Leveraging Qualitative Knowledge: A Comment

Bear F. Braumoeller
Yevgeniy Kirpichevsky

Harvard University
Department of Government
Littauer Center, North Yard
Cambridge, MA 02138

Draft, v. 2.2*

Abstract

Gordon and Smith (2004) do a great service by introducing innovative and creative quantitative methods that incorporate information from qualitative sources. It is nevertheless important to examine the conditions under which the proposed estimators will be useful in practice. These conditions prove to be surprisingly restrictive: virtually *all* of the cases of discernible causation must be coded as such, those codings must contain virtually no errors, and the process by which qualitative researchers produce evaluations of discernibility must conform to the authors' model of the qualitative data-generating process (QDGP) if the procedures are to retain any comparative advantage.

*The authors are grateful to Kevin Clarke, Sarah Croco, Brian Gaines, Gary Goertz, Sandy Gordon, Beth Simmons, Alastair Smith, Martin Steinwand, Tze Kwang Teo, and especially Kevin Quinn for comments on earlier drafts.

1 Introduction

In the twenty-five years since Poirier (1980) introduced the concept of partial observability in bivariate probit models, little has been done to address the discrepancy between the technique, which assumes that we know nothing about which causal mechanism is responsible for the realization of the dependent variable, and the reality, in which sometimes we do. Gordon and Smith (2004) do a great service by thinking through innovative methods, throbit and trubit, for incorporating additional information about causation from qualitative sources. This goal is a valuable one because, as the authors correctly point out, partial observability techniques are generally starved for information (2004, 234). For that reason, Gordon and Smith's article should be seen as a thoughtful solution to a long-standing problem of considerable importance.

However, it is crucial to highlight the conditions under which the general approach suggested by Gordon and Smith will prove useful. As we demonstrate below, these turn out to be rather restrictive. In particular, their approach rests on three key assumptions:

1. **No underreporting:** All events with discernible causes are in fact discerned and coded as such by experts.

2. **No coding error:** Expert coding of discernible causes contains no errors.

3. **Qualitative data-generating process correctly modeled:** Experts' judgments about the discernibility of causes are directly related to the values of the latent dependent variables: the conclusion that the cause is discernible is reached if, and only if, the value of only one of those variables exceeds a certain threshold.

We argue below that, in practice, these assumptions are likely to be violated; indeed, violation of the first two is nearly unavoidable, and it may be impossible to know whether the third condition is ever met.

In order to assess the extent to which violations of these assumptions is problematic, we run Monte Carlo simulations. First, the simulations suggest that, *even under ideal conditions,* i.e., when none of the assumptions is violated, the performance of trubit is no better (and is typically slightly worse) than that of the authors' baseline estimator, Boolean probit (Poirier 1980; Braumoeller 2003), which utilizes no qualitative information at all. Throbit, on the other hand, does outperform Boolean probit, but only if all assumptions are met: if any of the three is violated even trivially, substantial parameter inconsistency results and gains in efficiency are forfeited. Trubit is more robust (though not perfectly robust) to violations of the first and third assumptions than is throbit, while both are quite vulnerable to violations of the second assumption (coding error).

2 The Approach

Gordon and Smith's contribution to the literature on partial observability techniques lies in their approach to leveraging "qualitative information for quantitative inference":

> We propose two methods to overcome the problems of identification and labeling inherent in the partial observability context. ... Although for this class of problems information on which causal mechanism is responsible for an event is generally censored, we can occasionally discern that a cause was operative. Even if these instances are few and far between, we can use them as 'anchors' to greatly reduce the identification and labeling problems. (Gordon and Smith 2004, 239)

The information that the authors refer to would come from historians or other experts and would be generated as follows. Gordon and Smith posit that one of two or more mechanisms hypothesized to lead to an event $Y = 1$ is "discernible"—meaning that it will generate an expert claim, based on qualitative data, that the mechanism in question was the one responsible for the observation—if and only if the value of the latent dependent variable associated with this mechanism exceeds a certain threshold and is the only one to do so. If the values of several latent dependent variables associated with different causal mechanisms exceed their thresholds, or if none does, experts either remain silent or disagree on the causes of the event, and no coding is generated.

A simple analogy might be helpful. Imagine that a coroner is asked to assess the cause of death for a man who was known to have been a heavy smoker and drinker. If the coroner finds cirrhosis of the liver, advanced beyond a certain point, and little or no lung cancer, he will blame alcohol for the man's death. Conversely, advanced lung cancer and no cirrhosis would implicate cigarettes. Both lung cancer and cirrhosis, or neither of the two, would leave the coroner unable to adjudicate between the two causes.

The idea behind the procedure is that large-N researchers resemble biostatisticians who relate alcohol and cigarette intake to mortality in general, and case-study researchers (or historians) resemble coroners who evaluate the cause of death in particular cases. The goal is to incorporate the "coroners' " information into the "biostatisticians' " analysis in order to sharpen inferences about the causes of "mortality." That is, Gordon and Smith make use of information generated by qualitative experts to improve on simple partial-observability estimates of the coefficients that relate independent variables of interest to their associated dependent variables. (The model is formally restated in Section 3.)

Gordon and Smith focus exclusively on "substitutable" causation (e.g., X_1 *or* X_2 is responsible for event $Y = 1$, as in the example above), though they note that "the problem is isomorphic to one involving the conjunction of two causes (X_1 *and* X_2)" (Gordon and Smith 2004, 238). Although their claim is mathematically correct, the two problems are substantively quite different, and the applicability of their procedures to the latter case merits brief discussion. In the case of conjunctural causation, in which two causal mechanisms

produce a given event only in combination, it makes no sense to look at studies examining events ($Y = 1$) in order to label one or the other mechanism a "discernible cause"—both causes are present, so both underlying dependent variables exceeded their thresholds.[1] Instead, non-events ($Y = 0$) need to be examined, and expert judgments on which cause *of the non-event* was "discernably" present would need to be collected. The difficulty of getting qualitative data pertaining to non-events is daunting. One need only imagine trying to find a detailed analysis of the reasons why the United States did not request an IMF loan in every year during which it failed to do so to comprehend the magnitude of the problem. While some notable non-events have been studied in depth (e.g. intermissions in great-power conflict in the nineteenth and second half of the twentieth centuries), most simply are not.[2] This article will focus on substitutable causation to ensure full comparability with Gordon and Smith's arguments, but it is worth keeping this caveat in mind.

3 Method

To evaluate the robustness of this approach, we ran a series of Monte Carlo simulations. We chose the following parameters: 1,000 datasets, with 1,000 observations each, were created and analyzed for each simulation. Following the authors' procedures (which they were kind enough to provide in the form of GAUSS and Stata batch files), the exogenous variables (X) and the two uncorrelated error terms (ε) were drawn from a Normal(0,1) distribution; a constant was then subtracted from X_1 and X_2 to provide a more balanced distribution of Y.[3] The model was taken directly from Gordon and Smith (2004, 242):

$$
\begin{aligned}
Y_{1i}^{*} &= \beta_1 + \beta_2 X_{ci} + \beta_3 X_{1i} + \varepsilon_{1i} && (1)\\
Y_{2i}^{*} &= \beta_4 + \beta_5 X_{ci} + \beta_6 X_{2i} + \varepsilon_{2i} && (2)\\
Y_i &= \left\{ \begin{array}{l} 1 \text{ if } \max(Y_{1i}^{*}, Y_{2i}^{*}) > 0 \\ 0 \text{ otherwise} \end{array} \right\} && (3)
\end{aligned}
$$

—where X_c is an independent variable common to both "paths" and X_1 and X_2 are unique to Y_1^* and Y_2^*, respectively. Again following the authors' procedures,

[1] Or, rather, they clearly failed to exceed the thresholds of "absence" that would have produced a non-event.

[2] The growing literature on necessary conditions (Dion 1998; Braumoeller and Goertz 2000; Goertz and Starr 2002) is a notable exception: one of the main predictions to arise from necessary condition theories is that events will fail to occur when the condition in question is absent.

[3] In the simulation that was sent to us, the constant was 3.77; we found that 2.5 worked just as well if not better, providing a nearly 1:1 ratio of $Y_i = 1$ to $Y_i = 0$ cases in most datasets. The goal in seeking a balance between 1s and 0s was to ensure that the results are not polluted by the well-known parameter inconsistencies that plague all probit analyses in rare-events settings.

all parameters (β) were set to 1 and the dependent variable was generated. Finally, two variables representing qualitative evaluations were created:

$$Q_{1i} = \left\{ \begin{array}{l} 1 \text{ if } Y_{1i}^* > \tau_1 \text{ and } Y_{2i}^* < \tau_2 \\ 0 \text{ otherwise} \end{array} \right\} \quad (4)$$

$$Q_{2i} = \left\{ \begin{array}{l} 1 \text{ if } Y_{2i}^* > \tau_2 \text{ and } Y_{1i}^* < \tau_1 \\ 0 \text{ otherwise} \end{array} \right\} \quad (5)$$

It is worth noting, for the later discussion of the QDGP, that this is a rather strict requirement: a mechanism j is deemed discernible at $Y_{ji}^* = \tau_j + 0.01$ and not discernible at $Y_{ji}^* = \tau_j - 0.01$. As in the original simulations, τ_1 and τ_2 were both set to 0.53.[4]

The throbit likelihood function can be found in Gordon and Smith (2004, 240). The intuition behind it is complex but ultimately quite elegant: four different probabilities (the probability of a nonevent, the probability of an event arising discernibly from path 1, the probability of an event arising discernibly from path 2, and the probability of an event occurring but not being discernible) are calculated in such a way that they constitute different regions of a bivariate normal curve. In essence, the τ parameters are used as the intersections of these regions—the points at which they are connected to form a single density. The authors' Figure 2 provides the graphical intuition.

Likelihood functions derived from trubit, a Bayesian technique, were not presented in the original paper. Following the lead of the authors, who tested an ML version in their own Monte Carlos but did not present the results (fn. 20), we constructed them based both on the contents of their article and the code that they were kind enough to provide. Which likelihood to use depends critically on what we believe about the non-discernible cause: the authors distinguish between situations in which "[mechanism] 1 (and explicitly not [mechanism] 2) is responsible for causing success in a particular observation" and those in which "mechanism 1 causes the success but mechanism 2 cannot be ruled out." (247)[5] The two situations imply different truncations in the Bayesian model, which translate into different likelihood functions in the ML version. To be clear, we distinguish between trubit-c (in which we are certain that mechanism 2 did not cause success) and trubit-u (in which we are uncertain).[6]

[4] In terms of the earlier analogy, Y measures mortality, the Xs measure cigarette and alcohol consumption, Y_1^* and Y_2^* represent the extent to which cirrhosis and lung cancer have advanced, and τ represents the point beyond which the disease is considered advanced enough to have been fatal.

[5] The omitted words in the original of the first quote were "cause." As the authors seem to have intended the same thing in both passages, we have modified their wording slightly to avoid misunderstanding.

[6] Reasonable people may differ regarding which of the two variants is most appropriate to the situation described above. The authors' Monte Carlo code indicates that they utilized trubit-u in their simulations. The story behind the data-generating process—that an expert claim will be generated only if the value of one latent variable is high and the value of the other is low—lead us to believe that trubit-c is equally if not more appropriate to the task at hand. We therefore provide results for both variants.

Analogous to the authors' definitions for the throbit likelihood (2004: 240), let $\delta_{ji} = 1$ if mechanism j is a discernable cause, and zero otherwise. Constraining ρ, the correlation between the error terms, to zero, the trubit-u likelihood function is constructed from the underlying probabilities:

$$\begin{aligned} Pr(Y_i = 0) &= (1 - \Phi(X_{1i}\beta_1))(1 - \Phi(X_{2i}\beta_2)) && (6) \\ Pr(Y_i = 1, \delta_{1i} = \delta_{2i} = 0) &= 1 - Pr(Y_i = 0) && (7) \\ Pr(Y_i = 1, \delta_{1i} = 1, \delta_{2i} = 0) &= \Phi(X_{1i}\beta_1) && (8) \\ Pr(Y_i = 1, \delta_{1i} = 0, \delta_{2i} = 1) &= \Phi(X_{2i}\beta_2), && (9) \end{aligned}$$

where $\Phi(\cdot)$ is a cumulative Normal distribution. The trubit-c likelihood function is similar, save that equations (8) and (9) are replaced by

$$\begin{aligned} Pr(Y_i = 1, \delta_{1i} = 1, \delta_{2i} = 0) &= \Phi(X_{1i}\beta_1)(1 - \Phi(X_{2i}\beta_2)) && (10) \\ Pr(Y_i = 1, \delta_{1i} = 0, \delta_{2i} = 1) &= (1 - \Phi(X_{1i}\beta_1))\Phi(X_{2i}\beta_2) && (11) \end{aligned}$$

Once the data had been generated, they were analyzed with Boolean probit, throbit, and both ML trubit estimators. The goal was to evaluate the properties of the latter three estimators,[7] both under ideal conditions and in the case in which one of the three assumptions described below is violated.[8]

4 Assumptions

This section will describe the three conditions that must hold for this to be a viable approach. For each of the three, we explain why the approach requires these conditions to perform well and argue that these conditions are either rarely met in reality (in the case of the first two) or are impossible to verify (in the case of the third).

4.1 The Quantity of Qualitative Data: Underreporting

Qualitative judgments about which cause of an event is discernible come mostly from case studies conducted by experts in a particular academic field or by area specialists. Gordon and Smith note that "[u]nfortunately, such qualitative evidence is rarely available for all cases in a sample" (2004: 239). They proceed to ask: "even if such examples in which causes are revealed are few and far between, can we exploit that information in our large-N analysis?" (2004: 239).

[7] Boolean probit does not utilize the qualitative information that is brought to bear in both throbit and trubit, so it is robust to violations of these three assumptions by definition.

[8] The simulations were carried out in `R`, version 1.9.0. For Boolean probit simulations, the `boolean` package, version 1.04, was utilized. For throbit and trubit simulations, the GAUSS and Stata procedures utilized by the authors were translated into `R` to ensure comparability of results.

Their Monte Carlo studies demonstrate, clearly, that in the case of throbit the answer to this question is "yes."

The Monte Carlo simulations in the original article assume that every instance of a discernible cause will generate a report from a qualitative judge. Problems do not arise when events with discernible causes constitute a small proportion of relevant events. But—and this is a key distinction—they *do* arise when only a subset of the events that have discernible causes are actually coded by the experts. We refer to this phenomenon as "underreporting."

Because throbit's dimensions are essentially ordered probits (with some exceptions, as noted by the authors in their footnote 11) in which failure is coded as a 0, ambiguous success is coded as a 1, and success due to a discernible cause is coded as a 2, underreporting is not innocuous: rather, it is equivalent to coding a 1 rather than a 2 on one of the two dimensions. Accordingly, if all events with discernible causes are not coded as such by qualitative judges, the result is nonrandom error in the dependent variable—nonrandom because of the systematic conversion of 2s into 1s. In short, in this particular application, missing data produces inconsistency in coefficient estimates. Further, we cannot know how substantial this inconsistency is, as we cannot know what percentage of discernible causes have been identified.

As we will show below, if one is to exploit qualitative information effectively using throbit, virtually *all* $Y = 1$ events in question must be examined by experts, and opinions about the presence or absence of a discernible cause must be rendered in each case. Such structured abundance of qualitative evidence is rarely available for large-N analyses. For example, Gordon and Smith write the following about the source of their qualitative data: "for a limited number of cases Vreeland (2003) is able to differentiate between events where $Y_{1i} = 1$ and those where $Y_{2i} = 1$" (where Y_{1i} and Y_{2i} represent competing causal paths). "For the majority of cases this information is censored" (2004, 239). Underreporting is the norm: experts seldom produce, or even aim at producing, qualitative judgments about the whole universe of relevant events. Instead, their choice of case studies is dictated by how representative, important, or interesting they are. A quest for qualitative data on the causes of war, for instance, will produce more than enough case studies of World War I, but hardly any of the 1850s war between Brazil and Argentina—not because no cause of the latter event was discernible, but simply because the former event is more prominent.

4.1.1 Results

To simulate the effects of incomplete qualitative data on discernible causes, we drew a random number from a uniform distribution on the unit interval for each observation and coded $Q_{1i} = 0$ if that number exceeded some fraction γ; we then did the same for Q_{2i}. Varying γ permitted us to control the probability that a "discernible cause" as understood by throbit and trubit would be coded as such by the simulated qualitative judges. Tables 1 and 2 show the results.

As shown in Table 1, under ideal conditions throbit coefficients are on average less biased than Boolean probit coefficients, a result that reconfirms the

authors' original findings. The table also shows, however, that throbit's loss of efficiency relative to Boolean probit is substantial if even a relatively small percentage of discernible-cause cases fail to generate reports. Figure 1 illustrates the density of the parameter estimates when 50% of the discernible causes have generated expert claims that have been coded by the researcher: even under these relatively favorable conditions, the bias of the estimator is evident.

Table 2 and 3, on the other hand, call into question trubit's performance relative to Boolean probit even under ideal conditions: both sets of estimates exhibit some bias, but the upward bias in trubit-u's coefficients is substantially more pronounced than is the bias in Boolean probit's. Worse, the trubit-c coefficients on the "common" variable X_{ci}, $\hat{\beta}_2$ and $\hat{\beta}_5$, exhibit severe attenuation. These results are illustrated in Figure 2. The tables do demonstrate, however, that both variants of trubit perform much better than throbit when qualitative data suffers from underreporting.

The tables also contain information on coverage—the percentage of cases in which the 95% confidence interval covers the true population parameter. This figure is arguably of more interest to practitioners, as it reflects the ability of the test to reject (or fail to reject) a hypothesized parameter value correctly. With all discernible causes coded as such, throbit's coverage is, on average, nearly identical to Boolean probit's—95.6%, a slightly above-average performance for both. With 90% of definitive causes coded as such, average coverage drops to 94.5%. With only 50% of discernible causes coded as such, coverage drops to 80% on average. With all discernible causes coded as such and no coding error—that is, under ideal conditions—, trubit's coverage is, on average, 92.4%.

Though we restricted ourselves to simulations in which $N = 1,000$, it is worth emphasizing that the coverage problem becomes worse as N increases and the distribution of coefficients narrows accordingly. To illustrate this point, we ran a throbit simulation with $\gamma = 0.8$ and $N = 10,000$. While the mean coefficients were similar to those reported in Table 1, throbit's coverage dropped from an average of 93.47% to an average of 76.85%. Unlike disparities in efficiency, therefore, which can be mitigated by larger quantities of data, these disparities in coverage are only exacerbated by it.

The authors note in the abstract that "by anchoring 'discernible' causes for a handful of cases about which we possess qualitative information, we obtain greater efficiency." Clearly, the advantages in both efficiency and coverage depend on anchoring discernible causes for not just a handful but nearly *all* of the cases in question in which causes are discernible, at least when throbit is used. Even more daunting is the fact that one cannot know what percentage of discernible causes have been coded without coding, at a minimum, every case in which $Y_i = 1$.[9]

[9] Gordon and Smith assume that no qualitative judgments are ever rendered for nonevents. One might reasonably question this assumption—to take a straightforward example, "rally 'round the flag" effects could easily produce a crisis that does not escalate to war, and the path to war could be coded even though the war itself did not occur.

4.2 The Quality of Qualitative Data: Coding Error

History teaches us time and time again that what experts once thought to be true later turned out to be false or disputed. As one reviewer put it when discussing John Lewis Gaddis' "We Now Know" (Gaddis 1997), a book updating our understanding of the Cold War based on new archival sources, "Mr. Gaddis might have called his book 'What I Now Think'." (McMillan 1997, 20) If historians continue to change their minds and disagree on the causes of such large and salient events as the end of the Cold War, then, surely, less well-documented and more obscure events could produce some questionable evaluations on the part of single qualitative judges—or even questionable consensus on the part of multiple judges, as the fluctuation from traditional to revisionist to "post-revisionist" to corporatist schools of thought on the sources of American foreign policy demonstrates.[10] It is therefore worth evaluating the robustness of the approach to errors in qualitative evaluations, or more succinctly, "coding error."[11] Perhaps not surprisingly, the results below show that such errors, which by virtue of the logic of throbit and trubit are hardwired into the latent dependent variables, can make them produce coefficients that are severely asymptotically biased.

Unfortunately, qualitative data used in this manner are especially prone to coding error. Gordon and Smith note that disagreement among scholars is one reason to code causation as ambiguous (2004, 240). However, their qualitative data for the IMF example comes from a single source (Vreeland 2003). It is possible that other experts have disputed the relative importance of causes identified by Vreeland in some cases. This would mean that those causes have been wrongly coded as discernible.[12] It is also possible that the cause is *not* ambiguous, and that Vreeland, despite his attention to detail, is simply wrong about which of the two causes was operative. If so, the case is correctly believed to contain a discernible cause, but the wrong cause is implicated. Returning to the ordered-probit analogy as it pertains to throbit, if the nonreporting problem described in Section 4.1 is analogous to coding 1s rather than 2s, the former error corresponds to coding 2s rather than 1s and the latter corresponds to accidentally reversing a 1 and a 2 on different dimensions. If such errors are made in trubit, the wrong truncation is used,[13] and coefficient estimates suffer.

[10] See *inter alia* Williams (1972), Gaddis (1983), and Hogan (1990) for discussions.

[11] One might object that error in the independent variable always produces inconsistency, sometimes dramatically, in probit analyses in general, and, therefore, that we are holding Gordon and Smith to a standard that ordinary probit would fail. Such an objection would be misguided: we are examining robustness in the face of error in expert judgments, which—mathematically speaking—is incorporated as part of the *dependent* variable.

[12] This form of error could also be produced if some other variable, in addition to the one associated with the mechanism labeled "discernible cause," may have exceeded its threshold but gone unnoticed. (See Gordon and Smith's definition of "ambiguous cause" (2004, 240).) This possibility is simply noted here but not examined in the Monte Carlo results, as any number of arbitrarily horrifying scenarios could be concocted to little useful effect. This could happen, for example, when disagreement among scholars exist as to which mechanism caused a particular event, but only one source was consulted by a researcher, and so the ambiguity in causation was not properly noted.

[13] Equivalent, in this setting, to using equation (8) rather than equation (9) or vice-versa for trubit-u, or using equation (10) rather than equation (11) or vice-versa for trubit-c.

In short, even if qualitative data on all cases could be gathered, the data-gathering itself would be a daunting task that would require both extensive research and the ability to adjudicate among competing historians when a discernible cause is in fact present but disputed (or absent but purported).

4.2.1 Complications

To make matters worse, there is a tension among these sources of bias: in many cases eliminating one could produce another. Avoiding underreporting requires that data for every case in which $Y = 1$ be examined, but unless very few of these cases exist, examining all of them thoroughly would be difficult or impossible, and the reliability of the codings would suffer as a result. Ambiguity may be missed by experts in a case that has not undergone a thorough investigation—for example, when more than one mechanism exceeds its threshold, only one might be discerned by the expert(s)—or if only one study has been consulted for coding purposes. Similarly, cases that have received quite a bit of historical scrutiny and debate are, ideally, less likely to fall victim to coding error than those that have received little scrutiny. Unfortunately, the no-underreporting assumption requires that expert judgments be obtained even in cases that have not been very thoroughly scrutinized, so meeting the first assumption might increase the extent to which the second is violated.

By the same token, tradeoffs can exist between different forms of coding error. Imagine, for example, that a phenomenon under investigation is of great scholarly interest and contains relatively few events $Y = 1$ (e.g., the literature on the causes of great-power war). Historians have probably analyzed each event painstakingly and have come up with many contradictory theories. It is very likely that some historians are simply wrong, while others are correct. What is the analyst to do? If we code all causes of events over which there is disagreement as ambiguous, even though some are actually discernible and have been discerned by the "better" historians, then not all discernible cases are coded as such. If we attempt to avoid this outcome by turning some of these "uncoded" cases into coded ones (by, say, adjudicating among historians), then we run the risk that these contested codings will be erroneous, either because the case is ambiguous in reality or because the historians that we chose were mistaken. Either outcome corresponds to a form of coding error, and there is a clear tradeoff between the two. All in all, given the assumptions of the technique, the requirements of coding the qualitative data put the analyst on the uncomfortable horns of an ugly dilemma.

4.2.2 Results

To see how throbit and trubit perform when errors are present in qualitative data, we again drew a random number from a uniform distribution on the unit interval for each observation and reversed the codings of Q_{1i} and Q_{2i} prior to analysis if that number failed to exceed some fraction θ. By altering the value of θ we were able to control the probability that an observation would be miscoded.

Table 1 shows the results for throbit, which are again presented for an illustrative case ($\theta = 0.10$) in Figure 1. The results demonstrate that even a modest amount of error in the qualitative judgments can lead to severe inconsistency in the parameter estimates. Moreover, coverage drops quite rapidly: at $\theta = 0.15$, mean coverage drops below 50%, and at $\theta = 0.25$—not, arguably, an unreasonable error rate to expect in the social sciences—it dips into the single digits for four of the six parameters. Tables 2 and 3 show the results for both variants of trubit. Again, trubit-u performs better than throbit when errors are present, but the coverage does drop as the percentage of miscoded cases increases. At $\theta = 0.15$, trubit's mean coverage drops below 80%. The performance of trubit-c, on the other hand, is actually marginally *worse* than that of throbit, both in terms of both bias and coverage.

One exception to this generalization can clearly be seen, however: coefficient estimates $\hat{\beta}_2$ and $\hat{\beta}_5$ in Table 2 seem remarkably robust to violations of this assumption. Unfortunately, this fact is nothing more than an artifact of the parameters of the simulation: the population coefficients are all set to 1, the independent variables are uncorrelated, and β_2 and β_5 are both coefficients on the same variable, X_{ci}, which appears in both of the model's "causal paths." Therefore, when trubit-u mistakenly estimates $\Phi(\beta_1 + \beta_2 X_{ci} + \beta_3 X_{1i})$ rather than $\Phi(\beta_4 + \beta_5 X_{ci} + \beta_6 X_{2i})$, or vice-versa, it happens to produce the same coefficient.

4.3 The Qualitative Data-Generating Process

The Gordon-Smith approaches are based on a model of the process by which an expert makes a judgment about whether or not a given mechanism or path is discernible. It seems likely that the precise way in which data translates into a qualitative judgment differs from expert to expert and from case to case; in any event, it is most likely unknowable. This is not a criticism of the Gordon-Smith model of the process *per se*—but since the process is hidden inside a black box, we do need to ensure that their approach would not balk at qualitative data generated by a process no less reasonable than the one that they describe. We show that trubit is considerably more robust than throbit under these conditions, though neither is as robust as one would like.

Briefly, Gordon and Smith posit that for each causal mechanism j there exists a certain threshold τ_j, and that the experts implicitly use these thresholds when they make qualitative statements. If one, and only one, latent dependent variable exceeds its threshold, it is deemed to be the discernible cause of the event in question. If more than one, or none, does so, the cause of the event is deemed ambiguous. The experts need not examine the quantitative data, of course: the technique assumes only that they will code the cases as if they had done so. We propose two alternative qualitative data-generating processes (QDGPs) which we think more realistically represent the process by which qualitative judgments are rendered. We then analyze data created by these QDGPs and demonstrate that, unfortunately, throbit in particular is not robust to the inclusion of qualitative data generated by either one.

4.3.1 Results

To determine how throbit and trubit would perform if the process by which experts render qualitative judgments is in fact different from the one postulated by Gordon and Smith, we set up two alternative QDGPs. In the first, rather than making their judgments based on whether one latent variable or another has exceeded a certain threshold, qualitative experts make their judgments based on the difference between the values of the two latent dependent variables. If the value of Y_{1i}^* is very high and the value of Y_{2i}^* is very low, the judges are most likely to encounter qualitative evidence that backs up the conclusion that Y_{1i}^* is the discernible cause, regardless of whether either has crossed a certain threshold.

To capture this notion mathematically, we assumed that if $Y_i = 1$,[14] the probability that a mechanism will be coded as a "discernible cause" increases in direct proportion to the difference between predicted probabilities for the two unobservable latent variables. Simply put, to take the example from the original article, qualitative judges are most likely to conclude that IMF agreements are reached on the basis of economic need when the factors that predict need-based loans are present and those that predict domestic political loans are absent. Mathematically,

$$Q_{1i} = \left\{ \begin{array}{l} 1 \text{ if } \Phi(\beta_1 + \beta_2 X_{ci} + \beta_3 X_{1i}) - \\ \qquad \Phi(\beta_4 + \beta_5 X_{ci} + \beta_6 X_{2i}) > \kappa_1 \text{ and } Y_i = 1 \\ 0 \text{ otherwise} \end{array} \right\} \quad (12)$$

$$Q_{2i} = \left\{ \begin{array}{l} 1 \text{ if } \Phi(\beta_4 + \beta_5 X_{ci} + \beta_6 X_{2i}) - \\ \qquad \Phi(\beta_1 + \beta_2 X_{ci} + \beta_3 X_{1i}) > \kappa_2 \text{ and } Y_i = 1 \\ 0 \text{ otherwise} \end{array} \right\} \quad (13)$$

, where κ represents a random number drawn from a uniform distribution on the unit interval.

The results produced when trubit-u and trubit-c are applied to data created via this alternative QDGP (labeled "Alt. QDGP I" in Table 2 and 3) are encouraging: for the most part, the technique seems robust in the face of this alternative assumption. The results for throbit, on the other hand, are disheartening. The magnitude of the bias is more or less equivalent to an error rate of 10% or a qualitative-data sampling rate of 50%, but the bias is in the opposite direction; mean coverage is similar as well (80%).

One might reasonably object that this process still posits insufficient autonomy between qualitative and quantitative data—that archival research might produce strong evidence that the second mechanism is at work even though Y_{1i}^* is high and Y_{2i}^* is low. To take this objection into account, we created a second QDGP, one in which the probability that a mechanism will be flagged

[14] Though we question the logic of not coding nonevents (footnote 9), in the interest of exploring alternative assumptions within the context established by the original work we have assumed that only events generate qualitative judgments. Abandoning this assumption exacerbates the problems described here.

as the operative cause increases with the value of the latent dependent variable associated with it. Here,

$$Q_{1i} = \left\{ \begin{array}{l} 1 \text{ if } 2\kappa < \Phi(\beta_1 + \beta_2 X_{ci} + \beta_3 X_{1i}) \text{ and } Y_i = 1 \\ 0 \text{ otherwise} \end{array} \right\} \quad (14)$$

$$Q_{2i} = \left\{ \begin{array}{l} 1 \text{ if } 1 < 2\kappa < 1 + \Phi(\beta_4 + \beta_5 X_{ci} + \beta_6 X_{2i}) \text{ and } Y_i = 1 \\ 0 \text{ otherwise} \end{array} \right\} \quad (15)$$

, where κ is a single draw from a uniform distribution on the unit interval. To illustrate, if $Y^*_{1i} = 0.8$ and $Y^*_{2i} = 0.2$, the probability that mechanism 1 will be flagged is 0.4, the probability that mechanism 2 will be flagged is 0.1, and the probability that the cause will be deemed ambiguous is 0.5.

The results produced by this alternative QDGP ("Alt. QDGP II" in Tables 1-3) are more distressing. While trubit-u parameter estimates remain largely unchanged, throbit parameter estimates range from 0.65 to 0.92, and coverage ranges from 84% all the way down to 35.8%. Worse, trubit-c parameters range from 0.51 to 0.80, and coverage ranges from 43.6% to 3.7%, as illustrated in Figure 2. Clearly, throbit and trubit-c are not particularly robust to this alternative assumption about the QDGP either.

In short, while the results reconfirm the advantages of throbit over Boolean probit under ideal circumstances, they also demonstrate that those advantages hinge critically on a set of conditions—all, or virtually all, discernible causes coded as such; little to no coding error; and correct specification of the QDGP—which in practice may be difficult to achieve and even more difficult to verify.

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

[Figure 1 about here.]

[Figure 2 about here.]

Conclusion

Gordon and Smith offer a clever and badly needed solution to a problem with partial observability techniques. Incorporating qualitative information into quantitative data is a new, promising avenue of research, and the authors should be applauded for pursuing it. At the same time, it remains a first step. Future research should be oriented toward constructing an estimator that is more robust to heterogeneity in the QDGP and the amount and quality of qualitative data available.

No estimator, of course, is perfect. As Gordon and Smith pointed out in their original article, incorporating additional information via their estimators

sometimes permits those estimators to converge when the original data do not permit Boolean probit to do so—potentially an important advantage that could outweigh issues of relative efficiency. All probit-based techniques exhibit parameter inconsistency as the proportion of 1s or 0s becomes very small. Nonlinear models such as these can typically only promise asymptotic unbiasedness of coefficients, and researchers sometimes fail to realize just how much bias might exist even with a relatively large number of cases. Relevant omitted variables in probit models, unlike those in regression models, need not be correlated with included variables to induce asymptotic bias. And so forth. As a result, inferences, even in the ordinary probit model upon which all of these techniques are based, must be made with an ample admixture of humility and hope. Still, progress requires a thorough understanding of the characteristics of new techniques.

We have shown that Gordon and Smith's throbit estimator is sensitive to even minor violations of several key assumptions. Specifically, although throbit does convey advantages over Boolean probit under ideal conditions, the latter outperforms the former when even a fraction of the discernible causes are not identified by the experts, when even a very small fraction of the discernible causes are misidentified by the experts, or when a process by which experts identify a cause of an event is different from the one posited by the authors. Their trubit estimator generally performs better than throbit under less than ideal conditions, and the trubit-u variant generally outperforms the trubit-c variant, but both variants underperform relative to both throbit and Boolean probit under ideal conditions and have as much trouble as throbit when it comes to coding error. Currently, unless a researcher is quite certain that his or her qualitative data pass these three tests, he or she might be better advised to disregard the additional qualitative information rather than to incorporate it in this fashion.

References

Braumoeller, Bear F. 2003. "Causal Complexity and the Study of Politics." *Political Analysis* 11: 209-233.

Braumoeller, Bear F., and Gary Goertz. 2000. "The Methodology of Necessary Conditions." *American Journal of Political Science* 44: 844-858.

Dion, Douglas. 1998. "Evidence and Inference in the Comparative Case Study." *Comparative Politics* 30: 127-145.

Gaddis, John L. 1997. *We Now Know: Rethinking Cold War History.* New York: Oxford University Press.

Gaddis, John L. 1983. "The Emerging Post-Revisionist Synthesis on the Origins of the Cold War. *Diplomatic History* 7: 171-190.

Goertz, Gary, and Harvey Starr (eds.) 2002. *Necessary Conditions: Theory, Methodology, and Applications.* New York: Rowman and Littlefield.

Gordon, Sanford C., and Alastair Smith. 2004. "Quantitative Leverage through Qualitative Knowledge: Augmenting the Statistical Analysis of Complex Causes." *Political Analysis* 12: 233-255.

Hogan, Michael J. 1990. "Corporatism." *Journal of American History* 77: 153-160.

McMillan, Priscilla Johnson. 1997. "Review: We Now Know: Rethinking Cold War History." *The New York Times,* May 25, 1997, p. 20, col. 1.

Poirier, Dale J. 1980. "Partial Observability in Bivariate Probit Models." *Journal of Econometrics* 12: 209-217.

Vreeland, James. 2003. *The IMF and Economic Development.* New York: Cambridge University Press.

Williams, William Appleman. 1972. *The Tragedy of American Diplomacy.* New York: W.W. Norton and Co.

List of Figures

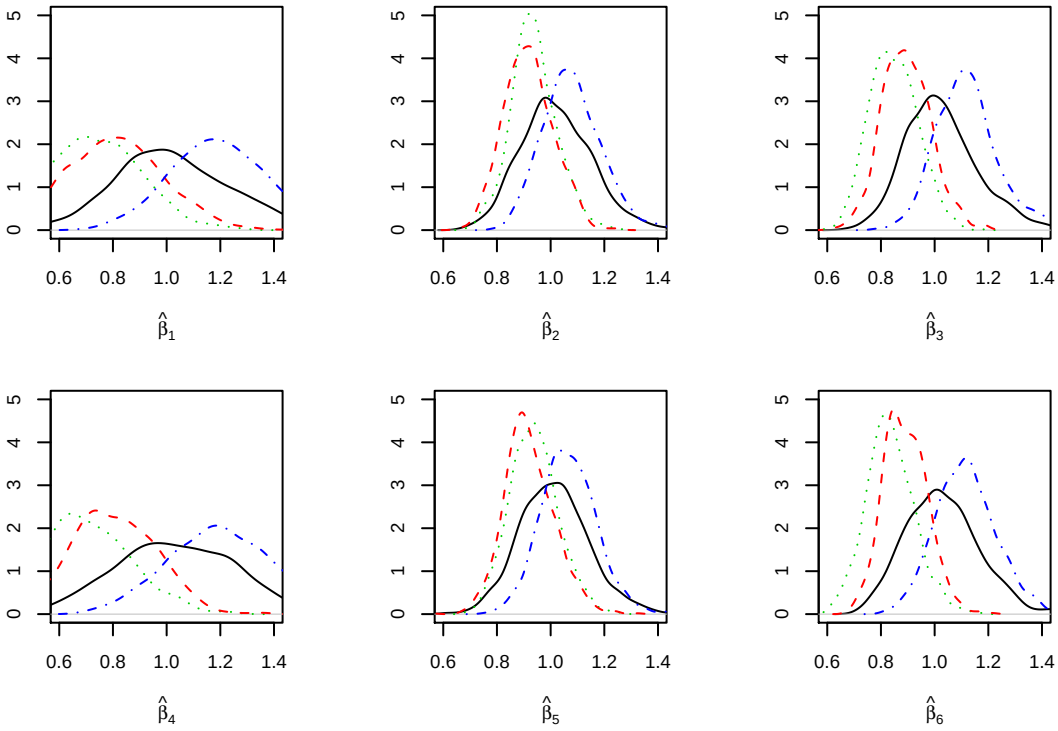

Figure 1: Distribution of Boolean probit coefficients (solid line), throbit coefficients with no error but data on only 50% of discernible causes (dashed), throbit coefficients with data on 100% of discernible causes but 10% error by coders of qualitative data (dotted), and throbit coefficients with error-free data on 100% of discernible causes but alternative QDGP I (dotted-dashed).

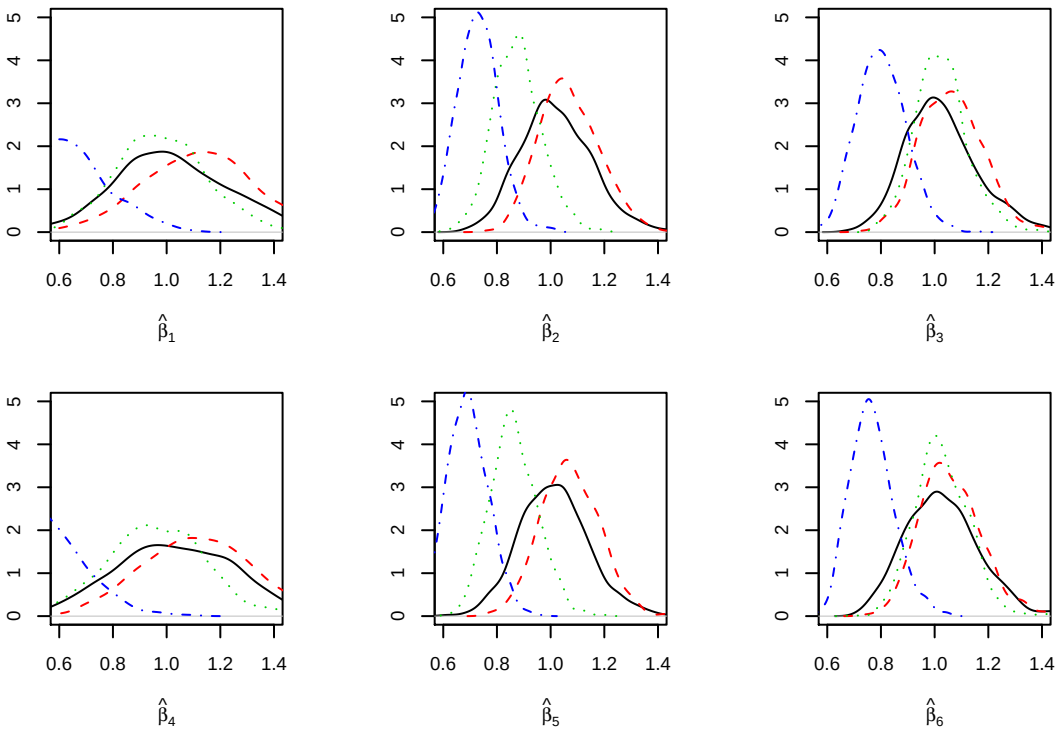

Figure 2: Distribution of Boolean probit coefficients (solid line), trubit-u coefficients with no error and data on 100% of discernible causes (dashed), trubit-c coefficients with no error and data on 100% of discernible causes (dotted), and trubit-c coefficients with alternative data-generating process QDGP II (dotted-dashed).

List of Tables

Estimator, Condition	Parameters	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
Boolean probit	N/A	1.027 (95.4)	1.020 (96.6)	1.025 (94.3)	1.037 (96.2)	1.013 (95.2)	1.030 (95.7)
Throbit 100% Data, No Errors	$\theta = 0$, $\gamma = 1$	1.013 (95.2)	1.010 (96.1)	1.010 (93.9)	1.005 (96.1)	1.008 (96.2)	1.009 (95.9)
Throbit with Underreporting	$\theta = 0$, $\gamma = 0.9$	0.967 (94.8)	1.011 (94.7)	0.990 (94.7)	0.957 (94.3)	1.002 (95.1)	0.978 (93.2)
	$\theta = 0$, $\gamma = 0.8$	0.920 (92.4)	0.996 (95.8)	0.960 (91.0)	0.929 (92.8)	1.002 (96.0)	0.967 (92.8)
	$\theta = 0$, $\gamma = 0.7$	0.873 (88.1)	0.980 (94.3)	0.933 (86.2)	0.890 (91.0)	0.977 (93.8)	0.941 (88.9)
	$\theta = 0$, $\gamma = 0.6$	0.835 (84.1)	0.948 (88.5)	0.912 (81.3)	0.821 (83.2)	0.963 (91.4)	0.906 (80.0)
	$\theta = 0$, $\gamma = 0.5$	0.806 (80.2)	0.917 (83.8)	0.895 (77.7)	0.802 (78.2)	0.921 (85.9)	0.891 (74.2)
Throbit with Coding Error	$\theta = 0.05$, $\gamma = 1$	0.833 (80.6)	0.966 (91.3)	0.909 (75.7)	0.864 (85.3)	0.973 (91.8)	0.926 (82.0)
	$\theta = 0.10$, $\gamma = 1$	0.728 (61.7)	0.937 (88.1)	0.848 (52.0)	0.690 (54.4)	0.938 (86.5)	0.835 (45.4)
	$\theta = 0.15$, $\gamma = 1$	0.609 (34.2)	0.898 (73.9)	0.785 (21.7)	0.631 (40.2)	0.915 (77.5)	0.799 (27.9)
	$\theta = 0.20$, $\gamma = 1$	0.454 (11.1)	0.890 (67.4)	0.704 (7.0)	0.462 (11.3)	0.882 (63.5)	0.705 (6.1)
	$\theta = 0.25$, $\gamma = 1$	0.356 (2.9)	0.876 (65.7)	0.648 (1.2)	0.341 (3.7)	0.866 (59.2)	0.647 (2.1)
Throbit with Alternative Data- Generating Processes	QDGP I	1.204 (83.3)	1.079 (91.3)	1.120 (84.8)	1.198 (83.8)	1.067 (92.2)	1.118 (85.1)
	QDGP II	0.686 (53.0)	0.903 (77.1)	0.837 (49.2)	0.653 (45.3)	0.923 (84.0)	0.807 (35.8)

Table 1: Sensitivity analysis of throbit. Numbers are mean coefficients; numbers in parentheses represent percentage of cases in which the 95% confidence intervals cover the population parameter.

Estimator, Condition	Parameters	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
Boolean probit	N/A	1.027 (95.4)	1.020 (96.6)	1.025 (94.3)	1.037 (96.2)	1.013 (95.2)	1.030 (95.7)
Trubit-u 100% Data, No Errors	$\theta = 0$, $\gamma = 1$	1.129 (91.7)	1.072 (91.8)	1.064 (93.2)	1.123 (91.1)	1.072 (93.0)	1.059 (93.6)
Trubit-u with Underreporting	$\theta = 0$, $\gamma = 0.9$	1.111 (94.1)	1.059 (94.4)	1.055 (95.0)	1.100 (93.4)	1.059 (93.4)	1.049 (95.0)
	$\theta = 0$, $\gamma = 0.8$	1.103 (94.6)	1.065 (94.2)	1.053 (94.3)	1.100 (93.5)	1.062 (94.4)	1.049 (95.2)
	$\theta = 0$, $\gamma = 0.7$	1.078 (94.5)	1.053 (95.2)	1.04 (95.8)	1.110 (94.8)	1.066 (93.0)	1.062 (94.3)
	$\theta = 0$, $\gamma = 0.6$	1.083 (94.4)	1.054 (93.7)	1.044 (94.2)	1.095 (93.6)	1.054 (95.3)	1.052 (94.5)
	$\theta = 0$, $\gamma = 0.5$	1.086 (94.3)	1.053 (94.8)	1.045 (95.9)	1.073 (95.7)	1.042 (95.8)	1.041 (96.9)
Trubit-u with Coding Error	$\theta = 0.05$, $\gamma = 1$	1.009 (95.1)	1.038 (94.7)	0.9874 (93.8)	1.026 (94.7)	1.051 (94.0)	1.001 (94.0)
	$\theta = 0.10$, $\gamma = 1$	0.909 (90.4)	1.021 (94.7)	0.928 (81.7)	0.902 (87.8)	1.035 (95.8)	0.921 (78.8)
	$\theta = 0.15$, $\gamma = 1$	0.827 (81.6)	1.004 (94.3)	0.874 (65.5)	0.811 (79.6)	1.007 (94.6)	0.871 (63.2)
	$\theta = 0.20$, $\gamma = 1$	0.729 (66.0)	0.985 (95.1)	0.820 (45.0)	0.748 (69.3)	1.000 (94.6)	0.831 (50.6)
	$\theta = 0.25$, $\gamma = 1$	0.658 (51.4)	1.000 (95.9)	0.788 (34.3)	0.619 (46.2)	0.977 (95.6)	0.756 (24.4)
Trubit-u with Alternative Data- Generating Process	QDGP I	1.120 (95.2)	1.047 (95.9)	1.071 (94.4)	1.133 (92.6)	1.050 (95.4)	1.079 (93.2)
	QDGP II	1.052 (94.0)	1.087 (92.3)	1.022 (94.1)	1.046 (96.0)	1.087 (92.2)	1.018 (95.1)

Table 2: Sensitivity analysis of ML version of trubit-u. Numbers are mean coefficients; numbers in parentheses represent percentage of cases in which the 95% confidence intervals cover the population parameter.

Estimator, Condition	Parameters	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$
Boolean probit	N/A	1.0274 (95.4)	1.020 (96.6)	1.025 (94.3)	1.037 (96.2)	1.013 (95.2)	1.030 (95.7)
Trubit-c 100% Data, No Errors	$\theta = 0$, $\gamma = 1$	0.984 (95)	0.876 (69.6)	1.024 (94.5)	0.966 (93.8)	0.864 (66.3)	1.026 (95.6)
Trubit-c with Underreporting	$\theta = 0$, $\gamma = 0.9$	0.973 (95.7)	0.911 (81)	1.022 (96.6)	0.960 (95.8)	0.861 (64.7)	1.024 (96.8)
	$\theta = 0$, $\gamma = 0.8$	0.928 (92.5)	0.912 (81.1)	0.997 (94.8)	0.947 (94.8)	0.866 (69.1)	1.005 (97.1)
	$\theta = 0$, $\gamma = 0.7$	0.982 (94)	0.908 (81.6)	1.022 (95.2)	0.965 (95.5)	0.909 (82.4)	1.008 (95.6)
	$\theta = 0$, $\gamma = 0.6$	0.995 (95.2)	0.923 (85.6)	1.025 (96.1)	1.021 (95.8)	0.912 (82)	1.032 (95.5)
	$\theta = 0$, $\gamma = 0.5$	0.993 (95.4)	0.922 (85.6)	1.020 (95.1)	0.983 (94.8)	0.937 (87.1)	1.015 (94.7)
Trubit-c with Coding Error	$\theta = 0.05$, $\gamma = 1$	0.786 (75.7)	0.812 (41.9)	0.919 (80.5)	0.814 (78)	0.832 (48)	0.926 (82.2)
	$\theta = 0.10$, $\gamma = 1$	0.650 (44.7)	0.800 (37.3)	0.836 (49.4)	0.652 (44.3)	0.805 (39.5)	0.838 (49.9)
	$\theta = 0.15$, $\gamma = 1$	0.504 (19.8)	0.781 (28.1)	0.759 (21.3)	0.524 (19.6)	0.764 (22.7)	0.769 (23)
	$\theta = 0.20$, $\gamma = 1$	0.428 (5.9)	0.747 (15)	0.721 (9)	0.349 (3.1)	0.727 (10.8)	0.668 (3.5)
	$\theta = 0.25$, $\gamma = 1$	0.279 (2.7)	0.733 (13.3)	0.646 (3)	0.272 (1.4)	0.729 (10.4)	0.643 (1.8)
Trubit-c with Alternative Data- Generating Process	QDGP I	1.159 (91.8)	0.972 (93.6)	1.111 (90.7)	1.126 (92.7)	0.968 (92.1)	1.098 (91.7)
	QDGP II	0.580 (33.5)	0.730 (7.9)	0.802 (43.6)	0.514 (16.7)	0.691 (3.7)	0.767 (24.8)

Table 3: Sensitivity analysis of ML version of trubit-c. Numbers are mean coefficients; numbers in parentheses represent percentage of cases in which the 95% confidence intervals cover the population parameter.