Mis-classified, Binary, Endogenous Regressors: Identification and Inference*

Francis J. DiTraglia[1] and Camilo García-Jimeno[2,3]

[1]Department of Economics, University of Pennsylvania
[2]Institute for Quantitative Theory and Methods, Emory University
[3]NBER

October 4, 2018

Abstract

This paper studies identification and inference for the effect of a mis-classified, binary, endogenous regressor when a discrete-valued instrumental variable is available. We begin by showing that the only existing point identification result for this model is incorrect. We go on to derive the sharp identified set under mean independence assumptions for the instrument and measurement error. The resulting bounds are novel and informative, but fail to point identify the effect of interest. This motivates us to consider alternative and slightly stronger assumptions: we show that adding second and third moment independence assumptions suffices to identify the model. We then turn our attention to inference. We show that both our model, and related models from the literature that assume regressor exogeneity, suffer from weak identification when the effect of interest is small. To address this difficulty, we exploit the inequality restrictions that emerge from our derivation of the sharp identified set under mean independence only. These restrictions remain informative irrespective of the strength of identification. Combining these with the moment equalities that emerge from our identification result, we propose a robust inference procedure using tools from the moment inequality literature. Our method performs well in simulations.

Keywords: Instrumental variables, Measurement error, Endogeneity, Weak identification, Moment inequalities

JEL Codes: C10, C25, C26

*We thank Daron Acemoglu, Manuel Arellano, Kristy Buzard, Xu Cheng, Bernardo da Silveira, Bo Honoré, Arthur Lewbel, Chuck Manski, Sophocles Mavroeidis, Francesca Molinari, and Yuya Takahashi for valuable comments and suggestions. This document supersedes an earlier version entitled "On Mis-measured Binary Regressors: New Results and Some Comments on the Literature."

1 Introduction

Measurement error and endogeneity are pervasive features of economic data. Conveniently, a valid instrumental variable corrects for both problems when the measurement error is classical, i.e. uncorrelated with the true value of the regressor. Many regressors of interest in applied work, however, are binary and thus cannot be subject to classical measurement error.[1] When faced with non-classical measurement error, the instrumental variables estimator can be severely biased. In this paper, we study an additively separable model of the form

$$y = c(\mathbf{x}) + \beta(\mathbf{x})T^* + \varepsilon \tag{1}$$

where ε is a mean-zero error term, T^* is a binary, potentially endogenous regressor of interest, and $\mathbf{x}$ is a vector of exogenous controls.[2] We ask whether, and if so under what conditions, a discrete instrumental variable z suffices to non-parametrically identify the causal effect $\beta(\mathbf{x})$ of T^*, when we observe not T^* but a mis-classified binary surrogate T.

We proceed under the assumption of non-differential measurement error. This condition has been widely used in the existing literature and imposes that T provides no additional information beyond that contained in $(T^*, \mathbf{x})$. Even in this fairly standard setting, identification remains an open question: we begin by showing that the only existing identification result for this model is incorrect. We then go on to derive the sharp identified set under the standard first-moment assumptions from the related literature. We show that regardless of the number of values that z takes on, the model is not point identified. This motivates us to consider alternative, and slightly stronger assumptions. We show that, given a binary instrument, the addition of a second moment independence assumption suffices to identify a model with one-sided mis-classification. Adding a second moment restriction on the measurement error along with a third moment independence assumption for the instrument suffices to identify the model in general. This result likewise requires only a binary z.

We then turn our attention to inference, showing that both our model and related models from the literature suffer from a weak identification problem. In essence, binary misclassification creates a mixture model and to correct the bias in the instrumental variables estimator, we must estimate the mixing probabilities. But when $\beta(\mathbf{x})$ is small the "mixture modes" are nearly indistinguishable, making it impossible to reliably estimate these probabilities. To address this difficulty, we exploit the inequality moment restrictions that emerge from our derivation of the sharp identified set. These restrictions remain informative even when $\beta(\mathbf{x})$ is small or zero. Combining them with the moment equalities that emerge from our identification result, we propose an identification robust procedure for uniformly valid inference using tools from the moment inequality literature. Our procedure is computationally attractive and performs well in simulations. Moreover, it can be used both in our model and related models from the literature that assume an exogenous T^*.

Our work relates to a large literature that considers departures from classical measurement error, by allowing the measurement error to be related to the true value of the un-

[1] The only way to mis-classify a true one is downwards, as a zero, while the only way to mis-classify a true zero is upwards, as a one. This creates negative dependence between the truth and measurement error.

[2] Because T^* is binary, there is no loss of generality from writing the model in this form rather than the more familiar $y = h(T^*, \mathbf{x}) + \varepsilon$. Simply define $\beta(\mathbf{x}) = h(1, \mathbf{x}) - h(0, \mathbf{x})$ and $c(\mathbf{x}) = h(0, \mathbf{x})$.

observed regressor. Chen et al. (2005) obtain identification in a general class of moment condition models with mis-measured data by relying on the existence of an auxiliary dataset from which they can estimate the measurement error process. In contrast, Hu and Shennach (2008) and Song (2015) rely on an instrumental variable and an additional conditional location assumption on the measurement error distribution. More recently, Hu et al. (2015) use a continuous instrument to identify the ratio of partial effects of two continuous regressors, one measured with error, in a linear single index model. Unfortunately, these approaches cannot be applied to the case of a mis-measured binary regressor.

A number of papers have studied models with an exogenous binary regressor subject to non-differential measurement error. One group of papers asks what can be learned without recourse to an instrumental variable. An early contribution by Aigner (1973) characterizes the asymptotic bias of OLS in this setting, and proposes a correction using outside information on the mis-classification process. Related work by Bollinger (1996) provides partial identification bounds. More recently, Chen et al. (2008a) use higher moment assumptions to obtain identification in a linear model, and Chen et al. (2008b) extend these results to the non-parametric setting. van Hasselt and Bollinger (2012) and Bollinger and van Hasselt (2015) provide additional partial identification results. For results on the partial identification of discrete probability distributions under mis-classification, see Molinari (2008).

Continuing under the assumption of exogeneity and non-differential measurement error, another group of papers relies on the availability of either an instrumental variable or a second measure of T^*. Black et al. (2000) and Kane et al. (1999) consider a linear model and show that when *two* alternative measures T_1 and T_2 of T^* are available, a non-linear GMM estimator can be used to recover the effect of interest. Subsequently, Frazis and Loewenstein (2003) note that an instrumental variable can take the place of one of the measures. Mahajan (2006) extends the results of Black et al. (2000) and Kane et al. (1999) to a more general setting using a binary instrument in place of one of the treatment measures, establishing non-parametric identification of the conditional mean function. When T^* is in fact exogenous, this coincides with the causal effect. Hu (2008) derives related results when the mis-classified discrete regressor may take on more than two values. Lewbel (2007) provides an identification result for the same model as Mahajan (2006) under different assumptions. In particular, his "instrument-like variable" need not satisfy the usual exclusion restriction so long as it does not interact with T^* and takes on three or more values.

Much less is known about the case in which a binary, or discrete, regressor is not only mis-classified but endogenous. The first paper to provide a formal result for this case is Mahajan (2006). He extends his main result to the case of an endogenous treatment, providing an explicit proof of identification under the usual IV assumption in a model with additively separable errors. As we show below, however, this result is false.[3] Several more recent papers also consider the case of a mis-classified, endogenous, binary regressor. Kreider et al. (2012), partially identify the effects of food stamps on health outcomes of children under weak measurement error assumptions by relying on auxiliary data. Similarly, Battistin et al. (2014) study the returns to schooling in a setting with multiple mis-reported measures of educational qualifications. Unlike these two papers, our approach does not depend on the availability of auxiliary data. In a different vein, Shiu (2016) uses an exclusion restriction

[3] Appendix B provides a detailed explanation of the error in Mahajan's proof.

for the participation equation and an additional valid instrument to identify the effect of a discrete, mis-classified endogenous regressor in a semi-parametric selection model. Similarly, Nguimkeu et al. (2016) use exclusion restrictions for both the participation equation and measurement error equation to identify a parametric model with endogenous participation and one-sided endogenous mis-reporting. Unlike those of the preceding two papers, our results rely neither on parametric assumptions nor additional exclusion restrictions. Other than Mahajan (2006), the paper most closely related to our own is that of Ura (Forthcoming), who derives partial identification results for a local average treatment effect without the non-differential assumption. In contrast, we study an additively separable model under non-differential measurement error and derive both partial and point identification results.

Our work also relates to a large literature on inference using inequality moment conditions. In particular, we adopt the generalized moment selection (GMS) approach of Andrews and Soares (2010) to construct a procedure for identification-robust inference that combines the moment equalities from our point identification results with inequalities from our partial identification results. Although the equalities alone globally identify our model, the inequalities turn out to be extremely valuable in settings where $\beta(\mathbf{x})$ may be small. Although our specific approach differs from theirs, the idea of including moment inequalities in a model that is already point identified by a collection of moment equalities relates to work by Moon and Schorfheide (2009). While the weak identification problem that we point out and address here also emerges in several closely related models, e.g. (Mahajan, 2006) and Frazis and Loewenstein (2003), we are unaware of any other work from the literature that explicitly acknowledges or addresses it. As shown in Appendix D, our inference procedure can be applied to the case of an exogenous regressor with only minor modifications.

The remainder of the paper is organized as follows. Section 2.1 describes our model and assumptions, Section 2.2 relates our results to existing work, and Sections 2.3–2.4 present our identification results. Section 3.1 points out the special inferential difficulties that arise in models with mis-classification while Section 3.2 gives a high-level overview of our proposed inference procedure. Full details of the procedure follow in Sections 3.3–3.5. Section 4 presents simulation results, and Section 5 concludes. Proofs appear in Appendix A, and we give a detailed explanation of the error in Mahajan (2006) in Appendix B. Appendix C explains how our partial identification bounds from Section 2.3 can be interpreted in a local average treatment effects (LATE) setting.

2 Identification

2.1 Baseline Assumptions

As defined in the preceding section, our model is $y = c(\mathbf{x}) + \beta(\mathbf{x})T^* + \varepsilon$, where ε is a mean-zero error term, and the parameter of interest is $\beta(\mathbf{x})$ – the effect of an unobserved, binary, endogenous regressor T^*. Suppose we observe a valid and relevant binary instrument z. In the discussion following Corollary 2.2 below, we explain how these results generalize to the case of an arbitrary discrete-valued instrument. We assume that the model and instrument satisfy the following conditions:

Assumption 2.1.

(i) $y = c(\mathbf{x}) + \beta(\mathbf{x})T^* + \varepsilon$ *where* $T^* \in \{0,1\}$ *and* $\mathbb{E}[\varepsilon] = 0$*;*

(ii) $z \in \{0,1\}$*, where* $0 < \mathbb{P}(z=1|\mathbf{x}) < 1$*, and* $\mathbb{P}(T^*=1|\mathbf{x},z=1) \neq \mathbb{P}(T^*=1|\mathbf{x},z=0)$*;*

(iii) $\mathbb{E}[\varepsilon|\mathbf{x},z] = 0$*.*

Assumption 2.1(i) is a restatement of the additively separable model from Equation 1, which includes as a special case the linear model $y = c + \beta T^* + \mathbf{x}'\boldsymbol{\gamma} + \varepsilon$ that is pervasive in empirical economics. Assumptions 2.1(ii) and (iii) are the textbook instrumental variable relevance and validity conditions, respectively. Under Assumption 2.1, the Wald estimator

$$\left[\mathbb{E}\left(y|z=1,\mathbf{x}\right) - \mathbb{E}\left(y|z=0,\mathbf{x}\right)\right] / \left[\mathbb{E}\left(T^*|z=1,\mathbf{x}\right) - \mathbb{E}\left(T^*|z=0,\mathbf{x}\right)\right]$$

identifies $\beta(\mathbf{x})$. Unfortunately this estimator is infeasible, as we observe not T^* but a mis-classified binary surrogate T.[4] To make further progress, we must impose conditions on the process that generates T. Accordingly, define the following mis-classification probabilities:

$$\alpha_0(\mathbf{x},z) = \mathbb{P}\left(T=1|T^*=0,\mathbf{x},z\right) \qquad \alpha_0(\mathbf{x}) = \mathbb{P}\left(T=1|T^*=0,\mathbf{x}\right)$$
$$\alpha_1(\mathbf{x},z) = \mathbb{P}\left(T=0|T^*=1,\mathbf{x},z\right) \qquad \alpha_1(\mathbf{x}) = \mathbb{P}\left(T=0|T^*=1,\mathbf{x}\right).$$

Assumption 2.2.

(i) $\alpha_0(\mathbf{x},z) = \alpha_0(\mathbf{x})$*,* $\alpha_1(\mathbf{x},z) = \alpha_1(\mathbf{x})$

(ii) $\alpha_0(\mathbf{x}) + \alpha_1(\mathbf{x}) < 1$

(iii) $\mathbb{E}[\varepsilon|\mathbf{x},z,T^*,T] = \mathbb{E}[\varepsilon|\mathbf{x},z,T^*]$

Assumption 2.2, or a variant thereof, is standard in the theoretical literature on mis-classification (Black et al., 2000; Frazis and Loewenstein, 2003; Hu, 2008; Lewbel, 2007; Mahajan, 2006) and in empirical studies that allow for measurement error in a binary or discrete variable (Battistin et al., 2014; Feng and Hu, 2013; Kane et al., 1999). Assumption 2.2 (i) states that the mis-classification probabilities do not depend on z. Assumption 2.2 (ii) restricts the extent of mis-classification and is equivalent to requiring that T and T^* be positively correlated. Assumption 2.2 (iii) is often referred to as "non-differential measurement error." Intuitively, it maintains that T provides no additional information about ε, and hence y, given knowledge of $(T^*, z, \mathbf{x})$. While Assumption 2.2(ii) is quite mild, Assumptions 2.2 (i) and (iii) are more restrictive, as discussed by Bound et al. (2001). To take a specific example, suppose that y is log wage and T^* is an indicator for college completion. If T is a potentially erroneous measure of college completion taken from a university's administrative records, then the assumption of non-differential measurement error is quite plausible. If, on the other hand, T is a self-report of college completion and there are "returns to lying" about college completion, i.e. employers only imperfectly observe worker ability, this assumption is less plausible.[5] Note, however, that our assumptions on the mis-classification process are *conditional* on $\mathbf{x}$: we place no restrictions on the relationship between observed covariates

[4] Although it involves T^*, Assumption 2.1(ii) is testable: see the discussion following Lemma 2.1.

[5] See Hu and Lewbel (2012) for a proposal to estimate the "returns to lying" in this context.

and the mis-classification errors. In contrast, Bound et al. (2001) considers *unconditional* versions of our Assumption 2.2. Instrument validity – Assumption 2.1 (iii) – is more plausible after conditioning on a rich set of exogenous controls, and the same is true of our mis-classification assumptions. For more discussion of settings in which the assumption of non-differential measurement error is warranted, see Carroll et al. (2006).

2.2 Point Identification Results from the Literature

Existing results from the literature – see for example Frazis and Loewenstein (2003) and Mahajan (2006) – establish that $\beta(\mathbf{x})$ is point identified if Assumptions 2.1–2.2 are augmented to include the following condition:

Assumption 2.3 (Joint Exogeneity). $\mathbb{E}[\varepsilon|\mathbf{x}, z, T^*] = 0$.

Assumption 2.3 strengthens the mean independence condition from Assumption 2.1 (iii) to hold *jointly* for T^* and z. By iterated expectations, this implies that T^* is exogenous, i.e. $\mathbb{E}[\varepsilon|\mathbf{x}, T^*] = 0$. If T^* is endogenous, Assumption 2.3 clearly fails. Mahajan (2006) argues, however, that the following restriction, along with our Assumptions 2.1–2.2, suffices to identify $\beta(\mathbf{x})$ when T^* may be endogenous:

Assumption 2.4 (Mahajan (2006) Equation 11). $\mathbb{E}[\varepsilon|\mathbf{x}, z, T^*, T] = \mathbb{E}[\varepsilon|\mathbf{x}, T^*]$.

Assumption 2.4 does not require $\mathbb{E}[\varepsilon|\mathbf{x}, T^*]$ to be zero, but maintains that it does not vary with z. We show in Appendix B, however, that under Assumptions 2.1–2.2, Assumption 2.4 can only hold if T^* is exogenous. If z is a valid instrument and T^* is endogenous, then Assumption 2.4 implies that there is no first-stage relationship between z and T^*. As such, identification in the case where T^* is endogenous is an open question.

2.3 Partial Identification

In this section we derive the sharp identified set under Assumptions 2.1–2.2 and show that $\beta(\mathbf{x})$ is not point identified. For a discussion of how our partial identification results can be interpreted in a local average treatment effects (LATE) setting, see Appendix C.

To simplify the notation, define the following shorthand for the unobserved and observed first stage probabilities

$$p_k^*(\mathbf{x}) = \mathbb{P}(T^* = 1|\mathbf{x}, z = k), \quad p_k(\mathbf{x}) = \mathbb{P}(T = 1|\mathbf{x}, z = k). \tag{2}$$

We first state two lemmas that that will be used repeatedly below.

Lemma 2.1. *Under Assumption 2.2 (i),*

$$\begin{aligned}[1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})]\, p_k^*(\mathbf{x}) &= p_k(\mathbf{x}) - \alpha_0(\mathbf{x}) \\ [1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})]\, [1 - p_k^*(\mathbf{x})] &= 1 - p_k(\mathbf{x}) - \alpha_1(\mathbf{x})\end{aligned}$$

where the first-stage probabilities $p_k^*(\mathbf{x})$ and $p_k(\mathbf{x})$ are as defined in Equation 2.

Lemma 2.2. *Under Assumptions 2.1 and 2.2 (i)–(ii),*

$$\beta(\mathbf{x})\,Cov(z,T|\mathbf{x}) = [1-\alpha_0(\mathbf{x})-\alpha_1(\mathbf{x})]\,Cov(y,z|\mathbf{x})$$

Lemma 2.1 relates the observed first-stage probabilities $p_k(\mathbf{x})$ to their unobserved counterparts $p_k^*(\mathbf{x})$ in terms of the mis-classification probabilities $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$. By Assumption 2.2 (ii), $1-\alpha_0(\mathbf{x})-\alpha_1(\mathbf{x})>0$ so that Lemma 2.1 bounds $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ in terms of the observed first-stage probabilities. Moreover, by taking differences evaluated at $k=1$ and $k=0$, this Lemma shows that $p_0^*(\mathbf{x})=p_1^*(\mathbf{x})$ if and only if $p_0(\mathbf{x})=p_1(\mathbf{x})$. In other words, Assumption 2.1 (ii) is testable under Assumption 2.2 (ii). Lemma 2.2 relates the instrumental variables (IV) estimand, $\text{Cov}(y,z|\mathbf{x})/\text{Cov}(z,T|\mathbf{x})$, to the mis-classification probabilities. Since $1-\alpha_0(\mathbf{x})-\alpha_1(\mathbf{x})>0$, IV is biased *upwards* in the presence of mis-classification. Together these lemmas bound the causal effect of interest: $\beta(\mathbf{x})$ lies between the reduced form and IV estimators. Without Assumption 2.2 (iii), non-differential measurement error, these bounds are sharp.

Theorem 2.1. *Under Assumptions 2.1 and 2.2 (i)–(ii),* $\alpha_0(\mathbf{x}) \le p_k(\mathbf{x}) \le 1-\alpha_1(\mathbf{x})$ *for* $k=0,1$ *and*

$$\mathbb{E}[y|\mathbf{x},z=k] = c(\mathbf{x}) + \beta(\mathbf{x})\left[\frac{p_k(\mathbf{x})-\alpha_0(\mathbf{x})}{1-\alpha_0(\mathbf{x})-\alpha_1(\mathbf{x})}\right]. \tag{3}$$

Provided that $p_0(\mathbf{x}) \neq p_1(\mathbf{x})$*, these expressions characterize the sharp identified set for* $c(\mathbf{x})$*,* $\beta(\mathbf{x})$*,* $\alpha_0(\mathbf{x})$*, and* $\alpha_1(\mathbf{x})$*.*

Corollary 2.1. *Under the conditions of Theorem 2.1, the sharp identified set for* $\beta(\mathbf{x})$ *is the closed interval between the reduced form estimand* $Cov(y,z|\mathbf{x})/Var(z|\mathbf{x})$ *and the IV estimand* $Cov(y,z|\mathbf{x})/Cov(z,T|\mathbf{x})$*.*

Corollary 2.1 follows by taking differences of the expression for $\mathbb{E}[y|\mathbf{x},z=k]$ across $k=1$ and $k=0$, and substituting the maximum and minimum value for $\alpha_0(\mathbf{x})+\alpha_1(\mathbf{x})$ consistent with the observed first-stage probabilities.[6] Note that the only role of the condition $p_0(\mathbf{x}) \neq p_1(\mathbf{x})$ in the preceding two results is to ensure that it is possible to satisfy Assumption 2.1 (ii). Frazis and Loewenstein (2003) point out that the IV estimand provides an upper bound for $\beta(\mathbf{x})$, and Lemmas 2.1–2.2 are well-known in the literature (see e.g. Frazis and Loewenstein, 2003; Mahajan, 2006). Nevertheless, we are unaware of any published result that explicitly states both bounds from Corollary 2.1 or proves that they are sharp under Assumptions 2.1 and 2.2 (i)–(ii).

Neither Theorem 2.1 nor Corollary 2.1 imposes Assumption 2.2 (iii) – non-differential measurement error. While this assumption plays an important role in existing identification results for an exogenous T^* (see Section 2.2), its identifying power under endogeneity has not been addressed in the literature.[7] We now show that this assumption in general yields further restrictions on probabilities $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$, but fails to point identify $\beta(\mathbf{x})$. To simplify the proof of sharpness, we assume that y is continuously distributed, which is natural in an additively separable model. Without this assumption, the bounds that we derive are still

[6] If *a priori* restrictions on α_0 and α_1 are available, e.g. $\alpha_0=0$, $\alpha_1=0$, or $\alpha_0=\alpha_1$, these bounds can be improved. For more discussion, see Corollary 2.2 of DiTraglia and García-Jimeno (2017).

[7] The only exception is the incorrect result of Mahajan (2006) described in Section 2.2 and Appendix B.

valid, but may not be sharp. Nevertheless, the reasoning from our proof can be generalized to cases in which y does not have a continuous support set.

Theorem 2.2. *Suppose that the conditional distribution of y given $(\mathbf{x}, T, z)$ is continuous. Further suppose that the conditions of Theorem 2.1 and Assumption 2.2 (iii) hold. For any k such that $\mathbb{E}[y|\mathbf{x}, T = 0, z = k] \neq \mathbb{E}[y|\mathbf{x}, T = 1, z = k]$, let $\mathcal{A}_k$ denote the set of pairs $\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x})\big)$ such that $\alpha_0(\mathbf{x}) < p_k(\mathbf{x}) < 1 - \alpha_1(\mathbf{x})$ and*

$$\underline{\mu}_{tk}\Big(\underline{q}_{tk}\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}\big),\, \mathbf{x}\Big) \leq \mu_k\big(\alpha_0(\mathbf{x}), \mathbf{x}\big) \leq \overline{\mu}_{tk}\Big(\overline{q}_{tk}\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}\big),\, \mathbf{x}\Big)$$

for all $t = 0, 1$ where

$$\underline{\mu}_{tk}\big(q, \mathbf{x}\big) = \mathbb{E}\left[y \mid y \leq q, \mathbf{x}, T = t, z = k\right], \qquad \overline{\mu}_{tk}\big(q, \mathbf{x}\big) = \mathbb{E}\left[y \mid y > q, \mathbf{x}, T = t, z = k\right]$$

$$\mu_k\big(\alpha_0(\mathbf{x}), \mathbf{x}\big) = \frac{p_k(\mathbf{x})\mathbb{E}[y|\mathbf{x}, z = k, T = 1] - \alpha_0(\mathbf{x})\mathbb{E}[y|\mathbf{x}, z = k]}{p_k(\mathbf{x}) - \alpha_0(\mathbf{x})}$$

and we define

$$\underline{q}_{tk}\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}\big) = F^{-1}_{tk}\Big(r_{tk}\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}\big) \,\Big|\, \mathbf{x}\Big)$$
$$\overline{q}_{tk}\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}\big) = F^{-1}_{tk}\Big(1 - r_{tk}\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}\big) \,\Big|\, \mathbf{x}\Big)$$

where $F^{-1}_{tk}(\cdot|\mathbf{x})$ is the conditional quantile function of y given $(\mathbf{x}, T = t, z = k)$,

$$r_{0k}\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}\big) = \frac{\alpha_1(\mathbf{x})}{1 - p_k(\mathbf{x})}\left[\frac{p_k(\mathbf{x}) - \alpha_0(\mathbf{x})}{1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})}\right]$$
$$r_{1k}\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x}), \mathbf{x}\big) = \frac{1 - \alpha_1(\mathbf{x})}{p_k(\mathbf{x})}\left[\frac{p_k(\mathbf{x}) - \alpha_0(\mathbf{x})}{1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})}\right]$$

and $p_k(\mathbf{x})$ is defined in Equation 2. The sharp identified set for $c(\mathbf{x})$, $\beta(\mathbf{x})$, $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ is characterized by Equation 3 and $\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x})\big) \in \mathcal{A}^*$ where

(i) $\mathcal{A}^* \equiv \mathcal{A}_0 \cap \mathcal{A}_1$ *if* $\mathbb{E}[y|\mathbf{x}, T = 0, z = k] \neq \mathbb{E}[y|\mathbf{x}, T = 1, z = k]$ *for all* $k = 0, 1$*;*

(ii) $\mathcal{A}^* \equiv \mathcal{A}_k$ *if* $\mathbb{E}[y|\mathbf{x}, T = 0, z = k] \neq \mathbb{E}[y|\mathbf{x}, T = 1, z = k]$ *and* $\mathbb{E}[y|\mathbf{x}, T = 0, z = \ell] = \mathbb{E}[y|\mathbf{x}, T = 1, z = \ell]$*;*

(iii) $\mathcal{A}^* \equiv \big\{\big(\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x})\big) \colon \alpha_0(\mathbf{x}) \leq p_k(\mathbf{x}) \leq 1 - \alpha_1(\mathbf{x})$ *for all* $k\big\}$ *if* $\mathbb{E}[y|\mathbf{x}, T = 0, z = k] = \mathbb{E}[y|\mathbf{x}, T = 1, z = k]$ *for all* $k = 0, 1$*.*

Imposing Assumption 2.2 (iii) strictly improves upon the identified set from Theorem 2.1 unless $\mathbb{E}[y|\mathbf{x}, T = 0, z = k] = \mathbb{E}[y|\mathbf{x}, T = 1, z = k]$ for all k. Even if $\beta(\mathbf{x}) = 0$, the

difference of these observable means is generically nonzero.[8] The intuition for Theorem 2.2 is as follows. For simplicity, suppress dependence on $\mathbf{x}$. Now, fix $(T = t, z = k)$ and (α_0, α_1). The observed distribution of y given $(T = t, z = k)$, call it F_{tk}, is a mixture of two unobserved distributions: the distribution of y given $(T = k, z = k, T^* = 1)$, call it F^1_{tk}, and the distribution of y given $(T = t, z = k, T^* = 0)$, call it F^0_{tk}. The mixing probabilities are r_{tk} and $1 - r_{tk}$ from the statement of Theorem 2.2 and are fully determined by (α_0, α_1) and p_k. Assumptions 2.1 (i) and 2.2 (ii) imply that the unobserved means $\mathbb{E}[y|T^*, T, z]$ are fully determined by (α_0, α_1) given the observed means $\mathbb{E}[y|T, z]$. The question is whether it is possible, given the observed distribution F_{tk}, to construct F^1_{tk} and F^0_{tk} with the required values for $\mathbb{E}[y|T^*, T, z]$ such that $F_{tk} = r_{tk}F^1_{tk} + (1 - r_{tk})F^0_{tk}$ for all combinations (t, k). If not, then (α_0, α_1) does not belong to the identified set. Our proof provides necessary and sufficient conditions for such a mixture to exist at a given point (α_0, α_1). We can then appeal to the reasoning from Theorem 2.1 to complete the argument. By ruling out values for α_0 and α_1, Theorem 2.2 restricts β via Lemma 2.2. While these restrictions can be very informative, they do not yield point identification.

Corollary 2.2. *Under Assumptions 2.1 and 2.2 the identified set for $\beta(\mathbf{x})$ contains both the IV estimand $Cov(y, z|\mathbf{x})/Cov(z, T|\mathbf{x})$ and the true coefficient $\beta(\mathbf{x})$.*

Corollary 2.2 follows by Lemma 2.2 because $\alpha_0(\mathbf{x}) = \alpha_1(\mathbf{x}) = 0$ always belongs to the sharp identified set from Theorem 2.2. Non-differential measurement error cannot exclude the possibility that there is no mis-classification because in this case it is trivial to construct the required mixtures. Although we focus throughout this paper on the case of a binary instrument, one might wonder whether point identification can be achieved by increasing the support of z, perhaps along the lines of Lewbel (2007). The answer turns out to be no. Suppose that we were to modify Assumptions 2.1 and 2.2 to hold for all values of z in some discrete support set. By Lemma 2.2, a binary instrument identifies $\beta(\mathbf{x})$ up to knowledge of the mis-classification probabilities $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$. It follows that *any* pair of values (k, ℓ) in the support set of z identifies the same object. Accordingly, to identify $\beta(\mathbf{x})$ it is necessary and sufficient to identify the mis-classification probabilities. A binary instrument fails to identify these probabilities because we can never exclude the possibility of zero mis-classification. The same is true of a discrete K-valued instrument. Increasing the support of z does, however, shrink the identified set by increasing the number of restrictions available: in this case Theorems 2.1–2.2 continue to apply replacing "$k = 0, 1$" with "for all k."

2.4 Point Identification

The results of the preceding section establish that $\beta(\mathbf{x})$ is not point identified under Assumptions 2.1 and 2.2. In light of this, there are two possible ways to proceed: either one can report partial identification bounds based on our characterization of the sharp identified set from Theorem 2.2, or one can attempt to impose stronger assumptions to obtain point identification. In this section we consider the second possibility. We begin by defining the

[8] Suppress dependence on $\mathbf{x}$ for simplicity. There are only two settings in which $\mathbb{E}[y|T = 0, z = k] = \mathbb{E}[y|T = 1, z = k]$. The first is if the true value of either α_0 or α_1 lies at the upper boundary of the identified set from Theorem 2.1. The second is if $\beta = \mathbb{E}[\varepsilon|T^* = 0, z = k] - \mathbb{E}[\varepsilon|T^* = 0, z = k]$.

following functions of the model parameters:

$$\theta_1(\mathbf{x}) = \beta(\mathbf{x})\left[1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})\right]^{-1} \tag{4}$$

$$\theta_2(\mathbf{x}) = \left[\theta_1(\mathbf{x})\right]^2\left[1 + \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})\right] \tag{5}$$

$$\theta_3(\mathbf{x}) = \left[\theta_1(\mathbf{x})\right]^3\left[\left\{1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})\right\}^2 + 6\alpha_0(\mathbf{x})\left\{1 - \alpha_1(\mathbf{x})\right\}\right] \tag{6}$$

Now consider the following additional assumption:

Assumption 2.5. $\mathbb{E}[\varepsilon^2|\mathbf{x}, z] = \mathbb{E}[\varepsilon^2|\mathbf{x}]$

Assumption 2.5 is a *second moment* version of the standard mean exclusion restriction for the instrument z – Assumption 2.1 (iii). It requires that the conditional variance of the error term given the covariates $\mathbf{x}$ does not depend on z, but does *not* require homoskedasticity with respect to $\mathbf{x}, T^*$ or T. Assumption 2.5 allows us to derive the following lemma:

Lemma 2.3. *Under Assumptions 2.1, 2.2 and 2.5,*

$$Cov(y^2, z|\mathbf{x}) = 2\,Cov(yT, z|\mathbf{x})\theta_1(\mathbf{x}) - Cov(T, z|\mathbf{x})\theta_2(\mathbf{x})$$

where $\theta_1(\mathbf{x})$ and $\theta_2(\mathbf{x})$ are defined in Equations 4–5.

Lemma 2.2 identifies $\theta_1(\mathbf{x})$. Since $\text{Cov}(z, T|\mathbf{x}) \neq 0$ by Assumption 2.1 (ii), we can solve for $\theta_2(\mathbf{x})$ in terms of observables only, using Lemma 2.3. Given knowledge of $\theta_1(\mathbf{x})$, we can solve Equation 5 for the difference of mis-classification rates so long as $\beta(\mathbf{x}) \neq 0$.

Corollary 2.3. *Under Assumptions 2.1–2.2 and 2.5, $\alpha_1(\mathbf{x}) - \alpha_0(\mathbf{x})$ is identified so long as $\beta(\mathbf{x}) \neq 0$.*

Corollary 2.3 identifies the difference of mis-classification error rates. Hence, under one-sided mis-classification, $\alpha_0(\mathbf{x}) = 0$ or $\alpha_1(\mathbf{x}) = 0$, augmenting our baseline Assumptions 2.1–2.2 with Assumption 2.5 suffices to identify $\beta(\mathbf{x})$. Notice that $\beta(\mathbf{x}) = 0$ if and only if $\theta_1(\mathbf{x}) = 0$. Thus, $\beta(\mathbf{x})$ is still identified in the case where Corollary 2.3 fails to apply.

Assumption 2.5 does not suffice to identify $\beta(\mathbf{x})$ without *a priori* restrictions on the mis-classification error rates. To achieve identification in the general case, we impose the following additional conditions:

Assumption 2.6.

(i) $\mathbb{E}[\varepsilon^2|\mathbf{x}, z, T^*, T] = \mathbb{E}[\varepsilon^2|\mathbf{x}, z, T^*]$

(ii) $\mathbb{E}[\varepsilon^3|\mathbf{x}, z] = \mathbb{E}[\varepsilon^3|\mathbf{x}]$

Assumption 2.6 (i) is a second moment version of the non-differential measurement error assumption, Assumption 2.2 (iii). It requires that, given knowledge of $(\mathbf{x}, T^*, z)$, T provides no additional information about the variance of the error term. Note that Assumption 2.6 (i) does not require homoskedasticity of ε with respect to $\mathbf{x}$ or T^*. Assumption 2.6 (ii) is a third moment version of Assumption 2.5. It requires that the conditional third moment of the error term given $\mathbf{x}$ does not depend on z. This condition neither requires nor excludes skewness

in the error term conditional on covariates: it merely states that the skewness is unaffected by the instrument. While Assumptions 2.5 and 2.6 may appear somewhat unusual, they are implied by the more intuitive independence conditions $\varepsilon \perp\!\!\!\perp z|\mathbf{x}$ and $\varepsilon \perp\!\!\!\perp T|(\mathbf{x}, T^*, z)$. Although $\mathbb{E}[\varepsilon|\mathbf{x}, z] = 0$ and $\mathbb{E}[\varepsilon|\mathbf{x}, z, T^*, T] = \mathbb{E}[\varepsilon|\mathbf{x}, z, T^*]$ are technically weaker than assuming full independence, we would be somewhat dubious of any supposed "natural experiment" that purportedly satisfied mean exclusion but not independence. Indeed, as discussed by Imbens and Rubin (1997), an instrument satisfying mean exclusion but not independence could become invalid if the outcome variable were transformed, for example by taking logs. As it is not uncommon for applied papers to report results in both logs and levels (e.g. Angrist, 1990), our view is that researchers *implicitly* assume more than mean exclusion in typical applications of instrumental variables. Analogous reasoning applies to the non-differential measurement error assumption.

Assumption 2.6 allows us to derive the following Lemma which, combined with Lemma 2.3, leads to point identification:

Lemma 2.4. *Under Assumptions 2.1–2.2 and 2.5–2.6,*

$$Cov(y^3, z|\mathbf{x}) = 3Cov(y^2T, z|\mathbf{x})\theta_1(\mathbf{x}) - 3Cov(yT, z|\mathbf{x})\theta_2(\mathbf{x}) + Cov(T, z|\mathbf{x})\theta_3(\mathbf{x})$$

where $\theta_1(\mathbf{x}), \theta_2(\mathbf{x})$ and $\theta_3(\mathbf{x})$ are defined in Equations 4–5.

Theorem 2.3. *Under Assumptions 2.1–2.2 and 2.5–2.6, $\beta(\mathbf{x})$ is identified. If $\beta(\mathbf{x}) \neq 0$, then $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ are likewise identified.*

Lemmas 2.2–2.4 yield a linear system of three equations in $\theta_1(\mathbf{x}), \theta_2(\mathbf{x})$ and $\theta_3(\mathbf{x})$. Under Assumption 2.1 (ii), the system has a unique solution so $\theta_1(\mathbf{x}), \theta_2(\mathbf{x})$ and $\theta_3(\mathbf{x})$ are identified. The proof of Theorem 2.3 shows that, so long as $\beta(\mathbf{x}) \neq 0$, Equations 4–6 can be solved for $\beta(\mathbf{x})$, $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$. In particular, using steps from the proof of Theorem 2.3

$$\beta(\mathbf{x}) = \text{sign}\big[\theta_1(\mathbf{x})\big]\sqrt{3\big[\theta_2(\mathbf{x})/\theta_1(\mathbf{x})\big]^2 - 2\big[\theta_3(\mathbf{x})/\theta_1(\mathbf{x})\big]}. \tag{7}$$

If we relax Assumption 2.2 (ii) and assume $\alpha_0(\mathbf{x}) + \alpha_1(\mathbf{x}) \neq 1$ only, $\beta(\mathbf{x})$ is only identified up to sign: in this case the sign of $\theta_1(\mathbf{x})$ need not equal that of $\beta(\mathbf{x})$.

3 Identification-Robust Inference

We now turn our attention to inference based on the identification results from above. We begin by expressing Lemmas 2.2, 2.3 and 2.4 as unconditional equality moment conditions, and describing the resulting just-identified GMM estimator. As we explain in Section 3.1, inference under binary mis-classification is complicated by problems of weak identification and parameters on the boundary. Section 3.2 provides an overview of our inference procedure. Full details appear in Sections 3.3–3.5. For simplicity we fix the exogenous covariates at some specified level and suppress dependence on $\mathbf{x}$ in the notation. This is appropriate if the covariates have a discrete support. We discuss how to incorporate covariates more generally in Section 3.6.

3.1 The Non-standard Inference Problem

Lemmas 2.2–2.4 yield the following system of linear moment equalities in the reduced form parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ from Equations 4–6:

$$\begin{aligned} \text{Cov}(y, z) - \text{Cov}(T, z)\theta_1 = 0 \\ \text{Cov}(y^2, z) - 2\text{Cov}(yT, z)\theta_1 + \text{Cov}(T, z)\theta_2 = 0 \\ \text{Cov}(y^3, z) - 3\text{Cov}(y^2T, z)\theta_1 + 3\text{Cov}(yT, z)\theta_2 - \text{Cov}(T, z)\theta_3 = 0 \end{aligned}$$

Non-linearity arises solely through the relationship between the reduced from parameters $\boldsymbol{\theta}$ and the structural parameters $(\alpha_0, \alpha_1, \beta)$. To convert the preceding moment equations into unconditional moment equalities, we define the additional reduced form parameters $\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \kappa_3)$ as follows:

$$\begin{aligned} \kappa_1 &= c - \alpha_0\theta_1 \\ \kappa_2 &= c^2 + \sigma_{\varepsilon\varepsilon} + \alpha_0(\theta_2 - 2c\theta_1) \\ \kappa_3 &= c^3 + 3\left(c - \theta_1\alpha_0\right)\sigma_{\varepsilon\varepsilon} + \mathbb{E}[\varepsilon^3] - \alpha_0\theta_3 - 3c\alpha_0\left[\theta_1\left(c + \beta\right) - 2\theta_1^2(1 - \alpha_1)\right] \end{aligned}$$

Building on this notation, let

$$\boldsymbol{\psi}_1' = (-\theta_1, 1, 0, 0, 0, 0), \quad \boldsymbol{\psi}_2' = (\theta_2, 0, -2\theta_1, 1, 0, 0), \quad \boldsymbol{\psi}_3' = (-\theta_3, 0, 3\theta_2, 0, -3\theta_1, 1) \tag{8}$$

and collect these in the matrix $\boldsymbol{\Psi} = \left[\begin{array}{ccc} \boldsymbol{\psi}_1 & \boldsymbol{\psi_2} & \boldsymbol{\psi_3} \end{array}\right]$. Defining the observed data vector $\mathbf{w}_i' = (T_i, y_i, y_iT_i, y_i^2, y_i^2T_i, y_i^3)$ for observation i, we can re-write the moment equations as:

$$\mathbb{E}\left[\left(\boldsymbol{\Psi}'(\boldsymbol{\theta})\mathbf{w}_i - \boldsymbol{\kappa}\right) \otimes \begin{pmatrix} 1 \\ z_i \end{pmatrix}\right] = \mathbf{0}. \tag{9}$$

Equation 9 is a just-identified, linear system of moment equalities in the reduced form parameters $(\boldsymbol{\theta}, \boldsymbol{\kappa})$ and yields explicit GMM estimators $(\widehat{\boldsymbol{\kappa}}, \widehat{\boldsymbol{\theta}})$. From Theorem 2.3, knowledge of $\boldsymbol{\theta}$ suffices to identify β. From the definitions of $\boldsymbol{\kappa}$ above and $\boldsymbol{\theta}$ in Equations 4–6, however, the moment equalities from Equation 9 do not depend on (α_0, α_1) if β equals zero. By continuity, they are *nearly* uninformative about the mis-classification probabilities if β is small. But unless $\beta = 0$, knowledge of (α_0, α_1) is necessary to recover β, via Lemma 2.2. Thus, we face a weak identification problem.[9] Indeed, the GMM estimator

$$\widehat{\beta} = \text{sign}\left(\widehat{\theta}_1\right)\sqrt{3\left(\widehat{\theta}_2/\widehat{\theta}_1\right)^2 - 2\left(\widehat{\theta}_3/\widehat{\theta}_1\right)}$$

corresponding to Equation 7 may even fail to exist. Under our assumptions, $3(\theta_2/\theta_1)^2 > 2(\theta_3/\theta_1)$ provided that $\beta \neq 0$, but this may not be true of the sample analogue. Indeed, because $\widehat{\theta}_1$ appears in the denominator, the terms within the square root will be highly variable if β is small. Even if the GMM estimator exists, it may violate the partial identification bounds for (α_0, α_1) from Theorem 2.2, or imply that (α_0, α_1) are not valid probabilities. Im-

[9] This is essentially equivalent to the problem of estimating mixture probabilities when the means of the component distributions are very similar to each other.

portantly, the partial identification bounds remain informative even if β is small or zero: so long as Assumption 2.1 (ii) holds, the first-stage probabilities bound α_0 and α_1 from above.

Exactly the same inferential difficulties arise in the case where T^* and z are assumed to be jointly exogenous, as in Black et al. (2000); Frazis and Loewenstein (2003); Kane et al. (1999); Lewbel (2007); Mahajan (2006).[10] This issue, however, has received little attention in the literature. Kane et al. (1999) ensure that (α_0, α_1) are valid probabilities by employing a logit specification. Frazis and Loewenstein employ a pseudo-Bayesian approach to ensure that α_0 and α_1 are valid probabilities, and to impose partial identification bounds related to those from our Theorem 2.1, i.e. without using the non-differential measurement error restrictions. Because they provide neither simulation evidence nor a theoretical justification for their procedure, however, it is unclear whether this method will yield valid Frequentist coverage. We are unaware of any papers in the related literature that discuss the weak identification problem arising when β is small.

3.2 Overview of the Inference Procedure

In the following sections we develop a procedure for uniformly valid inference in models with a mis-classified binary regressor. Our purpose is to construct a confidence interval for β that is robust to possible weak identification, respects the restricted parameter space for (α_0, α_1), and incorporates both the information in the equality moment conditions from Equation 9 along with the partial identification bounds from Theorem 2.2.[11] As argued in the preceding section, our partial identification bounds remain informative even when the equality moment conditions contain essentially no information about (α_0, α_1).

To carry out identification-robust inference combining equality and inequality moment conditions, we adopt the generalized moment selection (GMS) approach of Andrews and Soares (2010). This procedure provides a uniformly valid test of a *joint* null hypothesis for the full parameter vector. In our model, this includes the parameter of interest β along with various nuisance parameters: the mis-classification probabilities α_0 and α_1, the reduced form parameters $\boldsymbol{\kappa}$, defined in Section 3.1, and a vector $\mathbf{q}$ of parameters that enter the moment inequalities.[12] Under a given joint null hypothesis for $(\beta, \alpha_0, \alpha_1)$, however, $\boldsymbol{\kappa}$ and $\mathbf{q}$ are strongly identified and lie on the interior their respective parameter spaces. Accordingly, in Section 3.4 we explain how to concentrate these parameters out of the GMS procedure, by deriving an appropriate correction to the asymptotic variance matrix for the test.[13]

This leaves us with a uniformly valid test of any joint null hypothesis for $(\beta, \alpha_0, \alpha_1)$. To construct a marginal confidence interval for β we proceed as follows. Suppose that z is a strong instrument. Then the usual IV estimator provides a valid confidence interval for the reduced from parameter θ_1. By Lemma 2.2, knowledge of $(1 - \alpha_0 - \alpha_1)$ suffices to determine β from θ_1. Thus, a valid confidence interval for $(1 - \alpha_0 - \alpha_1)$ can be combined with the IV

[10] We provide details for Frazis and Loewenstein (2003) and Mahajan (2006) in Appendix D.

[11] Note that $\beta = 0$ if and only if $\theta_1 = 0$. Thus, if one is merely interested in testing $H_0 \colon \beta = 0$, one can ignore the mis-classification error problem and test $H_0 \colon \theta_1 = 0$ using the standard IV estimator and standard error, provided that z is a strong instrument.

[12] These are defined below in Section 3.3.

[13] Note that we cannot take the same approach to concentrate out α_0 and α_1 because the mis-classification probabilities may be weakly identified or lie on the boundary of their parameter space.

interval for θ_1 to yield a corresponding interval for β, via a Bonferroni-type correction. To construct the required interval for $(1-\alpha_0-\alpha_1)$, we note from Equations 4–6 that β only enters the moment equality conditions in Equation 9 through θ_1. But, again, inference for θ_1 is standard provided that z is a strong instrument. We can thus pre-estimate θ_1 along with $\boldsymbol{\kappa}$ and $\mathbf{q}$, yielding a uniformly valid GMS test of any joint null hypothesis for (α_0, α_1). By inverting this test, we construct a joint confidence set for (α_0, α_1) which we then project to obtain a confidence interval for $(1-\alpha_0-\alpha_1)$. Because the parameter space for (α_0, α_1) is bounded and two-dimensional, the projection step is computationally trivial.[14] If desired, one could also carry out a valid test of the null hypothesis that there is no mis-classification, $\alpha_0 = \alpha_1 = 0$, using the joint test for (α_0, α_1). In the following sections we provide full details of our Bonferroni-based confidence interval procedure for β.

3.3 Moment Inequalities

As noted above, the partial identification bounds from Theorems 2.1 and 2.2 remain informative about (α_0, α_1) even when β is small. To incorporate them in our inference procedure, we first express them as unconditional moment inequalities. The bounds from Theorem 2.1 are given by

$$p_k - \alpha_0 \geq 0, \qquad 1 - p_k - \alpha_1 \geq 0, \quad \text{for all } k$$

where the first-stage probabilities p_k are defined in Equation 2. We write these inequalities as

$$\mathbb{E}\left[m_1^I(\mathbf{w}_i, \boldsymbol{\vartheta})\right] \geq \mathbf{0}, \quad m_1^I(\mathbf{w}_i, \boldsymbol{\vartheta}) \equiv \begin{bmatrix} (1-z_i)(T_i - \alpha_0) \\ (1-z_i)(1-T_i-\alpha_1) \\ z_i(T-\alpha_0) \\ z_i(1-T_i-\alpha_1) \end{bmatrix} \tag{10}$$

The bounds derived in Theorem 2.2 by imposing assumption 2.2 (iii) are

$$\mu_k(\alpha_0) - \underline{\mu}_{tk}\big(\underline{q}_{tk}(\alpha_0, \alpha_1)\big) \geq 0, \qquad \overline{\mu}_{tk}\big(\overline{q}_{tk}(\alpha_0, \alpha_1)\big) - \mu_k(\alpha_0) \geq 0, \quad \text{for all } t, k$$

[14] We considered two alternatives to the Bonferroni-based inference procedure described here. The first constructs a marginal confidence interval for β by projecting a joint confidence set for $(\beta, \alpha_1, \alpha_0)$, i.e. *without* preliminary estimation of θ_1. This method is more computationally demanding than our two-dimensional projection and involves a parameter space that is unbounded along the β-dimension. From a practical perspective, the relevant question is whether the reduction in conservatism from projecting a lower dimensional set is outweighed by the additional conservatism induced by the Bonferroni correction. In our experiments, the full three-dimensional projection and Bonferroni procedure produced broadly similar results: neither reliably dominated in terms of confidence interval width. Given its substantially lower computational burden, we prefer the Bonferroni procedure. We also experimented with two recently proposed methods for sub-vector inference: Kaido et al. (2016) and Bugni et al. (2017). In both cases we obtained significant size distortions, suggesting that our model may not satisfy the regularity conditions required by these papers.

where $\mu_k, \underline{\mu}_{tk}, \overline{\mu}_{tk}, \underline{q}_{tk}$ and $\overline{q}_{tk}$ are defined in the statement of the Theorem. Expressing these as unconditional moment inequalities, we have

$$\mathbb{E}[m_2^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q})] \geq \mathbf{0}, \quad m_2^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) \equiv \begin{bmatrix} m_{2,00}^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) \\ m_{2,10}^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) \\ m_{2,01}^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) \\ m_{2,11}^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) \end{bmatrix} \tag{11}$$

where $\mathbf{q} \equiv (\underline{q}_{00}, \overline{q}_{00}, \underline{q}_{10}, \overline{q}_{10}, \underline{q}_{01}, \overline{q}_{01}, \underline{q}_{11}, \overline{q}_{11})$ and we define

$$m_{2,0k}^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) \equiv \begin{bmatrix} y_i \mathbf{1}(z_i = k) \left\{ (T_i - \alpha_0) - \mathbf{1}(y_i \leq \underline{q}_{0k})(1 - T_i) \left(\frac{1-\alpha_0-\alpha_1}{\alpha_1} \right) \right\} \\ -y_i \mathbf{1}(z_i = k) \left\{ (T_i - \alpha_0) - \mathbf{1}(y_i > \overline{q}_{0k})(1 - T_i) \left(\frac{1-\alpha_0-\alpha_1}{\alpha_1} \right) \right\} \end{bmatrix} \tag{12}$$

$$m_{2,1k}^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) \equiv \begin{bmatrix} y_i \mathbf{1}(z_i = k) \left\{ (T_i - \alpha_0) - \mathbf{1}(y_i \leq \underline{q}_{1k}) T_i \left(\frac{1-\alpha_0-\alpha_1}{1-\alpha_1} \right) \right\} \\ -y_i \mathbf{1}(z_i = k) \left\{ (T_i - \alpha_0) - \mathbf{1}(y_i > \overline{q}_{1k}) T_i \left(\frac{1-\alpha_0-\alpha_1}{1-\alpha_1} \right) \right\} \end{bmatrix}. \tag{13}$$

Finally we define $m^I = (m_1^{I\prime}, m_2^{I\prime})'$. Notice that the second set of inequalities, m_2^I, depends on the unknown parameter $\mathbf{q}$ which is in turn a function of (α_0, α_1). In the next section we discuss how $\mathbf{q}$ can be estimated under a given null hypothesis for (α_0, α_1).

3.4 Accounting for Preliminary Estimation

Let $\boldsymbol{\vartheta} = (\alpha_0, \alpha_1)$ and $\boldsymbol{\gamma} = (\boldsymbol{\kappa}, \theta_1)$ where θ_1 is defined in Equation 4 and $\boldsymbol{\kappa}$ in Section 3.1. Our moment conditions take the form

$$\mathbb{E}[m^I(\mathbf{w}_i, \boldsymbol{\vartheta}_0, \mathbf{q}_0)] \geq \mathbf{0}, \quad \mathbb{E}[m^E(\mathbf{w}_i, \boldsymbol{\vartheta}_0, \boldsymbol{\gamma}_0)] = \mathbf{0} \tag{14}$$

where $m^I = (m_1^{I\prime}, m_2^{I\prime})'$, defined in Section 3.3, and

$$m^E(\mathbf{w}_i, \boldsymbol{\vartheta}_0, \boldsymbol{\gamma}_0) = \begin{bmatrix} \{\boldsymbol{\psi}_2'(\theta_1, \alpha_0, \alpha_1)\mathbf{w}_i - \kappa_2\} z_i \\ \{\boldsymbol{\psi}_3'(\theta_1, \alpha_0, \alpha_1)\mathbf{w}_i - \kappa_3\} z_i \end{bmatrix}. \tag{15}$$

Notice that we now write $\boldsymbol{\psi}_2$ and $\boldsymbol{\psi}_3$, defined in Equation 8, as explicit functions of $(\theta_1, \alpha_0, \alpha_1)$, using the definitions of (θ_2, θ_3) from Equations 5–6. To construct a GMS test of the null hypothesis $H_0\colon \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_\mathbf{0}$ based on Equation 14, we require preliminary estimators $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\vartheta}_0)$ and $\widehat{\mathbf{q}}(\boldsymbol{\vartheta}_0)$ that are consistent and asymptotically normal *under the null*. We now provide full details of the construction and derive the associated adjustment to the asymptotic variance matrix.

Consider first the equality moment conditions m^E. For these we require preliminary estimators of θ_1, κ_2, and κ_3. Recall that θ_1 is simply the IV estimand: it can be consistently estimated directly from observations of (y, T, z) without knowledge of α_0 or α_1. Note, moreover, from Equation 9 that $\boldsymbol{\kappa}$ is simply a vector of *intercepts*. These can be directly estimated from observations of $\mathbf{w}$ because $\boldsymbol{\Psi}(\theta_1, \alpha_0, \alpha_1)$ is consistently estimable under the

null $H_0\colon \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0$: the hypothesis specifies α_0 and α_1 and IV provides a consistent estimator of θ_1. Accordingly, define

$$h^E(\mathbf{w}_i, \boldsymbol{\vartheta}, \boldsymbol{\gamma}) = \begin{bmatrix} \boldsymbol{\Psi}'(\theta_1, \alpha_0, \alpha_1)\mathbf{w}_i - \boldsymbol{\kappa} \\ \{\boldsymbol{\psi}_1'(\theta_1)\mathbf{w}_i - \kappa_1\} z_i \end{bmatrix}. \tag{16}$$

Under $H_0\colon \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0$, the just-identified GMM-estimator based on $\mathbb{E}[h^E(\mathbf{w}_i, \boldsymbol{\vartheta}_0, \boldsymbol{\gamma}_0)] = \mathbf{0}$ yields a consistent and asymptotically normal estimator of $\boldsymbol{\gamma}_0$ under the usual regularity conditions.

Now consider the inequality moment conditions m^I. From Section 3.3 we see that m_2^I depends on $\mathbf{q}$, the vector of conditional quantiles $\bar{q}_{tk}$ and $\underline{q}_{tk}$ defined in Theorem 2.2. Under the assumption that y follows a continuous distribution, as maintained in Theorem 2.2, these can be expressed as conditional moment equalities as follows:

$$\mathbb{E}\left[\mathbf{1}(y \leq \underline{q}_{tk}) | T = t, z = k\right] - r_{tk}(\alpha_0, \alpha_1) = 0 \tag{17}$$

$$\mathbb{E}\left[\mathbf{1}(y \leq \bar{q}_{tk}) | T = t, z = k\right] - \big(1 - r_{tk}(\alpha_0, \alpha_1)\big) = 0 \tag{18}$$

where r_{tk} is defined in Theorem 2.2 and $t, k = 0, 1$. Under $H_0\colon \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0$, a consistent estimator $\widehat{r}_{tk}$ of r_{tk} can be obtained directly from $\widehat{p}_k$, the sample analogue of p_k based on iid observations of $\mathbf{w}_i$. In turn, the $(\widehat{r}_{tk})^{\text{th}}$ and $(1-\widehat{r}_{tk})^{\text{th}}$ sample conditional quantiles of y provide consistent estimates of $\underline{q}_{tk}$ and $\bar{q}_{tk}$.[15] Collecting these for all (t, k) gives $\widehat{\mathbf{q}}(\boldsymbol{\vartheta}_0)$. Now, define

$$h^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) = \begin{bmatrix} h_0^I(\mathbf{w}, \boldsymbol{\vartheta}, \mathbf{q}) \\ h_1^I(\mathbf{w}, \boldsymbol{\vartheta}, \mathbf{q}) \end{bmatrix} \tag{19}$$

where

$$h_k^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) = \begin{bmatrix} \mathbf{1}(y_i \leq \underline{q}_{0k})\mathbf{1}(z_i = k)(1 - T_i) - \left(\frac{\alpha_1}{1-\alpha_0-\alpha_1}\right)\mathbf{1}(z_i = k)(T_i - \alpha_0) \\ \mathbf{1}(y_i \leq \bar{q}_{0k})\mathbf{1}(z_i = k)(1 - T_i) - \left(\frac{1-\alpha_0}{1-\alpha_0-\alpha_1}\right)\mathbf{1}(z_i = k)(1 - T_i - \alpha_1) \\ \mathbf{1}(y_i \leq \underline{q}_{1k})\mathbf{1}(z_i = k)T_i - \left(\frac{1-\alpha_1}{1-\alpha_0-\alpha_1}\right)\mathbf{1}(z_i = k)(T_i - \alpha_0) \\ \mathbf{1}(y_i \leq \bar{q}_{1k})\mathbf{1}(z_i = k)T_i - \left(\frac{\alpha_0}{1-\alpha_0-\alpha_1}\right)\mathbf{1}(z_i = k)(1 - T_i - \alpha_1) \end{bmatrix}. \tag{20}$$

Equation 20 gives the unconditional version of Equations 17–18. Now, under the null $\widehat{\mathbf{q}}(\boldsymbol{\vartheta}_0)$ converges in probability to $\mathbf{q}_0$, which satisfies the just-identified collection of moment equalities $\mathbb{E}[h^I(\mathbf{w}_i, \boldsymbol{\vartheta}_0, \mathbf{q}_0)] = \mathbf{0}$. Although h^I is a discontinuous function of $\mathbf{q}$, it is bounded for any fixed (α_0, α_1). Moreover, since $y|(T = t, z = k)$ is a continuous random variable, $\mathbb{E}[h^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q})]$ is continuously differentiable with respect to $\mathbf{q}$. Hence, $\widehat{\mathbf{q}}$ is asymptotically normal under mild regularity conditions.[16] To account for the effect of preliminary estimation of $\mathbf{q}$ and $\boldsymbol{\gamma}$ on the asymptotic variance matrix used in the GMS test, we rely on the following Lemma:

Lemma 3.1. *Let* $\bar{m}_{1,n}^I(\boldsymbol{\vartheta}) = n^{-1}\sum_{i=1}^n m_{1,n}^I(\mathbf{w}_i, \boldsymbol{\vartheta})$ *and define* $\bar{m}_{2,n}^I, \bar{m}_n^E, \bar{h}_n^I, \bar{h}_n^E$ *analogously.*

[15] Consistency of the sample quantiles requires $0 < r_{tk} < 1$. If $r_{tk} = 0$ or 1 for some (t, k), however, then the associated moment inequality is trivially satisfied and we no longer require estimates of $\underline{q}_{tk}, \bar{q}_{tk}$.

[16] For details, see Andrews (1994) and Newey and McFadden (1994) Section 7.

Further let $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\vartheta}_0) = \arg\min_{\boldsymbol{\gamma}} \|\bar{h}_n^E(\boldsymbol{\vartheta}_0, \boldsymbol{\gamma})\|$ *and* $\|h_n^I(\boldsymbol{\vartheta}_0, \widehat{\mathbf{q}}(\boldsymbol{\vartheta}_0))\| \leq \inf_{\mathbf{q}} \|h_n^I(\boldsymbol{\vartheta}_0, \mathbf{q})\| + o_p(1)$. *Then, under standard regularity conditions*

$$\sqrt{n} \begin{bmatrix} \bar{m}_{1,n}^I(\boldsymbol{\vartheta}_0) \\ \bar{m}_{2,n}^I\big(\boldsymbol{\vartheta}_0, \widehat{\mathbf{q}}(\boldsymbol{\vartheta}_0)\big) \\ \bar{m}_n^E\big(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\vartheta}_0)\big) \end{bmatrix} \to_p \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & B^I(\boldsymbol{\vartheta}_0, \boldsymbol{q}_0) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} & B^E(\boldsymbol{\vartheta}_0, \boldsymbol{\gamma}_0) \end{bmatrix} \sqrt{n} \begin{bmatrix} \bar{m}_{1,n}^I(\boldsymbol{\vartheta}_0) \\ \bar{m}_{2,n}^I(\boldsymbol{\vartheta}_0, \mathbf{q}_0) \\ \bar{m}_n^E(\boldsymbol{\vartheta}_0, \boldsymbol{\gamma}_0) \\ \bar{h}_n^I(\boldsymbol{\vartheta}_0, \mathbf{q}_0) \\ \bar{h}_n^E(\boldsymbol{\vartheta}_0, \boldsymbol{\gamma}_0) \end{bmatrix}$$

where we define $B^I(\boldsymbol{\vartheta}, \boldsymbol{q}) = (1 - \alpha_0 - \alpha_1)\,[diag(\mathbf{a})]^{-1}\,\mathbf{q}$ *and* $B^E(\boldsymbol{\vartheta}, \boldsymbol{\gamma}) = -M^E(H^E)^{-1}$ *with* $\mathbf{a}' = (\alpha_1, \alpha_1, 1-\alpha_1, 1-\alpha_1, \alpha_1, \alpha_1, 1-\alpha_1, 1-\alpha_1)$, *and*

$$M^E = \begin{bmatrix} 0 & -\mathbb{E}[z_i] & 0 & \left(\frac{\partial \psi_2}{\partial \theta_1}\right)' \mathbb{E}[\mathbf{w}_i z_i] \\ 0 & 0 & -\mathbb{E}[z_i] & \left(\frac{\partial \psi_3}{\partial \theta_1}\right)' \mathbb{E}[\mathbf{w}_i z_i] \end{bmatrix}, \; H^E = \begin{bmatrix} -1 & 0 & 0 & \left(\frac{\partial \psi_1}{\partial \theta_1}\right)' \mathbb{E}[\mathbf{w}_i] \\ 0 & -1 & 0 & \left(\frac{\partial \psi_2}{\partial \theta_1}\right)' \mathbb{E}[\mathbf{w}_i] \\ 0 & 0 & -1 & \left(\frac{\partial \psi_3}{\partial \theta_1}\right)' \mathbb{E}[\mathbf{w}_i] \\ -\mathbb{E}[z_i] & 0 & 0 & \left(\frac{\partial \psi_1}{\partial \theta_1}\right)' \mathbb{E}[\mathbf{w}_i z_i] \end{bmatrix}.$$

Lemma 3.1 relates the sample analogues $\bar{m}_{2,n}^I$ and $\bar{m}_n^E$ evaluated at the preliminary estimators $\widehat{\mathbf{q}}(\boldsymbol{\vartheta}_0)$ and $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\vartheta}_0)$ to their counterparts evaluated at the true parameter values $\mathbf{q}_0$ and $\boldsymbol{\gamma}_0$. The estimator $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\vartheta}_0)$ exactly solves $h_n^E(\boldsymbol{\vartheta}_0, \boldsymbol{\gamma}) = 0$ while $\widehat{q}(\boldsymbol{\vartheta}_0)$, constructed as described immediately before the statement of the Lemma, *approximately* solves $\bar{h}_n^I(\boldsymbol{\vartheta}_0, \mathbf{q}) = 0$. A few lines of matrix algebra show that the determinant of H^E equals $\mathrm{Cov}(z, T)$. Hence, B^E is well-defined if z is a relevant instrument. The matrix B^I is likewise well-defined provided that $\alpha_1 \neq 0$ and the elements of $\mathbf{q}_0$ are computed for probabilities strictly between zero and one. If either of these conditions fails, however, some of the moment inequalities in m_2^I are trivially satisfied and can be dropped (see Footnote 15). After removing the corresponding elements of $\mathbf{q}_0$ and $\mathbf{a}$, B^I becomes well-defined. The regularity conditions required for Lemma 3.1 are mild. The result relies on a number of mean-value expansions: $\bar{h}_n^E(\boldsymbol{\vartheta}_0, \boldsymbol{\gamma}_0)$ and $\bar{m}_n^E(\boldsymbol{\vartheta}_0, \boldsymbol{\gamma}_0)$ are expanded around $\boldsymbol{\gamma} = \widehat{\boldsymbol{\gamma}}(\boldsymbol{\vartheta}_0)$ while $\mathbb{E}[h^I(\mathbf{w}_i, \boldsymbol{\vartheta}_0, \mathbf{q}_0)]$ and $\mathbb{E}[m_2^I(\mathbf{w}_i, \boldsymbol{\vartheta}_0, \mathbf{q}_0)]$ are expanded around $\mathbf{q} = \widehat{\mathbf{q}}(\boldsymbol{\vartheta}_0)$. These expansions, in turn, rely on the fact that $\mathbf{q}$ and $\boldsymbol{\gamma}$ are interior to their respective parameter spaces and the relevant functions are all continuously differentiable in our example.

We now have all the ingredients required to construct an asymptotic variance matrix for the GMS test that accounts for preliminary estimation of $\boldsymbol{\gamma}$ and $\mathbf{q}$. Let $m' = (m_1^{I'}, m_2^{I'}, m^{E'})$, $h' = (h^{I'}, h^{E'})$, and define the shorthand $\boldsymbol{\tau}_0' = (\boldsymbol{\gamma}_0', \mathbf{q}_0')$ and $\widehat{\boldsymbol{\tau}}_0' = \big(\widehat{\boldsymbol{\gamma}}'(\boldsymbol{\vartheta}_0), \widehat{\mathbf{q}}'(\boldsymbol{\vartheta}_0)\big)$. Given a collection of iid observations $(\mathbf{w}_1, \ldots, \mathbf{w}_n)$, we have

$$\sqrt{n} \begin{bmatrix} \bar{m}_n(\boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0) \\ \bar{h}_n(\boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0) \end{bmatrix} \to_d N\big(\mathbf{0}, \mathcal{V}(\boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0)\big), \quad \mathcal{V}(\boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0) = \mathrm{Var} \begin{bmatrix} m(\mathbf{w}_i, \boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0) \\ h(\mathbf{w}_i, \boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0) \end{bmatrix} \tag{21}$$

under $H_0\colon \boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0$, by an appropriate central limit theorem. What we require for the test, however, is the asymptotic variance matrix of $\sqrt{n}\,\bar{m}_n(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)$. Combining Equation 21 with

Lemma 3.1, we obtain

$$\text{Avar}\big(\sqrt{n}\,\bar{m}_n(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)\big) = \Xi(\boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0)\,\mathcal{V}(\boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0)\,\Xi'(\boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0) \tag{22}$$

with

$$\Xi(\boldsymbol{\vartheta}, \boldsymbol{\tau}) = \left[\begin{array}{cc} \mathbf{I} & B(\boldsymbol{\vartheta}, \boldsymbol{\tau}) \end{array}\right], \quad B(\boldsymbol{\vartheta}, \boldsymbol{\tau}) = \left[\begin{array}{cc} \mathbf{0} & \mathbf{0} \\ B^I(\boldsymbol{\vartheta}, \mathbf{q}) & \mathbf{0} \\ \mathbf{0} & B^E(\boldsymbol{\vartheta}, \boldsymbol{\gamma}) \end{array}\right] \tag{23}$$

where $B^I(\cdot,\cdot)$ and $B^E(\cdot,\cdot)$ are defined in Lemma 3.1. Finally, we construct a consistent estimator $\widehat{\Sigma}_n(\boldsymbol{\vartheta}_0)$ of the asymptotic variance matrix of $\sqrt{n}\,\bar{m}_n(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)$ under the null:

$$\widehat{\Sigma}_n(\boldsymbol{\vartheta}_0) \equiv \Xi(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)\,\widehat{\mathcal{V}}_n(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}})\,\Xi'(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0) \tag{24}$$

where

$$\widehat{\mathcal{V}}_n(\boldsymbol{\vartheta}, \boldsymbol{\tau}) \equiv \frac{1}{n}\sum_{i=1}^{n} \left[\begin{array}{c} m(\mathbf{w}_i, \boldsymbol{\vartheta}, \boldsymbol{\tau}) - \bar{m}_n(\boldsymbol{\vartheta}, \boldsymbol{\tau}) \\ h(\mathbf{w}_i, \boldsymbol{\vartheta}, \boldsymbol{\tau}) - \bar{h}_n(\boldsymbol{\vartheta}, \boldsymbol{\tau}) \end{array}\right] \left[\begin{array}{c} m(\mathbf{w}_i, \boldsymbol{\vartheta}, \boldsymbol{\tau}) - \bar{m}_n(\boldsymbol{\vartheta}, \boldsymbol{\tau}) \\ h(\mathbf{w}_i, \boldsymbol{\vartheta}, \boldsymbol{\tau}) - \bar{h}_n(\boldsymbol{\vartheta}, \boldsymbol{\tau}) \end{array}\right]'. \tag{25}$$

In the following section we provide a step-by-step description of our inference procedure.

3.5 Details of the Inference Procedure

In this section we provide full details of our Bonferroni-based inference procedure. We begin by defining some notation. Let J denote the total number of inequality moment conditions, K denote the total number of equality moment conditions, and define

$$\bar{m}_n(\boldsymbol{\vartheta}, \boldsymbol{\tau}) = \left[\begin{array}{c} \bar{m}_n^I(\boldsymbol{\vartheta}, \mathbf{q}) \\ \bar{m}_n^E(\boldsymbol{\vartheta}, \boldsymbol{\gamma}) \end{array}\right] = \left[\begin{array}{c} \bar{m}_{n,1}^I(\boldsymbol{\vartheta}) \\ \bar{m}_{n,2}^I(\boldsymbol{\vartheta}, \mathbf{q}) \\ \bar{m}_n^E(\boldsymbol{\vartheta}, \boldsymbol{\gamma}) \end{array}\right] = \frac{1}{n}\sum_{i=1}^{n} \left[\begin{array}{c} m_1^I(\mathbf{w}_i, \boldsymbol{\vartheta}) \\ m_2^I(\mathbf{w}_i, \boldsymbol{\vartheta}, \mathbf{q}) \\ m^E(\mathbf{w}_i, \boldsymbol{\vartheta}, \boldsymbol{\gamma}) \end{array}\right] \tag{26}$$

with m_1^I as defined in Equation 10, m_2^I in Equations 11–13 and m^E in Equation 15.[17] Now let S be the function

$$S(\mathbf{x}, \mathbf{y}) = \sum_j \min\left\{0, x_j^2\right\} + \mathbf{y}'\mathbf{y} \tag{27}$$

where $\mathbf{x}, \mathbf{y}$ are two finite-dimensional real vectors and x_j denotes the j^{th} element of $\mathbf{x}$. This function will be used to calculate the modified method of moments (MMM) test statistic as part of the GMS test below. The argument $\mathbf{x}$ stands in for the moment inequalities, which only contribute to the test statistic when they are violated, i.e. take on a *negative* value. Using this notation, we now detail the first step of our inference procedure: a GMS test for $\boldsymbol{\vartheta} = (\alpha_0, \alpha_1)$ with preliminary estimation of $\mathbf{q}$ and $\boldsymbol{\gamma}$ under the null.

Algorithm 3.1 (GMS Test for α_0 and α_1).
Inputs: hypothesis $\boldsymbol{\vartheta}_0$; iid dataset $\{\mathbf{w}_i\}_{i=1}^n$; simulations $\{\boldsymbol{\zeta}^{(r)}\}_{r=1}^R \sim$ iid $N_{J+K}(\mathbf{0}, \mathbf{I})$.

[17] In our problem $K = 2$ and J is at most 12. Under certain nulls for (α_0, α_1), however, we drop components of m_2^I as they are trivially satisfied. See footnote 15 and Section 3.4 for further discussion.

1. Calculate the variance matrix estimator $\widehat{\Sigma}_n(\boldsymbol{\vartheta}_0)$.

   (i) Calculate $\widehat{\boldsymbol{\tau}}_0 = (\widehat{\mathbf{q}}_0', \widehat{\boldsymbol{\gamma}}_0')'$ where $\widehat{\boldsymbol{\gamma}}_0 = \widehat{\boldsymbol{\gamma}}(\boldsymbol{\vartheta}_0)$ and $\widehat{\mathbf{q}}_0 = \widehat{\mathbf{q}}(\boldsymbol{\vartheta}_0)$ from Section 3.4.

   (ii) Calculate $\Xi(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)$ using Equation 23.

   (iii) Calculate $\widehat{\mathcal{V}}_n(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)$ using Equation 25.

   (iv) Set $\widehat{\Sigma}_n(\boldsymbol{\vartheta}_0) = \Xi(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)\, \widehat{\mathcal{V}}_n(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}})\, \Xi'(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)$.

2. Calculate the test statistic $T_n(\boldsymbol{\vartheta}_0)$.

   (i) Calculate $\sqrt{n}\, \bar{m}_n(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)$ using Equation 26.

   (ii) Set $\boldsymbol{\nu}_n(\boldsymbol{\vartheta}_0) = \left[\text{diag}\left\{\widehat{\Sigma}_n(\boldsymbol{\vartheta}_0)\right\}\right]^{-1/2} [\sqrt{n}\, \bar{m}_n(\boldsymbol{\vartheta}_0, \widehat{\boldsymbol{\tau}}_0)]$.

   (iii) Let $\boldsymbol{\nu}_n^I(\boldsymbol{\vartheta}_0)$ denote the first J elements of $\boldsymbol{\nu}_n$ and $\boldsymbol{\nu}_n^E(\boldsymbol{\vartheta}_0)$ the last K elements.

   (iv) Set $T_n(\boldsymbol{\vartheta}_0) = S\big(\boldsymbol{\nu}_n^I(\boldsymbol{\vartheta}_0),\, \boldsymbol{\nu}_n^E(\boldsymbol{\vartheta}_0)\big)$ using Equation 27.

3. Construct the moment selection matrix Φ.

   (i) For $j = 1, \ldots, J$ set $\varphi_j^I = \mathbf{1}\left\{\nu_{n,j}^I(\boldsymbol{\vartheta}_0) \leq \sqrt{\log n}\right\}$ and let $\widetilde{J} = \sum_{j=1}^J \varphi_j^I$.

   (ii) For $j = 1, \ldots, K$ set $\varphi_j^E = 1$.

   (iii) Set $\boldsymbol{\varphi} = (\varphi_1^I, \ldots, \varphi_J^I, \varphi_1^E, \ldots, \varphi_K^E)'$.

   (iv) Let Φ be the $(\widetilde{J} + K) \times (J + K)$ of zeros and ones that selects those elements x_j of an arbitrary vector $\mathbf{x}$ that correspond to $\varphi_j = 1$.

4. Simulate the sampling distribution of $T_n(\boldsymbol{\vartheta}_0)$ under the null.

   (i) Let $\widehat{\Omega}$ be the correlation matrix that corresponds to $\widehat{\Sigma}_n(\boldsymbol{\vartheta}_0)$.

   (ii) For each $r = 1, \ldots, R$ set $\boldsymbol{\xi}^{(r)} = \left[\Phi\, \widehat{\Omega}\, \Phi'\right]^{1/2} \Phi \boldsymbol{\zeta}^{(r)}$.

   (iii) Let $\boldsymbol{\xi}_I^{(r)}$ denote the first $\widetilde{J}$ and $\boldsymbol{\xi}_E^{(r)}$ the last K elements of $\boldsymbol{\xi}^{(r)}$.

   (iv) For each $r = 1, \ldots, R$ set $T_n^{(r)}(\boldsymbol{\vartheta}_0) = S\big(\boldsymbol{\xi}_I^{(r)},\, \boldsymbol{\xi}_E^{(r)}\big)$ using Equation 27.

5. Calculate the p-value of the test: $\widehat{p}(\boldsymbol{\vartheta}_0) = \dfrac{1}{R} \displaystyle\sum_{r=1}^{R} \mathbf{1}\left\{T_n^{(r)}(\boldsymbol{\vartheta}_0) > T_n(\boldsymbol{\vartheta}_0)\right\}$.

Algorithm 3.1 corresponds to the asymptotic version of the GMS test from Andrews and Soares (2010), based on the MMM test statistic – S_1 in Andrews and Soares (2010) – and the "BIC choice" $\kappa_n = \sqrt{\log n}$ for the sequence of constants κ_n used for moment selection. The procedure is as follows. In Step 1, we compute a consistent estimator of the asymptotic variance matrix of the full set of moment conditions, under the null, accounting for preliminary estimation of $\mathbf{q}$ and $\boldsymbol{\gamma}$ as explained in Section 3.4. In step 2, we calculate the observed value of the MMM test statistic. Note that this test statistic uses only the diagonal elements of $\widehat{\Sigma}_n(\boldsymbol{\vartheta}_0)$. Moreover, the moment inequalities only contribute to T_n if

they are violated, i.e. if they take on a negative value. In step 3 we determine which moment inequalities are "far from binding," defined as having a t-ratio greater than $\sqrt{\log n}$. These moment inequalities will be excluded when approximating the large-sample distribution of the test statistic. The matrix Φ encodes the results of the moment selection step. Pre–multiplying a $(J+K)$–vector $\mathbf{x}$ by Φ results in a $(\widetilde{J} \times K)$–vector $\widetilde{\mathbf{x}}$ whose last K elements match the last K elements of $\mathbf{x}$ but whose first $\widetilde{J}$ elements contain the subset of $(x_1, \ldots, x_J)$ whose indices match those of the moment inequalities with t-ratios less than or equal to $\sqrt{\log n}$, i.e. those that are *not* far from binding.[18] Step 4 uses a collection of iid normal draws, $\{\boldsymbol{\zeta}^{(r)}\}_{r=1}^{R}$, to approximate the large-sample distribution of T_n under the null. The appropriate multiplications by Φ ensure that this approximation includes all moment equalities, but excludes any moment inequality judged to be far from binding in step 3. Finally, step 5 computes the p-value of the test by comparing the actual test statistic $T_n(\boldsymbol{\vartheta}_0)$ to the collection of simulated test statistics $\{T_n^{(r)}(\boldsymbol{\vartheta}_0)\}_{r=1}^{R}$ from step 4. We now detail our Bonferroni-based confidence interval for β.[19]

Algorithm 3.2 (Bonferroni-based Confidence Interval for β)**.**
Inputs: significance levels (δ_1, δ_2); iid dataset $\{\mathbf{w}_i\}_{i=1}^{n}$; simulations $\{\boldsymbol{\zeta}^{(r)}\}_{r=1}^{R} \sim$ iid $N_{J+K}(\mathbf{0}, \mathbf{I})$.

1. Construct a $(1-\delta_1) \times 100\%$ joint confidence set $\mathcal{C}(\delta_1)$ for $\boldsymbol{\vartheta} = (\alpha_0, \alpha_1)'$.
   (i) Let $\Lambda_N = \left\{0, \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-2}{N}, \frac{N-1}{N}\right\}$ where $N > 1$ is a natural number.
   (ii) Set $\Delta_N = \{(\alpha_0, \alpha_1) \in (\Lambda_N \times \Lambda_N) \colon \alpha_0 + \alpha_1 < 1\}$.
   (iii) For each $\boldsymbol{\vartheta} \in \Delta_N$ calculate $\widehat{p}(\boldsymbol{\vartheta})$ using Algorithm 3.1, holding $\{\boldsymbol{\zeta}^{(r)}\}_{r=1}^{R}$ *fixed*.
   (iv) Set $\mathcal{C}(\delta_1) = \{\boldsymbol{\vartheta} \in \Delta_N \colon \widehat{p}(\boldsymbol{\vartheta}) \geq \delta_1\}$.
2. Construct a $(1-\delta_1) \times 100\%$ confidence interval $[\underline{s}(\delta_1), \overline{s}(\delta_1)]$ for $s \equiv (1 - \alpha_0 - \alpha_1)$.
   (i) Set $\underline{s}(\delta_1) = \min\{(1-\alpha_0-\alpha_1) \colon (\alpha_0, \alpha_1) \in \mathcal{C}(\delta_1)\}$.
   (ii) Set $\overline{s}(\delta_1) = \max\{(1-\alpha_0-\alpha_1) \colon (\alpha_0, \alpha_1) \in \mathcal{C}(\delta_1)\}$.
3. Construct a $(1-\delta_2) \times 100\%$ confidence interval $\left[\underline{\theta}_1(\delta_2), \overline{\theta}_1(\delta_2)\right]$ for θ_1.
   (i) Use the standard IV interval from a regression of y on T with instrument z.
4. Construct the $(1-\delta) \times 100\%$ Bonferroni-based confidence interval $\left[\underline{\beta}(\delta), \overline{\beta}(\delta)\right]$ for β.
   (i) Let $\delta = \delta_1 + \delta_2$.
   (ii) Set $\underline{\beta}(\delta) = \min\{\underline{s}(\delta_1) \times \underline{\theta}_1(\delta_2),\ \overline{s}(\delta_1) \times \underline{\theta}_1(\delta_2)\}$.

[18] Although this does not affect the results of the procedure, notice that Algorithm 3.1 carries out moment selection in a slightly different way from the steps given by Andrews and Soares (2010). In particular, before carrying out any further calculations, we *subset* the correlation matrix $\widehat{\Omega}$ and normal vectors $\boldsymbol{\zeta}^{(r)}$ to remove elements corresponding to moment inequalities deemed far from binding. In contrast, Andrews and Soares (2010) carry along the full set of inequalities throughout, but add $+\infty$ to the appropriate elements when computing $T_n^{(r)}$ to ensure that only the moment inequalities that are not far from binding affect the results. Although it requires more notation to describe, sub-setting is substantially faster, as it avoids carrying out computations for inequalities that cannot affect the result.

[19] Code implementing this procedure is available at https://github.com/fditraglia/mbereg.

(iii) Set $\overline{\beta}(\delta) = \max\left\{\underline{s}(\delta_1) \times \overline{\theta}_1(\delta_2),\, \overline{s}(\delta_1) \times \overline{\theta}_1(\delta_2)\right\}$.

Step 1 of Algorithm 3.2 constructs a $(1-\delta_1) \times 100\%$ joint confidence set $\mathcal{C}(\delta_1)$ for $\boldsymbol{\vartheta} = (\alpha_0, \alpha_1)$ by inverting the GMS test from Algorithm 3.1 over a discretized parameter space Δ_N. Because the parameter space for (α_0, α_1) is bounded, this is computationally straightforward. Note that the *same* normal draws $\{\boldsymbol{\zeta}^{(r)}\}_{r=1}^R$ are used to test each null hypothesis contained in Δ_N. Step 2 projects $\mathcal{C}(\delta_1)$ to yield a $(1-\delta_1) \times 100\%$ confidence interval for $s \equiv (1 - \alpha_0 - \alpha_1)$, simply taking the maximum and minimum values of s in the discrete set $\mathcal{C}(\delta_1)$. Step 3 constructs the usual IV confidence interval for the reduced form parameter θ_1, and step 4 combines the results of steps 2–3 with Bonferroni's inequality to yield a $(1-\delta_1-\delta_2) \times 100\%$ confidence interval for β. For some discussion of alternatives to Algorithm 3.2, see Footnote 14. Notice that, by construction, the Bonferroni interval for β excludes zero if and only if the confidence interval for θ_1 from step 3 of Algorithm 3.2 excludes zero. Under mild regularity conditions, the confidence interval from Algorithm 3.2 is uniformly asymptotically valid.

Theorem 3.1. *Let $\mathbf{w}_1, \ldots, \mathbf{w}_n$ be an iid collection of observations satisfying the conditions of Theorems 2.2 and 2.3, and let z be a strong instrument. Then, under standard regularity conditions, the confidence interval for β from Algorithm 3.2 has asymptotic coverage probability no less than $1 - (\delta_1 + \delta_2)$ as $R, N, n \to \infty$ uniformly over the parameter space.*

Theorem 3.1 is effectively a corollary of Theorem 1 from Andrews and Soares (2010), which establishes the uniform asymptotic validity of the GMS test, and Lemma 3.1, which accounts for preliminary estimation of $\boldsymbol{\gamma}$ and $\mathbf{q}$. Given iid observations $\mathbf{w}_i$, the only substantive condition required for Theorem 3.1 is the joint asymptotic normality of $\sqrt{n}\,\bar{m}_n(\boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0)$ and $\sqrt{n}\bar{h}_n(\boldsymbol{\vartheta}_0, \boldsymbol{\tau}_0)$, where $\bar{h}_n$ denotes the sample analogues for the full set of auxiliary moment conditions (h^I, h^E) defined in Section 3.4. For further discussion of the regularity conditions required for the GMS procedure, see Appendix A3 of Andrews and Soares (2010). For some discussion of the regularity conditions required for Lemma 3.1, see Section 3.4.

Theorem 3.1 invokes the higher-moment assumptions (Assumptions 2.5–2.6) under which we establish global identification of β in Theorem 2.3, and Algorithm 3.1 likewise incorporates the higher-moment equality conditions that arise from this result. To proceed without these conditions, simply remove m^E from the set of moment conditions used in the algorithm and leave the steps unchanged. In this case β is no longer point identified but the inference procedure provides valid inference for the points in the sharp identified set from Theorem 2.2. Algorithm 3.2 can likewise be used in the case of an exogenous T^*, as in Mahajan (2006) and Frazis and Loewenstein (2003). As mentioned above in Section 3.1, the exogenous regressor case is subject to the same inferential difficulties as the endogenous case on which we focus in this paper. To accommodate an exogenous regressor, simply replace m^E with the moment equalities described in Appendix D.

3.6 Further Details Regarding Covariates

The inference procedure described in the preceding sections holds $\mathbf{x}$ fixed, and is thus appropriate for examples with discrete covariates. To accommodate covariates more generally,

there are several possible approaches. At one extreme, suppose one were willing to assume that (α_0, α_1) did not vary with $\mathbf{x}$ and that $y = c + \beta T^* + \mathbf{x}'\boldsymbol{\phi} + \varepsilon$, as in Frazis and Loewenstein (2003). In this case, the standard IV estimator identifies $\boldsymbol{\phi}$ and one could simply augment the moment equalities m^E from Equation 15 above to provide a preliminary estimator of $\boldsymbol{\phi}$ in Algorithm 3.1. At the other extreme, if one wished to remain fully non-parametric, one could adopt the approach of Andrews and Shi (2014), based on kernel averaging near a fixed covariate value $\mathbf{x} = \mathbf{x}_0$. As a compromise between these two extremes, one could alternatively specify a semi-parametric model, perhaps along the lines of Section 4 of Lewbel (2007), and follow the approach of Andrews and Shi (2013). Both of these latter possibilities could be an interesting extension of the method described above.

4 Simulation Study

In this section we present results from a simulation study using the inference procedure described in Section 3.5 above. Unless otherwise specified, all calculations are based on 2000 simulation replications with $n = 1000$ using Algorithm 3.2 with $R = 5000$ simulation draws. Supplementary simulation results appear in Appendix E.

4.1 Simulation DGP

Our simulation design generates n iid draws of the observables (y_i, T_i, z_i) as follows:

1. Generate the instrumental variable z.
   (i) For each $1 \leq i \leq n/2$ set $z_i = 0$.
   (ii) For each $n/2 < i \leq n$, set $z_i = 1$.
2. Generate the error terms:
$$\begin{bmatrix} \eta_i \\ \varepsilon_i \end{bmatrix} \sim \text{iid N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right).$$
3. Generate the unobserved regressor: $T_i^* = \mathbf{1}\{d_0 + d_1 z_i + \eta_i > 0\}$.
4. Generate the outcome: $y_i = c + \beta T_i^* + \varepsilon_i$.
5. Generate the observed, mis-classified regressor T.
   (i) For all i with $T_i^* = 0$ draw $T_i \sim$ iid Bernoulli(α_0).
   (ii) For all i with $T_i^* = 1$ draw $T_i \sim$ iid Bernoulli$(1 - \alpha_1)$.

This DGP generates random variables that satisfy the conditions of Theorems 2.2 and 2.3. Thus β is point identified, and all moment equalities and inequalities from Section 3 hold at the true parameter values of the DGP. Note from step 1 that we *condition* on the instrument z, holding it fixed in repeated samples. Our simulation varies the parameters $(\alpha_0, \alpha_1, \beta, n)$ over a grid. Because ε has unit variance, values for β are measured in standard deviations of

the error. For simplicity we present results for $c = 0, d_0 = \Phi^{-1}(0.15)$, and $d_1 = \Phi^{-1}(0.85) - \Phi^{-1}(0.15)$, where $\Phi^{-1}(\cdot)$ denotes the quantile function of a standard normal random variable. Using these values for (d_0, d_1) holds the unobserved first stage probabilities fixed: $p_0^* = 0.15$ and $p_1^* = 0.85$. In contrast the *observed* first-stage probabilities p_0 and p_1 vary with (α_0, α_1) according to Lemma 2.1.

4.2 Simulation Results

As explained in Section 3.1 above, the just-identified, unconstrained GMM estimator based on Equation 9 suffers from weak identification and boundary value problems. Moreover, the estimator may not even exist in finite samples. Even when the GMM estimator exists, its asymptotic variance matrix could be numerically singular, so that the standard GMM confidence interval is undefined. Table 1 reports the percentage of simulation draws for which the standard GMM confidence interval is undefined, while Table 2 reports the coverage probability of a nominal 95% GMM confidence interval, conditional on its existence.

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	27	33	30	14	1	0	0	0
	0.1	27	32	29	13	2	0	0	0
	0.2	26	33	32	15	4	0	0	0
	0.3	26	34	30	17	5	0	0	0
0.1	0.0	26	32	31	14	2	0	0	0
	0.1	26	36	32	16	4	0	0	0
	0.2	27	35	31	18	8	0	0	0
	0.3	25	35	32	21	11	1	0	0
0.2	0.0	26	33	30	15	3	0	0	0
	0.1	26	33	30	19	6	0	0	0
	0.2	26	35	33	22	12	1	0	0
	0.3	26	35	33	26	15	3	0	0
0.3	0.0	26	32	32	16	6	0	0	0
	0.1	24	35	33	21	11	1	0	0
	0.2	26	32	35	27	15	4	0	0
	0.3	26	35	35	28	21	7	2	0

Table 1: Percentage of replications for which the standard GMM confidence interval based on Equation 9 fails to exist, either because the point estimate is NaN or the asymptotic covariance matrix is numerically singular. Calculations are based on 2000 replications of the DGP from 4.1 with $n = 1000$.

We see from Table 1 that when β is small compared to the error variance, the GMM confidence interval fails to exist with high probability. When $\beta = 0.5$, for example, the interval is undefined approximately 30% of the time. As β increases, however, it becomes less likely that the GMM interval is undefined. All else equal, larger amounts of mis-classification, i.e. higher values for (α_0, α_1), increase the probability that the GMM interval fails to exist. Turning our attention to the simulation draws for which it is well-defined, we see from

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	72	62	62	80	92	95	94	95
	0.1	72	62	63	79	92	95	96	95
	0.2	73	61	61	77	90	96	96	96
	0.3	73	59	62	76	88	95	96	95
0.1	0.0	73	63	60	78	91	95	96	96
	0.1	73	58	59	77	90	95	95	94
	0.2	73	59	61	75	86	95	95	94
	0.3	74	59	58	71	82	94	96	96
0.2	0.0	74	62	60	78	91	95	96	96
	0.1	73	60	61	74	87	95	96	94
	0.2	73	58	57	70	81	93	95	95
	0.3	73	58	56	66	78	92	95	96
0.3	0.0	74	62	60	76	89	95	96	96
	0.1	75	59	58	71	82	93	96	95
	0.2	74	61	56	65	78	90	96	96
	0.3	73	58	55	64	71	88	93	96

Table 2: Coverage (%) of the standard nominal 95% GMM confidence interval for β based on Equation 9. Coverage is calculated only for those simulation draws for which the interval exists. (See Table 1.) Calculations are based on 2000 replications of the DGP from 4.1 with $n = 1000$.

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	19.07	3.44	1.86	1.32	0.87	0.47	0.37	0.35
	0.1	17.52	3.47	1.92	1.41	1	0.61	0.51	0.46
	0.2	17.41	3.51	1.9	1.45	1.1	0.76	0.65	0.58
	0.3	18.23	3.34	1.92	1.48	1.24	0.91	0.79	0.7
0.1	0.0	17.13	3.51	1.86	1.38	0.97	0.61	0.51	0.46
	0.1	17.88	3.33	1.85	1.45	1.13	0.78	0.67	0.6
	0.2	17.37	3.36	1.95	1.54	1.24	0.97	0.85	0.75
	0.3	18.07	3.33	1.98	1.63	1.41	1.17	1.04	0.92
0.2	0.0	17.79	3.39	1.92	1.45	1.11	0.75	0.65	0.58
	0.1	18.98	3.43	1.96	1.54	1.26	0.97	0.84	0.75
	0.2	18.25	3.26	1.92	1.64	1.45	1.2	1.06	0.95
	0.3	19.03	3.31	2.02	1.75	1.66	1.49	1.33	1.19
0.3	0.0	18.27	3.48	1.87	1.5	1.25	0.9	0.79	0.7
	0.1	19.4	3.41	1.96	1.63	1.43	1.18	1.04	0.92
	0.2	18.22	3.56	1.96	1.74	1.67	1.49	1.35	1.19
	0.3	17.56	3.55	2.13	1.96	1.86	1.86	1.74	1.55

Table 3: Median width of the standard nominal 95% GMM confidence interval for β based on Equation 9. Coverage is calculated only for those simulation draws for which the interval exists. Calculations are based on 2000 replications of the DGP from 4.1 with $n = 1000$.

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	97.7	97.7	97.6	97.7	98.0	98.0	97.4	97.9
	0.1	98.0	98.7	98.8	99.1	98.8	98.4	97.1	96.4
	0.2	98.4	98.5	98.9	98.9	98.8	98.6	98.0	97.0
	0.3	98.5	98.8	98.8	99.0	98.7	98.4	97.8	97.5
0.1	0.0	98.1	98.5	98.3	98.8	98.8	98.4	96.8	95.7
	0.1	98.6	99.1	99.5	99.6	99.6	98.8	97.7	95.2
	0.2	99.0	99.3	99.7	99.8	99.7	98.9	97.5	95.7
	0.3	99.4	99.7	99.8	99.8	99.6	99.0	98.2	96.7
0.2	0.0	98.6	98.5	98.6	98.9	98.7	98.2	97.7	97.0
	0.1	99.0	99.5	99.7	99.7	99.4	99.0	98.1	96.5
	0.2	99.5	99.7	99.8	99.7	99.4	99.0	97.8	96.8
	0.3	99.7	99.8	99.8	99.8	99.5	99.0	98.7	97.7
0.3	0.0	98.7	98.7	98.8	98.7	98.7	98.2	98.1	97.6
	0.1	99.4	99.6	99.6	99.7	99.4	98.9	98.3	96.8
	0.2	99.8	99.8	99.7	99.8	99.5	99.1	98.5	97.8
	0.3	100.0	99.9	99.9	99.8	99.6	99.5	99.1	98.8

Table 4: Coverage probability (1 - size) in percentage points of a 97.5% GMS joint test for α_0 and α_1 using Algorithm 3.1 with $n = 1000$. Calculations are based on 10,000 replications of the DGP from Section 4.1.

Tables 2 and 3 that the GMM confidence interval performs extremely poorly when β is small. Substantial size distortions persist until β is 1.5 or larger. All else equal, the size distortions are more severe the larger the amount of mis-classification error. For sufficiently large β, however, standard GMM inference performs well. As β grows, the weak identification problem vanishes. For large enough β the inference problem in effect becomes standard.

We now examine the performance of the Bonferroni-based confidence interval from Algorithm 3.2, beginning with its first step: a joint GMS confidence set for (α_0, α_1). Table 4 presents coverage probabilities for a nominal 97.5% GMS confidence set for (α_0, α_1). Because these results are extremely fast to compute, Table 4 is based on 10,000 simulation replications. Aside from some slight under-coverage at intermediate values of (α_0, α_1) when $\beta = 3$, the GMS interval makes good on its promise of uniformly valid inference. As shown in Appendix E, the under-coverage problem appears to be a finite-sample artifact: if we increase n to 2000, the maximum size distortion becomes negligible. The GMS test tends, however, to be fairly conservative, particularly for larger values of (α_0, α_1). When there is no mis-classification error, the GMS confidence sets are very nearly exact. Results for nominal 95% and 90% intervals are qualitatively similar: see Appendix E.

We now present results for the Bonferroni interval from Algorithm 3.2, setting $\delta_1 = \delta_2 = 0.025$ to yield an interval with asymptotic coverage no less that 95%.[20] Table 5 presents coverage probabilities in percentage points and Table 5 presents median widths.

[20]In principle, one could optimize the choice of δ_1 subject to the constraint $\delta_1 + \delta_2 = 0.95$ to reduce the width of the resulting interval. In our experiments, there was no choice of δ_1 that uniformly dominated for all values of $(\alpha_0, \alpha_1, \beta)$ so we report only results for $\delta_1 = \delta_2$ here.

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	96	97	97	96	97	97	95	96
	0.1	97	99	99	99	99	100	100	99
	0.2	98	99	99	100	100	100	100	100
	0.3	97	100	100	100	100	100	100	100
0.1	0.0	97	99	99	99	100	100	100	98
	0.1	98	100	100	100	100	100	100	100
	0.2	98	100	100	100	100	100	100	100
	0.3	97	100	100	100	100	100	100	100
0.2	0.0	97	99	99	100	100	100	100	100
	0.1	98	100	100	100	100	100	100	100
	0.2	98	100	100	100	100	100	100	100
	0.3	98	100	100	100	100	100	100	100
0.3	0.0	97	99	100	100	100	100	100	100
	0.1	97	100	100	100	100	100	100	100
	0.2	98	100	100	100	100	100	100	100
	0.3	98	100	100	100	100	100	100	100

Table 5: Coverage probability in percentage points of a nominal $> 95\%$ Bonferroni confidence interval for β using Algorithm 3.2 with $n = 1000, R = 5000$ and $\delta_1 = \delta_2 = 0.025$. Calculations are based on 2000 replications of the DGP from Section 4.1.

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	0.4	0.41	0.43	0.43	0.43	0.42	0.41	0.41
	0.1	0.45	0.47	0.54	0.59	0.63	0.7	0.75	0.86
	0.2	0.51	0.54	0.65	0.76	0.85	0.95	1.01	1.17
	0.3	0.58	0.62	0.79	0.95	1.07	1.17	1.24	1.48
0.1	0.0	0.45	0.47	0.54	0.59	0.63	0.7	0.76	0.88
	0.1	0.51	0.54	0.66	0.77	0.86	1.03	1.18	1.46
	0.2	0.58	0.63	0.8	0.98	1.12	1.38	1.55	1.88
	0.3	0.67	0.75	1	1.25	1.46	1.74	1.94	2.4
0.2	0.0	0.51	0.54	0.65	0.76	0.86	0.96	1.02	1.19
	0.1	0.58	0.63	0.81	0.99	1.14	1.42	1.64	2.08
	0.2	0.67	0.75	1.01	1.29	1.54	1.97	2.33	2.9
	0.3	0.81	0.91	1.3	1.7	2.09	2.73	3.13	3.9
0.3	0.0	0.58	0.62	0.8	0.95	1.09	1.18	1.25	1.5
	0.1	0.68	0.74	1.01	1.26	1.49	1.84	2.13	2.78
	0.2	0.81	0.91	1.3	1.7	2.11	2.8	3.4	4.48
	0.3	1.01	1.16	1.74	2.35	2.93	4.17	5.2	6.85

Table 6: Median width of a nominal $> 95\%$ Bonferroni confidence interval for β using Algorithm 3.2 with $n = 1000, R = 5000$ and $\delta_1 = \delta_2 = 0.025$. Calculations are based on 2000 replications of the DGP from Section 4.1.

The Bonferroni interval achieves its stated minimum coverage uniformly over the parameter space. When there is no mis-classification, $\alpha_0 = \alpha_1$, its actual coverage is close or equal to 95%. In the presence of mis-classification, however, the interval can be quite conservative, particularly for larger values of β. For smaller but nonzero values of β, this conservatism reflects the fact that the model is *effectively* partially identified: although Theorem 2.3 shows that (α_0, α_1) are point identified for any $\beta \neq 0$, the amount of data required to distinguish one pair of alphas from another when β is small would be astronomical.

In spite of its conservatism, the Bonferroni interval is informative, as we see from the median widths in Table 6. Because median widths provide only a limited picture of the behavior of a confidence interval, Figures 1–3 present further evidence in the form of coverage functions (1 - power) for $\beta = 0.5, 1, 3$. Coverage curves for additional values of β and n appear in Appendix E. Each figure holds the true value of β fixed and varies (α_0, α_1) over the grid $\{0, 0.1, 0.2\} \times \{0, 0.2, 0.2\}$. The plots within each Figure give coverage in percentage points as a function of the specified alternative for β. Solid curves are computed using the full set of inequality moment conditions from Section 3.3, while dashed curves use only m_1^I, i.e. they do not impose the restrictions implied by non-differential measurement error. In each figure, the dashed horizontal line gives the nominal coverage probability, 95%, while the dashed vertical lines are the reduced from and instrumental variables estimands: for $\beta \geq 0$ the reduced form is always smaller than the IV.

As seen from Figures 1–3, and their counterparts in Appendix E, the Bonferroni procedure has power against the alternative $\beta = 0$, even when the true value of β is small. As described in Section 3.5, the Bonferroni interval excludes zero if and only if the confidence interval for θ_1 from which it is constructed also excludes zero. These figures also indicate the gains from including m_2^I, the moment inequalities that emerge from assuming non-differential measurement error: substantial increases in power against alternatives between the true parameter value and zero, particularly for larger values of β. Note moreover that the excellent performance of Bonferroni in the zero mis-classification case (α_0, α_1) depends crucially on imposing the assumption of non-differential measurement error. As the true value of β increases, the Bonferroni interval begins to have power against both the reduced form and IV estimands.

A drawback of the identification-robust inference procedure from Algorithm 3.2 becomes apparent when both β and the mis-classification probabilities are large. In this case the confidence interval for β is excessively wide, as we see from Table 6 and Figure 3.[21] Note from Tables 1 and 2, that this is a region of the parameter space in which the plain-vanilla GMM confidence interval yields valid inference. Moreover, we see from Table 3 that the median width of the GMM interval is far more reasonable when β is large, even in the presence of large amounts of mis-classification. It is important to stress that the source of this excess width is *not* the Bonferroni correction: the same behavior emerges if one projects a joint GMS confidence set for $(\alpha_0, \alpha_1, \beta)$ to yield marginal inference for β. Rather, it is the inevitable cost of applying a robust inference procedure in a region of the parameter space where standard inference performs well. While a detailed theoretical investigation of this problem is beyond the scope of the present paper, we now explore the performance of a "hybrid" confidence interval that uses a simple heuristic to transition between robust and

[21] As expected, median widths decrease with sample size: see the results for $n = 2000$ in Appendix E.

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	96	97	97	96	97	97	95	93
	0.1	97	99	99	99	99	98	96	95
	0.2	98	99	99	100	100	97	96	96
	0.3	97	100	100	100	99	96	96	96
0.1	0.0	97	99	99	99	100	98	97	95
	0.1	98	100	100	100	100	96	96	96
	0.2	98	100	100	100	99	96	96	95
	0.3	97	100	100	100	97	95	96	96
0.2	0.0	97	99	99	100	100	96	96	96
	0.1	98	100	100	100	99	96	96	96
	0.2	98	100	100	100	96	95	95	96
	0.3	98	100	100	98	95	95	95	96
0.3	0.0	97	99	100	100	100	95	96	97
	0.1	97	100	100	100	97	94	96	96
	0.2	98	100	100	98	94	94	96	96
	0.3	98	100	99	96	92	94	95	96

Table 7: Coverage probabilities (%) of a hybrid confidence interval constructed from the nominal 95% standard GMM interval and the $> 95\%$ Bonferroni confidence interval for β using Algorithm 3.2 with $n = 1000, R = 5000$ and $\delta_1 = \delta_2 = 0.025$. The hybrid interval reports Bonferroni unless the GMM interval exists and is contained within the Bonferroni interval. Calculations are based on 2000 replications of the DGP from Section 4.1.

standard inference.[22] The procedure for constructing the hybrid interval is as follows. First compute the robust confidence interval based on Algorithm 3.2. Next, determine whether the GMM interval is well-defined: if so, determine whether it is contained within the robust interval. If the GMM interval exists and lies within the robust interval, report GMM; otherwise report the robust interval. Table 7 presents coverage probabilities (in percentage points) and Table 8 median widths for the resulting hybrid confidence interval. Coverage plots for $\beta = 1, 2, 3$ appear in Figures 4–6. Plots for additional values of β and n appear in Appendix E. The conventions of these figures are identical to those of Figures 1–3 with one exception: in Figures 4–6 the dashed curves correspond to the hybrid confidence interval.

The hybrid interval performs extremely well: with the exception of a slight size distortion at $(\alpha_0 = \alpha_1 = 0.3, \beta = 1)$ and $(\alpha_0 = \alpha_1 = 0, \beta = 3)$, it is effectively a free lunch.[23] Note in particular that the coverage curves for the hybrid interval from Figures 4–6 (dashed curves) lie uniformly below those of the Bonferroni interval (solid curves) while still maintaining correct coverage at the true value of β. It could be interesting to investigate this idea further in future work.

[22] This idea is related to Andrews (2016), although somewhat different in its details.

[23] The distortion at $(\alpha_0 = \alpha_1 = 0.3, \beta = 1)$ disappears when n increases to 2000: see Appendix E.

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	0.4	0.41	0.43	0.43	0.43	0.42	0.4	0.35
	0.1	0.45	0.47	0.54	0.59	0.63	0.67	0.52	0.46
	0.2	0.51	0.54	0.65	0.76	0.84	0.82	0.65	0.58
	0.3	0.58	0.62	0.79	0.95	1.05	0.96	0.79	0.7
0.1	0.0	0.45	0.47	0.54	0.59	0.63	0.67	0.51	0.46
	0.1	0.51	0.54	0.66	0.77	0.86	0.92	0.69	0.61
	0.2	0.58	0.63	0.8	0.97	1.11	1.17	0.87	0.75
	0.3	0.67	0.75	1	1.25	1.4	1.4	1.06	0.92
0.2	0.0	0.51	0.54	0.65	0.76	0.85	0.83	0.65	0.58
	0.1	0.58	0.63	0.81	0.99	1.12	1.18	0.86	0.75
	0.2	0.67	0.75	1.01	1.29	1.48	1.56	1.08	0.95
	0.3	0.81	0.91	1.3	1.67	1.95	1.77	1.35	1.2
0.3	0.0	0.58	0.62	0.8	0.95	1.07	0.95	0.8	0.7
	0.1	0.68	0.74	1.01	1.26	1.43	1.48	1.06	0.93
	0.2	0.81	0.91	1.3	1.66	1.98	1.94	1.37	1.19
	0.3	1.01	1.16	1.73	2.24	2.71	2.33	1.78	1.55

Table 8: Median width of a hybrid confidence interval constructed from the nominal 95% standard GMM interval and the $> 95\%$ Bonferroni confidence interval for β using Algorithm 3.2 with $n = 1000, R = 5000$ and $\delta_1 = \delta_2 = 0.025$. The hybrid interval reports Bonferroni unless the GMM interval exists and is contained within the Bonferroni interval. Calculations are based on 2000 replications of the DGP from Section 4.1.

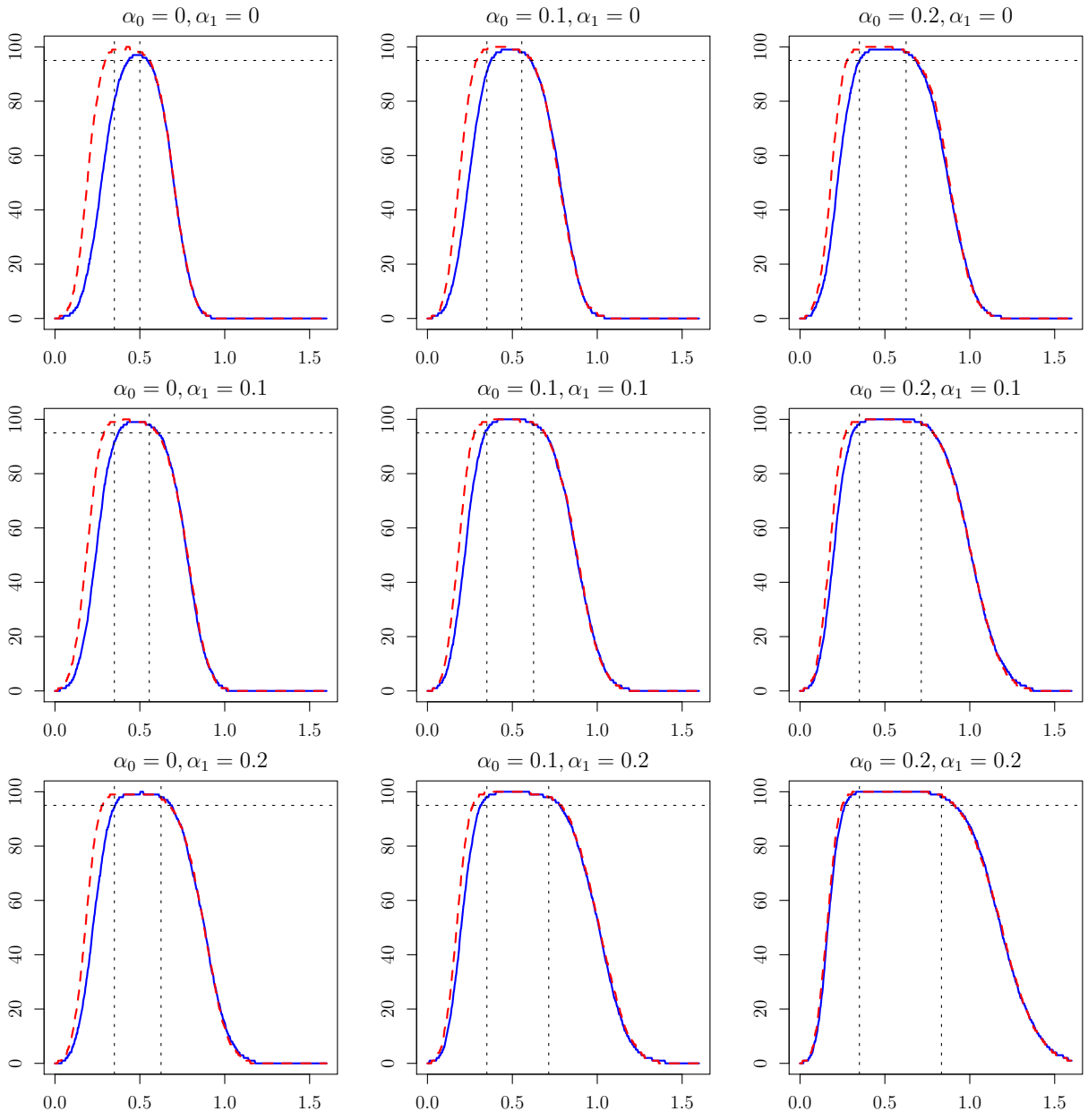

Figure 1: Coverage curves (1 - power) for β when the truth is $\beta = 0.5$, from a nominal $> 95\%$ Bonferroni confidence interval using Algorithm 3.2, with $n = 1000$ and $R = 5000$. The solid curve uses all moment inequalities from Section 3.3 in the GMS step, while the dashed curve excludes m_2^I, those implied by non-differential measurement error. The dashed horizontal line gives the nominal coverage (95%), while dashed vertical lines are the reduced form estimand (left) and the IV estimand (right). Calculations are based on 2000 replications of the DGP from Section 4.1.

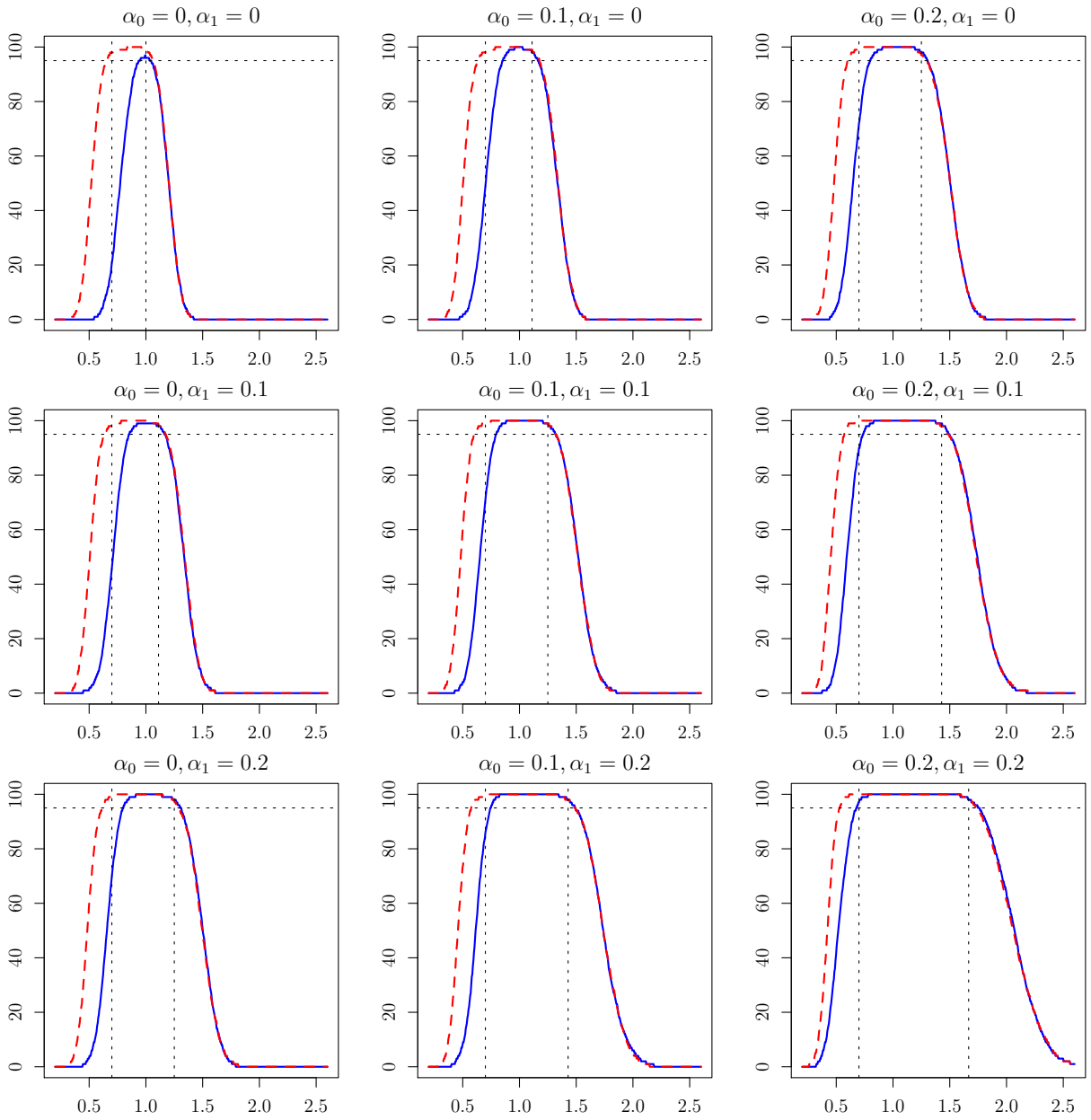

Figure 2: Coverage curves (1 - power) for β when the truth is $\beta = 1$, from a nominal $> 95\%$ Bonferroni confidence interval using Algorithm 3.2, with $n = 1000$ and $R = 5000$. The solid curve uses all moment inequalities from Section 3.3 in the GMS step, while the dashed curve excludes m_2^I, those implied by non-differential measurement error. The dashed horizontal line gives the nominal coverage (95%), while dashed vertical lines are the reduced form estimand (left) and the IV estimand (right). Calculations are based on 2000 replications of the DGP from Section 4.1.

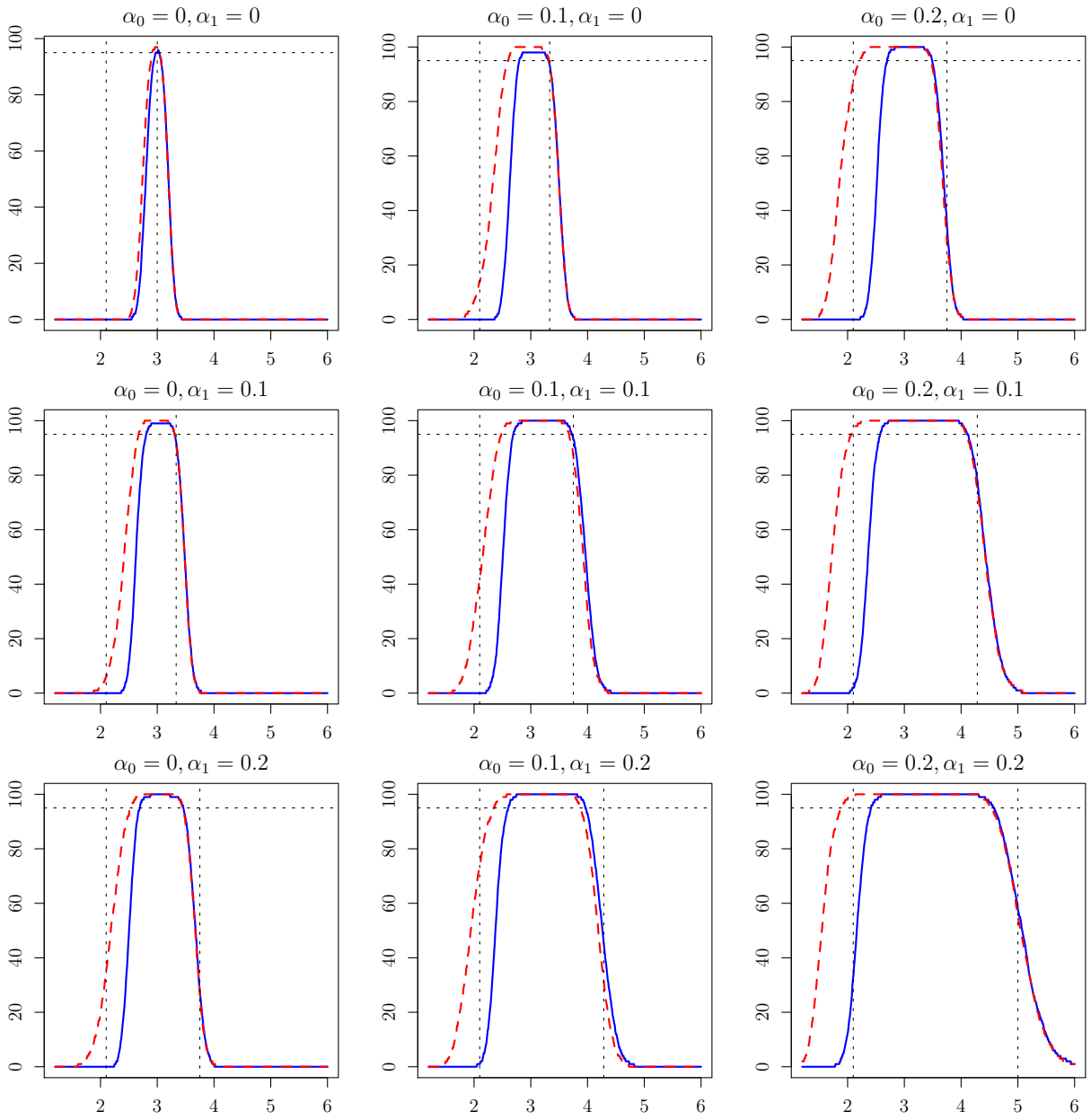

Figure 3: Coverage curves (1 - power) for β when the truth is $\beta = 3$, from a nominal $> 95\%$ Bonferroni confidence interval using Algorithm 3.2, with $n = 1000$ and $R = 5000$. The solid curve uses all moment inequalities from Section 3.3 in the GMS step, while the dashed curve excludes m_2^I, those implied by non-differential measurement error. The dashed horizontal line gives the nominal coverage (95%), while dashed vertical lines are the reduced form estimand (left) and the IV estimand (right). Calculations are based on 2000 replications of the DGP from Section 4.1.

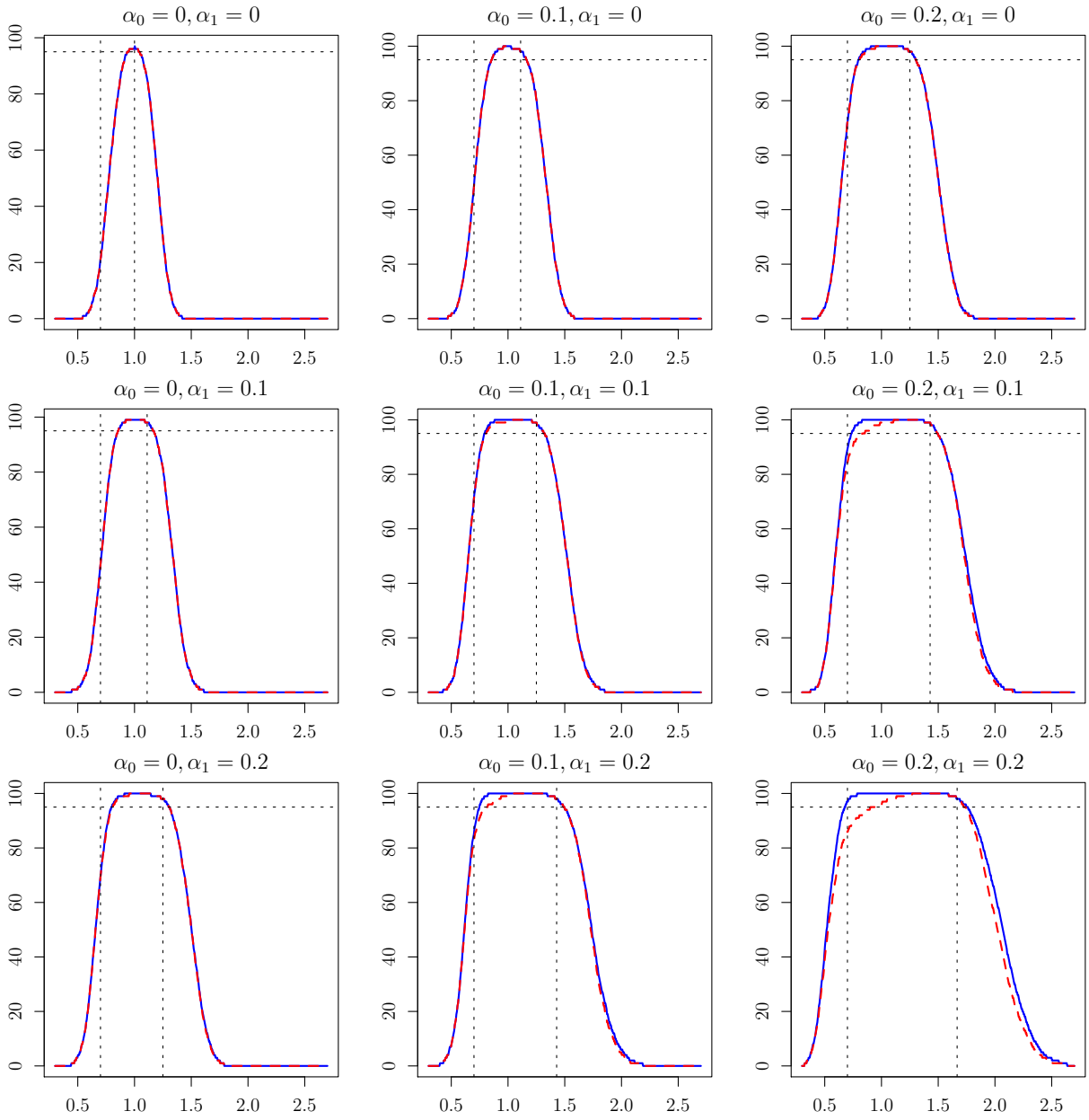

Figure 4: Comparison of Coverage curves (1 - power) for β when the truth is $\beta = 1$: the solid curve corresponds the Bonferroni nominal $> 95\%$ interval from Algorithm 3.2 and the dashed curve to the hybrid interval from Tables 7–8. The dashed horizontal line gives the nominal coverage (95%), while dashed vertical lines are the reduced form estimand (left) and the IV estimand (right). Results are based on 2000 simulation replications from the DGP in Section 4.1 with $n = 1000$.

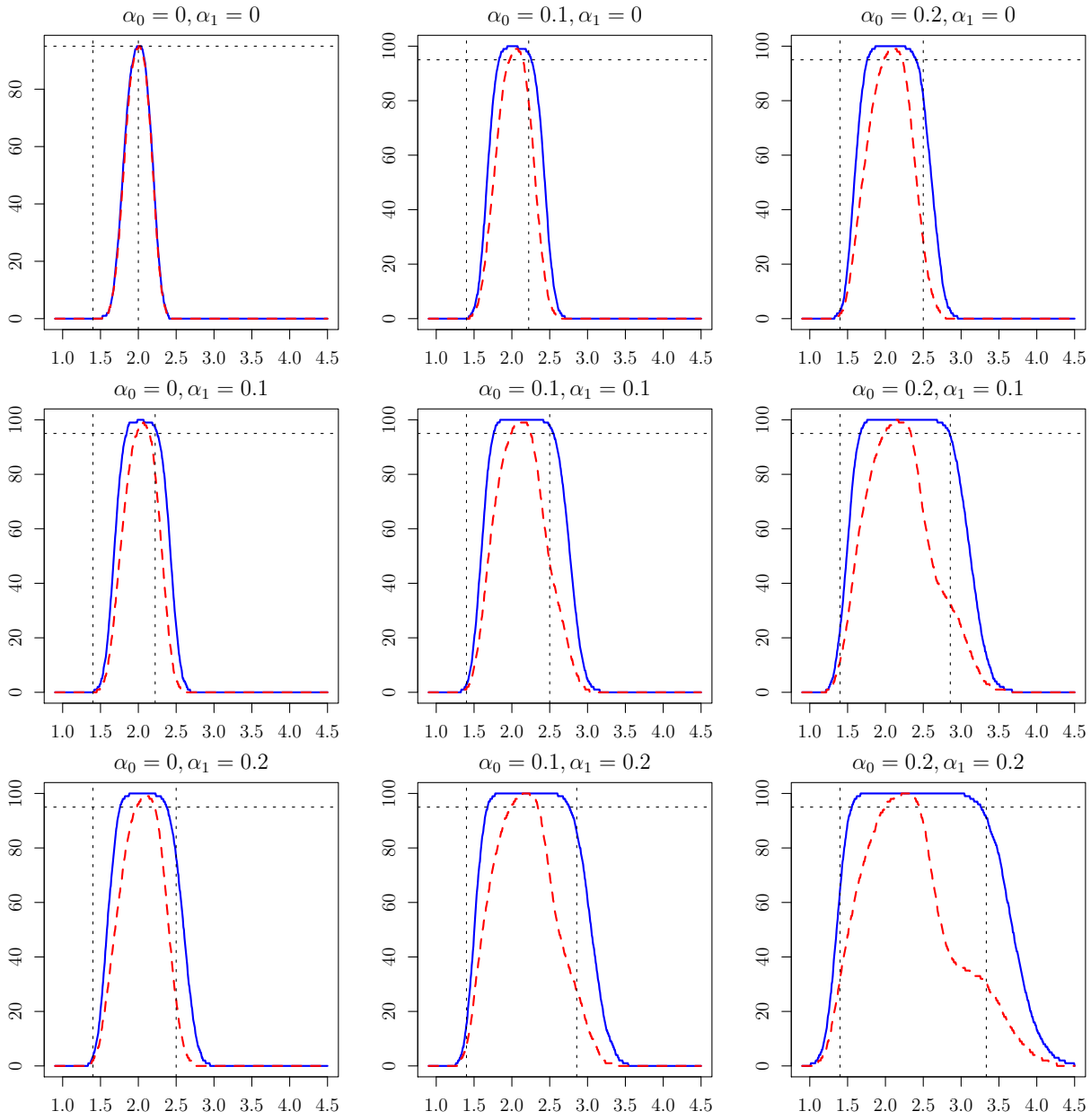

Figure 5: Comparison of Coverage curves (1 - power) for β when the truth is $\beta = 2$: the solid curve corresponds the Bonferroni nominal $> 95\%$ interval from Algorithm 3.2 and the dashed curve to the hybrid interval from Tables 7–8. The dashed horizontal line gives the nominal coverage (95%), while dashed vertical lines are the reduced form estimand (left) and the IV estimand (right). Results are based on 2000 simulation replications from the DGP in Section 4.1 with $n = 1000$.

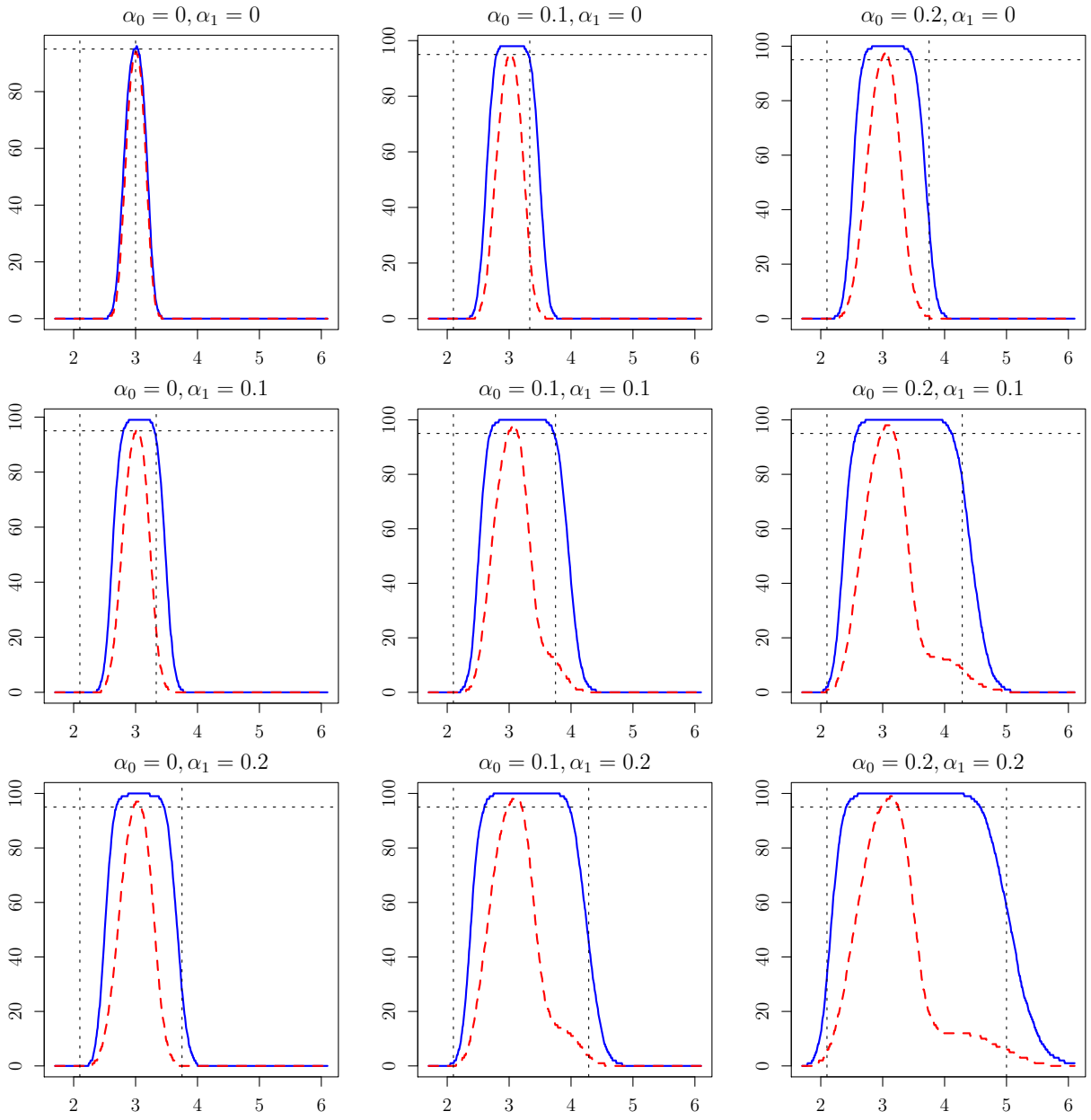

Figure 6: Comparison of Coverage curves (1 - power) for β when the truth is $\beta = 3$: the solid curve corresponds the Bonferroni nominal $> 95\%$ interval from Algorithm 3.2 and the dashed curve to the hybrid interval from Tables 7–8. The dashed horizontal line gives the nominal coverage (95%), while dashed vertical lines are the reduced form estimand (left) and the IV estimand (right). Results are based on 2000 simulation replications from the DGP in Section 4.1 with $n = 1000$.

5 Conclusion

This paper has studied identification and inference for a mis-classified, binary, endogenous regressor in an additively separable model using a discrete instrumental variable. We have shown that the only existing identification result for this model is incorrect, and gone on to derive the sharp identified set under standard first-moment assumptions from the literature. Strengthening these assumptions to hold for second and third moments, we have established point identification for the effect of interest. Inference in models with mis-classification error is complicated by problems of weak identification and parameters on the boundary. To address these challenges, we have proposed a Bonferroni-based procedure for identification robust inference, using both the moment equalities from our identification results and moment inequalities from our partial identification results. This procedure is computationally attractive and performs well in simulations. An interesting extension of the results presented here would be to explore the more general case of a discrete endogenous regressor subject to mis-classification error, possibly by combining our approach with the matrix factorization techniques from Hu (2008). Another interesting extension, inspired by our hybrid confidence interval heuristic from Section 4, would be to study the transition between robust and standard inference in moment condition models. It may be possible, for example, to adapt the techniques of Andrews (2016) in this direction to provide similar theoretical guarantees.

A Proofs

All of the results in this paper hold $\mathbf{x}$ *fixed*. This allows us to completely ignore the presence of covariates in the proofs that follow. Accordingly we work in terms of scalars $\alpha_0, \alpha_1, \beta, p_k$, etc. rather than functions $\alpha(\mathbf{x})$, $\alpha_1(\mathbf{x})$, $\beta(\mathbf{x}), p_k(\mathbf{x})$. The former should be understood as the value of the latter evaluated at some particular $\mathbf{x}$.

A.1 Partial Identification Results

Proof of Lemma 2.1. Follows from a simple calculation using the law of total probability. □

Proof of Lemma 2.2. Immediate since $\text{Cov}(z, T) = (1 - \alpha_0 - \alpha_1)\text{Cov}(z, T^*)$ by Lemma 2.1. □

Proof of Theorem 2.1. To show that $\alpha_0 \leq p_k \leq 1 - \alpha_1$, substitute $p_k^* = 0$ and $p_k^* = 1$, respectively, into Lemma 2.1 and rearrange. To show that $\mathbb{E}[y|z = k] = c + \beta(p_k - \alpha_0)/(1 - \alpha_0 - \alpha_1)$, take conditional expectations of Equation 1 and apply Assumption 2.1 (iii) and Lemma 2.1.

To prove sharpness we need to show that for any $(c, \beta, \alpha_0, \alpha_1)$ that satisfy $\alpha_0 \leq p_k \leq 1 - \alpha_1$ and $\mathbb{E}[y|z = k] = c + \beta(p_k - \alpha_0)/(1 - \alpha_0 - \alpha_1)$ we can construct a valid joint distribution for (y, T, T^*, z) that is compatible with the observed distribution of (y, T, z), provided that $p_1 \neq p_0$. To establish this result, we factorize the joint distribution of (y, T, T^*, z) into the product of a conditional $y|(T, T^*, z)$ and marginal (T, T^*, z). The argument proceeds in two steps. Our first step relies on the fact that Assumptions 2.1 (i) and (iii) do not constrain the distribution of (T, T^*, z) while 2.1 (ii) and 2.2 (i)–(ii) constrain *only* the distribution of (T, T^*, z). Under these latter three assumptions, we show how to construct a valid joint distribution for (T, T^*, z) that is compatible with the observed distribution of (T, z) for any (α_0, α_1) satisfying $\alpha_0 \leq p_k \leq 1 - \alpha_1$. Our second step shows how to construct a valid conditional distribution for y given (T, T^*, z) under Assumptions 2.1 (i) and (iii) that is compatible with the observed conditional distribution of y

given (T, z) for any $(c, \beta, \alpha_0, \alpha_1)$ satisfying $\mathbb{E}[y|z = k] = c + \beta(p_k - \alpha_0)(1 - \alpha_0 - \alpha_1)$. Combining the two steps gives the required joint distribution for (y, T^*, T, z).

For the first step, we need to construct a valid joint probability mass function $p(T^*, T, z)$ with support set $\{0, 1\} \times \{0, 1\} \times \{0, 1\}$. By Assumption 2.2 (i), $p(T|T^*, z) = p(T|T^*)$ and hence

$$p(T^*, T, z) = p(T|T^*)p(T^*|z)p(z).$$

Since $p(z)$ is observed, to construct a valid joint probability mass function $p(T^*, T, z)$ it suffices to construct valid *conditional* probability mass functions $p(T|T^*)$ and $p(T^*|z)$. Since $\alpha_0 \leq p_k \leq 1-\alpha_1$, both α_0 and α_1 are guaranteed to lie between zero and one. This gives a valid construction of $p(T|T^*)$. Moreover the corresponding values of p_k^* implied by Lemma 2.1 are also guaranteed to lie between zero and one. This gives a valid construction of $p(T^*|z)$ that satisfies Assumption 2.1 (ii), since $p_1 \neq p_0$ by assumption and $(p_1 - p_0) = (p_1^* - p_0^*)(1 - \alpha_0 - \alpha_1)$ by Lemma 2.1. Because our construction relies on Lemma 2.1, which is simply an application of the law of total probability, the resulting distribution $p(T, T^*, z)$ is automatically compatible with $p(T, z) = p(T|z)p(z)$.

For the second step, we need to construct a valid conditional distribution for y given (T, T^*, z). To begin we define the following notation:

$$\begin{aligned} r_{tk} &\equiv \mathbb{P}(T^* = 1|T = t, z = k) & F_t(\tau) &\equiv \mathbb{P}(y \leq \tau|z = k) \\ F_{tk}(\tau) &\equiv \mathbb{P}(y \leq \tau|T = t, z = k) & F_{tk}^{t^*}(\tau) &\equiv \mathbb{P}(y \leq \tau|T^* = t^*, T = t, z = k) \\ G_k(\tau) &\equiv \mathbb{P}(\varepsilon \leq \tau|z = k) & G_{tk}^{t^*}(\tau) &\equiv \mathbb{P}(\varepsilon \leq \tau|T^* = t^*, T = t, z = k). \end{aligned}$$

Assumption 2.1 (i) imposes a relationship between $G_{tk}^{t^*}$ and $F_{tk}^{t^*}$ for each t^*, namely

$$G_{tk}^0(\tau) = F_{tk}^0(\tau + c), \quad G_{tk}^1(\tau) = F_{tk}^1(\tau + c + \beta) \tag{A.1}$$

and thus we see that

$$\begin{aligned} G_k(\tau) = {} & r_{1k}p_kF_{1k}^1(\tau + c + \beta) + r_{0k}(1 - p_k)F_{0k}^1(\tau + c + \beta) \\ & + (1 - r_{1k})p_kF_{1k}^0(\tau + c) + (1 - r_{0k})(1 - p_k)F_{0k}^0(\tau + c) \end{aligned} \tag{A.2}$$

applying the law of total probability and Bayes' rule. Moreover,

$$F_{tk}(\tau) = r_{tk}F_{tk}^1(\tau) + (1 - r_{tk})F_{tk}^0(\tau) \tag{A.3}$$

for all $t, k \in \{0, 1\}$, and by Bayes' rule,

$$r_{1k} = (1 - \alpha_1)p_k^*/p_k, \quad r_{0k} = \alpha_1 p_k^*/(1 - p_k). \tag{A.4}$$

There are four cases, corresponding to different possibilities for the r_{tk}. The first case violates one of our model assumptions. For each of the remaining cases, we show that it is possible to construct the required distributions F_{tk}^0, F_{tk}^1 under Assumptions 2.1 (i) and (iii) for any $(c, \beta, \alpha_0, \alpha_1)$ such that $\mathbb{E}(y|z = k) = c + \beta(p_k - \alpha_0)/(1 - \alpha_0 - \alpha_1)$.

Case I: $r_{1k} = 0, r_{0k} \neq 0$ By Equation A.4 this requires $\alpha_1 = 1$, violating Assumption 2.2 (ii).

Case II: $r_{0k} = r_{1k} = 0$ By Equation A.4, this requires $p_k^* = 0$ which in turn requires $p_k = \alpha_0$. By Equation A.3 we have $F_{tk}^0 = F_{tk}$, while F_{tk}^1 is unrestricted. Substituting into A.2,

$$G_k(\tau) = p_k F_{1k}(\tau + c) + (1 - p_k) F_{0k}(\tau + c) = F_k(\tau + c)$$

Now, since $F_k(\tau + c)$ is the conditional CDF of $y - c$ given that $z = k$, and G_k is the conditional CDF of ε given $z = k$, we see that Assumption 2.1 (i) is satisfied if and only if $\mathbb{E}(y|z = k) = c$, which is equal to $c + \beta(p_k - \alpha_0)/(1 - \alpha_0 - \alpha_1)$ since $p_k - \alpha_0 = 0$.

Case III: $r_{1k} \neq 0, r_{0k} = 0$ By Equation A.4 this requires $\alpha_1 = 0$ and $p_k^* \neq 0$. By Equation A.3 we have $F_{0k}^0 = F_{0k}$ and since $r_{1k} \neq 1$, we can solve to obtain

$$F_{1k}^1(\tau) = \frac{1}{r_{1k}} \left[F_{1k}(\tau) - (1 - r_{1k}) F_{1k}^0(\tau) \right]$$

Substituting into Equation A.2, we obtain

$$\begin{aligned} G_k(\tau) = &\left[(1 - p_k) F_{0k}(\tau + c) + p_k F_{1k}(\tau + c + \beta) \right] \\ &+ p_k (1 - r_{1k}) \left[F_{1k}^0(\tau + c) - F_{1k}^0(\tau + c + \beta) \right] \end{aligned}$$

Now, $F_{0k}(\tau + c)$ is the conditional CDF of $(y - c)$ given $(T = 0, z = k)$ while $F_{1k}(\tau + c + \beta)$ is the conditional CDF of $(y - c - \beta)$ given $(T = 1, z = k)$. Similarly, $F_{1k}^0(\tau + c)$ is the conditional CDF of ε given $(T^* = 0, T = 1, z = k)$ while $F_{1k}^0(\tau + c + \beta)$ is the conditional CDF of $(\varepsilon - \beta)$ given $(T^* = 0, T = 1, z = k)$. Since $G_k(\tau)$ is the conditional CDF of ε given $z = k$, we see that Assumption 2.1 (iii) is satisfied if and only if

$$\begin{aligned} 0 = &(1 - p_k) \mathbb{E}(y - c | T = 0, z = k) + p_k \mathbb{E}(y - c - \beta | T = 1, z = k) \\ &+ p_k (1 - r_{1k}) \left[\mathbb{E}(\varepsilon | T^* = 0, T = 1, z = k) - \mathbb{E}(\varepsilon - \beta | T^* = 0, T = 1, z = k) \right] \end{aligned}$$

Rearranging, this is equivalent to

$$\mathbb{E}(y|z = k) = c + (1 - \alpha_1)\beta \left(\frac{p_k - \alpha_0}{1 - \alpha_0 - \alpha_1} \right) = c + \beta \left(\frac{p_k - \alpha_0}{1 - \alpha_0 - \alpha_1} \right)$$

since $\alpha_1 = 0$ in this case. As explained above, $F_{0k}^0 = F_{0k}$ in the present case while F_{0k}^1 is undefined. We are free to choose any distributions for F_{1k}^0 and F_{1k}^1 that satisfy Equation A.3, for example $F_{1k}^0 = F_{1k}^1 = F_{1k}$.

Case IV: $r_{1k} \neq 0, r_{0k} \neq 0$ In this case, we can solve Equation A.3 to obtain

$$F_{tk}^1(\tau) = \frac{1}{r_{tk}} \left[F_{tk}(\tau) - (1 - r_{tk}) F_{tk}^0(\tau) \right]$$

Substituting this into Equation A.2, we have

$$\begin{aligned} G_k(\tau) = F_k(\tau + c + \beta) &+ p_k (1 - r_{1k}) \left[F_{1k}^0(\tau + c) - F_{1k}^0(\tau + c + \beta) \right] \\ &+ (1 - p_k)(1 - r_{0k}) \left[F_{0k}^0(\tau + c) - F_{0k}^0(\tau + c + \beta) \right] \end{aligned}$$

using the fact that $F_k(\tau) = p_k F_{1k}(\tau) + (1 - p_k) F_{0k}(\tau)$. Now, $F_k(\tau + c + \beta)$ is the conditional CDF of $(y - c - \beta)$ given $z = k$, while $F_{tk}^0(\tau + c)$ is the conditional CDF of ε given $(T = t, z = k)$ and

$F^0_{tk}(\tau + c + \beta)$ is the conditional CDF of $(\varepsilon - \beta)$ given $(T = t, z = k)$. Since $G_k(\tau)$ is the conditional CDF of ε given $z = k$, we see that Assumption 2.1 (iii) is satisfied if and only if

$$\begin{aligned}
0 &= \mathbb{E}[y - c - \beta | z = k] + p_k(1 - r_{1k}) \left[\mathbb{E}(\varepsilon | T^* = 0, T = 1, z = k) - \mathbb{E}(\varepsilon - \beta | T^* = 0, T = 1, z = k)\right] \\
&\quad + (1 - p_k)(1 - r_{0k}) \left[\mathbb{E}(\varepsilon | T^* = 0, T = 0, z = k) - \mathbb{E}(\varepsilon - \beta | T^* = 0, T = 0, z = k)\right] \\
0 &= \mathbb{E}[y - c - \beta | z = k] + \beta \left[p_k(1 - r_{1k}) + (1 - p_k)(1 - r_{0k})\right]
\end{aligned}$$

But since $[p_k(1 - r_{1k}) + (1 - p_k)(1 - r_{0k})] = (1 - p^*_k)$ and $p^*_k = (p_k - \alpha_0)/(1 - \alpha_0 - \alpha_1)$, this becomes

$$\mathbb{E}[y | z = k] = c + \beta \left[(p_k - \alpha_0)(1 - \alpha_0 - \alpha_1)\right].$$

Thus, in this case we are free to choose *any* distributions for F^0_{tk} and F^1_{tk} that satisfy Equation A.3. For example we could take $F^0_{tk} = F^1_{tk} = F_{tk}$. □

Proof of Corollary 2.1. The result follows by substituting the largest and smallest possible values for $\alpha_0 + \alpha_1$ and taking the difference of the expressions for $\mathbb{E}[y | z = k]$. □

Proof of Theorem 2.2. The only difference between the conditions of Theorem 2.1 and those of 2.2 is that the latter imposes Assumption 2.2 (iii) while the former does not. Accordingly, the present argument builds on the proof of Theorem 2.1 and relies on the notation defined within it. Under Assumption 2.1 (i), Assumption 2.2 (iii) is equivalent to $\mathbb{E}[y | T, T^*, z] = \mathbb{E}[y | T^*, z]$. Hence, non-differential measurement error constrains only the conditional distribution of y given (T, T^*, z). For this reason, we need only revisit the second step of the proof of Theorem 2.1. Consider a point $(c, \beta, \alpha_0, \alpha_1)$ that satisfies Equation 3 and $\alpha_0 \leq p_k \leq 1 - \alpha_1$ for all k. Since this point lies in the identified set from Theorem 2.1, it suffices to determine whether there exist valid conditional CDFs F^0_{tk}, F^1_{tk} such that $F_{tk} = (1 - r_{tk})F^0_{tk} + r_{tk}F^1_{tk}$ for all t, k and $\mathbb{E}[y | T, T^*, z] = \mathbb{E}[y | T^*, z]$.

Let $\mu^{t^*}_{tk} \equiv \mathbb{E}[y | T = t, z = k, T^* = t^*]$, $\mu_{tk} \equiv \mathbb{E}[y | T = t, z = k]$, and $\mu^{t^*}_k \equiv \mathbb{E}[y | z = k, T^* = t^*]$. By Assumption 2.2 (iii) $\mu^{t^*}_{tk} = \mu^{t^*}_k$ for $t^* = 0, 1$. Hence, by iterated expectations,

$$\begin{aligned}
\mu_{0k} &= (1 - r_{0k})\mu^0_k + r_{0k}\mu^1_k \\
\mu_{1k} &= (1 - r_{1k})\mu^0_k + r_{1k}\mu^1_k.
\end{aligned}$$

Now, (μ_{0k}, μ_{1k}) are observed while r_{0k} and r_{1k} depend only on the observed first-stage probability p_k and the mis-classification probabilities (α_0, α_1). Thus, at a given point $(c, \beta, \alpha_0, \alpha_1)$ in the identified set from Theorem 2.1 the preceding equations form a linear system in μ^0_k and μ^1_k. After some algebra, we find that the determinant is

$$r_{1k} - r_{0k} = \left[\frac{p_k - \alpha_0}{1 - \alpha_0 - \alpha_1}\right]\left[\frac{1 - p_k - \alpha_1}{p_k(1 - p_k)}\right].$$

Suppose first that $r_{0k} = r_{1k} = r$ so the determinant condition fails. This occurs if and only if $\alpha_0 = p_k$ or $\alpha_1 = 1 - p_k$. If $\mu_{0k} \neq \mu_{1k}$, the system is inconsistent: no solution for (μ^0_k, μ^1_k) exists. Hence $\alpha_0 = p_k$ and $\alpha_1 = 1 - p_k$ are excluded from the identified set under non-differential measurement error so long as $\mu_{0k} \neq \mu_{1k}$. If instead $\mu_{0k} = \mu_{1k} = \mu$, the system is consistent but rank deficient: any pair (μ^0_k, μ^1_k) such that $\mu = (1 - r)\mu^0_k + r\mu^1_k$ is a solution and hence satisfies the assumption of non-differential measurement error. One such solution is $\mu^1_k = \mu^0_k = \mu$ so we are free to set $F^0_{0k} = F^1_{0k} = F_{0k}$ and $F^0_{1k} = F^1_{1k} = F_{1k}$. Hence, if $\mu_{0k} = \mu_{1k}$ then $\alpha_0 = p_k$ lies within the sharp identified set if $p_k < p_\ell$ and $\alpha_1 = 1 - p_k$ lies in the sharp identified set if $p_\ell < p_k$.

Now suppose that $r_{0k} \neq r_{1k}$, which occurs if and only if $\alpha_0 \neq p_k$ and $\alpha_1 \neq 1 - p_k$. In this case the system has a unique solution, namely

$$\mu_k^0 = \frac{r_{1k}\mu_{0k} - r_{0k}\mu_{1k}}{r_{1k} - r_{0k}} = \frac{(1-p_k)\mathbb{E}(y|T=0, z=k) - \alpha_1 \mathbb{E}(y|z=k)}{1 - p_k - \alpha_1}$$
$$\mu_k^1 = \frac{(\mu_{1k} - \mu_{0k}) + (r_{1k}\mu_{0k} - r_{0k}\mu_{1k})}{r_{1k} - r_{0k}} = \frac{p_k \mathbb{E}(y|T=1, z=k) - \alpha_0 \mathbb{E}(y|z=k)}{p_k - \alpha_0}.$$

Since $\mu_k^0 = \mu_{0k}^0 = \mu_{1k}^0$ and $\mu_k^1 = \mu_{0k}^1 = \mu_{1k}^1$ under non-differential measurement error, the misclassification probabilities (α_0, α_1) combined with the observable moments completely determine the means of F_{tk}^0 and F_{tk}^1 whenever the determinant condition holds. If $\mu_{0k} = \mu_{1k}$ then $\mu_k^0 = \mu_k^1$ so we are free to set $F_{0k}^0 = F_{0k}^1 = F_{0k}$ and $F_{1k}^0 = F_{1k}^1 = F_{1k}$. Combining this with the reasoning from the preceding paragraph, we see that Assumption 2.2 (iii) imposes *no additional restrictions* for any k such that $\mu_{0k} = \mu_{1k}$. Accordingly, for the remainder of the proof we consider only the case in which $\mu_{0k} \neq \mu_{1k}$. Given (α_0, α_1), r_{tk}, μ_k^0, and μ_k^1 are fixed. The question is whether, for a given pair (α_0, α_1) and observed CDFs F_{tk}, we can construct valid CDFs F_{tk}^0, F_{tk}^1 such that

$$\int_{\mathbb{R}} \tau F_{tk}^0(d\tau) = \mu_k^0, \quad \int_{\mathbb{R}} \tau F_{tk}^1(d\tau) = \mu_k^1, \quad F_{tk}(\tau) = r_{tk} F_{tk}^1(\tau) + (1 - r_{tk}) F_{tk}^0(\tau).$$

For a given pair (t, k), there are two cases: $0 < r_{tk} < 1$ and $r_{tk} \in \{0, 1\}$.

Case I: $r_{tk} \in \{0, 1\}$ If $r_{tk} = 1$ then $\mu_k^1 = \mu_{tk}$ so we can set $F_{tk}^1 = F_{tk}$. In this case F_{tk}^0 is unrestricted. Analogously, if $r_{tk} = 0$, $\mu_k^0 = \mu_{tk}$ so we can set $F_{tk}^0 = F_{tk}$ with F_{tk}^1 unrestricted.

Case II: $0 < r_{tk} < 1$ Define the function $\mu_{tk}(\xi) = \mathbb{E}[y | y \in I_{tk}(\xi), T = t, z = k]$ and the closed interval $I_{tk}(\xi) = \left[F_{tk}^{-1}(1 - \xi - r_{tk}), F_{tk}^{-1}(1 - \xi)\right]$ where $0 \leq \xi \leq 1 - r_{tk}$. The function μ_{tk} is decreasing in ξ, attaining its maximum $\overline{\mu}_{tk}$ at $\xi = 0$ and its minimum $\underline{\mu}_{tk}$ at $\xi = 1 - r_{tk}$.

Suppose first that μ_k^1 does *not* lie in the interval $[\underline{\mu}_{tk}, \overline{\mu}_{tk}]$. We show that it is impossible to construct valid CDFs F_{tk}^0 and F_{tk}^1 that satisfy $F_{tk}(\tau) = r_{tk} F_{tk}^1(\tau) + (1 - r_{tk}) F_{tk}^0(\tau)$. Since $r_{tk} \neq 1$, we can solve the expression for F_{tk} to yield $F_{tk}^0(\tau) = \left[F_{tk}(\tau) - r_{tk} F_{tk}^1(\tau)\right] / (1 - r_{tk})$. Hence, since $r_{tk} \neq 0$, the requirement that $0 \leq F_{tk}^0(\tau) \leq 1$ implies

$$\frac{F_{tk}(\tau) - (1 - r_{tk})}{r_{tk}} \leq F_{tk}^1(\tau) \leq \frac{F_{tk}(\tau)}{r_{tk}} \tag{A.5}$$

Now define $\underline{F}_{tk}^1(\tau) = \min\{1, F_{tk}(\tau)/r_{tk}\}$ and $\overline{F}_{tk}^1(\tau) = \max\{0, F_{tk}(\tau)/r_{tk} - (1 - r_{tk})/r_{tk}\}$. By combining Equation A.5 with $0 \leq F_{tk}^1(\tau) \leq 1$, we obtain $\overline{F}_{tk}^1(\tau) \leq F_{tk}^1(\tau) \leq \underline{F}_{tk}^1(\tau)$. Thus, $\overline{F}_{tk}^1$ first-order stochastically dominates F_{tk}^1 which first-order stochastically dominates $\underline{F}_{tk}^1$. Hence,

$$\int \tau \underline{F}_{tk}^1(d\tau) \leq \int \tau F_{tk}^1(d\tau) \leq \int \tau \overline{F}_{tk}^1(d\tau).$$

But notice that

$$\underline{\mu}_{tk} = \int \tau \underline{F}_{tk}^1(d\tau), \quad \mu_k^1 = \int \tau F_{tk}^1(d\tau), \quad \overline{\mu}_{tk} = \int \tau \overline{F}_{tk}^1(d\tau)$$

so we have $\underline{\mu}_{tk} \leq \mu_k^1 \leq \overline{\mu}_{tk}$ which contradicts $\mu_k^1 \notin [\underline{\mu}_{tk}, \overline{\mu}_{tk}]$.

Now suppose that $\mu_k^1 \in \left[\underline{\mu}_{tk}, \overline{\mu}_{tk}\right]$. We show how to construct densities f_{tk}^1 and f_{tk}^0 that yield CDFs F_{tk}^0 F_{tk}^1 satisfying the requirements described above. Since the conditional distribution of y given (T, z) is continuous, μ_{tk} is continuous on its domain and takes on all values in $\left[\underline{\mu}_{tk}, \overline{\mu}_{tk}\right]$ by the intermediate value theorem. Thus, there exists a ξ^* such that $\mu_{tk}(\xi^*) = \mu_k^1$. Let $f_{tk}(\tau) = dF_{tk}(\tau)/d\tau$ which is non-negative by the assumption that y is continuously distributed. Now, define

$$f_{tk}^1(\tau) = \frac{f_{tk}(\tau) \times \mathbf{1}\left\{\tau \in I_{tk}(\xi^*)\right\}}{r_{tk}}, \quad f_{tk}^0(\tau) = \frac{f_{tk}(\tau) \times \mathbf{1}\left\{\tau \in I_{tk}(\xi^*)\right\}}{1 - r_{tk}}.$$

Clearly $f_{tk}^1 \geq 0$ and $f_{tk}^0 \geq 0$. Integrating,

$$\int_{\mathbb{R}} f_{tk}^1(\tau)\, d\tau = \frac{1}{r_{tk}} \int_{I_{tk}(\xi^*)} f_{tk}(\tau)\, d\tau = 1, \quad \int_{\mathbb{R}} f_{tk}^0(\tau)\, d\tau = \frac{1}{1 - r_{tk}} \int_{I_{tk}^C(\xi^*)} f_{tk}(\tau)\, d\tau = 1$$

where I_{tk}^C is the complement of I_{tk}. By construction

$$r_{tk} \int_A f_{tk}^1(\tau)\, d\tau + (1 - r_{tk}) \int_A f_{tk}^0(\tau)\, d\tau = \int_A f_{tk}(\tau)\, d\tau$$

for any set A. Finally,

$$\int_{\mathbb{R}} \tau f_{tk}^1(\tau)\, d\tau = \frac{1}{r_{tk}} \int_{I_{tk}(\xi^*)} \tau f_{tk}(\tau)\, d\tau = \mu_{tk}(\xi^*) = \mu_k^1.$$

□

A.2 Point Identification Results

In the proofs of Lemma 2.3, Lemma 2.4, and Theorem 2.3, we employ the shorthand $\pi \equiv \text{Cov}(T, z)$, $\eta_j \equiv \text{Cov}(y^j, z)$, and $\tau_j \equiv \text{Cov}(Ty^j, z)$ for $j = 1, 2, 3$. Hence Lemma 2.2 becomes $\eta_1 = \pi\theta_1$, while Lemma 2.3 becomes $\eta_2 = 2\tau_1\theta_1 - \pi\theta_2$, and Lemma 2.4 becomes $\eta_3 = 3\tau_2\theta_1 - 3\tau_1\theta_2 + \pi\theta_3$.

Proof of Lemma 2.3. By Assumption 2.1 (i) and the basic properties of covariance,

$$\begin{aligned}
\eta_2 &= \beta^2 \text{Cov}(T^*, z) + 2\beta\left[c\,\text{Cov}(T^*, z) + \text{Cov}(T^*\varepsilon, z)\right] + 2c\,\text{Cov}(\varepsilon, z) + \text{Cov}(\varepsilon^2, z) \\
\tau_1 &= c\pi + \text{Cov}(T\varepsilon, z) + \beta\text{Cov}(TT^*, z)
\end{aligned}$$

using the fact that T^* is binary. Now, by Assumptions 2.1 (iii) and 2.5 we have $\text{Cov}(\varepsilon, z) = \text{Cov}(\varepsilon^2, z) = 0$. And, using Assumptions 2.2 (i) and (ii), one can show that $\text{Cov}(TT^*, z) = (1 - \alpha_1)\text{Cov}(T^*, z)$ and $\text{Cov}(T^*, z) = \pi/(1 - \alpha_0 - \alpha_1)$. Hence,

$$\begin{aligned}
\eta_2 &= \theta_1\left(\beta + 2c\right)\pi + 2\beta\text{Cov}(T^*\varepsilon, z) \\
2\tau_1\theta_1 - \pi\theta_2 &= \left[2\theta_1 c + 2\theta_1^2(1 - \alpha_1) - \theta_2\right]\pi + 2\theta_1\text{Cov}(T\varepsilon, z)
\end{aligned}$$

but since $\theta_2 = \theta_1^2\left[(1 - \alpha_1) + \alpha_0\right]$, we see that $\left[2\theta_1^2(1 - \alpha_1) - \theta_2\right] = \theta_1\beta$. Thus, it suffices to show that $\beta\text{Cov}(T^*\varepsilon, z) = \theta_1\text{Cov}(T\varepsilon, z)$. This equality is trivially satisfied when $\beta = 0$, so suppose that $\beta \neq 0$. In this case it suffices to show that $(1 - \alpha_0 - \alpha_1)\text{Cov}(T^*\varepsilon, z) = \text{Cov}(T\varepsilon, z)$. Define $m_{tk}^* = \mathbb{E}\left[\varepsilon | T^* = t, z = k\right]$ and $p_k^* = \mathbb{P}(T^* = 1 | z = k)$. Then, by iterated expectations, Bayes' rule,

and Assumption 2.2 (iii)

$$\begin{aligned}\text{Cov}(T^*\varepsilon, z) &= q(1-q)\left(p_1^* m_{11}^* - p_0^* m_{10}^*\right)\\ \text{Cov}(T\varepsilon, z) &= q(1-q)\left\{(1-\alpha_1)\left[p_1^* m_{11}^* - p_0^* m_{10}^*\right] + \alpha_0\left[(1-p_1^*)m_{01}^* - (1-p_0^*)m_{00}^*\right]\right\}\end{aligned}$$

But by Assumption 2.1 (iii), $\mathbb{E}[\varepsilon|z=k] = m_{1k}^* p_k^* + m_{0k}^*(1-p_k^*) = 0$ and thus we obtain $m_{0k}^*(1-p_k^*) = -m_{1k}^* p_k^*$. Therefore $(1-\alpha_0-\alpha_1)\text{Cov}(T^*\varepsilon, z) = \text{Cov}(T\varepsilon, z)$ as required. □

Proof of Lemma 2.4. Since T^* is binary, if follows from the basic properties of covariance that,

$$\begin{aligned}\eta_3 &= \text{Cov}\left[(c+\varepsilon)^3, z\right] + 3\beta\text{Cov}[(c+\varepsilon)^2 T^*, z] + 3\beta^2\text{Cov}[(c+\varepsilon)T^*, z] + \beta^3\text{Cov}(T^*, z)\\ \tau_2 &= \text{Cov}\left[(c+\varepsilon)^2 T, z\right] + 2\beta\text{Cov}\left[(c+\varepsilon)TT^*, z\right] + \beta^2\text{Cov}(TT^*, z)\end{aligned}$$

By Assumptions 2.1 (iii), 2.5, and 2.6 (ii) , $\text{Cov}\left[(c+\varepsilon)^3, z\right] = 0$. Expanding,

$$\begin{aligned}\eta_3 &= 3\beta\text{Cov}(T^*\varepsilon^2, z) + \left(3\beta^2 + 6c\beta\right)\text{Cov}(T^*\varepsilon, z) + \left(\beta^3 + 3c\beta^2 + 3c^2\beta\right)\text{Cov}(T^*, z)\\ \tau_2 &= c^2\text{Cov}(T, z) + \beta(\beta+2c)\text{Cov}(TT^*, z) + \text{Cov}(T\varepsilon^2, z) + 2c\text{Cov}(T\varepsilon, z) + 2\beta\,\text{Cov}(TT^*\varepsilon, z)\end{aligned}$$

Now, define $s_{tk}^* = \mathbb{E}[\varepsilon^2|T^*=t, z=k]$ and $p_k^* = \mathbb{P}(T^*=1|z=k)$. By iterated expectations, Bayes' rule, and Assumption 2.6 (i),

$$\begin{aligned}\text{Cov}(T^*\varepsilon^2, z) &= q(1-q)(p_1^* s_{11}^* - p_0^* s_{10}^*)\\ \text{Cov}(T\varepsilon^2, z) &= q(1-q)\left\{(1-\alpha_1)\left[p_1^* s_{11}^* - p_0^* s_{10}^*\right] + \alpha_0\left[(1-p_1^*)s_{01}^* - (1-p_0^*)s_{00}^*\right]\right\}\end{aligned}$$

By Assumption 2.5, $\mathbb{E}[\varepsilon^2|z=1] = \mathbb{E}[\varepsilon^2|z=0]$ and thus, by iterated expectations we have $p_1^* s_{11}^* - p_0^* s_{10}^* = -\left[(1-p_1^*)s_{01}^* - (1-p_0^*)s_{00}^*\right]$ which implies

$$\text{Cov}(T\varepsilon^2, z) = (1-\alpha_0-\alpha_1)\text{Cov}(T^*\varepsilon^2, z). \tag{A.6}$$

Similarly by iterated expectations and Assumptions 2.2 (i)–(ii)

$$\text{Cov}(TT^*\varepsilon, z) = q(1-q)(1-\alpha_1)(p_1^* m_{1k}^* - p_0^* m_{10}^*) = (1-\alpha_1)\text{Cov}(T^*\varepsilon, z) \tag{A.7}$$

where m_{tk}^* is defined as in the proof of Lemma 2.3. As shown in the proof of Lemma 2.3,

$$\text{Cov}(TT^*, z) = (1-\alpha_1)\text{Cov}(T^*, z), \quad \text{Cov}(T^*, z) = \frac{\pi}{1-\alpha_0-\alpha_1}, \quad \text{Cov}(T^*\varepsilon, z) = \frac{\text{Cov}(T\varepsilon, z)}{1-\alpha_0-\alpha_1}$$

and combining these equalities with Equations A.6 and A.7, it follows that

$$\begin{aligned}\tau_2 &= 2\left[(1-\alpha_1)(c+\beta) - c\alpha_0\right]\text{Cov}(T^*\varepsilon, z) + \left[(1-\alpha_1)(c+\beta)^2 - c^2\alpha_0\right]\text{Cov}(T^*, z)\\ &\quad + (1-\alpha_0-\alpha_1)\text{Cov}(T^*\varepsilon^2, z)\\ \tau_1 &= (1-\alpha_0-\alpha_1)\text{Cov}(T^*\varepsilon, z) + \left[(1-\alpha_1)(c+\beta) - c\alpha_0\right]\text{Cov}(T^*, z)\end{aligned}$$

using $\tau_1 = c\pi + \text{Cov}(T\varepsilon, z) + \beta\text{Cov}(TT^*, z)$ as shown in the proof of Lemma 2.3. Thus,

$$3\tau_2\theta_1 - 3\tau_1\theta_2 + \pi\theta_3 = K_1\text{Cov}(T^*\varepsilon^2, z) + K_2\text{Cov}(T^*\varepsilon, z) + K_3\text{Cov}(T^*, z)$$

where $K_1 \equiv 3\theta_1(1-\alpha_0-\alpha_1) = 3\beta$ and

$$K_2 \equiv 6\theta_1\left[(1-\alpha_1)(c+\beta) - c\alpha_0\right] - 3\theta_2(1-\alpha_0-\alpha_1)$$
$$K_3 \equiv 3\theta_1\left[(1-\alpha_1)(c+\beta)^2 - c^2\alpha_0\right] - 3\theta_2\left[(1-\alpha_1)(c+\beta) - c\alpha_0\right] + \theta_3(1-\alpha_0-\alpha_1)$$

Substituting the definitions of θ_1, θ_2, and θ_3 from Equations 4–6, tedious but straightforward algebra shows that $K_2 = 3\beta^2 + 6c\beta$ and $K_3 = \beta^3 + 3c\beta^2 + 3c^2\beta$. Therefore the coefficients of η_3 equal those of $3\tau_2 - 3\tau_1\theta_2 + \pi\theta_3$ and the result follows. □

Proof of Theorem 2.3. Collecting the results of Lemmas 2.2–2.4, we have

$$\eta_1 = \pi\theta_1, \quad \eta_2 = 2\tau_1\theta_1 - \pi\theta_2, \quad \eta_3 = 3\tau_2\theta_1 - 3\tau_1\theta_2 + \pi\theta_3$$

which is a linear system in $\theta_1, \theta_2, \theta_3$ with determinant $-\pi^3$. Since $\pi \neq 0$ by assumption 2.1 (ii), θ_1, θ_2 and θ_3 are identified. Now, so long as $\beta \neq 0$, we can rearrange Equations 5 and 6 to obtain

$$A = \theta_2/\theta_1^2 = 1 + (\alpha_0 - \alpha_1) \tag{A.8}$$
$$B = \theta_3/\theta_1^3 = (1-\alpha_0-\alpha_1)^2 + 6\alpha_0(1-\alpha_1) \tag{A.9}$$

Equation A.8 gives $(1-\alpha_1) = A - \alpha_0$. Hence $(1-\alpha_0-\alpha_1) = A - 2\alpha_0$ and $\alpha_0(1-\alpha_1) = \alpha_0(A-\alpha_0)$. Substituting into Equation A.9 and simplifying, $(A^2 - B) + 2A\alpha_0 - 2\alpha_0^2 = 0$. Substituting for α_0 analogously yields a quadratic in $(1-\alpha_1)$ with *identical* coefficients. It follows that one root of $(A^2 - B) + 2Ar - 2r^2 = 0$ is α_0 and the other is $1-\alpha_1$. Solving,

$$r = \frac{A}{2} \pm \sqrt{3A^2 - 2B} = \frac{1}{\theta_1^2}\left(\frac{\theta_2}{2} \pm \sqrt{3\theta_2^2 - 2\theta_1\theta_3}\right). \tag{A.10}$$

Substituting Equations 5 and 6, simple algebra shows that $3\theta_2^2 - 2\theta_1\theta_3 = \theta_1^2(1-\alpha_0-\alpha_1)^2$. This quantity is strictly greater than zero since $\theta_1 \neq 0$ and $\alpha_0 + \alpha_1 \neq 1$. It follows that both roots of the quadratic are real. Moreover, $3\theta_2^2/\theta_1^4 - 2\theta_3/\theta_1^3$ identifies $(1-\alpha_0-\alpha_1)^2$. Substituting into Equation 4, it follows that β is identified up to sign. If $\alpha_0 + \alpha_1 < 1$ then $\text{sign}(\beta) = \text{sign}(\theta_1)$ so that both the sign and magnitude of β are identified. If $\alpha_0 + \alpha_1 < 1$ then $1 - \alpha_1 > \alpha_0$ so $(1-\alpha_1)$ is the larger root of $(A^2 - B) + 2Ar - 2r^2 = 0$ and α_0 is the smaller root. □

B Comment on Mahajan (2006) A.2

Expanding on our discussion from Section 2.2 above, we now show that Mahajan's identification argument for an endogenous regressor in an additively separable model (A.2) is incorrect. Unless otherwise indicated, all notation used below is as defined in Section 2.

The first step of Mahajan (2006) A.2 argues (correctly) that under Assumptions 2.1 and 2.2 (i)–(ii), knowledge of $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ is sufficient to identify $\beta(\mathbf{x})$. This step is equivalent to our Lemma 2.2 above. The second step appeals to Mahajan (2006) Theorem 1 to argue that $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ are indeed point identified. To understand the logic of this second step, we first re-state Mahajan (2006) Theorem 1 in our notation. As in Section 2 above, T^* denotes an unobserved binary random variable, z is a instrument, T an observed binary surrogate for T^*, y an outcome of interest, and $\mathbf{x}$ a vector covariates.

Assumption B.1 (Mahajan (2006) Theorem 1)**.** *Define* $g(T^*, \mathbf{x}) \equiv \mathbb{E}[y|\mathbf{x}, T^*]$ *and* $v \equiv y - g(T^*, \mathbf{x})$. *Suppose that knowledge of* $(y, T^*, \mathbf{x})$ *is sufficient to identify* g *and that:*

(i) $\mathbb{P}(T^* = 1|\mathbf{x}, z = 0) \neq \mathbb{P}(T^* = 1|\mathbf{x}, z = 1)$.

(ii) T *is conditionally independent of* z *given* $(\mathbf{x}, T^*)$.

(iii) $\alpha_0(\mathbf{x}) + \alpha_1(\mathbf{x}) < 1$

(iv) $\mathbb{E}[v|\mathbf{x}, z, T^*, T] = 0$

(v) $g(1, \mathbf{x}) \neq g(0, \mathbf{x})$

Theorem B.1 (Mahajan (2006) Theorem 1)**.** *Under Assumption B.1,* $\alpha_0(\mathbf{x})$ *and* $\alpha_1(\mathbf{x})$ *are point identified, as is* $g(T^*, \mathbf{x})$.

Assumption B.1 (i) is equivalent to our Assumption 2.1 (ii), while Assumptions B.1 (ii)–(iii) are equivalent to our Assumptions 2.2 (i)–(ii). Assumption B.1 (v) serves the same purpose as $\beta(\mathbf{x}) \neq 0$ in our Theorem 2.3: unless T^* affects y, we cannot identify the mis-classification probabilities. The key difference between Theorem B.1 and the setting we consider in Section 2 comes from Assumption B.1 (iv). This is essentially a stronger version of our Assumptions 2.1 (iii) and 2.2 (iii) but applies to the *projection error* v, defined in Assumption B.1 rather than the structural error ε, defined in Assumption 2.1 (i). Accordingly, Theorem B.1 identifies the conditional mean function g rather than the causal effect $\beta(\mathbf{x})$.

Although the meaning of the error term changes when we move from a structural to a reduced form model, the meaning of the mis-classification error rates does not: $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ are simply conditional probabilities for T given $(T^*, \mathbf{x})$. Step 2 of Mahajan (2006) A.2 relies on this insight. The idea is to find a way to satisfy Assumption B.1 (iv) simultaneously with Assumptions 2.1 (iii) and 2.2 (iii), while allowing T^* to be endogenous. If this can be achieved, $\alpha_0(\mathbf{x}), \alpha_1(\mathbf{x})$ will be identified via Theorem B.1, and identification of $\beta(\mathbf{x})$ will follow from step 1 of A.2 (our Lemma 2.2). To this end, Mahajan (2006) invokes the condition

$$\mathbb{E}(y|\mathbf{x}, z, T^*, T) = \mathbb{E}(y|\mathbf{x}, T^*). \tag{B.1}$$

Because Mahajan (2006) A.2 assumes an additively separable model – our Assumption 2.1 (i) – we see that

$$\mathbb{E}(y|\mathbf{x}, z, T^*, T) = c(\mathbf{x}) + \beta(\mathbf{x})T^* + \mathbb{E}(\varepsilon|\mathbf{x}, z, T^*, T)$$

so Equation B.1 is equivalent to $\mathbb{E}(\varepsilon|\mathbf{x}, z, T^*, T) = \mathbb{E}(\varepsilon|\mathbf{x}, T^*)$. Note that this allows T^* to be endogenous, as it does not require $\mathbb{E}(\varepsilon|\mathbf{x}, T^*) = 0$. Now, applying Equation B.1 to the definition of v from Assumption B.1, we have

$$\mathbb{E}(v|\mathbf{x}, z, T^*, T) = \mathbb{E}\left[y - \mathbb{E}(y|\mathbf{x}, T^*) \mid \mathbf{x}, z, T^*, T\right] = 0$$

which satisfies Assumption B.1 (iv) as required. Based on this reasoning, Mahajan (2006) claims that Equation B.1 along with Assumptions B.1 (iv), 2.1, and 2.2 (i)–(ii) suffice to identify the effect $\beta(\mathbf{x})$ of an endogenous T^*, so long as $g(1, \mathbf{x}) \neq g(0, \mathbf{x})$. As we now show, however, these Assumptions are contradictory unless T^* is exogenous.

By Equation B.1 and Assumption 2.1 (i), $\mathbb{E}(\varepsilon|\mathbf{x}, z, T^*, T) = \mathbb{E}(\varepsilon|\mathbf{x}, T^*)$ and thus by iterated expectations, we obtain

$$\mathbb{E}(\varepsilon|\mathbf{x}, T^*, z) = \mathbb{E}_{T|\mathbf{x},T^*,z}\left[\mathbb{E}(\varepsilon|\mathbf{x}, T^*, T, z)\right] = \mathbb{E}_{T|\mathbf{x},T^*,z}\left[\mathbb{E}(\varepsilon|\mathbf{x}, T^*)\right] = \mathbb{E}(\varepsilon|\mathbf{x}, T^*). \tag{B.2}$$

Now, let $m^*_{tk}(\mathbf{x}) = \mathbb{E}(\varepsilon|\mathbf{x}, T^* = t, z = k)$. Using this notation, Equation B.2 is equivalent to $m^*_{t0}(\mathbf{x}) = m^*_{t1}(\mathbf{x})$ for $t = 0, 1$. Combining iterated expectations with Assumption 2.1 (iii),

$$\mathbb{E}(\varepsilon|\mathbf{x}, z = k) = [1 - p^*_k(\mathbf{x})]m^*_{0k}(\mathbf{x}) + p^*_k(\mathbf{x})m^*_{1k}(\mathbf{x}) = 0 \tag{B.3}$$

for $k = 0, 1$ where $p^*_k(\mathbf{x}) \equiv \mathbb{P}(T^* = 1|\mathbf{x}, z = k)$. But substituting $m^*_{t0}(\mathbf{x}) = m^*_{t1}(\mathbf{x})$ into Equation B.3 for $k = 0, 1$, we obtain

$$\begin{aligned}[1 - p^*_0(\mathbf{x})]m^*_{00}(\mathbf{x}) + p^*_0(\mathbf{x})m^*_{10}(\mathbf{x}) = 0\\ [1 - p^*_1(\mathbf{x})]m^*_{00}(\mathbf{x}) + p^*_1(\mathbf{x})m^*_{10}(\mathbf{x}) = 0\end{aligned}$$

The preceding two equalities are convex combinations of m^*_{00} and m^*_{10}. The only way that both can equal zero simultaneously is if either $p^*_0(\mathbf{x}) = p^*_1(\mathbf{x})$, contradicting Assumption 2.1 (ii), or if $m^*_{tk}(\mathbf{x}) = 0$ for all (t, k), which implies that T^* is exogenous. Hence Mahajan (2006) A.2 fails: given the assumption that z is a valid instrument for ε, Equation B.1 implies that either there is no first-stage relationship between z and T^* or that T^* is exogenous. The root of the problem with A.2 is the attempt to use *one* instrument to satisfy both the assumptions of Theorem B.1 and Lemma 2.2. If one had access to a second instrument w, or equivalently a second mis-measured surrogate for T^*, that satisfied Assumptions B.1, one could use w to recover $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ via Theorem B.1 and z to recover the IV estimand $\beta(\mathbf{x})/[1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})]$ via Lemma 2.2.

C Unobserved Heterogeneity

While allowing for arbitrary observed heterogeneity through the covariates $\mathbf{x}$, all of the results presented above assume an additively separable model – Assumption 2.1 (i). In this section we briefly discuss how our partial identification results can be interpreted in a local average treatment effects (LATE) setting. For simplicity, we suppress explicit conditioning on the covariates $\mathbf{x}$ throughout.

In lieu of Assumption 2.1 (i), consider a non-separable model of the form $y = h(T^*, z, \varepsilon)$. Let $T^*(z)$ denote an individual's potential treatment and $Y(t^*, z)$ denote her potential outcome, where $t^*, z \in \{0, 1\}$. Using this notation we can write $Y(t^*, z) = h(t^*, z, \varepsilon)$. Let $J \in \{a, c, d, n\}$ index the four LATE principal strata: a = always-taker, c = complier, d = defier, and n = never-taker. If $J = a$, then $T^*(z) = 1$; if $J = c$, then $T^*(z) = z$; if $J = d$, then $T^*(z) = 1 - z$; and if $J = n$, then $T^*(z) = 0$. In a LATE model, Assumption 2.1 (iii) is replaced by the standard LATE assumptions:

Assumption C.1 (Unconfounded Type)**.** $\mathbb{P}(J = j|z = 1) = \mathbb{P}(J = j|z = 0)$ *for all* $j \in \{a, c, d, n\}$.

Assumption C.2 (Mean Exclusion Restriction)**.** *For all* $t^* \in \{0, 1\}$ *and* $j \in \{a, c, d, n\}$,

$$\mathbb{E}\left[Y(t^*, 0)|T^* = t^*, z = 1\right] = \mathbb{E}\left[Y(t^*, 1)|T^* = t^*, z = 1\right] = \mathbb{E}[Y(t^*)|J = j].$$

Assumption C.3 (Monotonicity)**.** $\mathbb{P}\big(T^*(1) \geq T^*(0)\big) = 1$

As is well known, Assumption 2.1 (iii) combined with the preceding three conditions implies that the instrumental variables estimand based on T^* identifies the average treatment effect among compliers:

$$\frac{\mathbb{E}[y|z = 1] - \mathbb{E}[y|z = 0]}{p^*_1 - p^*_0} = \mathbb{E}[Y(1) - Y(0)|J = c].$$

The numerator of the preceding expression is observed, but under mis-classification the denominator is not. Notice, however, that Assumptions 2.2 (i)–(ii) only concern the joint distribution of T given

(T^*, z). As such, they have the same meaning in a LATE model as in an additively separable model. Imposing these conditions, Lemma 2.1 continues to hold in a LATE model. It follows that $p_1 - p_0 = (1 - \alpha_0 - \alpha_1)(p_1^* - p_0^*)$ so that

$$\frac{\mathbb{E}[y|z=1] - \mathbb{E}[y|z=0]}{p_1 - p_0} = \frac{\mathbb{E}[Y(1) - Y(0)|J = c]}{1 - \alpha_0 - \alpha_1}.$$

Moreover, $\alpha_0 \leq p_k \leq 1 - \alpha_1$ for all k. Thus, the bound from Corollary 2.1 remains valid in a LATE model: $\mathbb{E}[Y(1) - Y(0)|J = c]$ must lie between the IV and reduced form estimands.

Unlike Assumptions 2.2 (i)–(ii), Assumption 2.2 (iii), non-differential measurement error, is explicitly stated in terms of the unobservable error term in an additively separable model. Our derivation of the additional restrictions on (α_0, α_1) implied by non-differential measurement error in the proof of Theorem 2.2, however, does not use Assumption 2.2 (iii) directly. Rather, it uses a condition that is *equivalent* to it in an additively separable model, namely $\mathbb{E}[Y|T^*, T, z] = \mathbb{E}[Y|T^*, z]$. Hence, as long as this equality holds, regardless of whether one is in an additively separable model or a LATE model, the bounds on (α_0, α_1) from Theorem 2.2 remain valid. Since $Y = (1 - T^*)Y(0) + T^*Y(1)$, the appropriate modification of Assumption 2.2 (iii) is as follows.

Assumption C.4 (Non-differential Measurement Error)**.**

$$\mathbb{E}[Y(0)|T^*, T, z] = \mathbb{E}[Y(0)|T^*, z] \quad \textit{and} \quad \mathbb{E}[Y(1)|T^*, T, z] = \mathbb{E}[Y(1)|T^*, z]$$

To summarize, if one wishes to re-interpret our parameter β as a local average treatment effect, the partial identification bounds from Theorems 2.1 and 2.2 above remain valid. Assumption 2.1 (i) is replaced by $Y = h(T^*, z, \varepsilon)$, Assumption 2.1 (iii) is replaced by Assumptions C.1–C.3, and Assumption 2.2 (iii) is replaced by Assumption C.4. In a LATE model, however, our proofs of sharpness no longer apply, as they do not consider the testable implications of the LATE assumptions themselves. For partial identification results that consider these implications but do not impose non-differential measurement error, see Ura (Forthcoming). For discussion of the testable implications of a LATE model, see Kitagawa (2015).

D Moment Equalities Under Joint Exogeneity

In this Section we discuss the moment equalities that replace Equation 9 under joint exogeneity: Assumption 2.3. Because the moment inequalities from Section 3.3 are unchanged under this assumption, we do not discuss them further here. Define θ_1 as in Equation 4, κ_1 as in Section 3.1, and let $\rho = -\theta_1\alpha_0(1 - \alpha_1)$ and $\eta = \theta_1(1 + \alpha_0 - \alpha_1)$. Now, under Assumptions 2.1, 2.2, and 2.3:

$$\mathbb{E}\left\{ \begin{bmatrix} y - \kappa_1 - \theta_1 T \\ (y - \kappa_1)T - \rho - \eta T \end{bmatrix} \otimes \begin{bmatrix} 1 \\ z \end{bmatrix} \right\} = \mathbf{0}. \tag{D.1}$$

where the equalities involving ρ and η follow from an argument similar to one of the steps from the proof of Lemma 2.3 – see, e.g., Frazis and Loewenstein (2003) and Mahajan (2006). The moment equalities from D.1 point identify the reduced form parameters $(\theta_1, \kappa_1, \rho, \eta)$ and lead to a just-identified method of moments estimator of the same. To see why knowledge of $(\theta_1, \kappa_1, \rho, \eta)$ suffices to identify $(\beta, \alpha_0, \alpha_1)$, define

$$A \equiv \eta/\theta_1 = 1 + \alpha_0 - \alpha_1, \qquad B \equiv -\rho/\theta_1 = \alpha_0(1 - \alpha_1)$$

Eliminating $(1-\alpha_1)$ and α_0, respectively, we obtain:

$$\alpha_0^2 - A\alpha_0 + B = 0, \qquad (1-\alpha_1)^2 - A(1-\alpha_1) + B = 0$$

These are exactly the same quadratic, namely $x^2 - Ax + B = 0$. Hence one root is α_0 while the other is $(1-\alpha_1)$. The discriminant is

$$A^2 - 4B = [(1-\alpha_1) + \alpha_0]^2 - 4[\alpha_0(1-\alpha_1)] = (1-\alpha_0-\alpha_1)^2$$

so that both roots are real as long as $\alpha_0 + \alpha_1 \neq 0$. To solve for α_0 and α_1 we need to calculate the roots of $x^2 - Ax + B = 0$, namely $x = \frac{1}{2}\left(A \pm \sqrt{A^2 - 4B}\right)$. One of these roots is α_0 and the other is $1-\alpha_1$. By assumption, however, $\alpha_0 + \alpha_1 < 1$ and thus $\alpha_0 < 1 - \alpha_1$. It follows that the smaller of the two roots is α_0 and the larger is $1-\alpha_1$. Given that (α_0, α_1) are identified, identification of β follows by Lemma 2.2.

Inference based on the moment equalities from Equation D.1 suffers from the same difficulties as that based on Equation 9 above. First, note that, while $A^2 > 4B$ in population since $\alpha_0 + \alpha_1 < 1$ by assumption, the same may not hold in sample. In this case the GMM estimator of β will fail to exist. Second, notice that the moment equalities from Equation D.1 only depend on β through θ_1 and are completely uninformative about (α_0, α_1) if $\beta = 0$.

Substituting Equation D.1 for Equation 9 in Algorithm 3.1 requires some small changes. First, m^E and h^E from Equations 15–16 are replaced by

$$h^E = \begin{bmatrix} y - \kappa_1 - \theta_1 T \\ (y - \kappa_1 - \theta_1 T)z \end{bmatrix}, \quad m^E = \begin{bmatrix} (y-\kappa_1)T - \rho - \eta T \\ \{(y-\kappa_1)T - \rho - \eta T\}\, z \end{bmatrix}$$

where in this case we require preliminary estimators of κ_1 and θ_1. Accordingly, H^E and M^E from Lemma 3.1 become

$$H^E = \begin{bmatrix} -1 & -\mathbb{E}(T) \\ -\mathbb{E}(z) & -\mathbb{E}(Tz) \end{bmatrix}, \quad M^E = \begin{bmatrix} -\mathbb{E}[T] & 0 \\ -\mathbb{E}[Tz] & 0 \end{bmatrix}$$

and thus

$$B^E = -M^E (H^E)^{-1} = \frac{1}{\mathrm{Cov}(T,z)} \begin{bmatrix} -\mathbb{E}(T)\mathbb{E}(Tz) & \mathbb{E}(T)^2 \\ -\mathbb{E}(Tz)^2 & \mathbb{E}(Tz)\mathbb{E}(T) \end{bmatrix}$$

which is well-defined as long as T is correlated with z.

References

Aigner, D. J., 1973. Regression with a binary independent variable subject to errors of observation. Journal of Econometrics 1, 49–60.

Andrews, D. W., 1994. Empirical process methods in econometrics. Handbook of econometrics 4, 2247–2294.

Andrews, D. W., Shi, X., 2013. Inference based on conditional moment inequalities. Econometrica 81 (2), 609–666.

Andrews, D. W., Shi, X., 2014. Nonparametric inference based on conditional moment inequalities. Journal of Econometrics 179 (1), 31–45.

Andrews, D. W., Soares, G., 2010. Inference for parameters defined by moment inequalities using generalized moment selection. Econometrica 78 (1), 119–157.

Andrews, I., 2016. Valid two-step identification-robust confidence sets for GMM. Review of Economics and Statistics (Forthcoming).

Angrist, J. D., 1990. Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records. The American Economic Review, 313–336.

Battistin, E., Nadai, M. D., Sianesi, B., 2014. Misreported schooling, multiple measures and returns to educational qualifications. Journal of Econometrics 181 (2), 136–150.

Black, D. A., Berger, M. C., Scott, F. A., 2000. Bounding parameter estimates with nonclassical measurement error. Journal of the American Statistical Association 95 (451), 739–748.

Bollinger, C. R., 1996. Bounding mean regressions when a binary regressor is mismeasured. Journal of Econometrics 73, 387–399.

Bollinger, C. R., van Hasselt, M., 2015. Bayesian moment-based inference in a regression models with misclassification error, working Paper.

Bound, J., Brown, C., Mathiowetz, N., 2001. Measurement error in survey data. In: Handbook of econometrics. Vol. 5. Elsevier, pp. 3705–3843.

Bugni, F. A., Canay, I. A., Shi, X., 2017. Inference for subvectors and other functions of partially identified parameters in moment inequality models. Quantitative Economics 8 (1), 1–38.

Carroll, R. J., Ruppert, D., Crainiceanu, C. M., Stefanski, L. A., 2006. Measurement error in nonlinear models: a modern perspective. Chapman and Hall/CRC.

Chen, X., Hong, H., Tamer, E., 2005. Measurement error models with auxiliary data. The Review of Economic Studies 72 (2), 343–366.

Chen, X., Hu, Y., Lewbel, A., 2008a. Nonparametric identification of regression models containing a misclassified dichotomous regressor with instruments. Economics Letters 100, 381–384.

Chen, X., Hu, Y., Lewbel, A., 2008b. A note on the closed-form identification of regression models with a mismeasured binary regressor. Statistics & Probability Letters 78 (12), 1473–1479.

DiTraglia, F. J., García-Jimeno, C., 2017. Mis-classified, binary, endogenous regressors: Identification and inference. Tech. rep., NBER working paper #23814.

Feng, S., Hu, Y., 2013. Misclassification errors and the underestimation of the us unemployment rate. American Economic Review 103 (2), 1054–70.

Frazis, H., Loewenstein, M. A., 2003. Estimating linear regressions with mismeasured, possibly endogenous, binary explanatory variables. Journal of Econometrics 117, 151–178.

Hu, Y., 2008. Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. Journal of Econometrics 144 (1), 27–61.

Hu, Y., Lewbel, A., 2012. Returns to lying? identifying the effects of misreporting when the truth is unobserved. Frontiers of Economics in China 7 (2), 163–192.

Hu, Y., Shennach, S. M., January 2008. Instrumental variable treatment of nonclassical measurement error models. Econometrica 76 (1), 195–216.

Hu, Y., Shiu, J.-L., Woutersen, T., 2015. Identification and estimation of single-index models with measurement error and endogeneity. The Econometrics Journal 18 (3), 347–362.

Imbens, G. W., Rubin, D. B., 1997. Estimating outcome distributions for compliers in instrumental variables models. The Review of Economic Studies 64 (4), 555–574.

Kaido, H., Molinari, F., Stoye, J., 2016. Confidence intervals for projections of partially identified parameters. arXiv preprint arXiv:1601.00934.

Kane, T. J., Rouse, C. E., Staiger, D., July 1999. Estimating returns to schooling when schooling is misreported. Tech. rep., National Bureau of Economic Research, NBER Working Paper 7235.

Kitagawa, T., 2015. A test for instrument validity. Econometrica 83 (5), 2043–2063.

Kreider, B., Pepper, J. V., Gundersen, C., Jolliffe, D., 2012. Identifying the effects of SNAP (food stamps) on child health outcomes when participation is endogenous and misreported. Journal of the American Statistical Association 107 (499), 958–975.

Lewbel, A., March 2007. Estimation of average treatment effects with misclassification. Econometrica 75 (2), 537–551.

Mahajan, A., 2006. Identification and estimation of regression models with misclassification. Econometrica 74 (3), 631–665.

Molinari, F., 2008. Partial identification of probability distributions with misclassified data. Journal of Econometrics 144 (1), 81–117.

Moon, H. R., Schorfheide, F., 2009. Estimation with overidentifying inequality moment conditions. Journal of Econometrics 153 (2), 136–154.

Newey, W. K., McFadden, D., 1994. Large sample estimation and hypothesis testing. Handbook of econometrics 4, 2111–2245.

Nguimkeu, P., Denteh, A., Tchernis, R., 2016. On the estimation of treatment effects with endogenous misreporting. Working Paper.

Shiu, J.-L., 2016. Identification and estimation of endogenous selection models in the presence of misclassification errors. Economic Modelling 52 (Part B), 507–518.

Song, S., 2015. Semiparametric estimation of models with conditional moment restrictions in the presence of nonclassical measurement errors. Journal of Econometrics 185 (1), 95–109.

Ura, T., Forthcoming. Heterogeneous treatment effects with mismeasured endogenous treatment. Quantitative Economics.

van Hasselt, M., Bollinger, C. R., 2012. Binary misclassification and identification in regression models. Economics Letters 115, 81–84.

E Supplementary Simulation Results: Online Only

In this section we provide additional simulation results to supplement those from Section 4 above. For details of the simulation DGP, etc. see the discussion above.

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	90	90	90	91	90	91	90	90
	0.1	91	93	94	94	94	94	90	89
	0.2	92	93	94	94	94	94	92	90
	0.3	93	93	94	94	94	93	92	91
0.1	0.0	92	93	93	94	94	93	90	87
	0.1	93	95	96	97	97	96	92	87
	0.2	95	96	97	98	97	96	92	87
	0.3	96	98	98	98	98	95	92	88
0.2	0.0	93	93	93	93	93	93	92	89
	0.1	95	96	98	98	97	95	93	89
	0.2	97	97	98	98	97	95	92	89
	0.3	98	98	98	98	97	95	93	91
0.3	0.0	93	94	94	94	94	93	92	91
	0.1	97	97	98	98	97	95	93	89
	0.2	98	98	98	98	97	94	93	91
	0.3	99	99	99	98	98	96	95	94

Table E.1: Coverage (1 - size) of 90% GMS joint test for α_0 and α_1: $n = 1000$.

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	90	91	91	90	90	90	90	90
	0.1	91	92	92	93	94	94	92	90
	0.2	91	92	93	93	93	94	93	91
	0.3	92	93	93	93	94	93	93	91
0.1	0.0	90	92	93	94	93	94	92	89
	0.1	92	93	95	96	97	97	94	90
	0.2	92	94	96	97	97	96	95	89
	0.3	94	95	97	98	98	96	94	90
0.2	0.0	91	93	93	93	93	94	92	90
	0.1	92	95	96	97	97	96	94	90
	0.2	94	96	97	97	97	95	93	90
	0.3	96	97	98	98	97	95	93	90
0.3	0.0	92	92	93	93	93	93	92	91
	0.1	93	96	97	97	97	96	93	90
	0.2	96	97	97	97	96	95	93	90
	0.3	98	98	98	98	97	95	94	92

Table E.2: Coverage (1 - size) of 90% GMS joint test for α_0 and α_1: $n = 2000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	95	95	95	96	96	96	95	95
	0.1	96	97	97	97	97	97	95	94
	0.2	96	97	98	98	97	97	96	95
	0.3	97	97	97	98	97	97	96	95
0.1	0.0	96	97	97	97	97	97	95	93
	0.1	97	98	99	99	99	98	96	92
	0.2	98	99	99	99	99	98	96	93
	0.3	99	99	99	99	99	98	96	94
0.2	0.0	97	97	97	97	97	96	96	94
	0.1	98	99	99	99	99	98	96	94
	0.2	99	99	99	99	99	98	96	94
	0.3	99	100	100	99	99	98	97	95
0.3	0.0	97	97	97	97	97	96	96	95
	0.1	99	99	99	99	99	98	97	94
	0.2	99	99	99	99	99	98	97	96
	0.3	100	100	100	99	99	98	98	97

Table E.3: Coverage (1 - size) of 95% GMS joint test for α_0 and α_1: $n = 1000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	95	95	96	95	95	95	95	95
	0.1	96	96	96	97	97	97	96	95
	0.2	96	96	97	97	97	97	96	95
	0.3	96	97	97	97	97	97	97	95
0.1	0.0	95	96	97	97	97	97	96	94
	0.1	96	97	98	98	99	99	97	94
	0.2	96	98	98	99	99	98	97	94
	0.3	97	98	99	99	99	98	97	95
0.2	0.0	96	96	97	97	97	97	96	95
	0.1	96	98	98	99	99	98	97	94
	0.2	97	98	99	99	99	98	97	95
	0.3	98	99	99	99	99	98	97	94
0.3	0.0	96	96	97	97	97	97	96	95
	0.1	97	98	99	99	99	98	96	94
	0.2	98	99	99	99	98	98	96	95
	0.3	99	99	99	99	99	98	97	96

Table E.4: Coverage (1 - size) of 95% GMS joint test for α_0 and α_1: $n = 2000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	97.7	97.7	97.6	97.7	98.0	98.0	97.4	97.9
	0.1	98.0	98.7	98.8	99.1	98.8	98.4	97.1	96.4
	0.2	98.4	98.5	98.9	98.9	98.8	98.6	98.0	97.0
	0.3	98.5	98.8	98.8	99.0	98.7	98.4	97.8	97.5
0.1	0.0	98.1	98.5	98.3	98.8	98.8	98.4	96.8	95.7
	0.1	98.6	99.1	99.5	99.6	99.6	98.8	97.7	95.2
	0.2	99.0	99.3	99.7	99.8	99.7	98.9	97.5	95.7
	0.3	99.4	99.7	99.8	99.8	99.6	99.0	98.2	96.7
0.2	0.0	98.6	98.5	98.6	98.9	98.7	98.2	97.7	97.0
	0.1	99.0	99.5	99.7	99.7	99.4	99.0	98.1	96.5
	0.2	99.5	99.7	99.8	99.7	99.4	99.0	97.8	96.8
	0.3	99.7	99.8	99.8	99.8	99.5	99.0	98.7	97.7
0.3	0.0	98.7	98.7	98.8	98.7	98.7	98.2	98.1	97.6
	0.1	99.4	99.6	99.6	99.7	99.4	98.9	98.3	96.8
	0.2	99.8	99.8	99.7	99.8	99.5	99.1	98.5	97.8
	0.3	100.0	99.9	99.9	99.8	99.6	99.5	99.1	98.8

Table E.5: Coverage (1 - size) of 97.5% GMS joint test for α_0 and α_1: $n = 1000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	97.7	97.7	97.6	97.6	97.6	97.5	97.4	97.5
	0.1	98.0	98.1	98.4	98.3	98.8	98.6	97.8	97.0
	0.2	98.1	98.2	98.8	98.6	98.9	98.6	98.3	97.3
	0.3	98.2	98.5	98.6	98.6	98.8	98.4	98.2	97.4
0.1	0.0	97.4	98.1	98.3	98.8	98.5	98.5	97.9	96.9
	0.1	98.0	98.6	99.1	99.4	99.5	99.3	98.4	96.8
	0.2	98.2	98.9	99.4	99.6	99.7	99.3	98.8	96.8
	0.3	98.6	99.1	99.6	99.8	99.6	99.2	98.4	97.0
0.2	0.0	97.8	98.1	98.5	98.6	98.5	98.4	98.0	97.6
	0.1	98.3	98.9	99.2	99.6	99.5	99.1	98.6	97.0
	0.2	98.7	99.4	99.7	99.6	99.5	99.0	98.4	96.9
	0.3	99.1	99.6	99.7	99.7	99.5	99.0	98.2	97.0
0.3	0.0	98.2	98.3	98.7	98.5	98.6	98.5	98.0	97.7
	0.1	98.6	99.3	99.4	99.6	99.5	99.2	98.1	97.0
	0.2	99.2	99.7	99.7	99.6	99.4	98.8	98.4	97.4
	0.3	99.6	99.8	99.8	99.7	99.4	99.1	98.8	98.2

Table E.6: Coverage (1 - size) of 97.5% GMS joint test for α_0 and α_1: $n = 2000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	27	33	30	14	1	0	0	0
	0.1	27	32	29	13	2	0	0	0
	0.2	26	33	32	15	4	0	0	0
	0.3	26	34	30	17	5	0	0	0
0.1	0.0	26	32	31	14	2	0	0	0
	0.1	26	36	32	16	4	0	0	0
	0.2	27	35	31	18	8	0	0	0
	0.3	25	35	32	21	11	1	0	0
0.2	0.0	26	33	30	15	3	0	0	0
	0.1	26	33	30	19	6	0	0	0
	0.2	26	35	33	22	12	1	0	0
	0.3	26	35	33	26	15	3	0	0
0.3	0.0	26	32	32	16	6	0	0	0
	0.1	24	35	33	21	11	1	0	0
	0.2	26	32	35	27	15	4	0	0
	0.3	26	35	35	28	21	7	2	0

Table E.7: Percentage of simulation replications for which the standard GMM confidence interval fails to exist, either becuase the point estimate is NaN or the asymptotic covariance matrix is numerically singular ($n = 1000$)

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	25	36	29	7	0	0	0	0
	0.1	28	36	29	7	0	0	0	0
	0.2	28	37	28	10	1	0	0	0
	0.3	27	36	28	12	2	0	0	0
0.1	0.0	27	36	27	10	0	0	0	0
	0.1	26	36	29	9	1	0	0	0
	0.2	28	38	29	13	2	0	0	0
	0.3	24	36	31	15	5	0	0	0
0.2	0.0	26	36	30	9	1	0	0	0
	0.1	25	37	29	12	2	0	0	0
	0.2	27	38	32	17	4	0	0	0
	0.3	25	39	34	20	9	1	0	0
0.3	0.0	26	37	30	10	2	0	0	0
	0.1	25	38	31	16	4	0	0	0
	0.2	27	38	34	19	9	0	0	0
	0.3	27	36	36	23	13	2	0	0

Table E.8: Percentage of simulation replications for which the standard GMM confidence interval fails to exist, either becuase the point estimate is NaN or the asymptotic covariance matrix is numerically singular ($n = 2000$)

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	72	62	62	80	92	95	94	95
	0.1	72	62	63	79	92	95	96	95
	0.2	73	61	61	77	90	96	96	96
	0.3	73	59	62	76	88	95	96	95
0.1	0.0	73	63	60	78	91	95	96	96
	0.1	73	58	59	77	90	95	95	94
	0.2	73	59	61	75	86	95	95	94
	0.3	74	59	58	71	82	94	96	96
0.2	0.0	74	62	60	78	91	95	96	96
	0.1	73	60	61	74	87	95	96	94
	0.2	73	58	57	70	81	93	95	95
	0.3	73	58	56	66	78	92	95	96
0.3	0.0	74	62	60	76	89	95	96	96
	0.1	75	59	58	71	82	93	96	95
	0.2	74	61	56	65	78	90	96	96
	0.3	73	58	55	64	71	88	93	96

Table E.9: Coverage of nominal 95% GMM Intervals with $n = 1000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	19.07	3.44	1.86	1.32	0.87	0.47	0.37	0.35
	0.1	17.52	3.47	1.92	1.41	1	0.61	0.51	0.46
	0.2	17.41	3.51	1.9	1.45	1.1	0.76	0.65	0.58
	0.3	18.23	3.34	1.92	1.48	1.24	0.91	0.79	0.7
0.1	0.0	17.13	3.51	1.86	1.38	0.97	0.61	0.51	0.46
	0.1	17.88	3.33	1.85	1.45	1.13	0.78	0.67	0.6
	0.2	17.37	3.36	1.95	1.54	1.24	0.97	0.85	0.75
	0.3	18.07	3.33	1.98	1.63	1.41	1.17	1.04	0.92
0.2	0.0	17.79	3.39	1.92	1.45	1.11	0.75	0.65	0.58
	0.1	18.98	3.43	1.96	1.54	1.26	0.97	0.84	0.75
	0.2	18.25	3.26	1.92	1.64	1.45	1.2	1.06	0.95
	0.3	19.03	3.31	2.02	1.75	1.66	1.49	1.33	1.19
0.3	0.0	18.27	3.48	1.87	1.5	1.25	0.9	0.79	0.7
	0.1	19.4	3.41	1.96	1.63	1.43	1.18	1.04	0.92
	0.2	18.22	3.56	1.96	1.74	1.67	1.49	1.35	1.19
	0.3	17.56	3.55	2.13	1.96	1.86	1.86	1.74	1.55

Table E.10: Median Width of nominal 95% GMM Intervals with $n = 1000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	74	54	63	87	95	94	96	95
	0.1	72	54	62	86	94	95	95	96
	0.2	72	53	64	85	94	95	95	94
	0.3	73	54	64	81	94	95	95	94
0.1	0.0	73	54	65	83	94	95	94	96
	0.1	74	55	64	84	93	95	95	95
	0.2	72	52	63	80	93	96	95	95
	0.3	75	53	59	77	90	95	95	95
0.2	0.0	74	54	61	84	93	96	95	94
	0.1	74	54	63	81	92	96	95	96
	0.2	73	52	60	75	90	96	96	95
	0.3	74	50	57	72	86	95	96	96
0.3	0.0	74	53	61	83	92	97	95	95
	0.1	75	52	60	78	90	95	96	96
	0.2	73	52	57	73	85	95	96	96
	0.3	73	53	54	69	80	93	96	96

Table E.11: Coverage of nominal 95% GMM Intervals with $n = 2000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	17.4	2.42	1.47	1	0.62	0.33	0.27	0.24
	0.1	16.56	2.51	1.49	1.06	0.7	0.43	0.36	0.33
	0.2	16.33	2.4	1.53	1.13	0.81	0.53	0.46	0.41
	0.3	17.06	2.52	1.57	1.19	0.91	0.65	0.56	0.5
0.1	0.0	17.2	2.5	1.53	1.05	0.71	0.43	0.36	0.33
	0.1	17.48	2.5	1.53	1.15	0.83	0.56	0.48	0.43
	0.2	16.32	2.45	1.57	1.2	0.97	0.69	0.6	0.53
	0.3	18.37	2.43	1.51	1.3	1.1	0.84	0.73	0.65
0.2	0.0	17.64	2.5	1.49	1.13	0.8	0.54	0.46	0.41
	0.1	18.25	2.47	1.58	1.22	0.96	0.69	0.6	0.54
	0.2	17.02	2.4	1.57	1.31	1.13	0.86	0.76	0.67
	0.3	18.05	2.39	1.61	1.43	1.33	1.09	0.95	0.85
0.3	0.0	17.72	2.43	1.53	1.19	0.91	0.65	0.56	0.5
	0.1	18.8	2.46	1.55	1.32	1.11	0.84	0.74	0.65
	0.2	18.24	2.45	1.61	1.45	1.3	1.08	0.96	0.85
	0.3	17.43	2.55	1.67	1.62	1.57	1.4	1.24	1.1

Table E.12: Median Width of nominal 95% GMM Intervals with $n = 2000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	96	97	97	96	97	97	95	96
	0.1	97	99	99	99	99	100	100	99
	0.2	98	99	99	100	100	100	100	100
	0.3	97	100	100	100	100	100	100	100
0.1	0.0	97	99	99	99	100	100	100	98
	0.1	98	100	100	100	100	100	100	100
	0.2	98	100	100	100	100	100	100	100
	0.3	97	100	100	100	100	100	100	100
0.2	0.0	97	99	99	100	100	100	100	100
	0.1	98	100	100	100	100	100	100	100
	0.2	98	100	100	100	100	100	100	100
	0.3	98	100	100	100	100	100	100	100
0.3	0.0	97	99	100	100	100	100	100	100
	0.1	97	100	100	100	100	100	100	100
	0.2	98	100	100	100	100	100	100	100
	0.3	98	100	100	100	100	100	100	100

Table E.13: Coverage of nominal $> 95\%$ Bonferroni Intervals with $n = 1000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	96	97	96	97	96	96	95	95
	0.1	97	98	99	100	100	100	100	99
	0.2	97	99	99	100	100	100	100	100
	0.3	97	99	100	100	100	100	100	100
0.1	0.0	97	99	99	99	100	100	100	99
	0.1	98	100	100	100	100	100	100	100
	0.2	98	100	100	100	100	100	100	100
	0.3	98	100	100	100	100	100	100	100
0.2	0.0	97	99	99	100	100	100	100	99
	0.1	98	100	100	100	100	100	100	100
	0.2	98	100	100	100	100	100	100	100
	0.3	98	100	100	100	100	100	100	100
0.3	0.0	97	100	100	100	100	100	100	100
	0.1	97	100	100	100	100	100	100	100
	0.2	97	100	100	100	100	100	100	100
	0.3	97	100	100	100	100	100	100	100

Table E.14: Coverage of nominal $> 95\%$ Bonferroni Intervals with $n = 2000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	0.4	0.41	0.43	0.43	0.43	0.42	0.41	0.41
	0.1	0.45	0.47	0.54	0.59	0.63	0.7	0.75	0.86
	0.2	0.51	0.54	0.65	0.76	0.85	0.95	1.01	1.17
	0.3	0.58	0.62	0.79	0.95	1.07	1.17	1.24	1.48
0.1	0.0	0.45	0.47	0.54	0.59	0.63	0.7	0.76	0.88
	0.1	0.51	0.54	0.66	0.77	0.86	1.03	1.18	1.46
	0.2	0.58	0.63	0.8	0.98	1.12	1.38	1.55	1.88
	0.3	0.67	0.75	1	1.25	1.46	1.74	1.94	2.4
0.2	0.0	0.51	0.54	0.65	0.76	0.86	0.96	1.02	1.19
	0.1	0.58	0.63	0.81	0.99	1.14	1.42	1.64	2.08
	0.2	0.67	0.75	1.01	1.29	1.54	1.97	2.33	2.9
	0.3	0.81	0.91	1.3	1.7	2.09	2.73	3.13	3.9
0.3	0.0	0.58	0.62	0.8	0.95	1.09	1.18	1.25	1.5
	0.1	0.68	0.74	1.01	1.26	1.49	1.84	2.13	2.78
	0.2	0.81	0.91	1.3	1.7	2.11	2.8	3.4	4.48
	0.3	1.01	1.16	1.74	2.35	2.93	4.17	5.2	6.85

Table E.15: Median Width of nominal $> 95\%$ Bonferroni Intervals with $n = 1000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	0.29	0.3	0.31	0.31	0.31	0.3	0.29	0.29
	0.1	0.32	0.35	0.4	0.44	0.48	0.53	0.55	0.61
	0.2	0.36	0.41	0.51	0.59	0.65	0.67	0.69	0.81
	0.3	0.41	0.48	0.64	0.76	0.81	0.8	0.85	1.01
0.1	0.0	0.32	0.35	0.4	0.44	0.48	0.53	0.56	0.62
	0.1	0.36	0.41	0.51	0.6	0.69	0.82	0.88	1.02
	0.2	0.41	0.48	0.64	0.79	0.91	1.04	1.08	1.27
	0.3	0.48	0.59	0.82	1.02	1.16	1.25	1.33	1.61
0.2	0.0	0.36	0.41	0.51	0.59	0.65	0.67	0.7	0.82
	0.1	0.41	0.48	0.65	0.79	0.92	1.09	1.21	1.52
	0.2	0.48	0.59	0.83	1.05	1.24	1.49	1.61	1.96
	0.3	0.57	0.73	1.09	1.43	1.69	1.9	2.08	2.6
0.3	0.0	0.41	0.48	0.64	0.77	0.82	0.78	0.84	1.02
	0.1	0.48	0.59	0.83	1.03	1.18	1.36	1.57	2.06
	0.2	0.57	0.73	1.1	1.43	1.71	2.11	2.45	3.18
	0.3	0.72	0.95	1.5	2.03	2.53	3.15	3.56	4.56

Table E.16: Median Width of nominal $> 95\%$ Bonferroni Intervals with $n = 2000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	96	97	97	96	97	97	95	93
	0.1	97	99	99	99	99	98	96	95
	0.2	98	99	99	100	100	97	96	96
	0.3	97	100	100	100	99	96	96	96
0.1	0.0	97	99	99	99	100	98	97	95
	0.1	98	100	100	100	100	96	96	96
	0.2	98	100	100	100	99	96	96	95
	0.3	97	100	100	100	97	95	96	96
0.2	0.0	97	99	99	100	100	96	96	96
	0.1	98	100	100	100	99	96	96	96
	0.2	98	100	100	100	96	95	95	96
	0.3	98	100	100	98	95	95	95	96
0.3	0.0	97	99	100	100	100	95	96	97
	0.1	97	100	100	100	97	94	96	96
	0.2	98	100	100	98	94	94	96	96
	0.3	98	100	99	96	92	94	95	96

Table E.17: Coverage of hybrid CI constructed from nominal 95% GMM and nominal $> 95\%$ Bonferroni intervals: $n = 1000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	96	97	96	97	96	96	95	93
	0.1	97	98	99	100	100	98	97	96
	0.2	97	99	99	100	100	97	96	95
	0.3	97	99	100	100	99	96	96	96
0.1	0.0	97	99	99	99	100	98	96	95
	0.1	98	100	100	100	100	96	96	97
	0.2	98	100	100	100	99	96	96	97
	0.3	98	100	100	99	97	95	96	96
0.2	0.0	97	99	99	100	100	97	96	95
	0.1	98	100	100	100	98	96	96	97
	0.2	98	100	100	100	96	96	96	96
	0.3	98	100	100	97	95	95	96	96
0.3	0.0	97	100	100	100	99	98	97	96
	0.1	97	100	100	100	96	95	96	97
	0.2	97	100	100	97	94	96	96	97
	0.3	97	100	100	94	94	95	96	96

Table E.18: Coverage of hybrig CI constructed from nominal 95% GMM and nominal > 95% Bonferroni intervals: $n = 2000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	0.4	0.41	0.43	0.43	0.43	0.42	0.4	0.35
	0.1	0.45	0.47	0.54	0.59	0.63	0.67	0.52	0.46
	0.2	0.51	0.54	0.65	0.76	0.84	0.82	0.65	0.58
	0.3	0.58	0.62	0.79	0.95	1.05	0.96	0.79	0.7
0.1	0.0	0.45	0.47	0.54	0.59	0.63	0.67	0.51	0.46
	0.1	0.51	0.54	0.66	0.77	0.86	0.92	0.69	0.61
	0.2	0.58	0.63	0.8	0.97	1.11	1.17	0.87	0.75
	0.3	0.67	0.75	1	1.25	1.4	1.4	1.06	0.92
0.2	0.0	0.51	0.54	0.65	0.76	0.85	0.83	0.65	0.58
	0.1	0.58	0.63	0.81	0.99	1.12	1.18	0.86	0.75
	0.2	0.67	0.75	1.01	1.29	1.48	1.56	1.08	0.95
	0.3	0.81	0.91	1.3	1.67	1.95	1.77	1.35	1.2
0.3	0.0	0.58	0.62	0.8	0.95	1.07	0.95	0.8	0.7
	0.1	0.68	0.74	1.01	1.26	1.43	1.48	1.06	0.93
	0.2	0.81	0.91	1.3	1.66	1.98	1.94	1.37	1.19
	0.3	1.01	1.16	1.73	2.24	2.71	2.33	1.78	1.55

Table E.19: Median width of hybrid CI constructed from nominal 95% GMM and nominal > 95% Bonferroni intervals: $n = 1000$

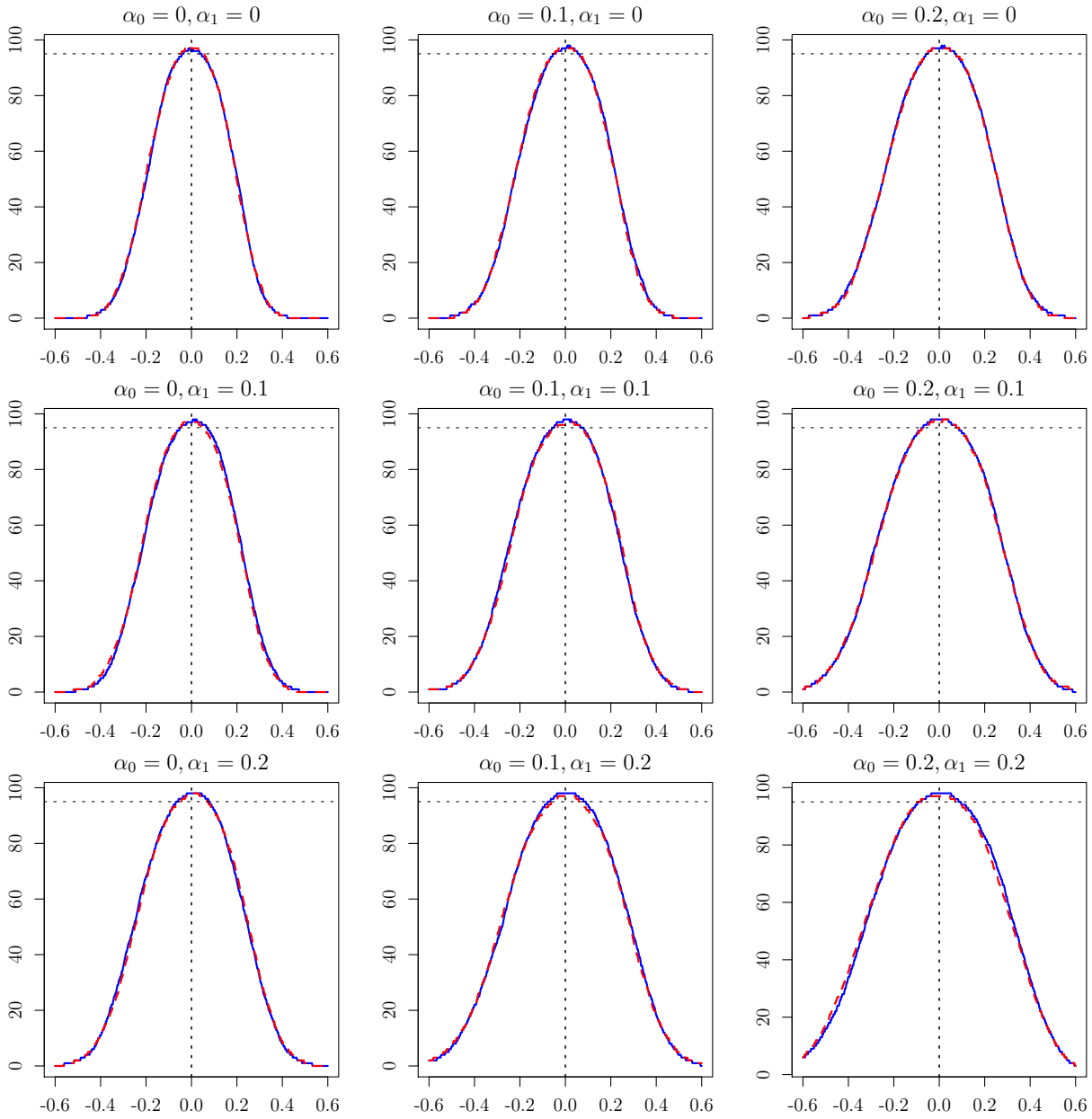

Figure E.1: Coverage Curves for Bonferroni with and without Non-differential Bounds: $\beta = 0, n = 1000$

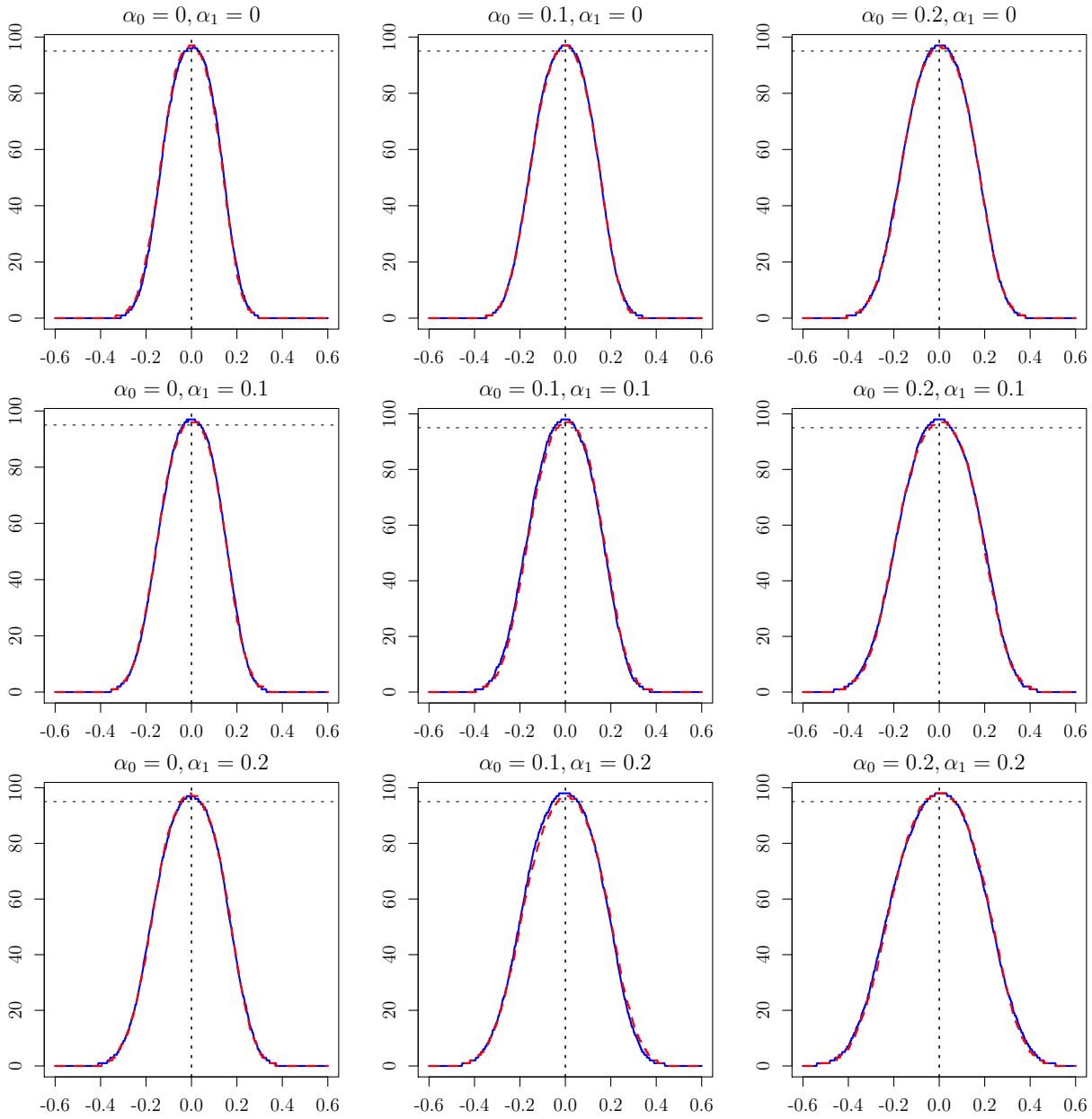

Figure E.2: Coverage Curves for Bonferroni with and without Non-differential Bounds: $\beta = 0, n = 2000$

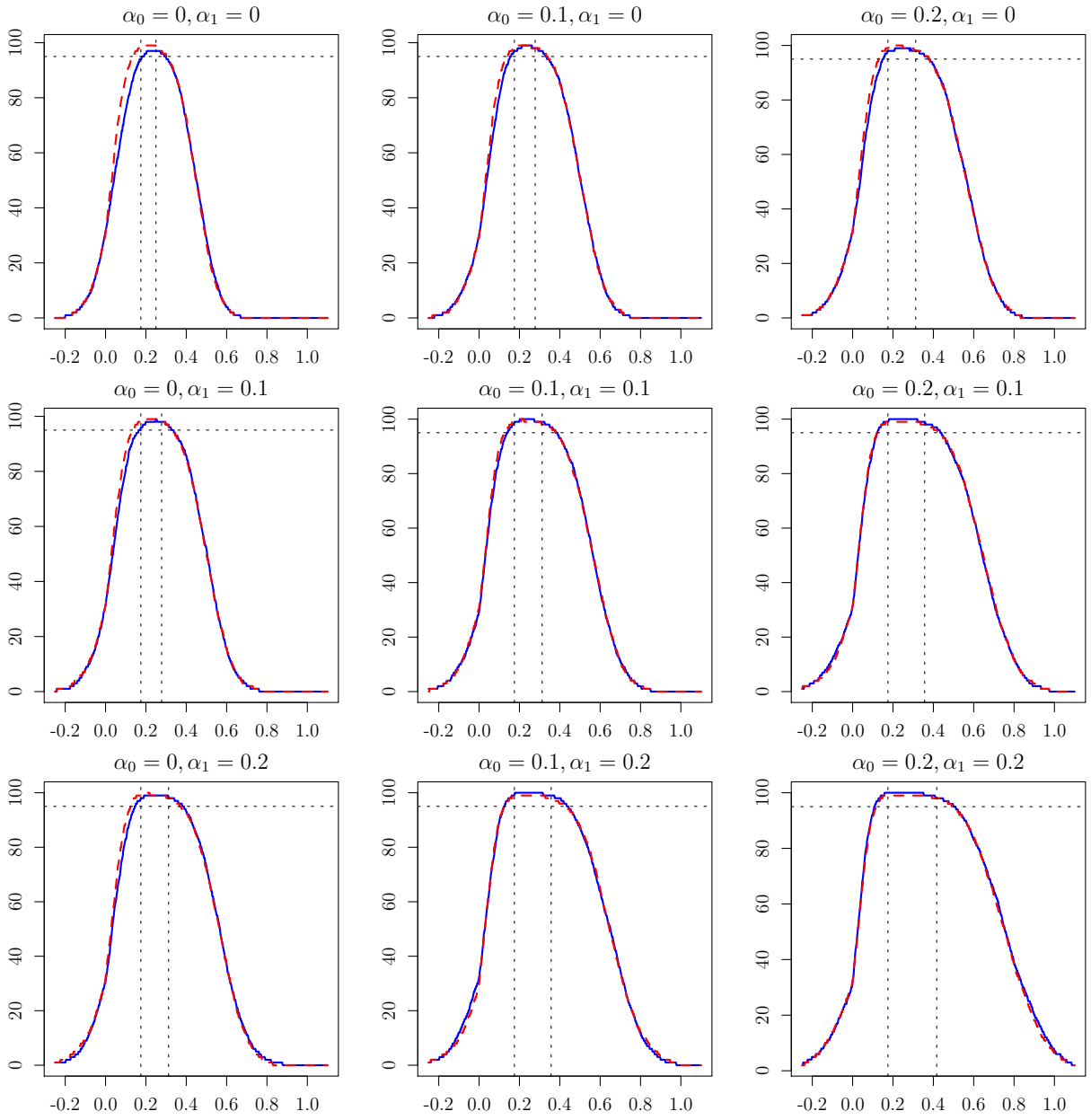

Figure E.3: Coverage Curves for Bonferroni with and without Non-differential Bounds: $\beta = 0.25, n = 1000$

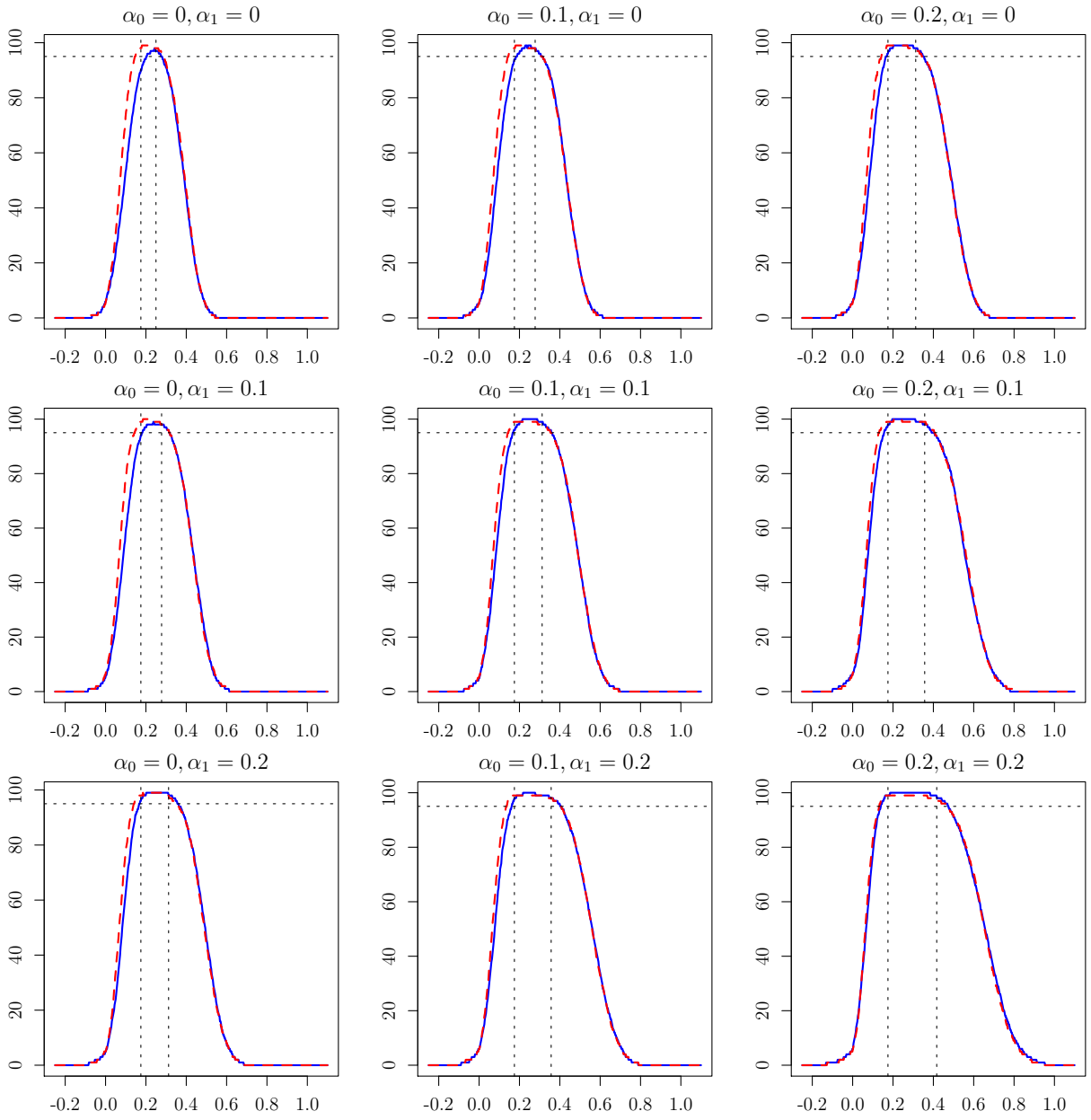

Figure E.4: Coverage Curves for Bonferroni with and without Non-differential Bounds: $\beta = 0.25, n = 2000$

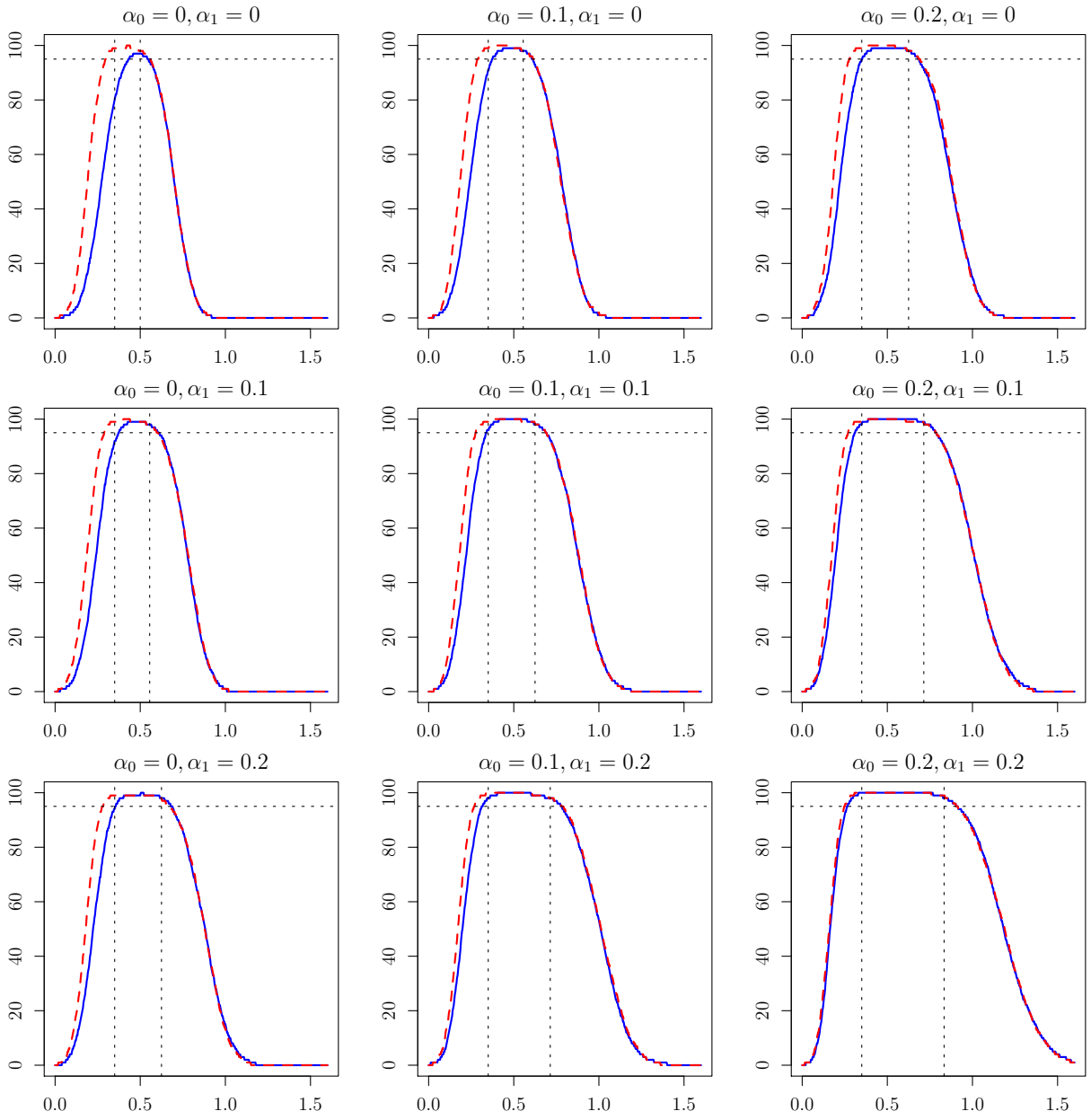

Figure E.5: Coverage Curves for Bonferroni with and without Non-differential Bounds: $\beta = 0.5, n = 1000$

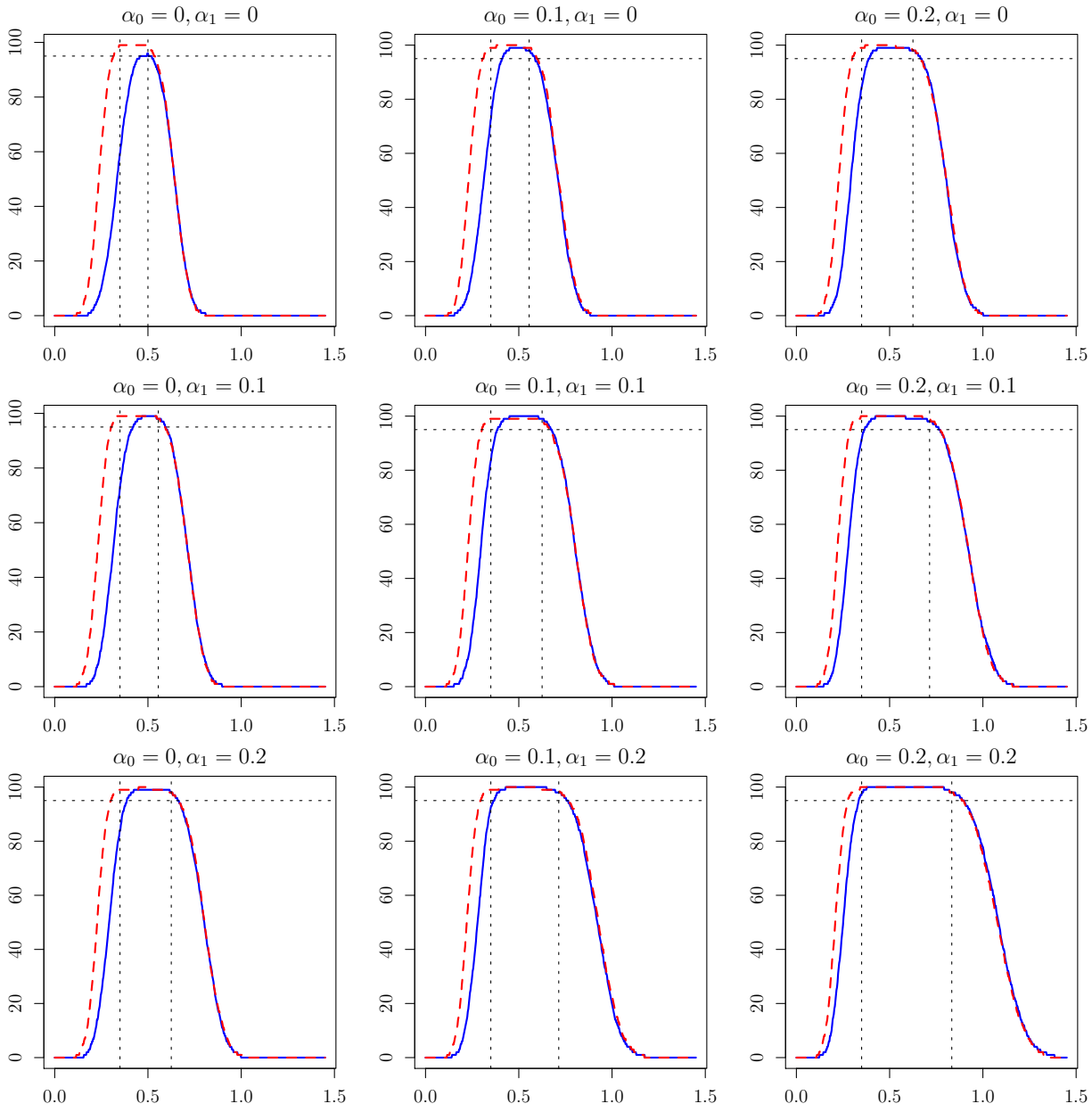

Figure E.6: Coverage Curves for Bonferroni with and without Non-differential Bounds: $\beta = 0.5, n = 2000$

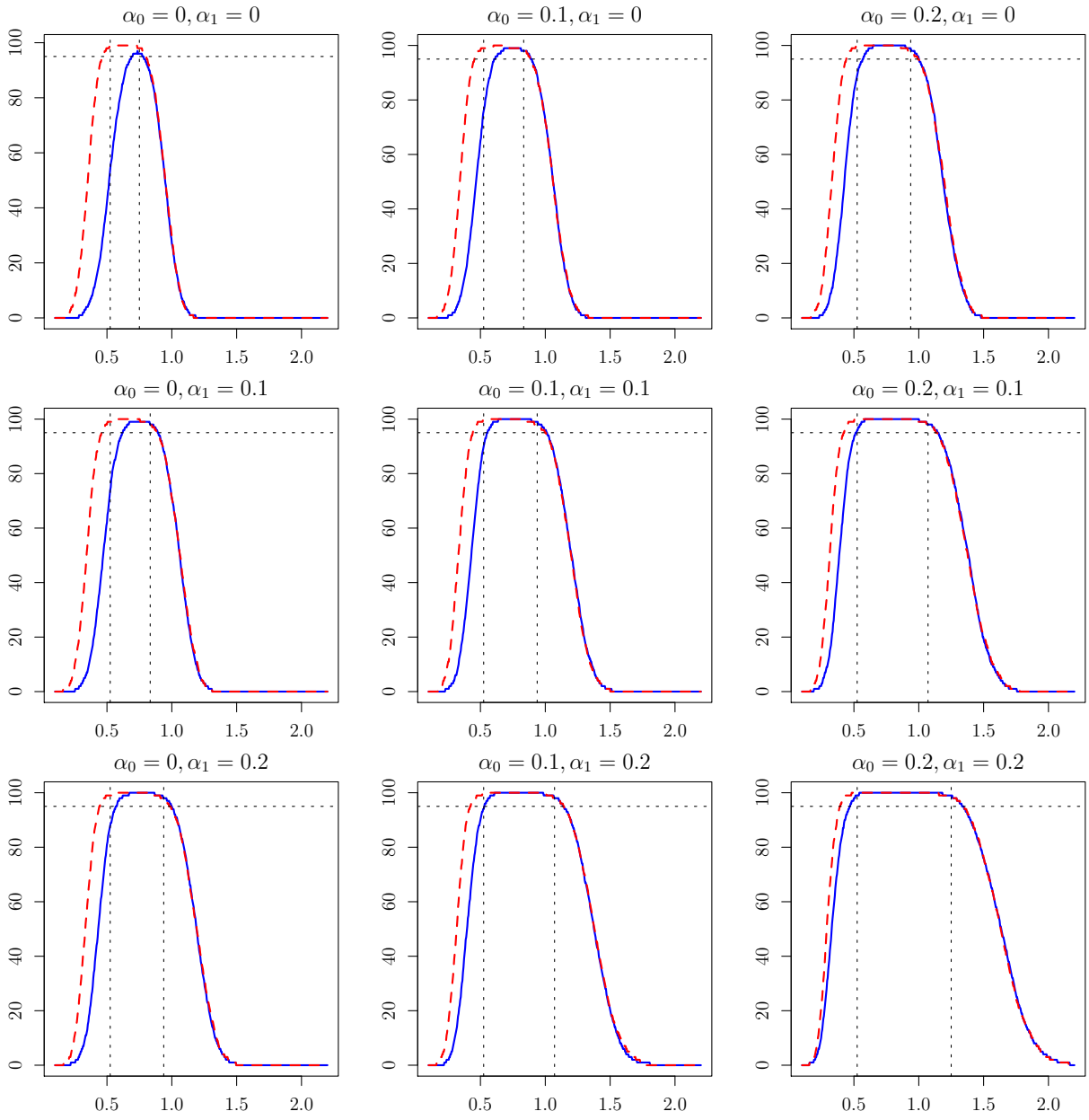

Figure E.7: Coverage Curves for Bonferroni with and without Non-differential Bounds: $\beta = 0.75, n = 1000$

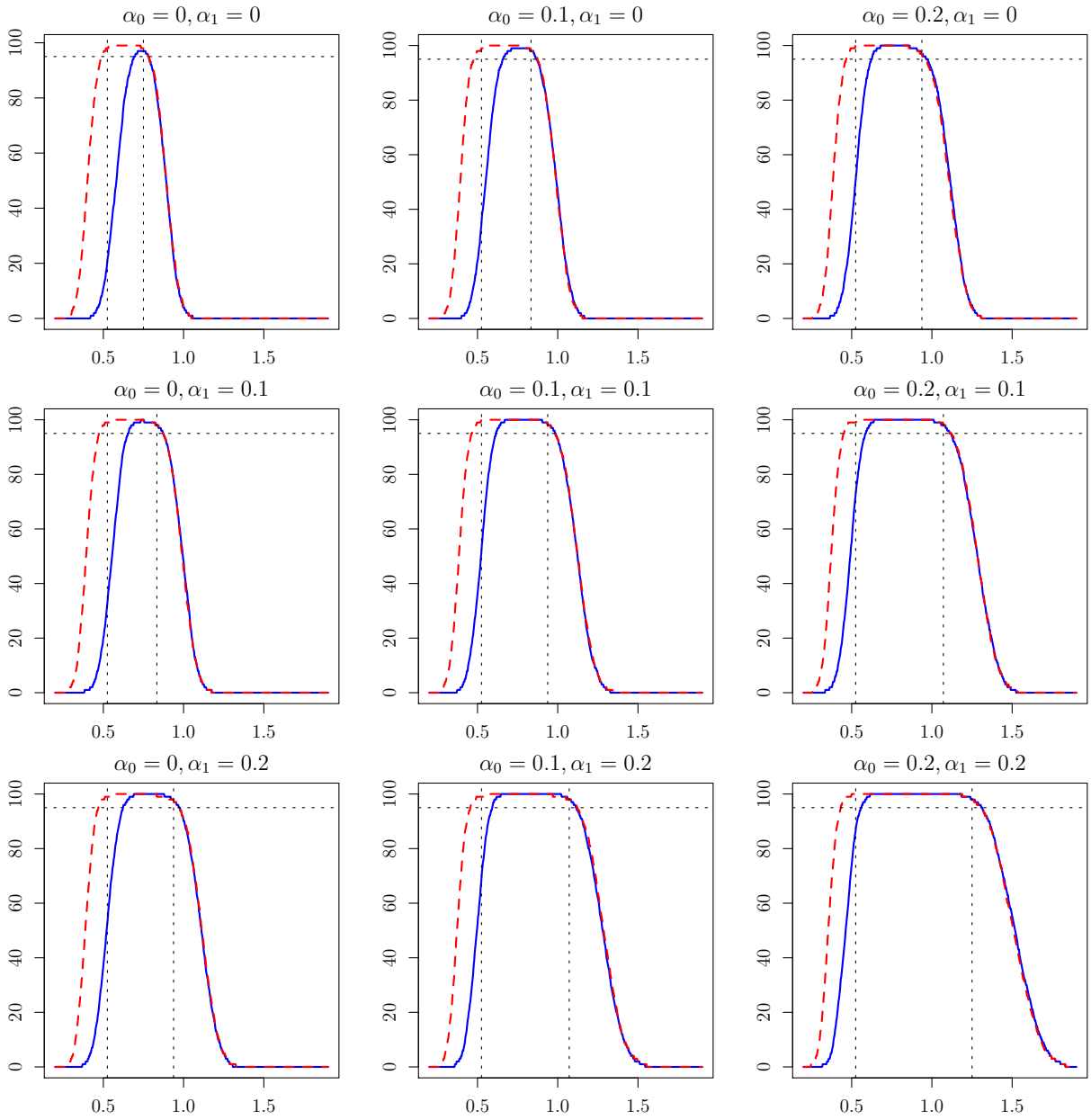

Figure E.8: Coverage Curves for Bonferroni with and without Non-differential Bounds: $\beta = 0.75, n = 2000$

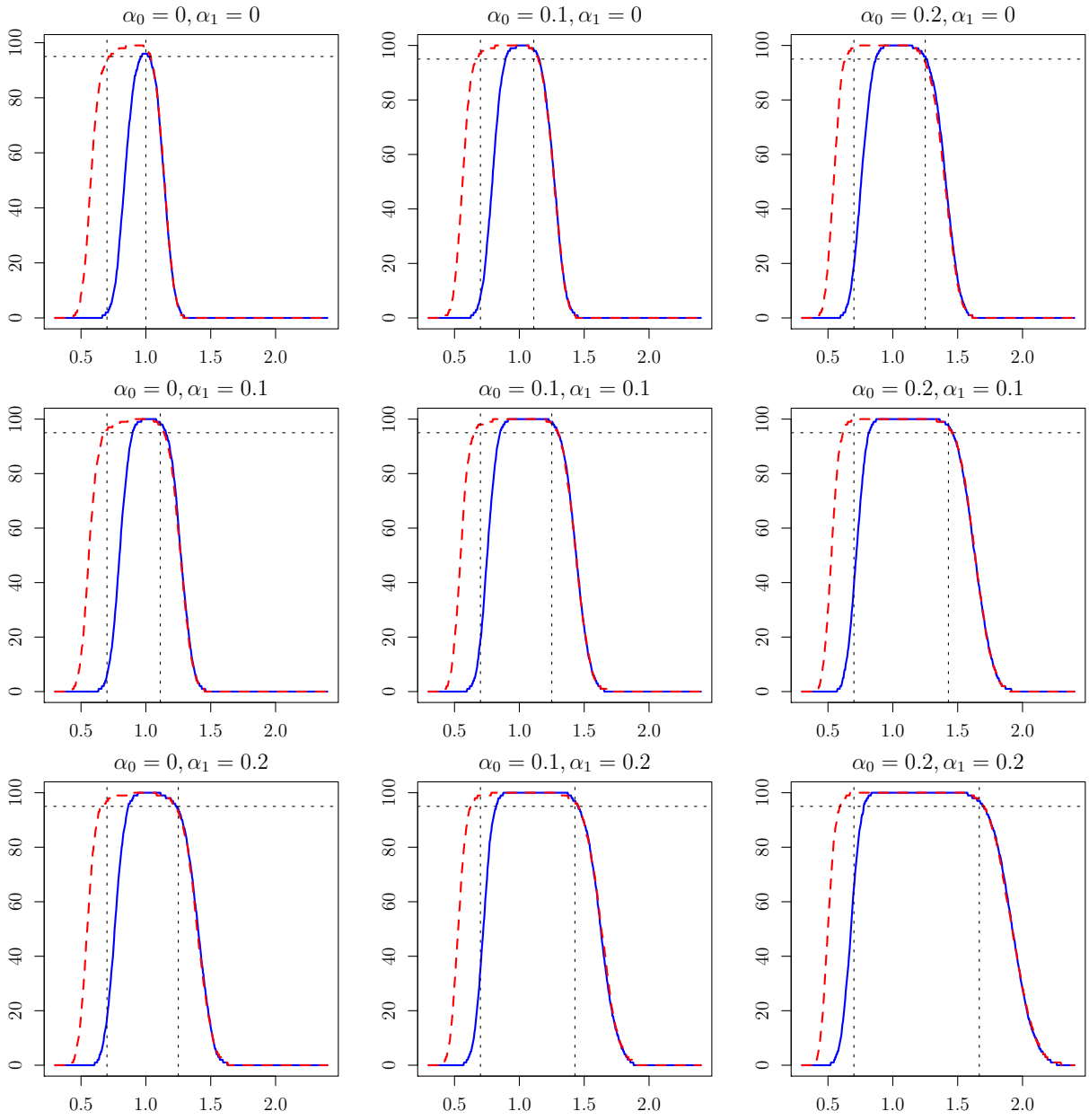

Figure E.9: Coverage Curves for Bonferroni with and without Non-differential Bounds: $\beta = 1, n = 1000$

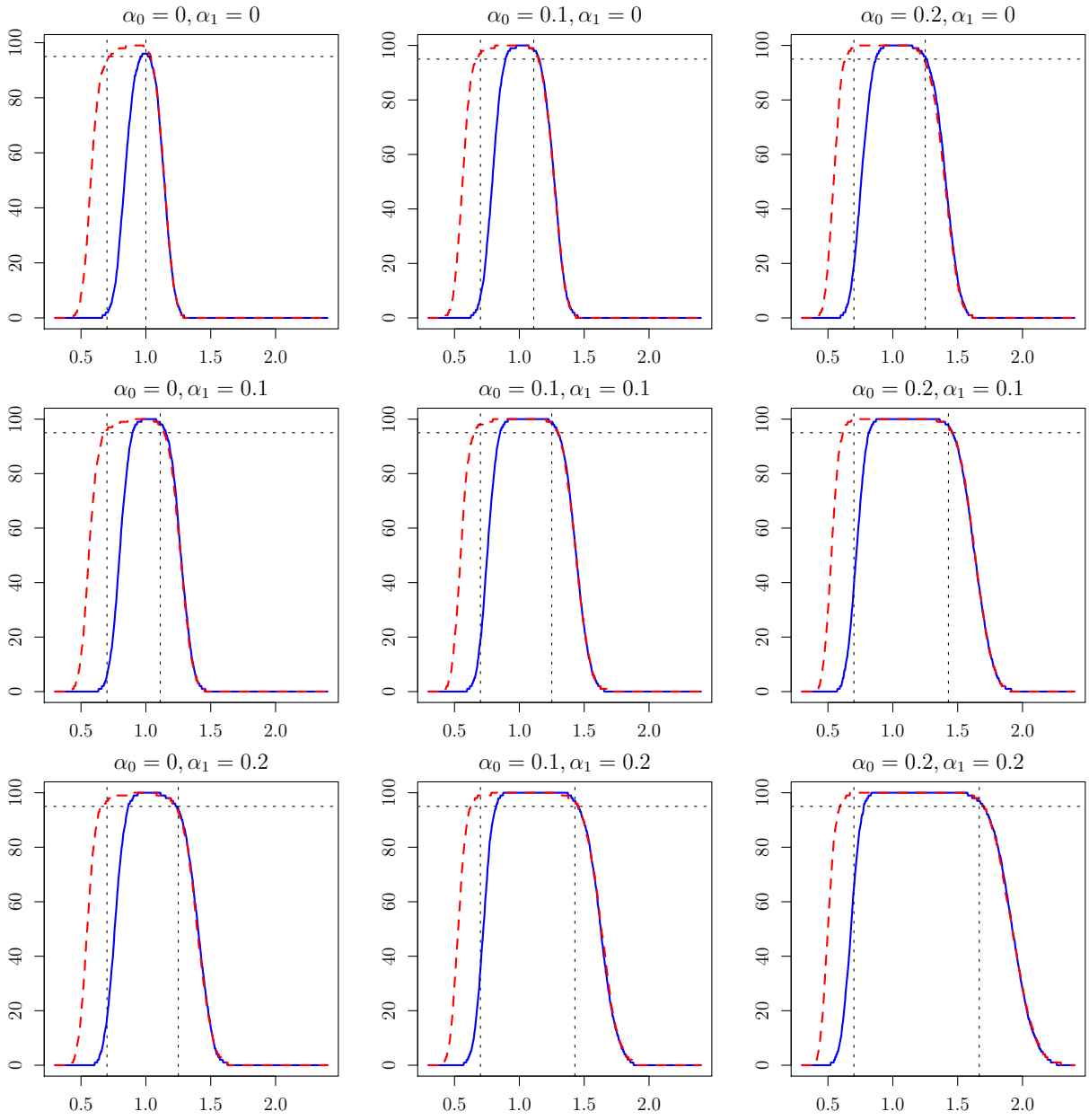

Figure E.10: Coverage Curves for Bonferroni with and without Non-differential Bounds: $\beta = 1, n = 2000$

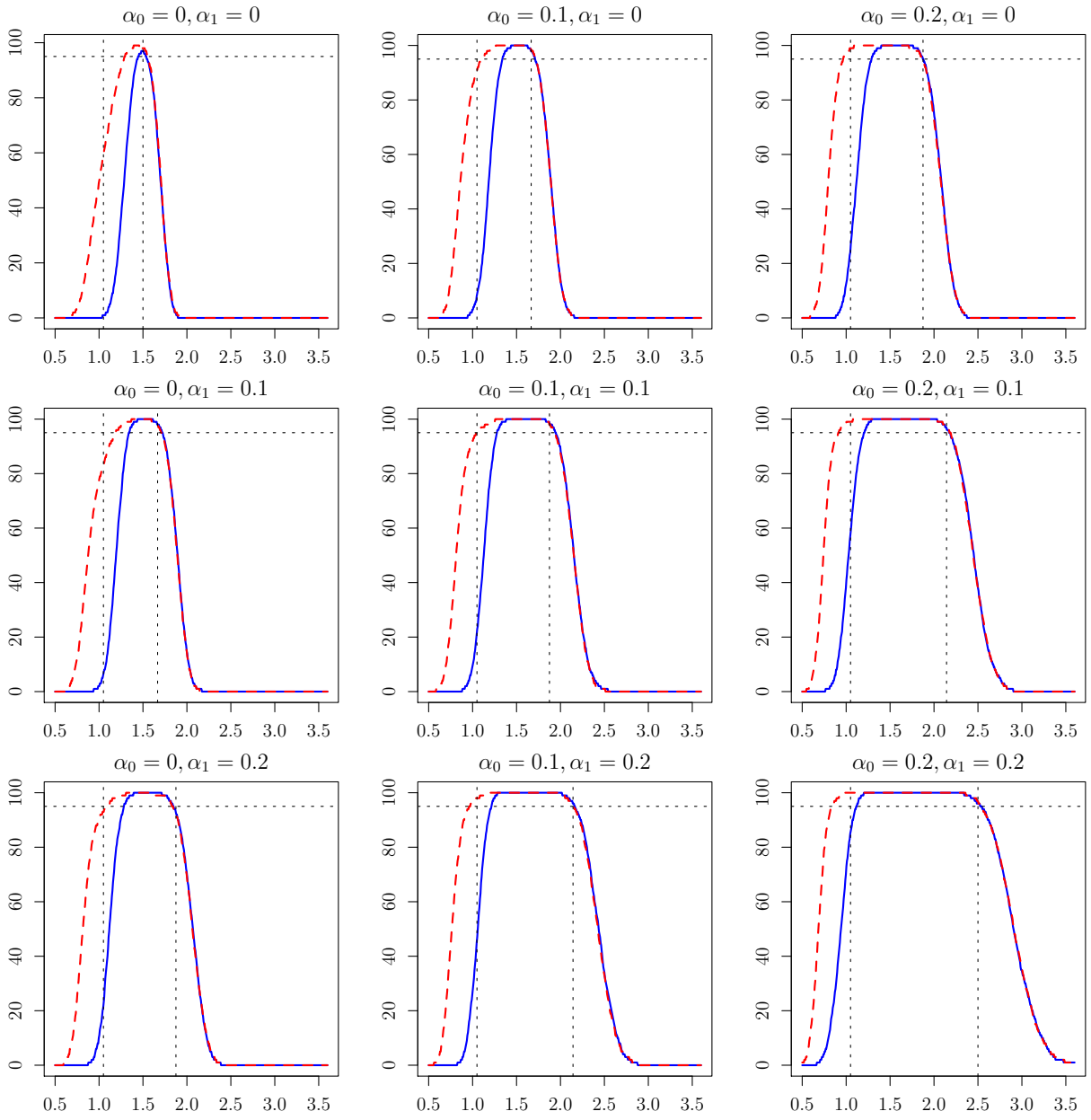

Figure E.11: Coverage Curves for Bonferroni with and without Non-differential Bounds: $\beta = 1.5, n = 1000$

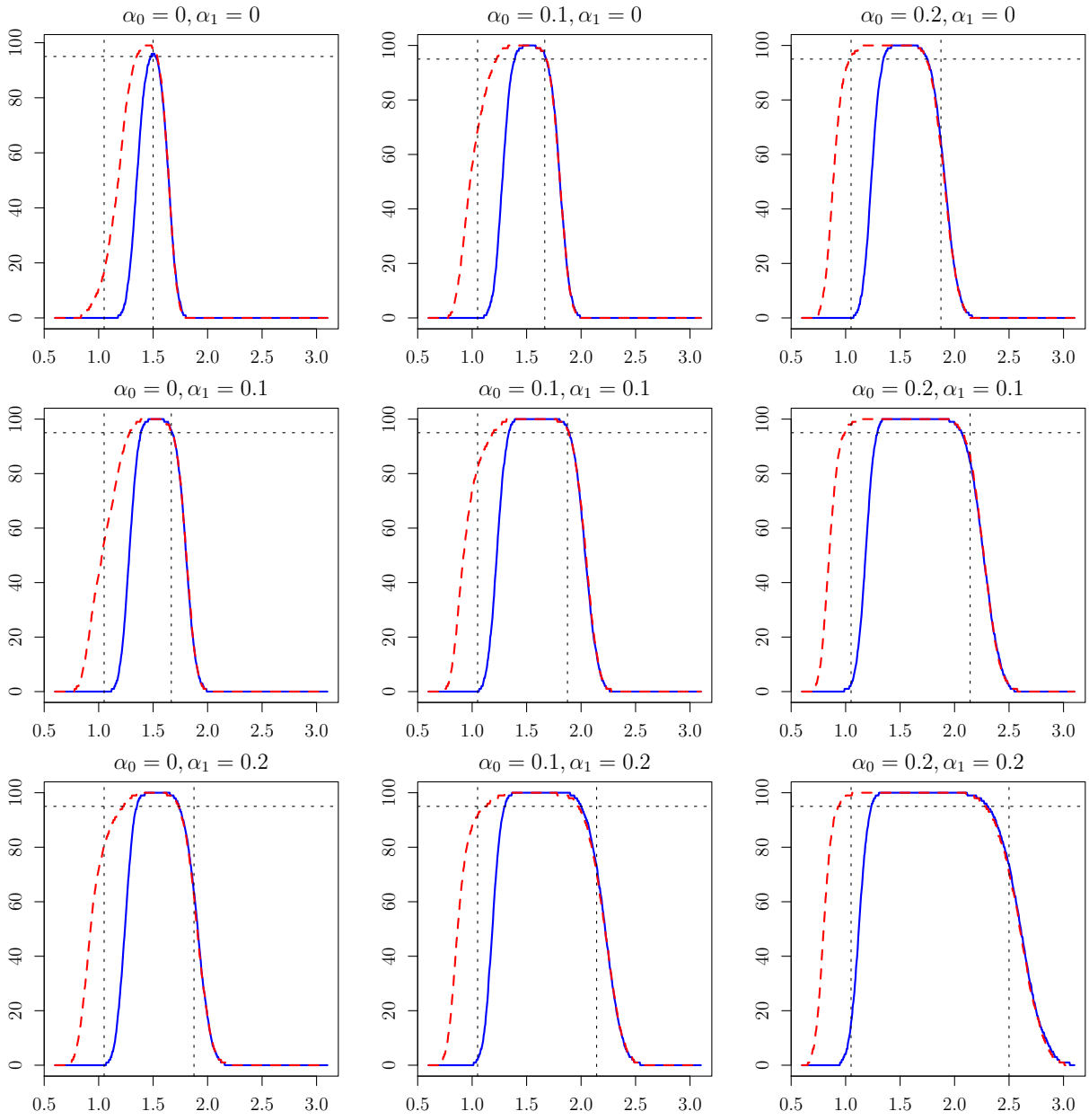

Figure E.12: Coverage Curves for Bonferroni with and without Non-differential Bounds: $\beta = 1.5, n = 2000$

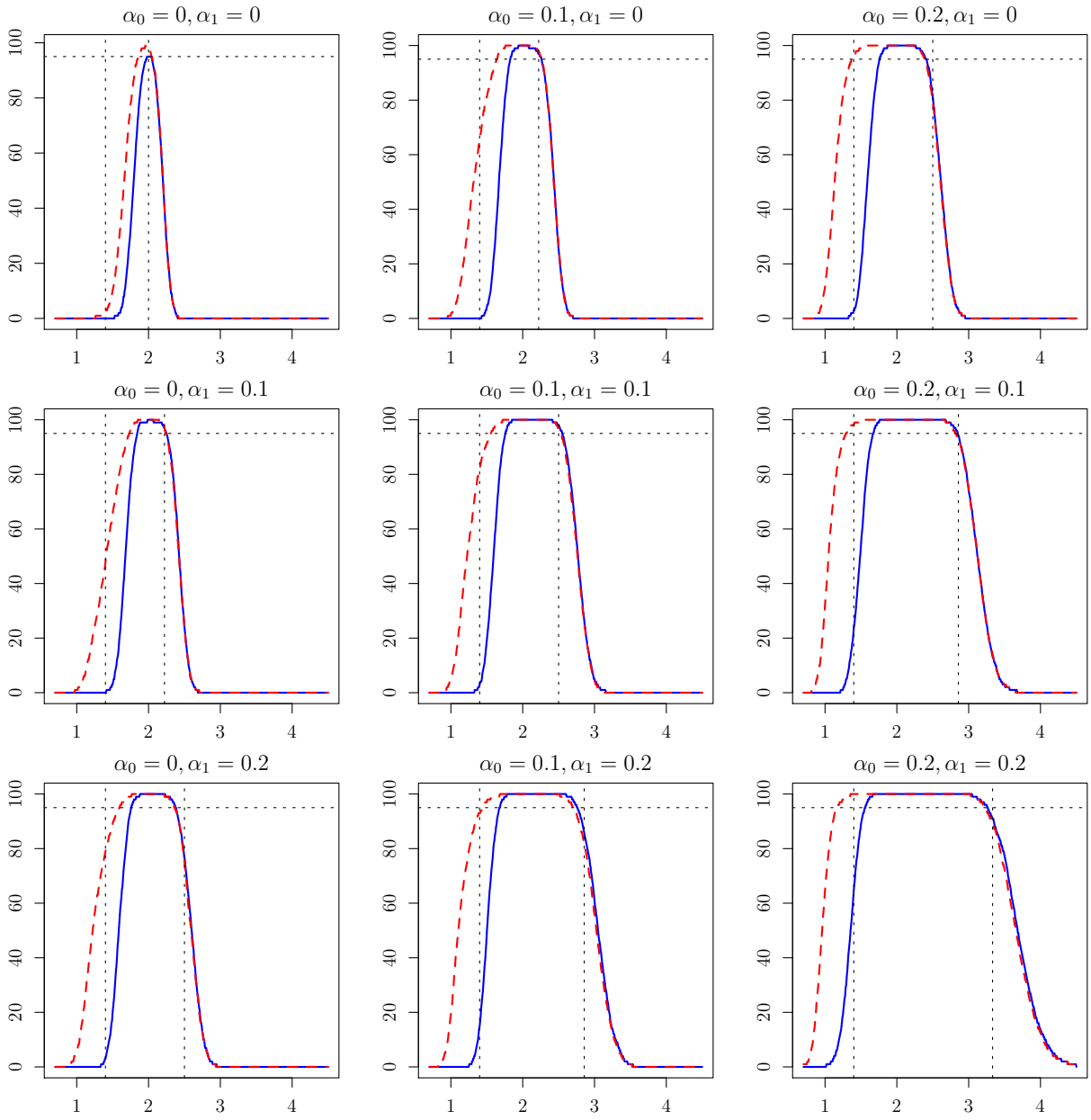

Figure E.13: Coverage Curves for Bonferroni with and without Non-differential Bounds: $\beta = 2, n = 1000$

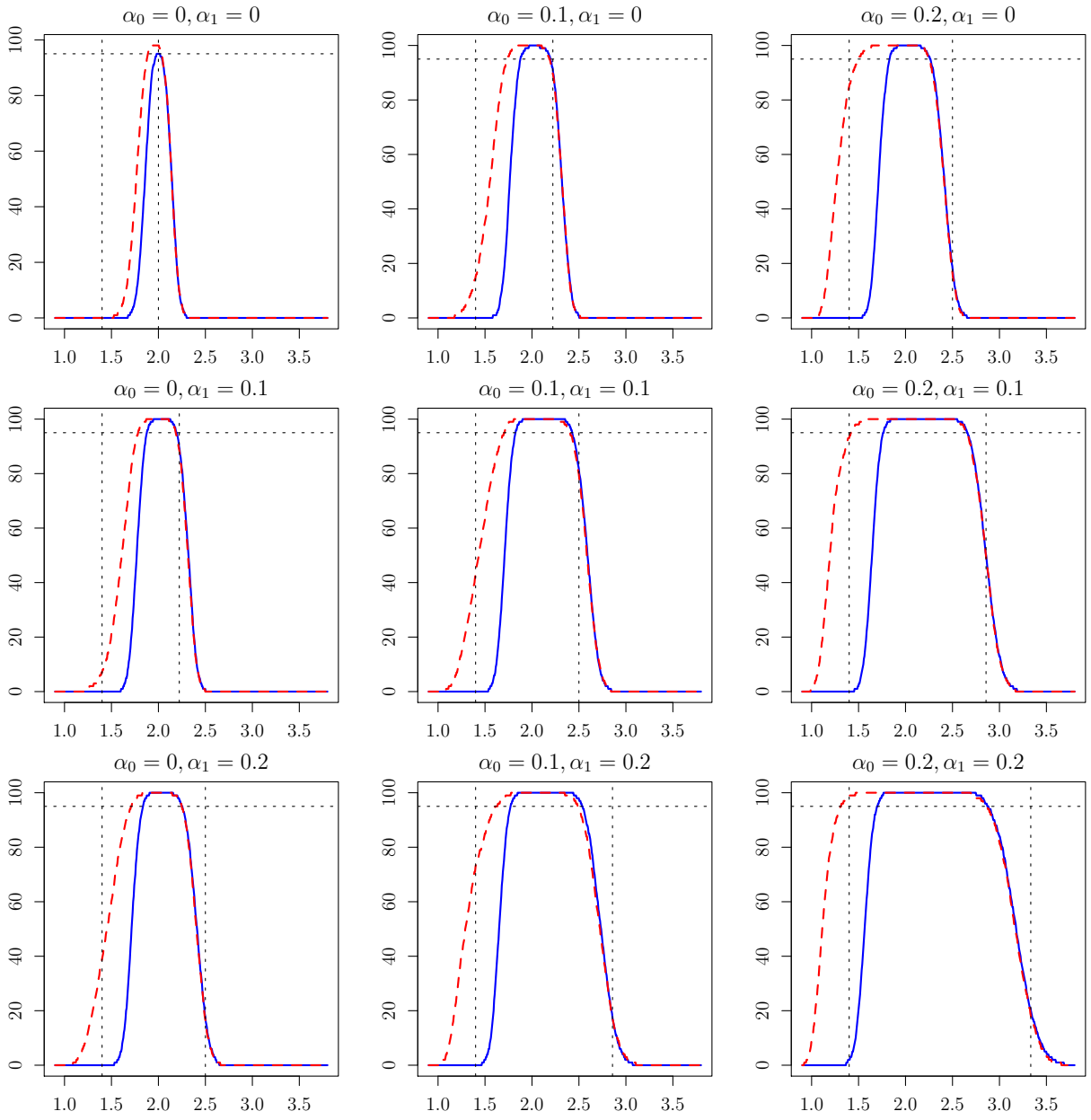

Figure E.14: Coverage Curves for Bonferroni with and without Non-differential Bounds: $\beta = 2, n = 2000$

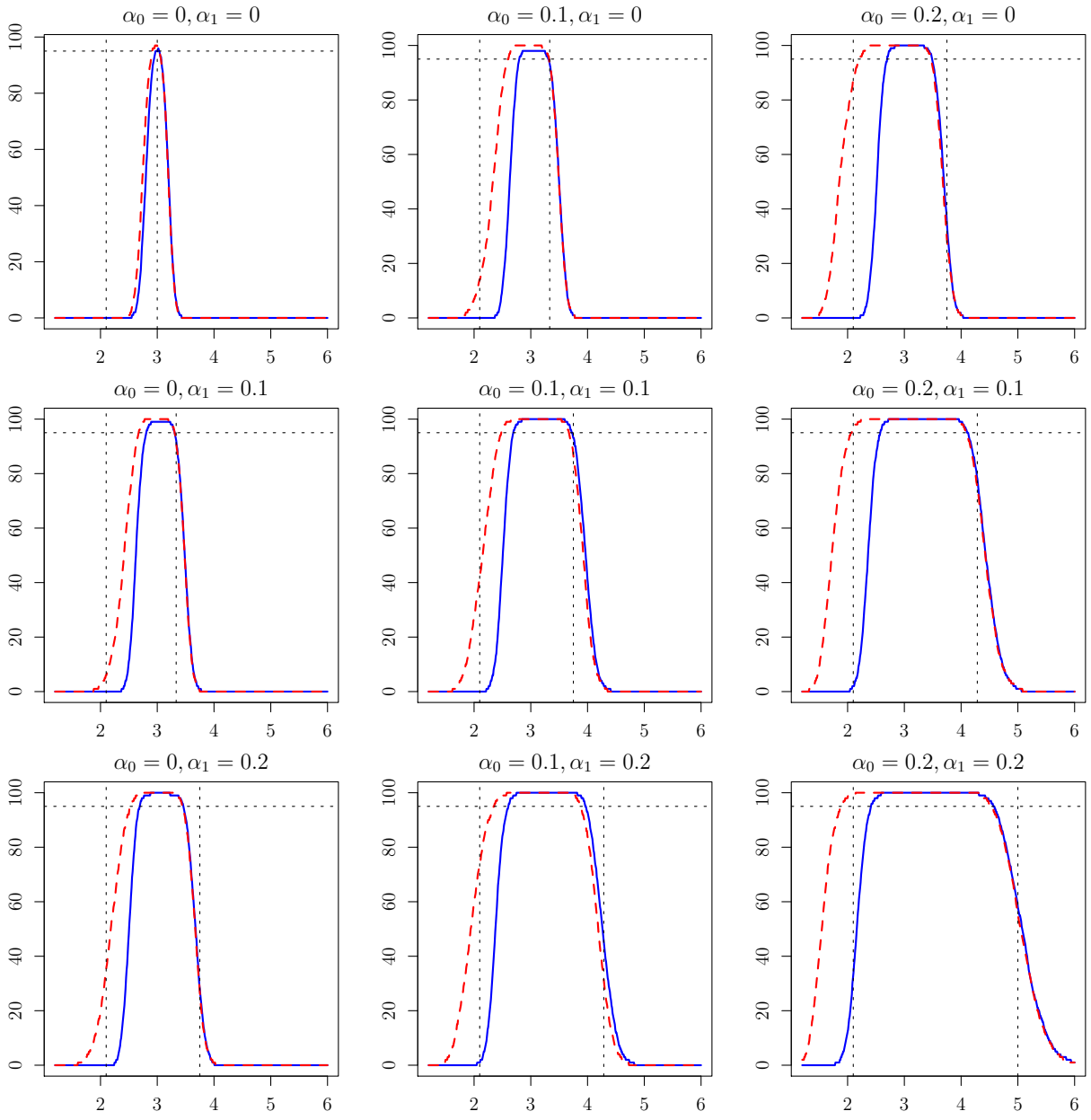

Figure E.15: Coverage Curves for Bonferroni with and without Non-differential Bounds: $\beta = 3, n = 1000$

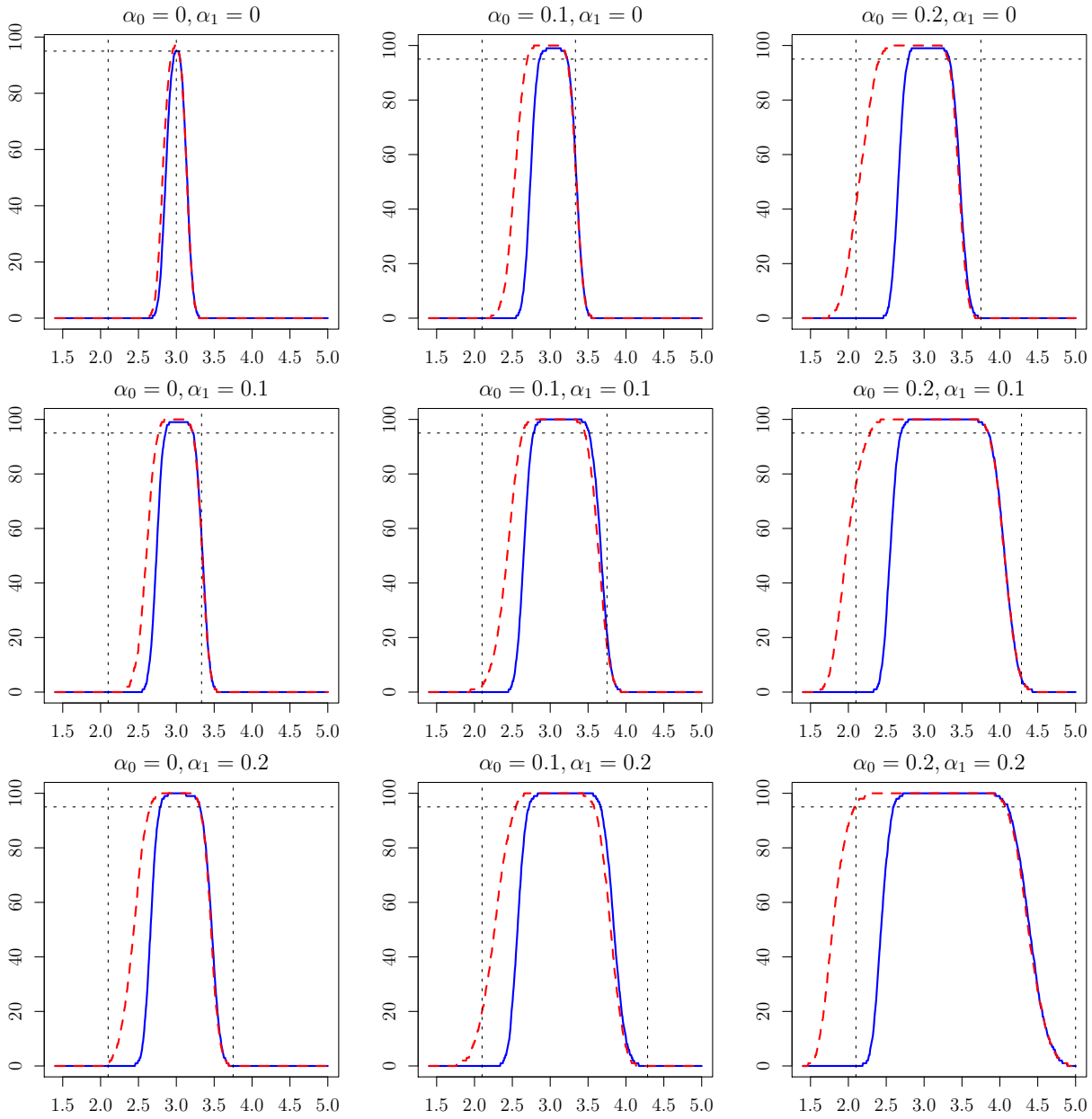

Figure E.16: Coverage Curves for Bonferroni with and without Non-differential Bounds: $\beta = 3, n = 2000$

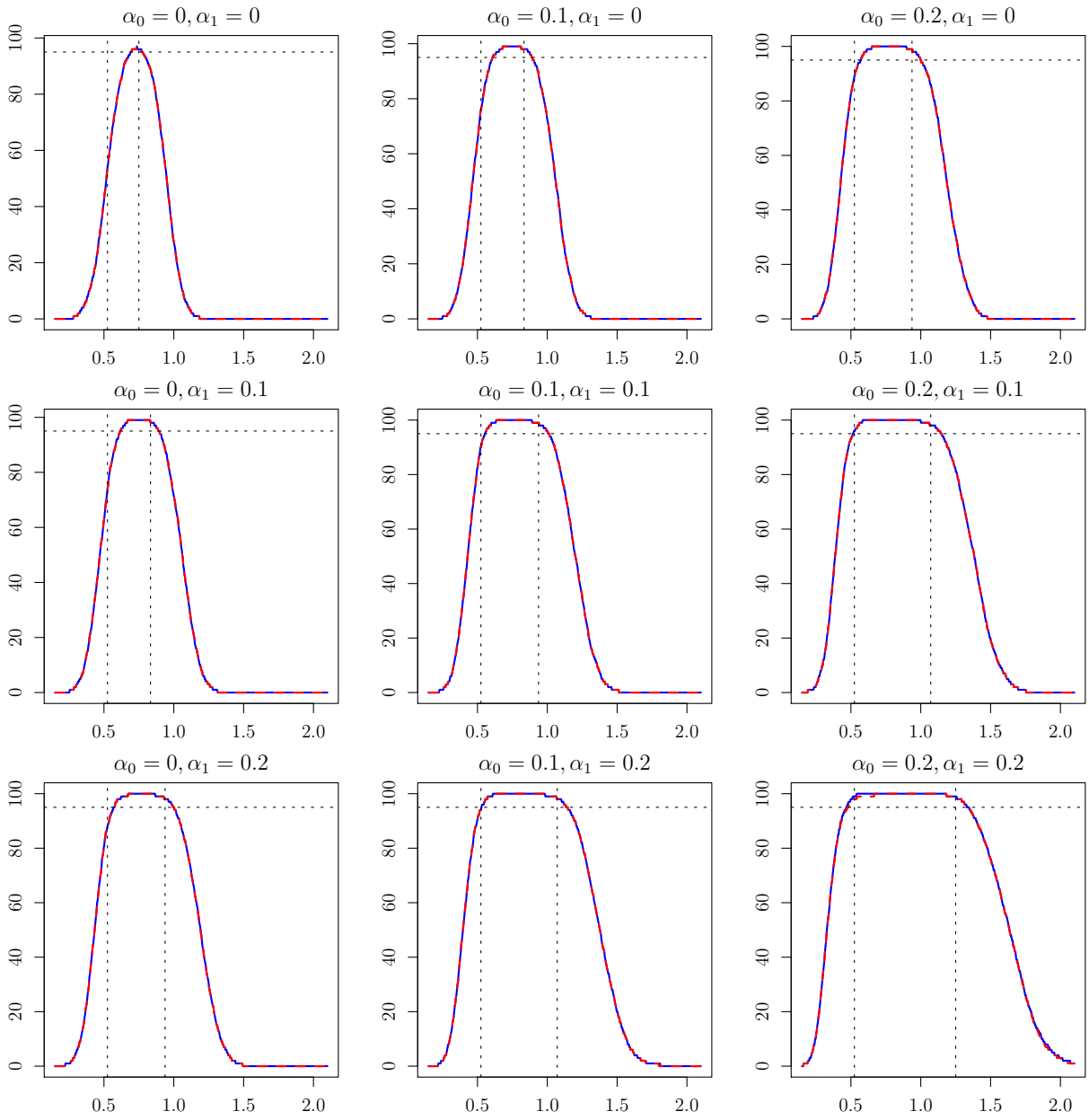

Figure E.17: Coverage Curves for Bonferroni versus Hybrid CIs: $\beta = 0.75, n = 1000$

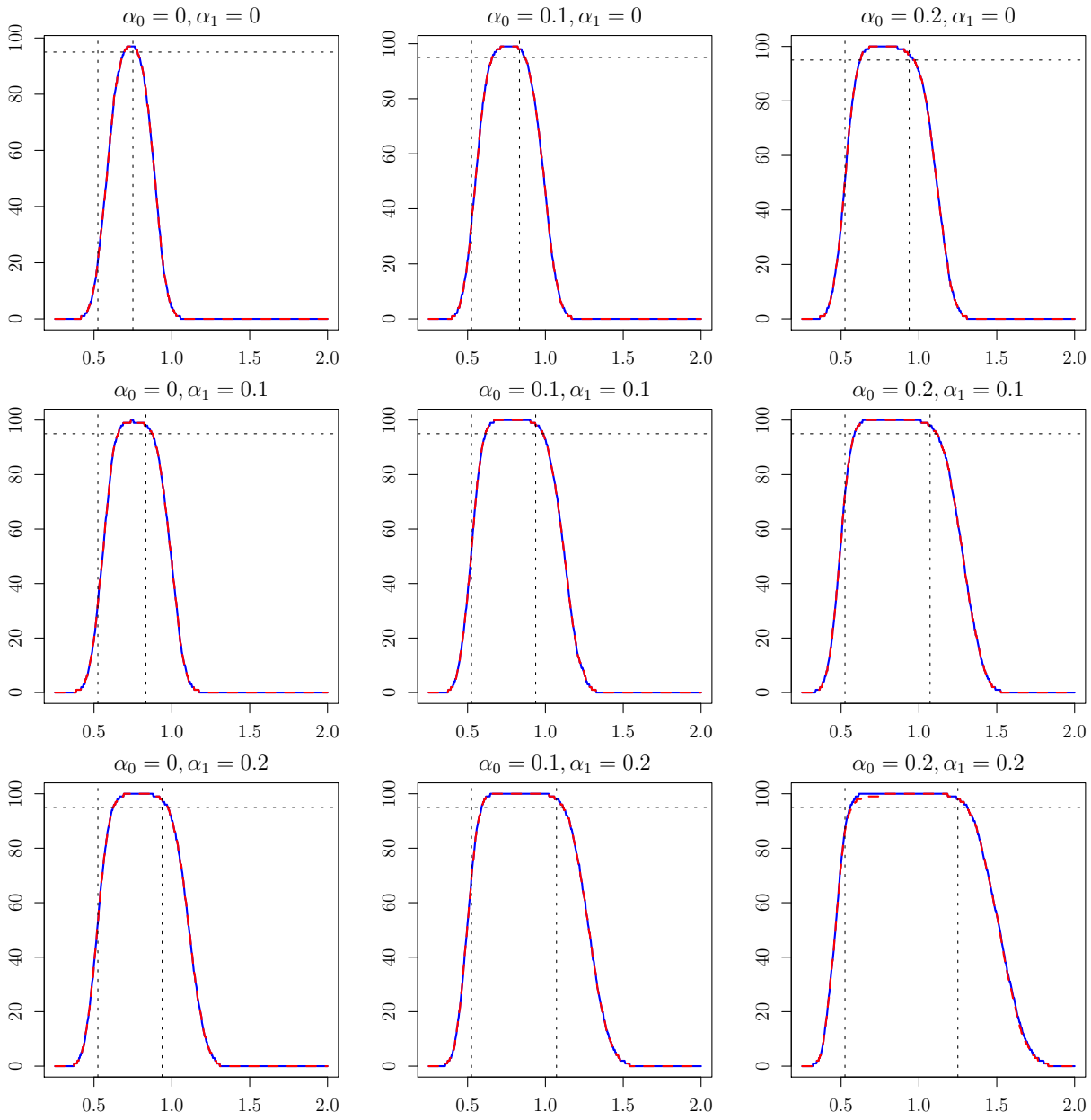

Figure E.18: Coverage Curves for Bonferroni versus Hybrid CIs: $\beta = 0.75, n = 2000$

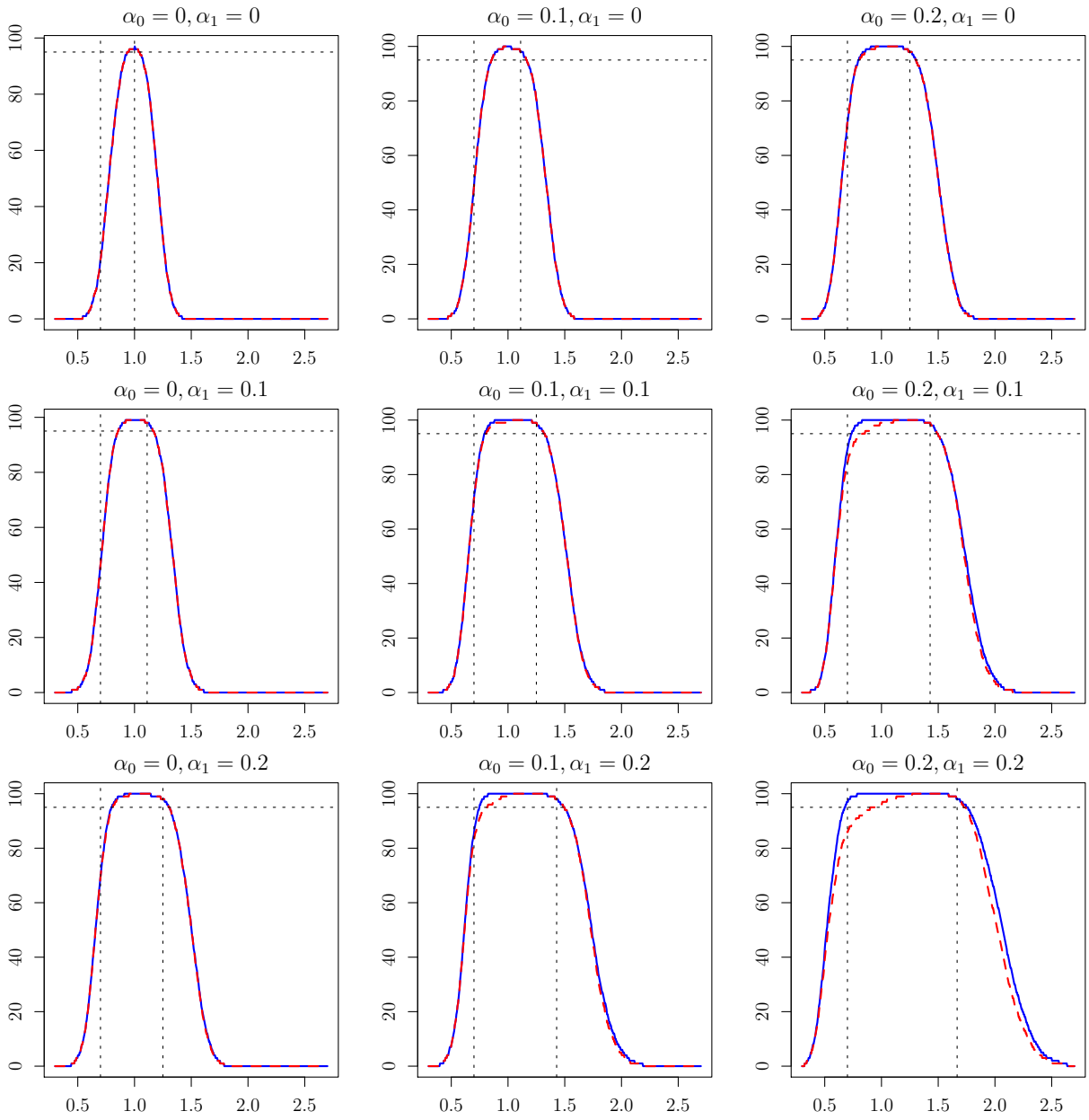

Figure E.19: Coverage Curves for Bonferroni versus Hybrid CIs: $\beta = 1, n = 1000$

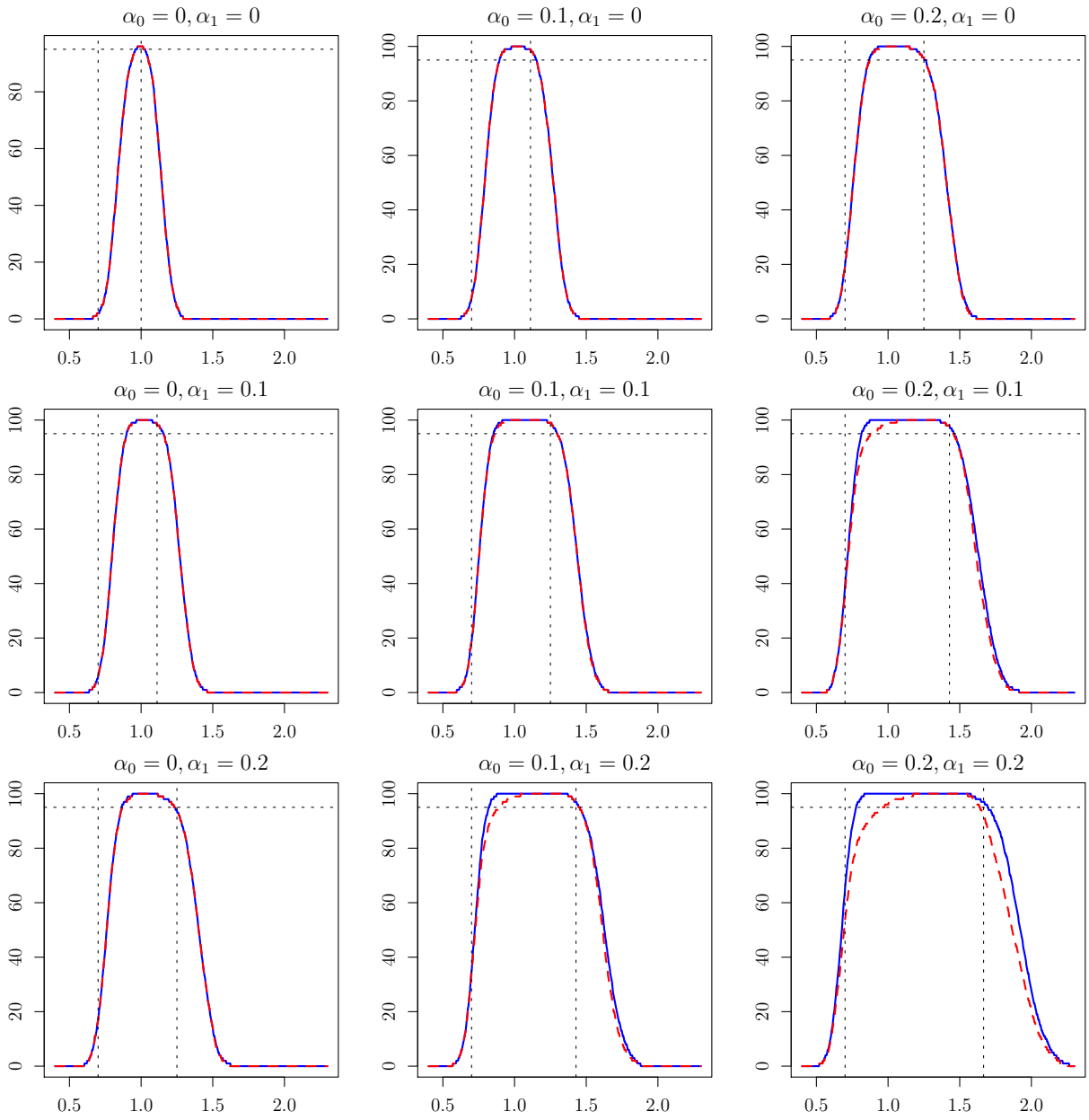

Figure E.20: Coverage Curves for Bonferroni versus Hybrid CIs: $\beta = 1, n = 2000$

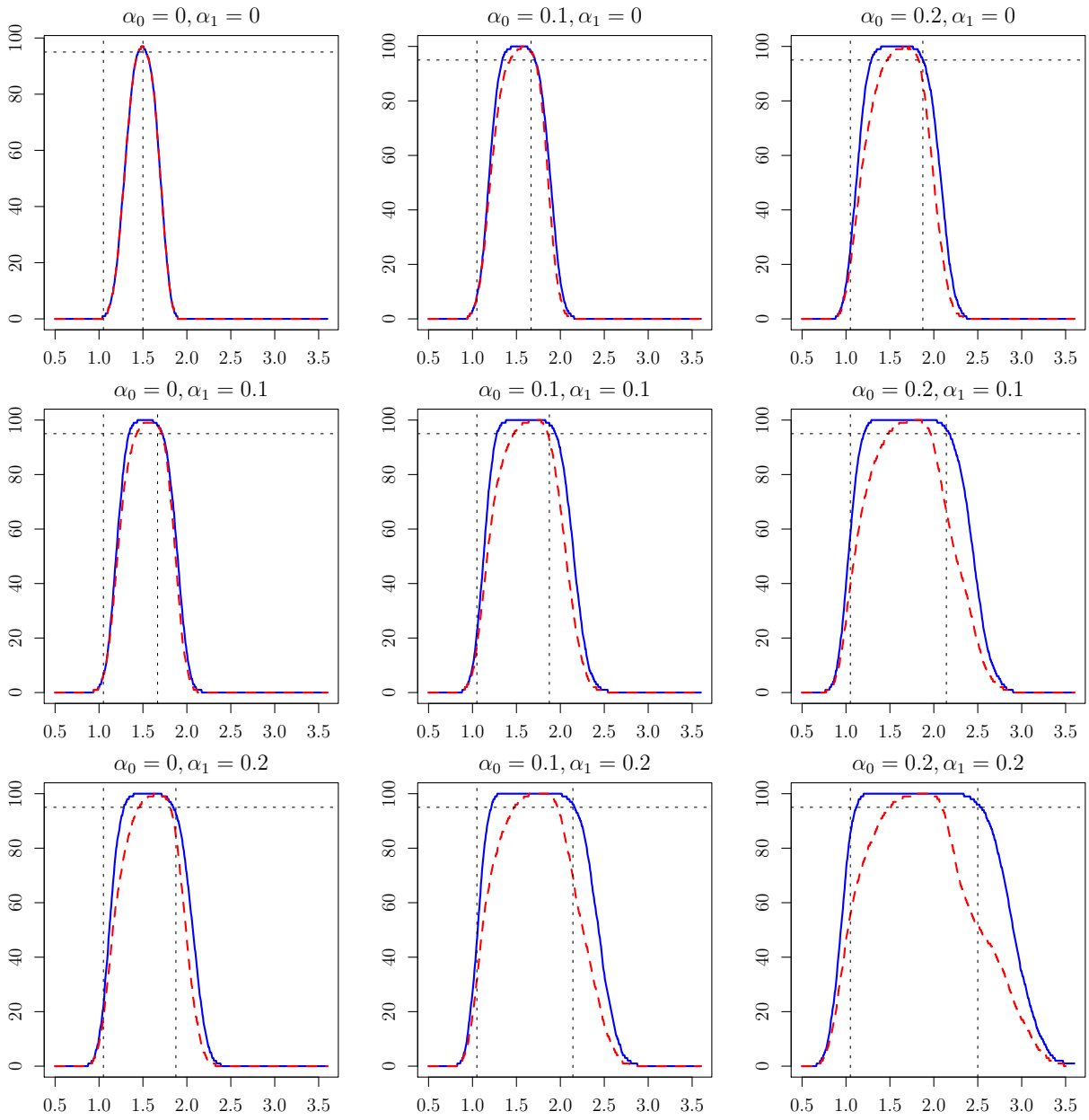

Figure E.21: Coverage Curves for Bonferroni versus Hybrid CIs: $\beta = 1.5, n = 1000$

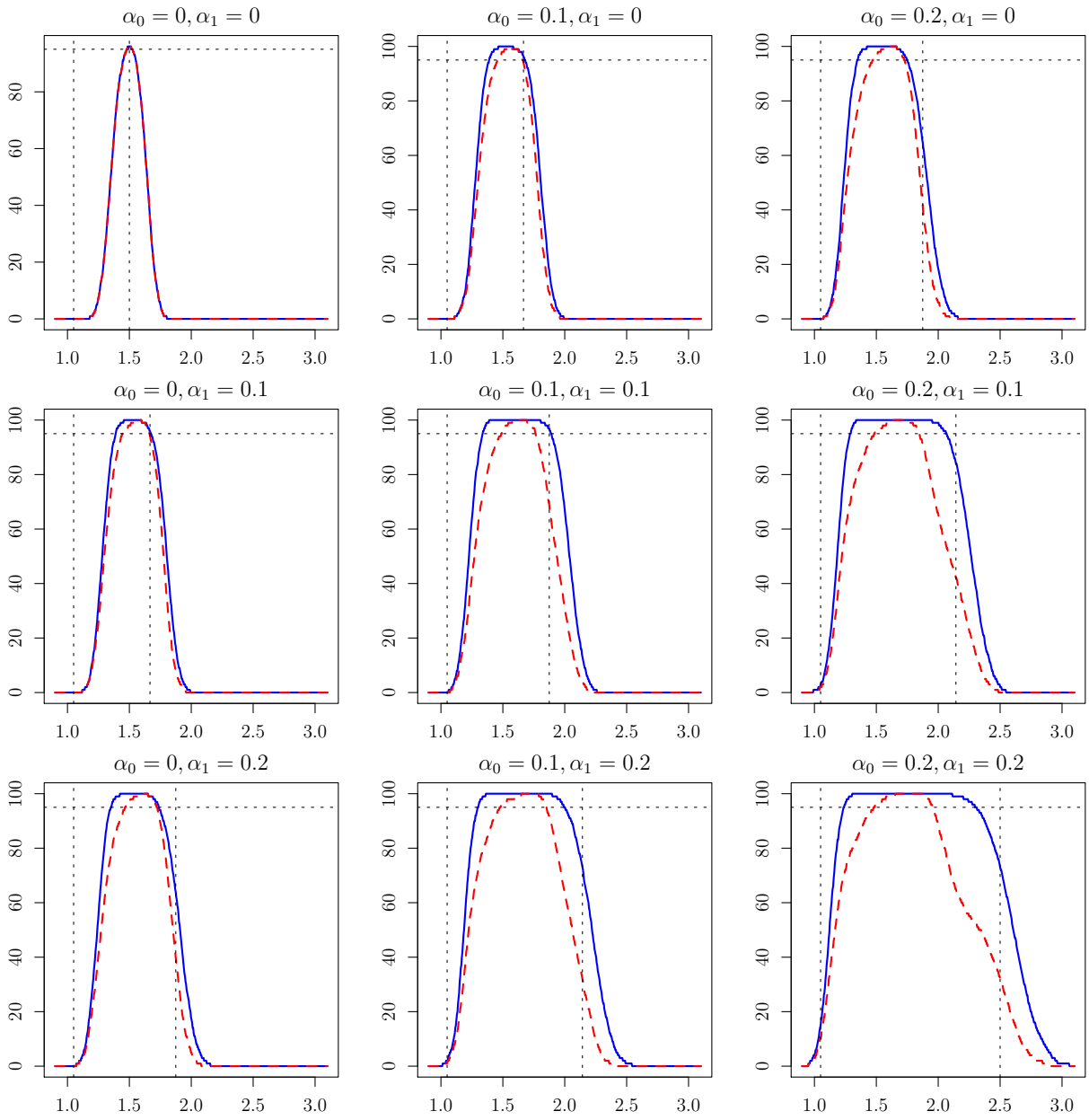

Figure E.22: Coverage Curves for Bonferroni versus Hybrid CIs: $\beta = 1.5, n = 2000$

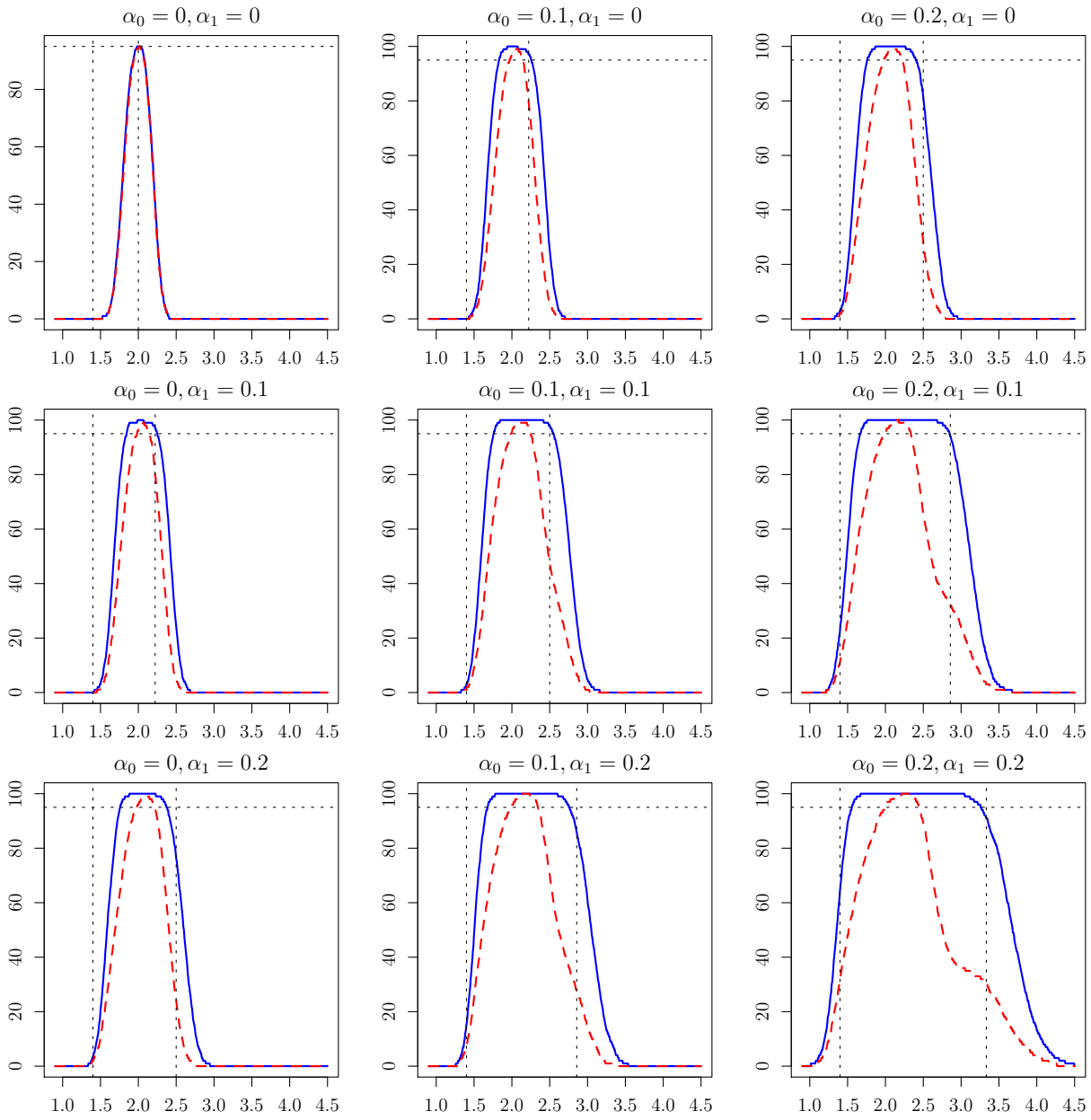

Figure E.23: Coverage Curves for Bonferroni versus Hybrid CIs: $\beta = 2, n = 1000$

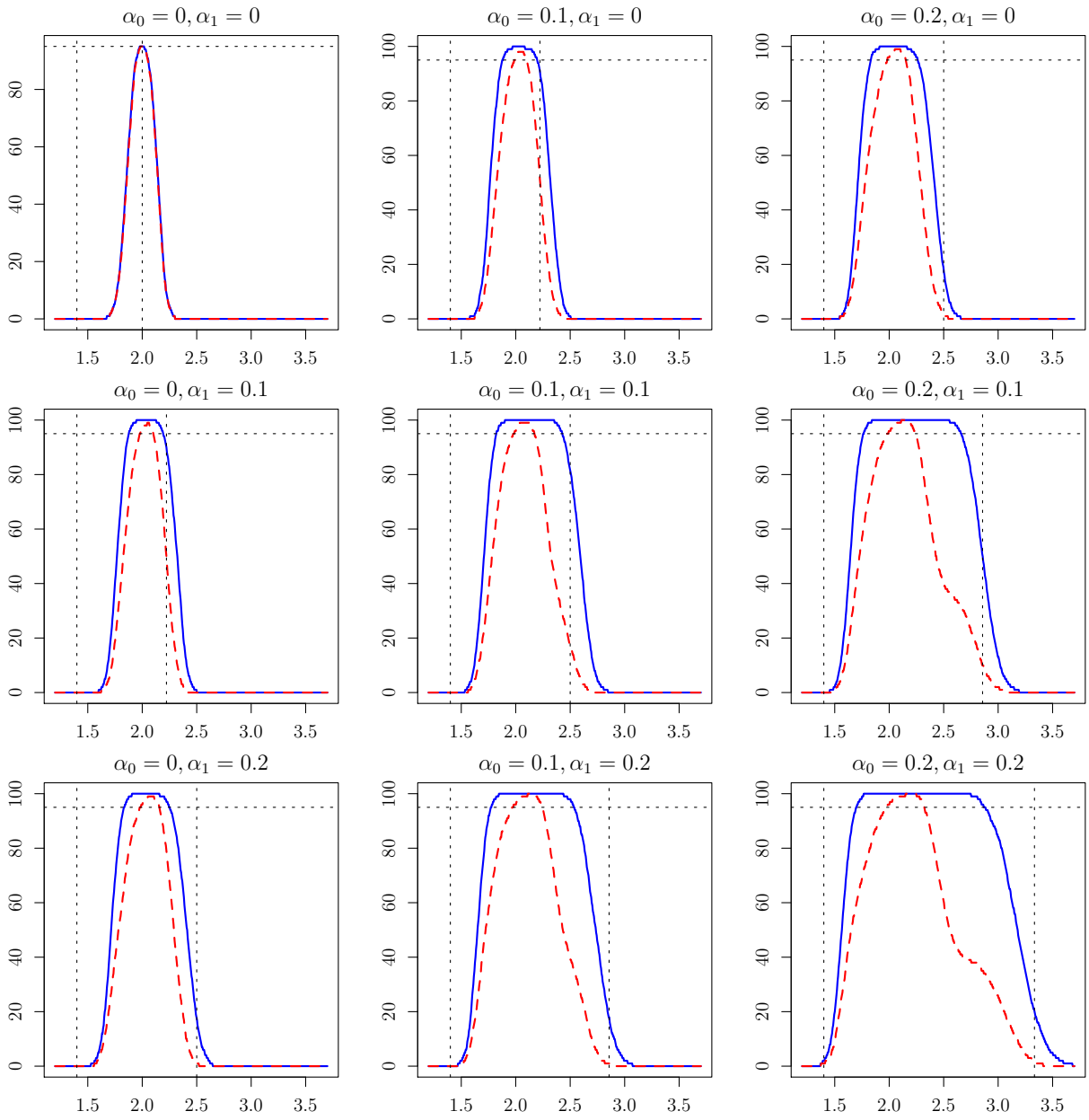

Figure E.24: Coverage Curves for Bonferroni versus Hybrid CIs: $\beta = 2, n = 2000$

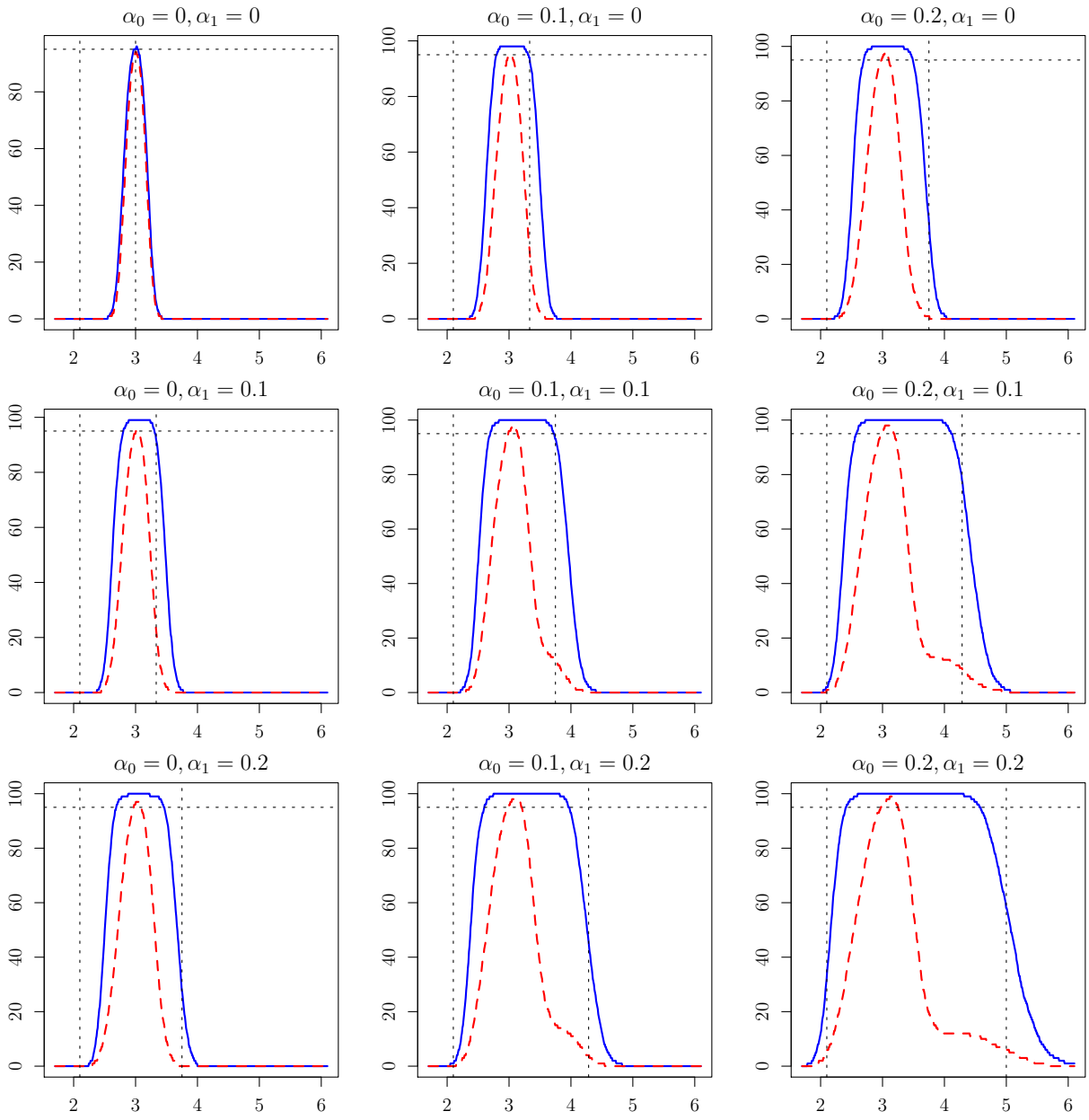

Figure E.25: Coverage Curves for Bonferroni versus Hybrid CIs: $\beta = 3, n = 1000$

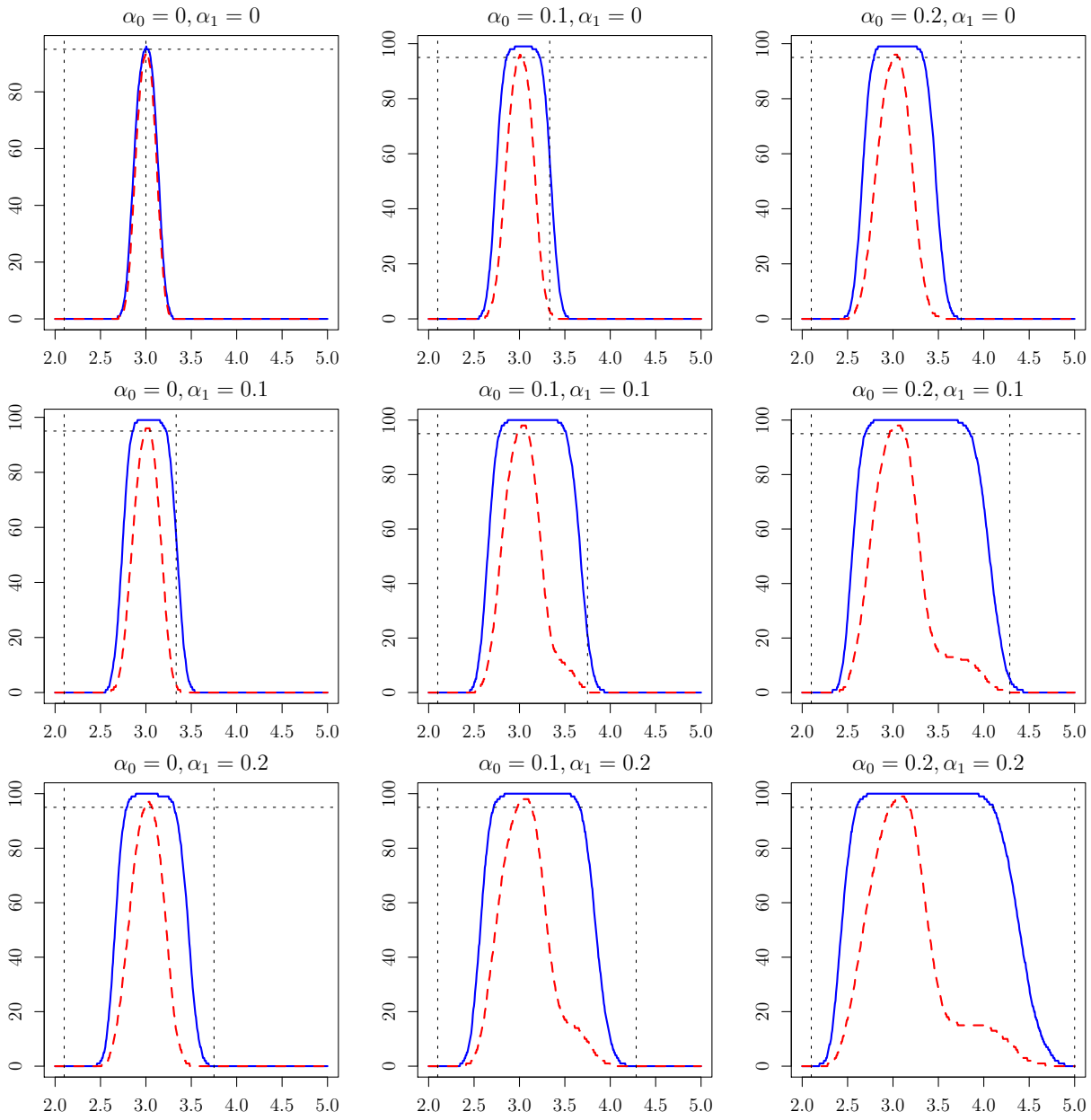

Figure E.26: Coverage Curves for Bonferroni versus Hybrid CIs: $\beta = 3, n = 2000$

		β							
α_0	α_1	0	0.25	0.5	0.75	1	1.5	2	3
0.0	0.0	0.29	0.3	0.31	0.31	0.31	0.3	0.29	0.25
	0.1	0.32	0.35	0.4	0.44	0.48	0.48	0.36	0.33
	0.2	0.36	0.41	0.51	0.59	0.65	0.57	0.46	0.41
	0.3	0.41	0.48	0.64	0.76	0.79	0.68	0.56	0.5
0.1	0.0	0.32	0.35	0.4	0.44	0.48	0.48	0.37	0.33
	0.1	0.36	0.41	0.51	0.6	0.68	0.65	0.48	0.43
	0.2	0.41	0.48	0.64	0.78	0.89	0.83	0.61	0.54
	0.3	0.48	0.59	0.82	1.02	1.09	0.98	0.75	0.65
0.2	0.0	0.36	0.41	0.51	0.59	0.65	0.58	0.46	0.41
	0.1	0.41	0.48	0.65	0.79	0.9	0.89	0.61	0.54
	0.2	0.48	0.59	0.83	1.05	1.2	1.22	0.77	0.67
	0.3	0.57	0.73	1.09	1.4	1.58	1.53	0.97	0.85
0.3	0.0	0.41	0.48	0.64	0.77	0.8	0.69	0.56	0.5
	0.1	0.48	0.59	0.83	1.02	1.13	1.19	0.75	0.65
	0.2	0.57	0.73	1.1	1.4	1.62	1.79	0.97	0.85
	0.3	0.72	0.95	1.49	1.93	2.36	1.58	1.25	1.1

Table E.20: Median width of hybrid CI constructed from nominal 95% GMM and nominal $> 95\%$ Bonferroni intervals: $n = 2000$