Received: 17 April 2020 | Accepted: 7 April 2021

DOI: 10.1111/rssb.12429

ROYAL STATISTICAL SOCIETY
Series B Statistical Methodology

ORIGINAL ARTICLE

Leveraging the Fisher randomization test using confidence distributions: Inference, combination and fusion learning

Xiaokang Luo | **Tirthankar Dasgupta** | **Minge Xie** | **Regina Y. Liu**

Department of Statistics, Rutgers University, New Brunswick, New Jersey, USA

Correspondence
Tirthankar Dasgupta, Department of Statistics, Rutgers University, 110 Frelinghuysen Road, Piscataway, NJ 08854, USA.
Email: td370@stat.rutgers.edu

Funding information
National Science Foundation (NSF) Division of Mathematical Sciences (DMS), Grant/Award Number: #1451817, #1737857, #1812048, #2015373 and #2027855

Abstract

The flexibility and wide applicability of the Fisher randomization test (FRT) make it an attractive tool for assessment of causal effects of interventions from modern-day randomized experiments that are increasing in size and complexity. This paper provides a theoretical inferential framework for FRT by establishing its connection with confidence distributions. Such a connection leads to development's of (i) an unambiguous procedure for inversion of FRTs to generate confidence intervals with guaranteed coverage, (ii) new insights on the effect of size of the Monte Carlo sample on the estimation of a *p*-value curve and (iii) generic and specific methods to combine FRTs from multiple independent experiments with theoretical guarantees. Our developments pertain to finite sample settings but have direct extensions to large samples. Simulations and a case example demonstrate the benefit of these new developments.

KEYWORDS

exact test, fiducial, model-free, monte carlo, p-value, sharp nulle

1 | INTRODUCTION

Fisher randomization tests (FRT) are flexible tools because they are model free, permit assessment of causal effects of interventions on *any* type of response for *any* assignment mechanism using *any* test statistic, and can be easily extended to model-based inference (Rubin, 1980, 1984). The tremendous recent development of computing resources has sparked much interest in using FRT to test complex

causal hypotheses that can arise from modern-day randomized experiments (e.g. Athey et al., 2017; Basse & Feller, 2018; Basse et al., 2019; Hennessy et al., 2016) in social, biomedical, educational and behavioural sciences. The work by Morgan and Rubin (2012) has shown how randomization tests can be applied to design and analyse randomized experiments with several pre-treatment covariates. As modern experiments continue to grow in *size* (in terms of number of experimental units, interventions, covariates and as combinations of several independent sub-experiments) and *complexity* (e.g. non-standard randomized assignment mechanisms), the flexibility and wide applicability of FRT make it a particularly promising tool to analyse such experiments.

However, there are three aspects of FRT that can arguably be made more transparent and thus more appealing to scientists. The first concern is related to the theoretical and implementation aspects of inverting FRTs to generate interval estimators of treatment effects, because interval estimates are typically more appealing than a p-value or an acceptance–rejection decision. This inversion is done by testing a sequence of sharp null hypotheses of constant treatment effects, and using the curve of the resulting p-values. The first original reference of a similar inversion procedure appears in Pitman (1937). Whereas proposed procedures and algorithms appear to work well in large sample settings (Ding, 2017; Garthwaite, 1996), it is somewhat surprising that the theoretical properties of this inversion procedure, especially in a finite population setting, have been scantily discussed in causal inference literature and apparently counter-intuitive simulation results have sometimes been difficult to explain. See, for example, discussion in Section 7.3 of Ding (2017) on the intervals for factorial effects obtained in Dasgupta et al. (2015). The research in this paper reveals how the discrete nature of the p-value statistic poses complexities associated with the inversion procedure in a finite population setting and proposes a viable solution.

The second aspect is computational. The FRT is a computation-intensive procedure, as its classical form involves generating all possible permutations of the observed assignment vector that are consistent with the assignment mechanism. The total number of such permutations in a balanced completely randomized design increases from 252 to 10^{29} as the number of units increases from 10 to 100. A common way to get around this issue is to generate a sample of all possible permutations, say 1000 or 5000, and use it to obtain a Monte Carlo estimate of the p-value. However, to the best of our knowledge, there does not exist any insights or theoretical results about how large a sample size is needed to guarantee acceptable inferential properties. This computational complexity increases manyfold when we consider the problem of interval estimation, because it entails computing the p-values at several values of the treatment effect.

The third aspect, related to the broader subject of fusion learning, is performing meta-analysis using FRT. This entails combining results from independently conducted randomized experiments, possibly with different assignment mechanisms, to draw sharper inference on a common treatment effect. Whereas there exist several methods in literature to combine p-values from independent tests of hypotheses, obtaining a composite interval with the desired coverage poses additional challenges, especially in the finite population case when the p-value function is discrete.

This paper aims to address the three issues mentioned above by providing a new theoretical perspective of FRT using the concept of confidence distributions (CDs), which will be formally introduced in Section 2.2. Specifically, the paper makes the following contributions: (i) Drawing inspiration from the concept of CDs, it provides the *first formal definitions* of a class of p-value functions in the context of FRTs. It is noteworthy that these definitions are not direct implications of the existing CD literature, considering the discrete nature of the randomization distribution of any test statistic. (ii) It identifies specific mathematical conditions that guarantees inversion of the p-value functions to generate confidence intervals with desired coverage. (iii) It provides a precise algorithm for computing confidence intervals that is more robust than the traditional approach, because it does not depend on the choice of

discrete levels of treatment effects. (iv) It addresses the computational complexity associated with (iii) by providing a novel result on the impact of Monte Carlo sample size on the accuracy of estimation of the *entire p*-value function, when such estimation is based on a *single* Monte Carlo sample. (v) It provides a general procedure for combining inferences from similar and dissimilar experiments by extending methods for combining CDs (again, such an extension is non-trivial due to the discreteness of the randomization distribution of the test statistic).

In Section 2, we introduce the basic notions and concepts of FRT and CD. Section 3 establishes the bridge between FRT and CD by defining five different *p*-value functions, examining theoretical properties of these *p*-value functions and showcasing their applications in the context of hypothesis testing. Section 4 identifies conditions that are necessary for inverting FRTs to generate confidence intervals with guaranteed coverage and provides an algorithm to do this inversion. Section 5 investigates the effect of size of the Monte Carlo sample on the estimation of *p*-value functions and provides a result that quantifies the estimation accuracy of the entire *p*-value function based on a single Monte Carlo sample. Section 6 develops efficient methods to combine FRTs from independent experiments. Using a real-life example, Section 8 demonstrates the usefulness of the *p*-value functions in drawing inference. Section 9 contains some concluding remarks.

2 | FUNDAMENTALS

2.1 | The Fisher randomization test understood through the potential outcomes model

Consider a finite population of N experimental units, each of which can be exposed to either a treatment (denoted by 1) or a control (denoted by 0). For unit i, let $Y_i(1)$ and $Y_i(0)$, respectively, denote the potential outcomes (Neyman, 1923; Rubin, 1974) under treatment and control. We define the unit-level causal effect of the treatment on unit i as $\theta_i = Y_i(1) - Y_i(0)$, and the finite-population level average causal effect

$$\theta = N^{-1}\sum_{i=1}^{N}\theta_i = N^{-1}\sum_{i=1}^{N}Y_i(1) - N^{-1}\sum_{i=1}^{N}Y_i(0).$$

In a randomized design, the N units are assigned to the two treatment groups using a known randomized assignment mechanism. Let $\mathbf{W} = (W_1, \ldots, W_N)^\top$ denote a binary random vector whose ith element W_i equals one or zero according as unit i is assigned to treatment or control. The assignment mechanism is defined as the probability distribution of the random vector $\mathbf{W}$ and dictates all inference statements. In a completely randomized design with N_1 and N_0 units assigned to treatment and control, respectively, where N_1 and N_0 are predetermined, the assignment mechanism is:

$$P(W_1 = w_1, \ldots, W_N = w_N) = \left(\frac{N!}{N_0!N_1!}\right)^{-1}\mathbb{I}\left(\sum\nolimits_{i=1}^{N} w_i = N_1\right),$$

where $\mathbb{I}$ (A) is the indicator function for event A. The development in this paper covers any randomized assignment mechanism as long as the assignment probability $P(W_1 = w_1, \ldots, W_N = w_N)$ is fully specified. The observed outcome for the ith unit is denoted by $Y_i^{\text{obs}} = W_iY_i(1) + (1 - W_i)Y_i(0),\ i = 1, \ldots, N$. Thus, only one of the two potential outcomes for each unit is observed and the other is missing.

Consider testing the sharp null hypothesis

$$H_0^\theta: Y_i(1) - Y_i(0) = \theta, \quad \text{for all } i = 1, \ldots, N, \tag{1}$$

that is, all units have an identical treatment effect θ. A special case of this hypothesis is $H_0^0: \theta = 0$, Fisher's sharp null hypothesis of no treatment effect on any unit (Fisher, 1935; Rubin, 1980). The hypothesis H_0^θ can be tested by considering a suitable test statistic T, and comparing its observed value T^{obs} with the randomization distribution of T under the null hypothesis. This randomization distribution of T is generated by imputing the missing outcomes under H_0^θ and repeatedly generating values of T by drawing from the known probability distribution of the assignment vector $\mathbf{W}$. The p-value is the tail probability measuring the extremeness of the test statistic with respect to its randomization distribution. Rejection of H_0^θ if the p-value is less than or equal to $\alpha \in (0,1)$ leads to a test procedure with level α, that is, the probability of Type-I error not exceeding α. The beauty of this procedure is, it can be tested with any reasonable test statistic that is capable of summarizing the difference between the treatment and control groups.

By varying θ and testing a set of sharp null hypotheses H_0^θ, it is possible to obtain a 'p-value function' of θ, which is a step-value function. This step function can be inverted to generate an interval estimator for the true additive effect θ. As we shall see in Section 3, most of the subsequent developments will be based on this p-value function and its variants. A toy example presented in the supplementary material demonstrates each step involved in conducting a randomization test, generating a p-value function and inverting it to obtain a confidence interval for θ.

2.2 | A brief overview of confidence distributions and confidence curves

The idea of a *confidence distribution* is to use *a sample-dependent distribution function* defined on the parameter space to estimate a fixed but unknown (scalar/vector) parameter (Cox, 1958; Efron, 1993, 1998; Schweder & Hjort, 2016; Xie & Singh, 2013). Such a practice elevates one point (point estimator using the single value of a sample statistic) and two points (confidence interval using a lower limit and an upper limit) to a full function that can be used to draw inference on the parameter of interest. Similar to a Bayesian posterior, a CD contains rich inferential information and can yield all forms of inference, including the classical point and interval estimators.

For ease of illustration, consider the simple case of a scalar parameter $\theta \in \Theta$ with sample data $\mathbf{Y}_n = (Y_1, \ldots, Y_n) \in \mathcal{Y}$. A function $H_n(\cdot) \equiv H(\cdot, \mathbf{Y}_n)$ on $\Theta \times \mathcal{Y}$ is called a *CD function* for θ, if (i) given $\mathbf{Y}_n$, $H_n(\cdot)$ is a cumulative distribution function (CDF) on Θ; and (ii) at the true parameter value $\theta = \theta_0$, $H_n(\theta_0) = H(\theta_0, \mathbf{Y}_n)$, as a function of the sample $\mathbf{Y}_n$, follows a Uniform[0, 1] distribution (Schweder & Hjort, 2002; Singh et al., 2005). In other words, (i) requires that a CD is a sample-dependent distribution function on Θ. Requirement (ii) ensures that the CD function can be used to obtain confidence intervals and test hypotheses. For example, by (ii), $(-\infty, H_n^{-1}(\alpha))$ is a $100\alpha\%$ confidence interval for θ, and $H_n(b)$ provides a p-value function for testing the hypothesis $\Omega_0 : \theta \leq b$ versus $\Omega_1 : \theta > b$. This shows that a one-sided p-value function is a special case of a CD. Corresponding to a CD function $H_n(\theta)$, one can obtain a *confidence curve* (CV)

$$CV(\theta) = 2\min\{H_n(\theta), 1 - H_n(\theta)\},$$

which can also be used to draw similar inferences (Birnbaum, 1961).

Due to the discrete nature of the FRT in which the p-value is a step function as in the last panel of Figure 3 (supplementary material), the following new definition will be useful for this paper. Note that, in the existing literature (e.g. Schweder & Hjort, 2016; Xie & Singh, 2013), CDs

in discrete sample distributions are handled by asymptotics. Finite sample performance is not investigated.

Definition 1 (Upper and Lower CDs)A function $H_n^{\mathrm{L}}(\cdot) = H^{\mathrm{L}}(\cdot, \mathbf{Y}_n)$ mapping $\Theta \times \mathcal{Y}$ to [0,1] is said to be a lower CD for a parameter θ if at the true parameter value $\theta{=}\theta_0$, $H_n^{\mathrm{L}}(\theta_0) \equiv H^{\mathrm{L}}(\theta_0, \mathbf{Y}_n)$, as a function of the sample $\mathbf{Y}_n$ is stochastically larger than a Uniform[0, 1] random variable, that is,

$$P\left[H^{\mathrm{L}}(\theta_0, \mathbf{Y}_n) \leq \alpha\right] \leq \alpha \quad \text{for all} \quad \alpha \in (0, 1). \tag{2}$$

An upper CD $H_n^{\mathrm{U}}(\cdot) = H^{\mathrm{U}}(\cdot, \mathbf{Y}_n)$ for parameter θ can be defined similarly but with Equation (2) replaced by $P\left[H^{\mathrm{U}}(\theta_0, \mathbf{Y}_n) \leq \alpha\right] \geq \alpha$ for all $\alpha \in (0, 1)$.

3 | BRIDGING FRT AND CD THROUGH *p*-VALUE FUNCTIONS

We note that both FRT and CD historically have an implicit 'fiducial' flavour, although in recent developments (Schweder & Hjort, 2016; Xie & Singh, 2013), the concept of CD has been developed without any fiducial interpretation or reasoning. Some researchers consider a CD as 'a frequentist analogue of a Bayesian posterior' (Schweder & Hjort, 2003). On the other hand, Rubin (1984) provided the following Bayesian justification of the FRT: it gives the posterior predictive distribution of the estimand of interest under a model of constant treatment effects and fixed units with fixed responses. These connections motivate us to better understand the properties of FRT by connecting it to CD and exploiting recent results on CD. It should be pointed out that this connection is non-trivial because the theory of CD primarily revolves around parametric models, whereas FRT is essentially a model-free procedure. Obviously, the discrete nature of the distribution of the *p*-value in FRT also adds further complication.

We first extend the notion of the *p*-value for the FRT to a *p*-value function along the lines of that introduced in Section 2.2. To do this, we start with a more careful handling of the notations involved. Let $\mathbf{Y}^{\text{true}}$ denote the true $N \times 2$ matrix of potential outcomes and $\mathbf{Y}_\theta^{\text{imp}}$ the $N \times 2$ imputed matrix consisting of the observed outcomes and imputed missing outcomes under the null hypothesis H_0^θ. Let $\mathbf{W}^{\text{obs}}$ denote the $N \times 1$ observed assigned vector and $\mathbf{Y}^{\text{obs}}$ the $N \times 1$ observed vector of responses. Then the observed data from the experiment can be denoted by $\mathbf{D}^{\text{obs}} = (\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}})$. Also, let $\mathbf{W}^{\text{rep}}$ denote any repeated draw from the distribution of $\mathbf{W}$ while generating the randomization distribution of T. Such a repeated draw generates repeated data $\mathbf{D}_\theta^{\text{rep}} = (\mathbf{Y}_\theta^{\text{rep}}, \mathbf{W}^{\text{rep}})$, where $\mathbf{Y}_\theta^{\text{rep}}$ is the vector of observed outcomes generated from $\mathbf{Y}_\theta^{\text{imp}}$ by assignment vector $\mathbf{W}^{\text{rep}}$.

Let T be any test statistic and T_θ^{rep} denote the discrete random variable having the randomization distribution of T under the null hypothesis H_0^θ. Then the distribution of T_θ^{rep} depends on the imputed potential outcomes matrix $\mathbf{Y}_\theta^{\text{imp}}$ and $\mathbf{W}^{\text{rep}}$. Consequently, we can write

$$T_\theta^{\text{rep}} = T(\mathbf{D}_\theta^{\text{rep}}) = T(\mathbf{Y}_\theta^{\text{imp}}, \mathbf{W}^{\text{rep}}). \tag{3}$$

Finally, note that the observed value of the test statistic T^{obs} depends on $\mathbf{D}^{\text{obs}}$, and consequently on $\mathbf{Y}^{\text{true}}$ and $\mathbf{W}^{\text{obs}}$. This allows us to write

$$T^{\text{obs}} = T(\mathbf{D}^{\text{obs}}) = T(\mathbf{Y}^{\text{true}}, \mathbf{W}^{\text{obs}}). \tag{4}$$

3.1 | *p*-value functions for one-sided alternatives of the sharp null

Whereas the sharp null hypothesis has been widely discussed in literature, the alternative hypothesis against which the sharp null is tested has seldom been mentioned. In this paper, we will keep our alternatives restricted to the class of sharp-nulls to make the interval estimation problem readily interpretable. A violation of the sharp null can be one-sided or two-sided. Below, we define *p*-value functions for one-sided alternative hypotheses.

Definition 2 Consider the one-sided alternative

$$H_1^{\theta^+} : Y_i(1) - Y_i(0) = \phi(> \theta), \tag{5}$$

for all $i = 1, \ldots, N$. Assuming that larger values of the test statistic T indicate departure from the sharp null in favour of $H_1^{\theta^+}$, we define the following p-value functions for testing H_0^{θ} against alternatives $H_1^{\theta^+}$ as:

$$p^{L+}(\mathbf{D}^{\text{obs}}, \theta) = P(T_\theta^{\text{rep}} \geq T^{\text{obs}}) = P\left(T(\mathbf{Y}_\theta^{\text{imp}}, \mathbf{W}^{\text{rep}}) \geq T(\mathbf{D}^{\text{obs}})\right), \tag{6}$$

$$p^{U+}(\mathbf{D}^{\text{obs}}, \theta) = P(T_\theta^{\text{rep}} > T^{\text{obs}}) = P\left(T(\mathbf{Y}_\theta^{\text{imp}}, \mathbf{W}^{\text{rep}}) > T(\mathbf{D}^{\text{obs}})\right). \tag{7}$$

Definition 3

$$H_1^{\theta^-} : Y_i(1) - Y_i(0) = \psi(< \theta), \tag{8}$$

for all $i = 1, \ldots, N$. Assuming that smaller values of the test statistic T indicate departure from the sharp null in favour of $H_1^{\theta^-}$, we define the p-value function for testing H_0^{θ} against alternatives $H_1^{\theta^-}$ as

$$p^{L-}(\mathbf{D}^{\text{obs}}, \theta) = P(T_\theta^{\text{rep}} \leq T^{\text{obs}}) = P\left(T(\mathbf{Y}_\theta^{\text{imp}}, \mathbf{W}^{\text{rep}}) \leq T(\mathbf{D}^{\text{obs}})\right), \tag{9}$$

$$p^{U-}(\mathbf{D}^{\text{obs}}, \theta) = P(T_\theta^{\text{rep}} < T^{\text{obs}}) = P\left(T(\mathbf{Y}_\theta^{\text{imp}}, \mathbf{W}^{\text{rep}}) < T(\mathbf{D}^{\text{obs}})\right). \tag{10}$$

Note that the p-value functions defined in Equations (6)–(10) are random variables under the random mechanism of $\mathbf{W}^{\text{obs}}$, because of their dependence on $\mathbf{D}^{\text{obs}} = (\mathbf{Y}^{\text{obs}}, \mathbf{W}^{\text{obs}})$. However, conditional on $\mathbf{D}^{\text{obs}}$ (i.e., when $\mathbf{W}^{\text{obs}}$ is realized), they are functions of θ only.

Proposition 1 *For any test statistic T, the p-value functions defined in Equations (6)–(10) satisfy the following properties:*

1. Both $p^{L+}(\mathbf{D}^{\text{obs}}, \theta)$ in Equation (6) and $p^{L-}(\mathbf{D}^{\text{obs}}, \theta)$ in Equation (9) are lower CDs as per Definition 1, which means they both stochastically dominate the Uniform[0, 1] random variable at the true value θ_0 of θ and satisfy

$$P\left(p^{L+}(\mathbf{D}^{\text{obs}}, \theta_0) \leq \alpha\right) \leq \alpha, \quad \text{and} \quad P\left(p^{L-}(\mathbf{D}^{\text{obs}}, \theta_0) \leq \alpha\right) \leq \alpha,$$

for $\alpha \in (0, 1)$.

2. Both $p^{U+}(\mathbf{D}^{\text{obs}}, \theta)$ in Equation (7) and $p^{U-}(\mathbf{D}^{\text{obs}}, \theta)$ in Equation (10) are upper CDs in the sense that at $\theta = \theta_0$ they are both stochastically dominated by the Uniform[0, 1] random variable and satisfy

$$P\left(p^{U+}(\mathbf{D}^{\text{obs}}, \theta_0) \leq \alpha\right) \geq \alpha, \quad \text{and} \quad P\left(p^{U-}(\mathbf{D}^{\text{obs}}, \theta_0) \leq \alpha\right) \geq \alpha,$$

for $\alpha \in (0, 1)$.

3. Let $T_{(1)} < T_{(2)} < \cdots < T_{(m)}$ be the m unique ordered values of T for $\theta = \theta_0$ and $\gamma_i = P\left(T(\mathbf{Y}^{\text{true}}, \mathbf{W}) = T_{(i)}\right) > 0$ for $i = 1, 2, \ldots, m$. Then, for any $\alpha \in (0, 1)$,

$$\begin{aligned} &P\left(p^{L+}(\mathbf{D}^{\text{obs}}, \theta_0) \leq \alpha\right) \geq \alpha - \gamma^*, \quad P\left(p^{L-}(\mathbf{D}^{\text{obs}}, \theta_0) \leq \alpha\right) \geq \alpha - \gamma^*, \\ &P\left(p^{U+}(\mathbf{D}^{\text{obs}}, \theta_0) \leq \alpha\right) \leq \alpha + \gamma^*, \quad P\left(p^{U-}(\mathbf{D}^{\text{obs}}, \theta_0) \leq \alpha\right) \leq \alpha + \gamma^*, \end{aligned} \tag{11}$$

where $\gamma^* = \max\{\gamma_1, \gamma_2, \ldots, \gamma_m\}$.

Implications of Proposition 1 and some remarks

1. Consider testing the sharp null hypothesis (1) against one-sided alternatives (5) or (8) using a test statistic whose large or small values indicate departure from the null in favour of Equation (5) or (8), respectively. By part (a) of Proposition 1, the test procedure that rejects the sharp null if the observed value of $p^{L+}(\mathbf{D}^{\text{obs}}, \theta) \leq \alpha$ is *valid* in the sense that the probability of Type-I error does not exceed α. However, by part(b), the rejection rule $p^{U+}(\mathbf{D}^{\text{obs}}, \theta) \leq \alpha$ is not valid. Similarly, use of $p^{L-}(\mathbf{D}^{\text{obs}}, \theta)$ for the one-sided alternative (8) leads to a valid test, while use of $p^{U-}(\mathbf{D}^{\text{obs}}, \theta)$ does not.
2. Equation (11) provide a set of theoretical upper bounds for the discrepancies between the empirical CDFs of the four p-value functions given by Equations (6)–(10) from the CDF of a Uniform[0, 1] variable. However, in practice, these upper bounds will typically be unknown to an analyst because both m and the γ_j's depend on the unknown matrix of potential outcomes and the true parameter value θ_0. An illustration with a toy example is given in the supplementary material.

3.2 | Two-sided alternatives

We now consider testing the sharp null H_0^θ against a two-sided alternative hypotheses

$$H_1^{\theta\pm}: Y_i(1) - Y_i(0) = \eta \ (\neq \theta), \quad \text{for all } i = 1, \ldots, N. \tag{12}$$

Definition 4 The p-value function for testing H_0^θ against alternatives $H_1^{\theta\pm}$ is

$$p^L(\mathbf{D}^{\text{obs}}, \theta) = 2\min\left\{p^{L+}(\mathbf{D}^{\text{obs}}, \theta), p^{L-}(\mathbf{D}^{\text{obs}}, \theta)\right\}, \tag{13}$$

where $p^{L+}(\mathbf{D}^{\text{obs}}, \theta)$ and $p^{L-}(\mathbf{D}^{\text{obs}}, \theta)$ are defined in Equations (6) and (9), respectively.

The function $p^L(\mathbf{D}^{\text{obs}}, \theta)$ can be considered a discrete version of a CV function. By part (a) of Proposition 1, $p^{L+}(\mathbf{D}^{\text{obs}}, \theta)$ and $p^{L-}(\mathbf{D}^{\text{obs}}, \theta)$ stochastically dominate a Uniform[0, 1] random variable

when $\theta = \theta_0$ and thus is a valid p-value function to test H_0^{θ} against $H_1^{\theta_\pm}$. Note that, if the p-value function for this two-sided testing problem had been constructed along the lines of the CV function introduced in Section 2.2 as

$$2\min\{p^{L+}(\mathbf{D}^{\text{obs}}, \theta),\ 1 - p^{L+}(\mathbf{D}^{\text{obs}}, \theta)\} = 2\min\left\{p^{L+}(\mathbf{D}^{\text{obs}}, \theta),\ p^{U-}(\mathbf{D}^{\text{obs}}, \theta)\right\},$$

then it would not have dominated a Uniform[0, 1] random variable by part (b) of Proposition 1.

Figure 1 illustrates a $p^L(\mathbf{D}^{\text{obs}}, \theta)$ function based on the toy example in the supplementary materials.

4 | INVERTING THE FRT TO OBTAIN CONFIDENCE INTERVALS

As briefly mentioned in the introductory section, the procedure of inverting FRTs to obtain intervals for treatment effects has been described rather loosely in literature as one obtained by 'inversion' of the p-value function. In Section 3, we have defined five p-value functions, but the inversion procedure leading to construction of valid confidence intervals is not obvious from these definitions. Furthermore, the definitions and results stated so far do not guarantee monotonicity of the p-value functions (see Section 3 of the Supplementary materials for an example). Non-monotonic p-value functions will not produce confidence intervals for the treatment effect at all levels of significance. In this section, we explore conditions that guarantee monotonicity of p-value functions, and then provide a concrete algorithm for constructing valid confidence intervals.

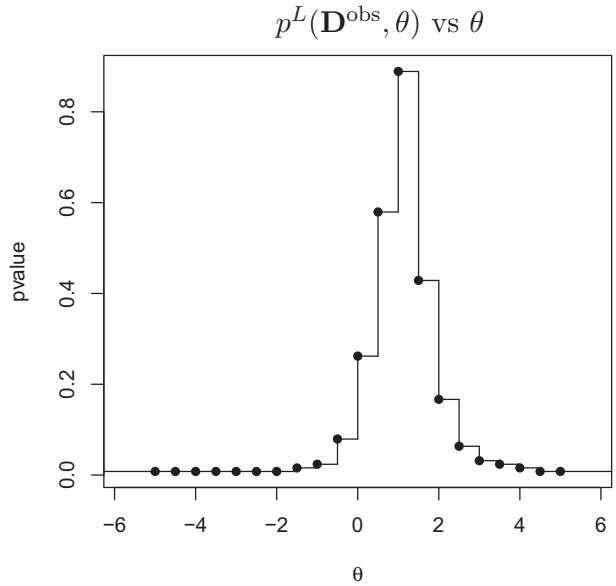

FIGURE 1 $p^L(\mathbf{D}^{\text{obs}}, \theta)$ vs. θ

4.1 | Monotonicity of the p-value functions

We now aim at providing a set of sufficient conditions to guarantee that $p^{L+}(\mathbf{D}^{\text{obs}}, \theta)$ 'behaves' like a CDF in the sense that is monotonically non-decreasing and right continuous. We first introduce the following definitions along the lines of Caughey et al. (2017).

Definition 5 (Ordered vectors of potential outcomes)Two vectors of potential outcomes under treatment $\mathbf{Y}(1) = \big(Y_1(1), \ldots, Y_N(1)\big)$ and $\mathbf{Y}'(1) = \big(Y'_1(1), \ldots, Y'_N(1)\big)$ are ordered as $\mathbf{Y}(1) \leq \mathbf{Y}'(1)$ if $Y_i(1) \leq Y'_i(1)$ for all $i = 1, \ldots, N$. An order between two vectors of potential outcomes under control $\mathbf{Y}(0)$ and $\mathbf{Y}'(0)$ is similarly defined.

Caughey et al. (2017) introduced the notion of an 'effect increasing' (EI) statistic in the context of testing null hypotheses that are weaker than the sharp null. A definition of an EI test statistic is given below.

Definition 6 (Effect increasing (EI) test statistic)A test statistic $T(\mathbf{Y}, \mathbf{W}) = T(\mathbf{Y}(1), \mathbf{Y}(0), \mathbf{W})$ is said to possess the EI property if it is non-decreasing in $\mathbf{Y}(1)$ and non-increasing in $\mathbf{Y}(0)$.

Examples of EI statistics include difference in means or Wilcoxon rank sum statistic. On the other hand, the commonly used Studentized Fisher–Behren-type statistic in the example given in Section 3 of the supplementary material does not satisfy the EI property. Caughey et al. (2017) pointed out the important role of EI statistics in constructing valid tests for null hypothesis that are weaker than the sharp null. Theorem 1 stated below relates the test statistic to the properties of the p-value functions, and establishes that the EI condition is sufficient for monotonicity of p-value functions in FRT.

Theorem 1

1. If the test statistic T is EI, then the p-value function $p^{L+}(\mathbf{D}^{\text{obs}}, \theta)$ defined in Equation (6) is non-decreasing in θ for fixed $\mathbf{D}^{\text{obs}}$.
2. For fixed $\mathbf{W}$, if $T(\mathbf{Y}_\theta^{\text{imp}}, \mathbf{W})$ is right continuous as a function of θ, then $p^{L+}(\mathbf{D}^{\text{obs}}, \theta)$ is right continuous in θ.
3. Furthermore, for fixed $\tilde{\mathbf{W}} \neq \mathbf{W}^{\text{obs}}$, if $T(\mathbf{Y}_\theta^{\text{imp}}, \tilde{\mathbf{W}})$ approaches $-\infty$ and $+\infty$ as $\theta \to -\infty$ and $\theta \to +\infty$, respectively, then $p^{L+}(\mathbf{D}^{\text{obs}}, \theta) \to 1$ as $\theta \to \infty$ and $p^{L+}(\mathbf{D}^{\text{obs}}, \theta) \to P(\mathbf{W} = \mathbf{W}^{\text{obs}})$ as $\theta \to -\infty$.

Similar results also hold for $p^{L-}(\mathbf{D}^{\text{obs}}, \theta)$ defined in Equation (9), which is non-increasing if T is EI.

4.2 | An Algorithm for generating confidence intervals with coverage at least $1 - \alpha$

From the foregoing discussion, it is clear that the 'traditional' approach of inverting just *one* p-value function based on an *arbitrary* test statistic does not yield one or two-sided intervals with the desired coverage. Based on (i) the properties of the p-value functions in Proposition 1, (ii) the description of valid procedures for testing the sharp null against one- or two-sided hypotheses in Section 3, and (iii) the conditions required to guarantee that inversion of p-value functions will generate intervals as stated in Theorem 1,

we now arrive at the following proposition that provides a rule to generate confidence intervals with the desired coverage.

Proposition 2 *Assume that for fixed* $\mathbf{D}^{\text{obs}}$*, the p-value functions* $p^{L+}(\mathbf{D}^{\text{obs}}, \theta)$ *and* $p^{L-}(\mathbf{D}^{\text{obs}}, \theta)$ *are (i), respectively, non-decreasing and non-increasing and (ii) right continuous functions of* θ*.*

1. Define $\theta_\ell(\alpha) = \sup_\theta \{\theta : p^{L+}(\mathbf{D}^{\text{obs}}, \theta) \leq \alpha\}$. Then the one-sided interval $\left[\theta_\ell(\alpha), \infty\right)$ covers the true value of θ with probability of at least $1 - \alpha$.
2. Define $\theta_u(\alpha) = \inf_\theta \{\theta : p^{L-}(\mathbf{D}^{\text{obs}}, \theta) \leq \alpha\}$. Then the one-sided interval $\left(-\infty, \theta_u(\alpha)\right)$ covers the true value of θ with probability of at least $1 - \alpha$.
3. For $0 < \alpha_1, \alpha_2 < 1$ and $\alpha_1 + \alpha_2 = \alpha$, define $\theta_\ell(\alpha_1) = \sup_\theta \{\theta : p^{L+}(\mathbf{D}^{\text{obs}}, \theta) \leq \alpha_1\}$ and $\theta_u(\alpha_2) = \inf_\theta \{\theta : p^{L-}(\mathbf{D}^{\text{obs}}, \theta) \leq \alpha_2\}$. Then the two-sided interval $[\theta_\ell(\alpha_1), \theta_u(\alpha_2))$ covers the true value of θ with probability of at least $1 - \alpha$.

Proposition 2 provides methods to construct confidence intervals for the treatment effect with the desired coverage. The most straightforward approach is to obtain the interval $[\theta_\ell(\alpha/2),\theta_u(\alpha/2))$ where $\theta_\ell(\alpha/2)$ and $\theta_u(\alpha/2)$ are obtained by substituting $\alpha_1 = \alpha_2 = \alpha/2$ in part (c) of Proposition 2 and solving the equations $p^{L+}(\mathbf{D}^{\text{obs}}, \theta) = \alpha/2$, $p^{L-}(\mathbf{D}^{\text{obs}}, \theta) = \alpha/2$, which are equivalent to solving $p^{L}(\mathbf{D}^{\text{obs}}, \theta) = \alpha$. Because $p^{L+}(\mathbf{D}^{\text{obs}}, \theta)$ and $p^{L-}(\mathbf{D}^{\text{obs}}, \theta)$ are stepwise functions that are monotonic in θ, we propose Algorithm 1 that is based on the conventional bisection method to find the left endpoint $\theta_\ell(\alpha/2)$ of the interval. The right endpoint $\theta_u(\alpha/2)$ can be obtained similarly.

Remark 1 The final θ_l and θ_r in Algorithm 1 still satisfy $p^{L+}(\mathbf{D}^{\text{obs}}, \theta_l) \leq \alpha/2$ and $p^{L+}(\mathbf{D}^{\text{obs}}, \theta_r) > \alpha/2$. These inequalities, the monotonicity of $p^{L+}(\mathbf{D}^{\text{obs}}, \theta)$ and definition of $\theta_\ell(\alpha/2)$ collectively imply that $\theta_l \leq \theta_\ell(\alpha/2) \leq \theta_r$ and consequently $\theta_\ell(\alpha/2) - \varepsilon < \theta_l \leq \theta_\ell(\alpha/2)$. It should be stressed that such a result is independent of the choice of initial input $[\theta_l^*, \theta_r^*]$.

Algorithm 1 Bisection method for estimating $\theta_\ell(\alpha/2)$

Inputs: An initial interval $[\theta_l^*, \theta_r^*]$ with $p^{L+}(\mathbf{D}^{\text{obs}}, \theta_l^*) \leq \alpha/2$ and $p^{L+}(\mathbf{D}^{\text{obs}}, \theta_r^*) > \alpha/2$; a specified error level $\epsilon > 0$

Procedure:

1. Fix $\theta_l \leftarrow \theta_l^*$ and $\theta_u \leftarrow \theta_u^*$.
2. Calculate the p-value function $p^{L+}(\mathbf{D}^{\text{obs}}, \theta)$ at $\theta = \frac{\theta_l+\theta_r}{2}$;
3. If $p^{L+}(\mathbf{D}^{\text{obs}}, \frac{\theta_l+\theta_r}{2}) \leq \alpha/2$, then $\theta_l \leftarrow \frac{\theta_l+\theta_r}{2}$. Otherwise, $\theta_r \leftarrow \frac{\theta_l+\theta_r}{2}$;
4. Repeat steps 1, 2 and 3 until $\theta_r - \theta_l < \epsilon$.

Output: θ_l.

Remark 2 All existing procedures (e.g. Dasgupta et al., 2015) of obtaining confidence intervals by inverting FRTs essentially involve a grid-search procedure that entails (i) choosing a sequence of parameter values $\{\theta_1, \theta_2, \ldots\}$, (ii) calculating $\widehat{p}(\theta_j)$, an estimator of the p-value function $p(\theta_j)$ at $\theta = \theta_j$ by testing the sharp null H_0:$\theta = \theta_j$, $j = 1, \ldots$, (iii) fitting a p-value function using the $\left(\theta_j, \widehat{p}(\theta_j)\right)$ pairs and (iv) inverting the fitted function to obtain the confidence interval for θ.

However, the interval obtained by inverting the p-value function fitted with these chosen parameter values depends on the choice of the sequence $\{\theta_1, \theta_2, \ldots\}$. The algorithm proposed above is more robust compared to the traditional approach because it does not depend on the choice of discrete levels of θ.

Remark 3 Although the proposed algorithm can be more robust compared to a grid search, calculating $p^{L+}(\mathbf{D}^{\text{obs}}, \theta)$ for several values of θ while searching for the lower and upper limits is still computationally challenging even when Monte Carlo estimates (obtained by randomly drawing assignment vectors $\mathbf{W}$) are used to estimate the p-values. However, the computational load can be considerably reduced if we can estimate the *entire* p-value function using a *single* Monte Carlo sample. Thus we arrive at an important question that connects statistical and computational efficiency: how to efficiently estimate the entire p-value function for an infinite number of parameter values in a computationally viable manner? We address this important question in the following section.

5 | EFFICIENT ESTIMATION OF THE p-VALUE FUNCTION: COMPUTATIONAL VIABILITY AND THEORETICAL GUARANTEE

The p-value functions defined in Sections 3.1 and 3.2 can be computed for any given value of θ if all possible realizations $\mathbf{W}^{\text{rep}}$ of the assignment vector $\mathbf{W}$ can be obtained and used to generate the exact randomization distribution of the test statistic $T(\mathbf{Y}_\theta^{\text{imp}}, \mathbf{W}^{\text{rep}})$. However, even for a moderate population size the total number of possible realizations of $\mathbf{W}$ is typically computationally prohibitive. The common solution to this problem is to draw, repeatedly and independently, randomized treatment assignment vectors $\mathbf{W}_1^{\text{rep}}, \ldots, \mathbf{W}_K^{\text{rep}}$, and obtain a Monte Carlo estimate of the p-value function based on the values of the test statistic computed from these K draws. Consider specifically the estimation of $p^{L+}(\mathbf{D}^{\text{obs}}, \theta)$ defined in Equation (6). The Monte Carlo estimator of $p^{L+}(\mathbf{D}^{\text{obs}}, \theta)$ based on K draws is given by

$$\widehat{p}_K^{L+}(\mathbf{D}^{\text{obs}}, \theta) = \frac{1}{K}\sum_{k=1}^{K} \mathbb{I}\left(T(\mathbf{Y}_\theta^{\text{imp}}, \mathbf{W}_k^{\text{rep}}) \geq T(\mathbf{D}^{\text{obs}})\right), \tag{14}$$

where $\mathbb{I}(A)$ is the indicator function for event A. All other p-value functions can be estimated similarly. Although this estimator has long been used since the times of Fisher, the effect of the Monte Carlo sample size K on the accuracy of the estimator $\widehat{p}^{L+}(\mathbf{D}^{\text{obs}}, \theta)$ for fixed θ remains largely unexplored. The current problem is even more challenging, because the goal is to estimate the entire p-value function, and not only values at specific values of θ. As mentioned in Remark 3, the computational viability of our proposed algorithm hinges crucially on the ability to estimate the p-value function using *only one* Monte Carlo sample. The question is, how large should such a sample be to ensure that estimated p-value function has a desired level of precision. Below, we provide a new result in the form of a concentration inequality to shed light on this question.

Theorem 2 *Let K denote the size of the Monte Carlo sample drawn from the distribution of $\mathbf{W}$ and let $\widehat{p}_K^{L+}(\mathbf{D}^{\text{obs}}, \theta)$ be as defined in Equation (14), where the underlying test statistic T satisfies the conditions in parts (1) and (2) of Theorem 1, that is, it is EI and a right continuous function of θ for fixed $\mathbf{W}$. Fix $\varepsilon > 0$. Then,*

$$P\left(\sup_{\theta}\left|\widehat{p}_K^{L+}(\mathbf{D}^{\text{obs}},\theta)-p^{L+}(\mathbf{D}^{\text{obs}},\theta)\right|>\epsilon\right)\leq\min\left\{1,\,4e^{-\frac{K\epsilon^2}{8}}\right\}. \tag{15}$$

It is important to note that the bound (15) does not depend on N, making it particularly useful for cases when the total number of possible assignments M is large. In our numerical studies in Section 7, for experiments in which $M \leq 10{,}000$, we have used a complete enumeration of all M assignments to compute the p-value function. If $M > 10{,}000$, 10000 Monte Carlo draws have been used to estimate the p-value functions. Such a choice would estimate the p-value functions without error if $M \leq 10{,}000$. When $M > 10{,}000$, Theorem 2 guarantees that the maximum probability of making a 5% estimation error does not exceed 0.18, and that of making a 10% estimation error is bounded above by approximately 1.5×10^{-5}. This guarantee holds however large M might be.

6 | CONFIDENCE DISTRIBUTION AS A TOOL FOR COMBINING FISHER RANDOMIZATION TESTS FROM INDEPENDENT STUDIES

Studies with a large number of experimental units now frequently arise from aggregation of information from multiple independent sources (e.g. Hemkens et al., 2017) and require strategies for efficient meta-analysis. Several researchers (e.g. Bareinboim & Pearl, 2016; Liu et al., 2020) have emphasized on the importance of development of new methodologies for combining information from multiple sources, stating that the objective of such fusion inference is 'to combine results from many experimental and observational studies, each conducted on a different population and under a different set of conditions in order to synthesize an aggregate measure of targeted effect size that is *better*, in some sense, than any one study in isolation'. In this section, we use CD to develop efficient and robust approaches to synthesizing an aggregate measure of effect size from difference sources.

There exist several classical methods in literature to combine p-values from independent tests of hypotheses, for example, Fisher's method (Fisher, 1932) and Stouffer's method (Stouffer et al., 1949). See Marden (1991) for a detailed review of these and other methods. However, while it is straightforward to combine p-values from multiple independent tests, it is not obvious how to combine the results into a *composite p-value function* from which a composite interval estimator for θ can be obtained. Singh et al. (2005) and Xie et al. (2011) proposed a general approach to combine CDs, and specifically p-value functions, that encompass all the classical methods for combining p-values as special cases. We describe their approach before explaining why it cannot be directly adopted to combine FRTs.

Xie et al. (2011) combined a sequence of CDs $H_1(\theta), \ldots, H_m(\theta)$ to obtain a combined CD

$$H_c(\theta) = G_c\left(g_c\left(H_1(\theta), \ldots, H_m(\theta)\right)\right), \tag{16}$$

where $g_c : [0,1]^m \to \mathbb{R}$ is a continuous function that is non-decreasing in each coordinate, $G_c : \mathbb{R} \to [0,1]$ is the continuous CDF of $g_c(U_1, \ldots, U_m)$ where $U_1, \ldots, U_m$ are independently and identically distributed (iid) Uniform[0, 1] random variables. This combined function $H_c(\theta)$, is non-decreasing, right continuous and a lower CD as per Definition 1. Consequently, it is used to obtain the left endpoint of a $100(1-\alpha)\%$ CI for θ_0, the true value of θ. To obtain the right endpoint of a $(1-\alpha)\%$ CI, the function $1 - H_c(\theta)$ is induced. This function is also a lower CD, right continuous, but non-increasing in θ. These three properties of this induced function $1 - H_c(\theta)$ are guaranteed by two conditions: (i) For $i = 1, \ldots, m$ all $H_i(\theta)$, are CDs and (ii) $G_c(t)$ is a continuous function of t.

In the context of FRT, while attempting to combine p-value functions from m independent experiments, we can combine the m lower CDs $p_i^{L+}(\theta) = p_i^{L+}(\mathbf{D}_i^{\text{obs}}, \theta)$, $i = 1, \ldots, m$, defined by Equation (6) obtained from these experiments in a manner similar to Equation (16) to obtain a combined lower CD function $p_c^{L+}(\theta)$ and use it to generate the lower endpoint of a confidence interval. However, this combination does not automatically generate the counterpart of $1 - H_c(\theta)$ because $1 - p_c^{L+}(\theta)$ is not a lower CD.

Thus we face the following theoretical questions: (a) Which p-value functions (e.g. $p^{L-}(\cdot)$ or $p^{U+}(\cdot)$) should be combined to mimic $1 - H_c(\cdot)$ of Xie et al. (2011) and (b) What conditions are needed to guarantee that the combined function is a lower CD, right continuous and non-increasing in θ. Proposition 3 provides an answer to these questions and also gives rise to an unambiguous procedure for combining results from m independent experiments.

Proposition 3 *For* i = *1, …, m, let* $p_i^{L+}(\theta) = p_i^{L+}(\mathbf{D}_i^{\text{obs}}, \theta)$ *defined by Equation (6) and* $p_i^{L-}(\theta) = p_i^{L-}(\mathbf{D}_i^{\text{obs}}, \theta)$ *defined by Equation (9) denote one-sided p-value functions obtained from m independent randomized experiments. Define the combined p-value functions:*

$$p_c^{L+}(\theta) = G_c\left(g_c\left(p_1^{L+}(\theta), \ldots, p_m^{L+}(\theta)\right)\right), \quad p_c^{L-}(\theta) = G_c\left(g_c\left(p_1^{L-}(\theta), \ldots, p_m^{L-}(\theta)\right)\right), \tag{17}$$

where $g_c : [0,1]^m \to \mathbb{R}$ is a continuous function that is non-decreasing in each coordinate, $G_c : \mathbb{R} \to [0, 1]$ is the CDF of $g_c(U_1, \ldots, U_m)$ where $U_1, \ldots, U_m$ are iid Uniform[0, 1] random variables. Then the combined p-value functions $p_c^{L+}(\theta)$ and $p_c^{L-}(\theta)$ are both lower CDs as per Definition 1.

Remark 4 The proof of the result that $H_c(\theta)$ in Xie et al. (2011) is a CD and thus $1-H_c(\theta)$ is a lower CD relies on the continuity of the function $G_c(\cdot)$. This condition is not necessary in Proposition 3. Thus, although the discrete nature of the p-value functions in FRT entails combining two different sets of functions, ultimately they can be used to generate valid confidence intervals under conditions weaker than those in Xie et al. (2011).

As a consequence of Proposition 3, the combined p-value functions $p_c^{L+}(\theta)$ and $p_c^{L-}(\theta)$ can be inverted to generate CIs for θ with guaranteed coverages. For $0 < \alpha_1, \alpha_2, \alpha < 1$ and $\alpha_1 + \alpha_2 = \alpha$, define $\theta_{\ell,c} = \sup_\theta \{\theta : p_c^{L+}(\theta) \le \alpha_1\}$ and $\theta_{u,c} = \inf_\theta \{\theta : p_c^{L-}(\theta) \le \alpha_2\}$. Then arguing along lines similar to that in part (3) of Proposition 2, the interval $[\theta_{\ell,c}, \theta_{u,c})$ is a $100(1 - \alpha)\%$ interval for θ obtained by combining the m studies.

To implement the steps described above, we need to choose specific forms of the function $g_c(\cdot)$. Xie et al. (2011) showed that the form $g_c(u_1, \ldots, u_m) = \sum_{i=1}^m w_i F_0^{-1}(u_i)$, where $F_0(\cdot)$ is a CDF of a random variable X, F_0^{-1} refers to the quantile function associate with X, that is, for $p \in [0, 1]$, $F_0^{-1}(p) = \min\{x \in \mathbb{R} : p \le F_X(x)\}$ and $w_1, \ldots, w_m$ are non-negative weights with at least one $w_i \neq 0$, generates most classical methods for combining p-values. Two examples are given below.

1. With $w_i = 1$ for all $i = 1, \ldots, m$ and negative exponential CDF $F_0(x) = e^x$ for $x \le 0$ generates Fisher's method, in which

$$p_c^{L+}(\theta) = P\left[\chi_{2m}^2 \ge -2\sum_{i=1}^m \log\left(p_i^{L+}(\theta)\right)\right], \quad p_c^{L-}(\theta) = P\left[\chi_{2m}^2 \ge -2\sum_{i=1}^m \log\left(p_i^{L-}(\theta)\right)\right]. \tag{18}$$

2. Again taking $w_i = 1$ for all $i = 1, \ldots, m$ and $F_0(x) = \frac{1}{2}e^x \mathbb{I}_{(x\le 0)} + (1 - \frac{1}{2}e^{-x})\mathbb{I}_{(x>0)}$, that is, the double exponential or Laplace CDF instead of the negative exponential CDF leads to the double exponential (DE) method for combining p-values.

The proposed approach for combining FRT-based inference from independent randomized experiments will be demonstrated in Section 7 using Fisher's and the DE methods described above. A theoretical comparison of the two methods performed by Singh et al. (2005) in the context of combining CDs established the superiority of the DE method over Fisher's method in terms of Bahadur efficiency. Because the p-value functions considered here are lower CDs, and not CDs, such superiority of the DE method, while intuitive, is not immediate. However, an empirical comparison of the two methods in terms of width of the generated confidence intervals (Section 7) suggests a similar phenomenon when combining p-value functions from randomized experiments.

Our simulations also suggest that taking equal weights $w_i = 1$ is not necessarily a good strategy for combining experiments, especially when the experiments have highly unbalanced sample sizes. Some empirical investigation along these lines is performed in Section 7. The results are interesting and open up possibilities of further theoretical investigations.

Remark 5 An intuitive approach for combining results of FRT from different studies is to treat them as a unique experiment, where each individual experiment constitutes a block (or a group of blocks) and the joint assignment mechanism follows the joint distribution assembled from the individual mechanisms. For this new block design experiment, we define the one-sided p-value functions as

$$p_{IT}^{L-}(\theta) = P\left(\sum_{i=1}^{m} w_i T_i(\mathbf{Y}_{\theta,i}^{\text{imp}}, \mathbf{W}_i^{\text{rep}}) \leq \sum_{i=1}^{m} w_i T_i(\mathbf{D}_i^{\text{obs}}) \right), \tag{19}$$

$$p_{IT}^{L+}(\theta) = P\left(\sum_{i=1}^{m} w_i T_i(\mathbf{Y}_{\theta,i}^{\text{imp}}, \mathbf{W}_i^{\text{rep}}) \geq \sum_{i=1}^{m} w_i T_i(\mathbf{D}_i^{\text{obs}}) \right), \tag{20}$$

where $w_1, \ldots, w_m$ the weights for the m blocks and $\sum_{i=1}^{m} w_i T_i(\mathbf{Y}_{\theta,i}^{\text{imp}}, \mathbf{W}_i^{\text{rep}})$ represents the combined test statistic. Applying Algorithm 1 to $p_{IT}^{L-}(\theta)$ and $p_{IT}^{L+}(\theta)$, we can obtain a valid confidence interval for the treatment effect. We will refer to this approach as the 'block-randomization inspired (BRI) approach'. Our preliminary empirical investigation (reported in Section 5.1 of the supplementary material) shows that the proposed approach described above is superior to the BRI approach when combining a large number of small size experiments. A more comprehensive comparison between the two approaches is left as future research. In the following remark, we define a generalization of the p-value combination approach that is closely related but superior to the BRI approach in terms of width of generated confidence intervals.

Remark 6 Two different g_c functions, say g_c^+ and g_c^-, can be employed in Equation (17) to define the p-value functions p_c^{L+} and p_c^{L-}. The functions g_c's can also be θ-dependent, and their continuity requirements can be dropped as well. The relaxation on g_c can further increase the flexibility and expand the reach of the proposed combined p-value framework. However, it may increase computational complexity and thus may not be preferred in practice. One such example is to take $g_{\theta,c}^+$ and $g_{\theta,c}^-$ in Equation (17) with

$$g_{\theta,c}^+(u_1, \ldots, u_m) = \sum_{i=1}^{m} w_i (F_{\theta,i}^+)^{-1}(u_i) \quad \text{and} \quad g_{\theta,c}^-(u_1, \ldots, u_m) = \sum_{i=1}^{m} w_i (F_{\theta,i}^-)^{-1}(u_i), \tag{21}$$

where for $i = 1, \ldots, m$, $(F^{+}_{\theta,i})^{-1}$ and $(F^{-}_{\theta,i})^{-1}$ are quantile functions of the random variables with CDFs $F^{+}_{\theta,i}(t) = P\left(-T_i(\mathbf{Y}^{\text{imp}}_{\theta,i}, \mathbf{W}^{\text{rep}}_{i}) \leq t\right)$ and $F^{-}_{\theta,i}(t) = P\left(T_i(\mathbf{Y}^{\text{imp}}_{\theta,i}, \mathbf{W}^{\text{rep}}_{i}) \leq t\right)$, respectively. We refer to this method of combining p-value functions using these choices of $g^{+}_{\theta,c}$ and $g^{-}_{\theta,c}$ as in Equation (21) as the 'Parameter-dependent double g_c' (PDD-g_c) approach. The PDD-g_c approach is closely related to the BRI approach. In fact, if the functions $F^{+}(\cdot)$ and $F^{-}(\cdot)$ in Equation (21) are continuous and monotonic, so that $(F^{+}(\cdot))^{-1}$ and $(F^{-}(\cdot))^{-1}$ represent inverse functions, then the PDD-g_c and BRI can be shown to be exactly equivalent along the lines of argument in Xie et al. (2011). In the case of FRT, while the two approaches are not exactly similar due to discreteness of $F^{+}(\cdot)$ and $F^{-}(\cdot)$, discreteness turns out to be advantage for the PDD-g_c approach. We prove (Section 5.2 of Supplementary materials) that confidence intervals generated by inverting the combined p-value function obtained through the PDD-g_c approach cannot be wider than those generated by inverting the p-value function obtained through the BRI approach.

7 | SIMULATIONS

In this section, we conduct simulations to establish that our proposed guidelines and algorithms for (a) estimating the p-value functions, (b) inverting them to obtain confidence intervals and (c) combining inferences across multiple independent experiments to produce the desired results. We consider two types of randomized experiments: the completely randomized design (CRD) and the randomized block design (RBD). In the former, an even number N of experimental units are equally split into treatment (denoted by 1) and control (denoted by 0) groups at random. In the latter, we consider b blocks of experimental units with an equal even number (k) of units in each block (block size), so that $N = bk$ is the total number of units. The k units within each block are equally split into treatment and control groups at random. Note that $b = 1$ for an RBD is equivalent to a CRD.

We consider several scenarios shown in Table 1, in each of which we consider combining results from two experiments with design parameters (b_1, k_1) and (b_2, k_2), where for $j = 1, 2$, b_j and k_j denote the number of blocks and the block size, respectively. The two individual experiments are either CRD or RBD.

For each individual experiment across all scenarios, the potential outcomes under control, $Y_i(0)$, $i = 1, \ldots, N$ are generated from a lognormal distribution with parameters 0 and 1. The true additive effect is assumed to be zero, so that $Y_i(1) = Y_i(0)$ for $i = 1, \ldots, N$. Potential outcomes once generated are kept fixed. The units are assigned to treatments in a manner described earlier, depending on whether the design is CRD or RBD.

Next, for each experiment, FRT is conducted using the difference of averages between treatment and control groups as the test statistic. Either a complete enumeration all M assignments for $M \leq 10{,}000$, or a set of 10,000 random permutations when $M > 10{,}000$, is used to calculate or estimate the p-value functions $p^{L+}(\mathbf{D}^{\text{obs}}, \theta)$ and $p^{L-}(\mathbf{D}^{\text{obs}}, \theta)$. A justification for this choice of K was provided in the last paragraph of Section 5. For each individual experiment, 95% confidence intervals are obtained using the method described in Algorithm 1 with $\alpha_1 = \alpha_2 = 0.025$. Finally, the p-value functions from the two experiments in each scenario are combined using Fisher's method given by Equation (18) and the double exponential (DE) method, and the 95% confidence intervals are generated using the combined p-value functions $p^{L+}_{c}(\theta)$ and $p^{L-}_{c}(\theta)$, again using Algorithm 1.

The simulation for each scenario is repeated 5000 times to calculate the coverage of the 95% intervals generated from the individual and combined experiments. The results are shown in Table 1. The simulations provide empirical evidence of the theoretical result that the proposed method for inverting FRT to obtain confidence intervals produces intervals with the desired coverage for individual as well as combined experiments.

TABLE 1 Coverage of 95% CIs obtained using Fisher's and DE p-value combination methods

Scenario					Coverage			
Designs 1 & 2	b_1	k_1	b_2	k_2	**Exp 1**	**Exp 2**	**Fisher**	**DE**
CRD & CRD	1	10	1	10	0.954	0.952	0.961	0.961
	1	16	1	16	0.951	0.949	0.952	0.953
	1	24	1	24	0.956	0.947	0.954	0.953
	1	30	1	30	0.946	0.944	0.948	0.947
	1	10	1	16	0.952	0.950	0.956	0.955
	1	16	1	24	0.947	0.956	0.948	0.949
	1	24	1	30	0.954	0.951	0.952	0.952
	1	10	1	30	0.953	0.954	0.955	0.943
RBD & RBD	2	8	2	8	0.953	0.949	0.950	0.949
	10	2	10	2	0.948	0.949	0.951	0.947
	4	4	4	4	0.956	0.952	0.949	0.949
	2	8	4	4	0.947	0.950	0.952	0.952
	4	4	10	2	0.952	0.948	0.950	0.950
	2	10	10	2	0.949	0.953	0.950	0.950
CRD & RBD	1	10	10	2	0.950	0.950	0.959	0.958
	1	10	2	10	0.949	0.948	0.954	0.956
	1	16	2	10	0.950	0.951	0.952	0.951
	1	24	2	10	0.953	0.947	0.947	0.950
	1	30	10	2	0.944	0.952	0.951	0.952

Table 2 provides a summary of comparison of widths of confidence intervals of the individual experiments and the combined experiments. It is natural to expect that the width of the interval generated by combining the two individual experiments would be shorter than the width of the interval obtained from *each* individual experiment with a high probability, as such a fusion should increase the precision of inference. To check if simulation results are consistent with these expectations, we compute (a) the percentages of cases in which the widths of the intervals obtained from combined experiments (using Fisher's and the DE method) are shorter than the lengths of intervals obtained from individual experiments, and (b) the median width of intervals obtained from individual as well as combined experiments. To compare the performance of the two combining methods, we also compute the proportion of cases in which The DE method produces shorter intervals than Fisher's method.

The results suggest that, as expected, combining experiments using Fisher's or DE methods always results in reducing median width of confidence intervals. Further, in almost all settings, combining experiments using either method reduces the width of confidence intervals in a very high percentage (90–100%) of cases. Only in one situation, where the number of units in the two experiments vary the most (CRDs with 10 and 30 units), this percentage reduces to 0.77 for the Fisher method and 0.79 for the DE method. However, it was interesting to note that once Fisher's method and the DE method were slightly modified by taking weights w_i proportional to the sample sizes in the two experiments, the percentages again were higher and consistent with other settings.

TABLE 2 A comparison between widths l of CIs: l_{Exp1}, l_{Exp2}, l_{Fisher}, l_{DE} and $l_m = \min\{l_{Exp1}, l_{Exp2}\}$

Scenario					The percentage among 5000 repetitions			Medians of widths of CIs			
Designs 1 & 2	b_1	k_1	b_2	k_2	$l_{Fisher} < l_m$	$l_{DE} < l_m$	$l_{DE} < l_{Fisher}$	l_{Exp1}	l_{Exp2}	l_{Fisher}	l_{DE}
CRD & CRD	1	10	1	10	1	1	0.954	4.158	4.249	1.994	1.871
	1	16	1	16	1	1	0.896	2.724	3.644	2.186	2.163
	1	24	1	24	0.995	0.999	0.918	3.121	1.906	1.762	1.740
	1	30	1	30	1	1	0.951	2.622	2.392	1.895	1.872
	1	10	1	16	0.988	0.996	0.967	4.163	3.478	2.528	2.499
	1	16	1	24	1	1	0.919	2.723	2.474	1.760	1.740
	1	24	1	30	1	1	0.919	3.116	2.342	2.020	1.998
	1	10	1	30	0.772	0.789	0.969	4.158	2.015	1.700	1.680
RBD & RBD	2	8	2	8	1	1	0.881	2.793	3.799	2.242	2.218
	10	2	10	2	0.806	0.934	0.922	3.944	2.479	2.174	2.146
	4	4	4	4	0.999	0.999	0.892	2.924	4.091	2.331	2.307
	2	8	4	4	0.999	1	0.886	2.792	4.090	2.307	2.280
	4	4	10	2	1	1	0.872	2.921	3.801	2.257	2.235
	2	10	10	2	0.964	0.980	0.909	3.554	2.479	2.094	2.070
CRD & RBD	1	10	10	2	0.995	0.998	0.939	4.158	3.829	2.708	2.675
	1	10	2	10	0.918	0.932	0.967	4.160	3.005	2.314	2.291
	1	16	2	10	1	1	0.909	2.724	3.000	2.003	1.983
	1	24	2	10	0.964	1	0.504	3.117	2.233	1.926	1.923
	1	30	10	2	0.999	0.999	0.949	2.591	1.892	1.627	1.604

Looking at the percentage of cases in which the DE method produced shorter intervals than Fisher's method, and the median widths produced by the two methods, it is obvious that the former method performs uniformly better than the latter across all the settings. As explained in Section 6, this observation is consistent with the theoretical comparison of the two methods performed by Singh et al. (2005), although such a theoretical extension of such a comparison to our case involving discrete p-value functions is not immediate.

8 | REAL DATA EXAMPLE

Using data from a randomized experiment reported in Shadish et al. (2008), we demonstrate the proposed approach of estimating the p-value function and its usefulness for testing any sharp null hypothesis and obtaining a confidence interval for the average treatment effect. In this experiment, 235 undergraduate students from introductory psychology classes at a large mid-southern public university received either the treatment (vocabulary training) or control (math training) through a completely randomized treatment assignment. The number of students assigned to treatment and control were 116 and 119, respectively. The outcome was the vocabulary test score after the experiment.

The p-value functions $p^{L+}(\mathbf{D}^{\text{obs}}, \theta)$, $p^{L-}(\mathbf{D}^{\text{obs}}, \theta)$ and $p^{L}(\mathbf{D}^{\text{obs}}, \theta)$ based on the test statistic $T = \overline{Y}^{\text{obs}}(1) - \overline{Y}^{\text{obs}}(0)$ are shown in Figure 2. These p-value functions are estimated from Equation (14) by drawing a single Monte Carlo sample of size $K = 10^6$ from the distribution of the treatment assignment vector. Each draw in the sample is essentially a permutation of a binary vector consisting of 116 ones and 119 zeros.

We can test the sharp null hypothesis of no treatment effect on any student against the two-sided alternative. The p-value for such a test can be obtained from the function $p^{L}(\mathbf{D}^{\text{obs}}, \theta)$ for $\theta = 0$, and turns out to be zero, indicating presence of a treatment effect. Suppose one is interested in testing the sharp null hypothesis that the treatment effect is 6 versus the alternative that it is greater than 6. The p-value for such a test can be obtained from $p^{L+}(\mathbf{D}^{\text{obs}}, \theta)$, and is also zero providing strong evidence against the null hypothesis. Next, using Algorithm 1, we obtain 95% confidence intervals for the average treatment effect as [7.24, 10.72]. Note that the lower and upper limits can be individually obtained from the one-sided p-value functions $p^{L+}(\mathbf{D}^{\text{obs}}, \theta)$ and $p^{L-}(\mathbf{D}^{\text{obs}}, \theta)$, respectively, and they can also be obtained from the two-sided p-value function $p^{L}(\mathbf{D}^{\text{obs}}, \theta)$, as shown by the superimposed red dotted lines in all the three curves in Figure 2.

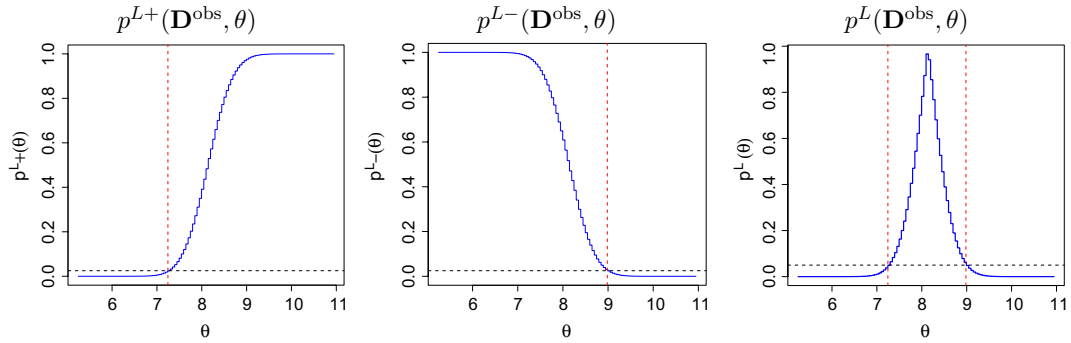

FIGURE 2 p-value functions for the Shadish experiment

9 | DISCUSSION AND FUTURE DIRECTIONS

In most scientific studies, assessing causal relationships among variables is considered more important with more practical implications than merely studying associations. The distinct difference between association and causality is now well understood: causality can only be determined by utilizing known or assumed knowledge about how the data were collected and consequently is more difficult to establish than association. However, technology has now created a perfect platform to design and analyse large studies conducted to assess causal effects of interventions and the FRT, with its unique ability to facilitate model-free assessment of causal effects is expected to have tremendous potential in modern day experiments. In this article, we attempt to address some long standing unclear aspects associated with the methodology for computation, inversion and principled aggregation of FRTs by developing a unified and comprehensive framework based on the versatile inferential tool confidence distribution.

One of the main criticisms of FRT has been the sharp null hypothesis, that many, including Neyman had considered overly strong, triggering the infamous Neyman–Fisher debate in 1935 (Sabbaghi & Rubin, 2014). However, the possibility of applying FRT to assess weaker null hypothesis has been explored and identified by a few researchers - see for example, Ding and Dasgupta (2018), Caughey et al. (2017), Wu and Ding (2019) and Cohen and Fogarty (2021). Caughey et al. (2017) showed that the interval estimators obtained by inverting FRT can be interpreted more meaningfully under a bounded null hypotheses if EI test statistics are used. Ding and Dasgupta (2018) derived a statistic that is asymptotically valid while testing Neyman's null hypothesis on average treatment effects. It will be interesting to extend our results to such weaker hypotheses and consequently have broader interpretations of the interval estimators.

We believe this article will open up a number of research possibilities. First, all our results pertain to finite samples. Exploring asymptotic properties of the interval estimators for individual and combined experiments using finite population asymptotics (Li & Ding, 2017) and borrowing relevant literature from CD literature will be a useful direction. Second, exploring ways to optimally combine experiments, as discussed in the last paragraph of Section 6 will be an interesting line of investigation. Third, extending the FRT-CD framework for analysis of data from observational studies and conducting sensitivity analysis is an interesting possibility. Finally, combining experiments and observational studies is an area of growing interest, and the FRT-CD may provide an excellent foundation for this area of research.

ACKNOWLEDGEMENTS

We are grateful to two reviewers, whose insightful comments resulted in significant improvement to this manuscript. This research was partially supported by National Science Foundation Grant Number DMS 1451817, DMS 1737857, DMS 1812048, DMS 2015373 and DMS 2027855.

ORCID

Tirthankar Dasgupta http://orcid.org/0000-0001-6590-6936

REFERENCES

Athey, S., Eckles, D. & Imbens, G.W. (2017) Exact *p*-values for network interference. *Journal of the American Statistical Association*, forthcoming.

Bareinboim, E. & Pearl, J. (2016) Causal inference and the data-fusion problem. *PNAS*, 113, 7345–7352.

Basse, G. & Feller, A. (2018) Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association*, 113, 41–55.

Basse, G., Feller, A. & Toulis, P. (2019) Randomization tests of causal effects under interference. *Biometrika*, 106(2), 487–494.

Birnbaum, A. (1961) Confidence curves: an omnibus technique for estimation and testing statistical hypotheses. *Journal of the American Satistical Association*, 56(294), 246–249.

Caughey, D., Dafoe, A. & Miratrix, L. (2017) Beyond the sharp null: randomization inference, bounded null hypotheses, and confidence intervals for maximum effects. Available from: https://arxiv.org/abs/1709.07339.

Cohen, P.L. & Fogarty, C.B. (2021) Gaussian prepivoting for finite population causal inference. Available from: https://arxiv.org/pdf/2002.06654.pdf.

Cox, D.R. (1958) Some problems connected with statistical inference. *The Annals of Mathematical Statistics*, 29, 357–372.

Dasgupta, T., Pillai, N.S. & Rubin, D.B. (2015) Causal inference from 2^K factorial designs by using potential outcomes. *Journal of the Royal Statistical Society Series B*, 77, 717–753.

Ding, P. (2017) A paradox from randomization-based causal inference (with discussion). *Statistical Science*, 32, 331–345.

Ding, P. & Dasgupta, T. (2018) A randomization-based perspective on analysis of variance: a test statistic robust with respect to treatment effect heterogeneity. *Biometrika*, 105, 45–56.

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika*, 80, 3–26.

Efron, B. (1998). R. A. Fisher in the 21st century. *Statistical Science*, 13, 95–122.

Fisher, R.A. (1932) *Statistical methods for research workers*, 4th edn. Edinburgh, Scotland: Oliver & Boyd.

Fisher, R.A. (1935) *The design of experiments*. Oxford: Oliver & Boyd.

Garthwaite, P.H. (1996) Confidence intervals from randomization tests. *Biometrics*, 52, 1387–1393.

Hemkens, L.G., Contopoulos-Ioannidis, D.G. & Ioannidis, J.P. (2017) Agreement of treatment effects for mortality from routinely collected data and subsequent randomized trials: metaepidemiological survey. *BMJ*, 352, i493.

Hennessy, J., Dasgupta, T., Miratrix, L., Pattanayak, C. & Sarkar, P. (2016) A conditional randomization test to account for covariate imbalance in randomized experiments. *Journal of Causal Inference*, 4, 61–80.

Li, X. & Ding, P. (2017) General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112, 1759–1769.

Liu, D., Liu, R.Y. & Xie, M. (2020). Nonparametric fusion learning: synthesize inferences from diverse sources using depth confidence distribution. *arXiv preprint* arXiv:2011.07047.

Marden, J.I. (1991). Sensitive and sturdy *p*-values. *The Annals of Statistics*, 19, 918–934.

Morgan, K.L. & Rubin, D.B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2), 1263–1282.

Neyman, J. (1923) On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4), 465–472. Translated by Dabrowska, DM and Speed, TP (1990).

Pitman, E.J.G. (1937) Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4, 119–130.

Rubin, D.B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.

Rubin, D.B. (1980) Comment on "Randomization analysis of experimental data: the Fisher randomization test" by D. Basu. *Journal of the American Statistical Association*, 75, 591–593.

Rubin, D.B. (1984) Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics*, 12, 1151–1172.

Sabbaghi, A. & Rubin, D.B. (2014) Comments on the Neyman-Fisher controversy and its consequences. *Statistical Science*, 29, 267–284.

Schweder, T. & Hjort, N.L. (2002). Confidence and likelihood. *Scandinavian Journal of Statistics*, 29, 309–332.

Schweder, T. & Hjort, N.L. (2003). Frequentist analogues of priors and posteriors. In *Econometrics and the Philosophy of Economics*, Ed.B.P. Stigum, pp. 285–317. Princeton, New Jersey: Princeton University Press.

Schweder, T. & Hjort, N. (2016) *Confidence, likelihood and probability*. Cambridge, UK: Cambridge University Press.

Shadish, W.R., Clark, M.H. & Steiner, P.M. (2008) Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association*, 103, 1334–1343.

Singh, K., Xie, M. & Strawderman, W.E. (2005) Combining information from independent sources through confidence distributions. *The Annals of Statistics*, 33, 159–183.

Stouffer, S.A., Suchman, E.A., DeVinney, L.C., Star, S.A. & Williams, R.M. Jr. (1949) *Adjustment During Army Life*, Princeton, NJ: Princeton University Press.

Wu, J. & Ding, P. (2019) Randomization tests for weak null hypotheses. *arXiv preprint*, arXiv:1809.07419.

Xie, M. & Singh, K. (2013) Confidence distribution, the frequentist distribution estimator of a parameter (with discussions). *International Statistical Review*, 81, 3–39.

Xie, M., Singh, K. & Strawderman, W.E. (2011) Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association*, 106(493), 320–333.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Luo X, Dasgupta T, Xie M. & Liu RY. Leveraging the Fisher randomization test using confidence distributions: Inference, combination and fusion learning. *J R Stat Soc Series B*. 2021;00:1–21. https://doi.org/10.1111/rssb.12429

Preflight Results

Document Information

Title: Leveraging the Fisher randomization test using confidence distributions: Inference, combination and fusion learning
Author: mxie
Creator: com.apple.Preview 944.6.16.1
Producer: macOS Version 10.14.2 (Build 18C54) Quartz PDFContext

Preflight Information

Profile: Convert to PDF/A-1b
Version: Qoppa jPDFPreflight v2021R1.00
Date: Aug 16, 2021 9:19:18 PM

Legend: (X) - Can NOT be fixed by PDF/A-1b conversion.
(!X) - Could be fixed by PDF/A-1b conversion. User chose to be warned in PDF/A settings.

Page 3 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file. - 7 more not displayed

Page 4 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 5 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 6 Results

Page 6 Results (contd.)

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file. - 3 more not displayed

Page 7 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 8 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 9 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 10 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 10 Results (contd.)

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 11 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 12 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 13 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file. - 24 more not displayed

Page 14 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 14 Results (contd.)

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file. - 30 more not displayed

Page 15 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file. - 1 more not displayed

Page 16 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 17 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 18 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 19 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Preflight Results

Document Information

Title: Leveraging the Fisher randomization test using confidence distributions: Inference, combination and fusion learning
Author: mxie
Creator: com.apple.Preview 944.6.16.1
Producer: macOS Version 10.14.2 (Build 18C54) Quartz PDFContext

Preflight Information

Profile: Convert to PDF/A-3b
Version: Qoppa jPDFPreflight v2021R1.00
Date: Aug 16, 2021 9:22:18 PM

Legend: (X) - Can NOT be fixed by PDF/A-3b conversion.
(!X) - Could be fixed by PDF/A-3b conversion. User chose to be warned in PDF/A settings.

Page 3 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file. - 7 more not displayed

Page 4 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 5 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 6 Results

Page 6 Results (contd.)

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file. - 3 more not displayed

Page 7 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 8 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 9 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 10 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 10 Results (contd.)

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 11 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 12 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 13 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file. - 24 more not displayed

Page 14 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 14 Results (contd.)

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file. - 30 more not displayed

Page 15 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file. - 1 more not displayed

Page 16 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 17 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 18 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.

Page 19 Results

(X) Font widths must be the same in both the font dictionary and the embedded font file.
(X) Font widths must be the same in both the font dictionary and the embedded font file.