitation of the balanced block design is that it requires observations from all possible combinations of the instruments of interest, and is thus infeasible if the candidate instruments are nested. For the use of evidence factor analysis with nested instruments, we propose to stratify by all the instruments, except the one used for making a comparison, and other observed covariates. Then, under certain conditions, the resulting p-values are shown to be stochastically larger than the uniform under the null. This stratification method will be particularly useful when we have an ordinal instrument of which different levels may invalidate the instrument. In this case, we can enhance a causal comparison by dividing a single instrument into several nested instruments and investigating the sensitivity to the invalidity at each level.

However, dividing a single instrument into multiple instruments is likely to diminish power. We may consider additional adjustments for improving power under certain distributional assumptions (Zhao, Small and Su, 2019; Tian and Ramdas, 2019). In addition, without a priori knowledge of the instrument, it is unclear how to determine the cut points that divide an ordinal instrument into a few nested instruments. These two are in our future work agenda to use possibly invalid instrumental variables for evidence factor analysis.

We considered the randomization-based inference under assignment model (2). The findings provide insights for more complex generalizations. As commented by Imbens (2003), "often most of the insights of a sensitivity analysis can be obtained with relatively simple models." We conjecture that the theory extends to assignment mechanisms with more complex dependence structures as well. We leave the technical details to future work. In addition,

our approach is limited to testing the sharp null hypothesis of no treatment effect for any unit. Future work is needed to apply the evidence factor analysis with multiple instruments to test other null hypotheses, such as the weak null hypothesis of no average treatment effect or bounded nulls (Caughey et al., 2021).

The output of our proposed statistical analysis is a comprehensive set of results that quantify (a) separate pieces of evidence for a causal effect from different candidates instruments, (b) evidence that remains valid if there is some amount of bias from unmeasured confounding, and (c) combined evidence if it is guessed that the number of invalid instruments among the candidate instruments is less than some upper bound.

Funding. Bikram Karmakar was supported by NSF Grant DMS-2015250.

REFERENCES

ANGRIST, J. D., IMBENS, G. W. and RUBIN, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91** 444–455.

ANGRIST, J. D. and KRUEGER, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics* **106** 979–1014.

ATEBA, F. F., DOUMBIA, S., TER KUILE, F. O., TERLOUW, D. J., LEFEBVRE, G., KARIUKI, S. and SMALL, D. S. (2021). The effect of malaria on stunting: an instrumental variables approach. *Transactions of the Royal Society of Tropical Medicine and Hygiene, in press*.

BECKER, B. J. (1994). Combining significance levels. In *The Handbook of Research Synthesis* (H. Cooper and L. V. Hedges, eds.) 215–230.

BENJAMINI, Y. and HELLER, R. (2008). Screening for partial conjunction hypotheses. *Biometrics* **64** 1215-1222.

BOWDEN, J., DAVEY SMITH, G. and BURGESS, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44** 512–525.

BOWDEN, J., DAVEY SMITH, G., HAYCOCK, P. C. and BURGESS, S. (2016). Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. *Genetic Epidemiology* **40** 304–314.

BRESLOW, N. (1970). A generalized Kruskal-Wallis test for comparing K samples subject to unequal patterns of censorship. *Biometrika* **57** 579–594.

BURGESS, S., BUTTERWORTH, A. and THOMPSON, S. G. (2013). Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology* **37** 658–665.

BURGESS, S., DUDBRIDGE, F. and THOMPSON, S. G. (2016). Combining information on multiple instrumental variables in Mendelian randomization: Comparison of allele score and summarized data methods. *Statistics in Medicine* **35** 1880–1906.

BURGESS, S., BOWDEN, J., FALL, T., INGELSSON, E. and THOMPSON, S. G. (2017). Sensitivity analyses for robust causal inference from Mendelian randomization analyses with multiple genetic variants. *Epidemiology* **28** 30.

CARD, D. (1993). Using geographic variation in college proximity to estimate the return to schooling. *NBER working paper* **w4483**.

CARD, D. (1999). The causal effect of education on earnings. (O. C. Ashenfelter and D. Card, eds.). *Handbook of Labor Economics* **3** 1801–1863. Elsevier.

CAUGHEY, D., DAFOE, A., LI, X. and MIRATRIX, L. (2021). Randomization inference beyond the sharp null: Bounded null hypotheses and quantiles of individual treatment effects. *arXiv preprint arXiv:2101.09195*.

DAVIES, N. M., VON HINKE KESSLER SCHOLDER, S., FARBMACHER, H., BURGESS, S., WINDMEIJER, F. and SMITH, G. D. (2015). The many weak instruments problem and Mendelian randomization. *Statistics in Medicine* **34** 454–468.

DEATON, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature* **48** 424-55.

FISHER, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain* **33** 503–513.

FOGARTY, C. B. and HASEGAWA, R. B. (2019). Extended sensitivity analysis for heterogeneous unmeasured confounding with an application to sibling studies of returns to education. *The Annals of Applied Statistics* **13** 767–796.

GAIL, M. H., TAN, W. Y. and PIANTADOSI, S. (1988). Tests for no treatment effect in randomized clinical trials. *Biometrika* **75** 57–64.

GRECO M, F. D., MINELLI, C., SHEEHAN, N. A. and THOMPSON, J. R. (2015). Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Statistics in Medicine* **34** 2926–2940.

GUO, Z., KANG, H., TONY CAI, T. and SMALL, D. S. (2018). Confidence intervals for causal effects with invalid instruments by using two-stage hard thresholding with voting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 793–815.

HADLEY, J., POLSKY, D., MANDELBLATT, J. S., MITCHELL, J. M., WEEKS, J. C., WANG, Q., HWANG, Y.-T. and TEAM, O. R. (2003). An exploratory instrumental variable analysis of the outcomes of localized breast cancer treatments in a medicare population. *Health Economics* **12** 171–186.

HALSEY, A. H., HALSEY, A. H., ALBERT HENRY, H., HEATH, A. F., RIDGE, J. M. et al. (1980). *Origins and destinations: Family, class, and education in modern Britain*. Oxford: Clarendon Press; New York: Oxford University Press.

HAN, C. (2008). Detecting invalid instruments using L1-GMM. *Economics Letters* **101** 285–287.

HARMON, C. and WALKER, I. (1995). Estimates of the economic return to schooling for the United Kingdom. *The American Economic Review* **85** 1278–1286.

HASEGAWA, R. and SMALL, D. S. (2017). Sensitivity analysis for matched pair analysis of binary data: From worst case to average case analysis. *Biometrics* **73** 1424-1432.

HENG, S., SMALL, D. S. and ROSENBAUM, P. R. (2020). Finding the strength in a weak instrument in a study of cognitive outcomes produced by Catholic high schools. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **183** 935–958.

HENG, S. and SMALL, D. S. (2021). Sharpening the Rosenbaum sensitivity bounds to address concerns about interactions between observed and unobserved covariates. *Statistica Sinica* to appear.

HSU, J. Y., SMALL, D. S. and ROSENBAUM, P. R. (2013). Effect modification and design sensitivity in observational studies. *Journal of the American Statistical Association* **108** 135–148.

ICHINO, A. and WINTER-EBMER, R. (2004). The long-run educational cost of World War II. *Journal of Labor Economics* **22** 57–87.

IMBENS, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review* **93** 126-132.

IMBENS, G. W. (2010). Better LATE than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of economic literature* **48** 399-423.

JACKSON, B. D. and BLACK, R. E. (2017). A literature review of the effect of malaria on stunting. *The Journal of nutrition* **147** 2163S–2168S.

KANG, H., ZHANG, A., CAI, T. T. and SMALL, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to Mendelian randomization. *Journal of the American statistical Association* **111** 132–144.

KANG, H., LEE, Y., CAI, T. T. and SMALL, D. S. (2021). Two robust tools for inference about causal effects with invalid instruments. *Biometrics (In press)*.

KARMAKAR, B., DOUBENI, C. A. and SMALL, D. S. (2020). Evidence factors in a case-control study with application to the effect of flexible sigmoidoscopy screening on colorectal cancer. *Annals of Applied Statistics* **14** 829–849.

KARMAKAR, B., FRENCH, B. and SMALL, D. (2019). Integrating the evidence from evidence factors in observational studies. *Biometrika* **106** 353–367.

KARMAKAR, B. and SMALL, D. S. (2020). Assessment of the extent of corroboration of an elaborate theory of a causal hypothesis using partial conjunctions of evidence factors. *The Annals of Statistics* **48** 3283 – 3311.

KARMAKAR, B., SMALL, D. S. and ROSENBAUM, P. R. (2020). Using evidence factors to clarify exposure biomarkers. *American Journal of Epidemiology* **189** 243–249.

KARMAKAR, B., SMALL, D. S. and ROSENBAUM, P. R. (2021). 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** 82–92.

KOLESÁR, M., CHETTY, R., FRIEDMAN, J., GLAESER, E. and IMBENS, G. W. (2015). Identification and inference with many invalid instruments. *Journal of Business & Economic Statistics* **33** 474–484.

LORCH, S. A., BAIOCCHI, M., AHLBERG, C. E. and SMALL, D. S. (2012). The differential impact of delivery hospital on the outcomes of premature infants. *Pediatrics* **130** 270–278.

NATTINO, G., LU, B., SHI, J., LEMESHOW, S. and XIANG, H. (2021). Triplet matching for estimating causal effects with three treatment arms: A comparative study of mortality by trauma center level. *Journal of the American Statistical Association* **116** 44–53.

ROSENBAUM, P. R. (2001). Replicating effects and biases. *The American Statistician* **55** 223–227.

ROSENBAUM, P. R. (2002). Covariance adjustment in randomized experiments and observational studies. *Statistical Science* **17** 286–327.

ROSENBAUM, P. R. (2010a). *Design of Observational Studies*. Springer, New York.

ROSENBAUM, P. R. (2010b). Evidence factors in observational studies. *Biometrika* **97** 333–345.

ROSENBAUM, P. R. (2011). Some approximate evidence factors in observational studies. *Journal of the American Statistical Association* **106** 285–295.

ROSENBAUM, P. R. (2015). Two R packages for sensitivity analysis in observational studies. *Observational Studies* **1** 1–17.

ROSENBAUM, P. R. et al. (2017). The general structure of evidence factors in observational studies. *Statistical Science* **32** 514–530.

ROSENBAUM, P. R., ROSS, R. N. and SILBER, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *Journal of the American Statistical Association* **102** 75-83.

RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* **74** 318–328.

SANDER, W. (1995). Schooling and quitting smoking. *The Review of Economics and Statistics* 191–199.

SIMES, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73** 751–754.

SMALL, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association* **102** 1049–1058.

SMALL, D. S. and ROSENBAUM, P. R. (2008). War and wages: the strength of instrumental variables and their sensitivity to unobserved biases. *Journal of the American Statistical Association* **103** 924–933.

SOCIO-ECONOMIC PANEL (SOEP) (2018). Data for years 1984–2018, v35i, SOEP.

SPIEKER, A. J., GREEVY, R. A., NELSON, L. A. and MAYBERRY, L. S. (2020). Bounding the local average treatment effect in an instrumental variable analysis of engagement with a mobile intervention. *arXiv preprint arXiv:2008.06473*.

TAN, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association* **101** 1607–1618.

TIAN, J. and RAMDAS, A. (2019). ADDIS: an adaptive discarding algorithm for online FDR control with conservative nulls. In *Advances in Neural Information Processing Systems* 9388–9396.

TUKEY, J. W. (1993). Tightening the clinical trial. *Controlled Clinical Trials* **14** 266–285.

VOORS, M. J., NILLESEN, E. E., VERWIMP, P., BULTE, E. H., LENSINK, R. and VAN SOEST, D. P. (2012). Violent conflict and behavior: a field experiment in Burundi. *American Economic Review* **102** 941–64.

WAGNER, G. G., FRICK, J. R. and SCHUPP, J. (2007). The German Socio-Economic Panel study (SOEP)-evolution, scope and enhancements. SOEP papers on Multidisciplinary Panel Data Research No. 1, DIW Berlin, The German Socio-Economic Panel (SOEP).

WALKER, V. M., DAVIES, N. M., MARTIN, R. M. and KEHOE, P. G. (2020). Comparison of antihypertensive drug classes for dementia prevention. *Epidemiology (Cambridge, Mass.)* **31** 852.

WANG, X., JIANG, Y., ZHANG, N. R. and SMALL, D. S. (2018). Sensitivity analysis and power for instrumental variable studies. *Biometrics* **74** 1150–1160.

WILSON, A. L., DHIMAN, R. C., KITRON, U., SCOTT, T. W., VAN DEN BERG, H. and LINDSAY, S. W. (2014). Benefit of insecticide-treated nets, curtains and screening on vector borne diseases, excluding malaria: a systematic review and meta-analysis. *PLoS neglected tropical diseases* **8** e3228.

WINDMEIJER, F., FARBMACHER, H., DAVIES, N. and DAVEY SMITH, G. (2019). On the use of the lasso for instrumental variables estimation with some invalid instruments. *Journal of the American Statistical Association* **114** 1339–1350.

WU, J. and DING, P. (2020). Randomization tests for weak null hypotheses in randomized experiments. *Journal of the American Statistical Association*. (in press).

ZAYKIN, D. V., ZHIVOTOVSKY, L. A., WESTFALL, P. H. and WEIR, B. S. (2002). Truncated product method for combining P-values. *Genetic Epidemiology* **22** 170–185.

ZENG, S., LI, F. and DING, P. (2020). Is being the only child harmful to psychological health?: Evidence from an instrumental variable analysis of China's One-Child Policy. *Journal of the Royal Statistical Society - Series A (Statistics in Society)* **183** 1615–1635.

ZENG, D., THOMSEN, M. R., NAYGA JR, R. M. and ROUSE, H. L. (2019). Neighbourhood convenience stores and childhood weight outcomes: an instrumental variable approach. *Applied Economics* **51** 288–302.

ZHANG, K., SMALL, D. S., LORCH, S., SRINIVAS, S. and ROSENBAUM, P. R. (2011). Using split samples and evidence factors in an observational study of neonatal outcomes. *Journal of the American Statistical Association* **106** 511–524.

ZHAO, Q., SMALL, D. S. and SU, W. (2019). Multiple testing when many p-values are uniformly conservative, with application to testing qualitative interaction in educational interventions. *Journal of the American Statistical Association* **114** 1291–1304.

ZUBIZARRETA, J. R., NEUMAN, M., SILBER, J. H. and ROSENBAUM, P. R. (2012). Contrasting evidence within and between institutions that provide treatment in an observational study of alternate forms of anesthesia. *Journal of the American Statistical Association* **107** 901-915.