Sande, J. B., & Ghosh, M. (2018). Endogeneity in survey research. International *Journal of Research in Marketing, 35*(2), 185-204

Doi: https://doi.org/10.1016/j.ijresmar.2018.01.005

ENDOGENEITY IN SURVEY RESEARCH

Published in IJRM – International Journal of Research in Marketing
Volume 35, Issue 2, June 2018, Pages 185-204

Article website: https://doi.org/10.1016/j.ijresmar.2018.01.005

Companion website with data, Stata code, and Stata output:
http://www.runmycode.org/companion/view/2958

Jon Bingen Sande
Associate Professor of Marketing
BI Norwegian Business School
NO-0442 Oslo, Norway
Tel: +47 46 41 06 48
E-mail: jon.b.sande@bi.no

and

Mrinal Ghosh
Eller Professor of Marketing
Eller College of Management
University of Arizona, United States
1130 E Helen Street,
320P McClelland Hall
Tucson, AZ 85721
Ph: (520) 626 7353
Email: mghosh@eller.arizona.edu

The authors thank Matilda Dorotic, Jan Heide, Walter Herzog, George John, Silja Korhonen-Sande, Ragnhild Silkoset, Kenneth Wathne, as well as participants in a seminar at Department of Strategic Management and Globalization at Copenhagen Business School in March 2015, for their valuable comments on earlier versions of this paper. This paper has in part been funded by C3 Centre for Connected Care, Ullevaal Hospital, Kirkeveien 166, building 2 H in Oslo, Norway.

Highlights

- We provide advice for handling six “painful” decisions when facing endogeneity:
- Do you have an endogeneity problem? What technique/estimator is appropriate?
- What instrumental variables (IVs) should be chosen?
- How should IVs be evaluated empirically?
- How should you interpret and evaluate the results? What results should you report?

ENDOGENEITY IN SURVEY RESEARCH

Abstract

Endogeneity is a crucial problem in survey-based empirical research on marketing strategy (MS) and inter-organizational relationships (IORs); if not addressed, it can cause researchers to arrive at flawed conclusions and to offer poor advice to practitioners. Although the field is increasingly cognizant of endogeneity-related issues, many authors fail to properly address it, particularly in survey-based research. Emphasizing the role of *essential heterogeneity*, this article develops an *overarching framework* to help improve the understanding of endogeneity problems and how to tackle them when researchers use cross-sectional survey-based data. The authors provide explanations of and advice for how MS and IOR researchers can address six "painful" and sometimes hidden decisions: 1) Do you have an endogeneity problem? 2) What technique/estimator is appropriate? 3) What instrumental variables (IVs) should be chosen? 4) How should IVs be evaluated empirically? 5) How should the results be interpreted and evaluated? and 6) What results should you report? The authors provide a practical flowchart to guide researchers in their efforts to address endogeneity-related concerns.

Keywords: *Endogeneity; Essential Heterogeneity; Marketing Strategy; Inter-Organizational Relationships; Surveys; Cross-sectional research*

1 INTRODUCTION

Empirical research in marketing strategy (MS) and inter-organizational relationships (IORs) is primarily interested in the decisions and behaviors of firms, their employees, their interactions with customers, and the impact of such decisions on various outcome variables. Consider, for instance, a sampling of some of these questions: How do firms design sales force incentive plans that facilitate the selection and retention of salespeople with characteristics desirable to the firm? When does bundling products with associated services benefit both the firm and its customers? When should a firm use detailed formal contracts to govern its customer relationships and what is its impact on relationship performance? To analyze these type of questions, researchers should ideally seek a research design that randomly assigns firms to the relevant strategy alternatives and then test for differences. In reality, few firms willingly permit such a random assignment of key strategic decisions. Consequently, researchers have had to rely on observational data obtained from secondary sources or surveys. Indeed, as Rindfleisch et al. (2008) note, survey instruments are often necessary because key variables with appropriate nuances are not available elsewhere.

The fundamental problem with observational data is its vulnerability to a broad class of problems that econometricians call *endogeneity*. Endogeneity means that an explanatory variable correlates with the disturbance term of the regression equation and not accounting for it will likely result in biased parameter estimates that undermine the validity of the findings obtained from regression-type analyses of observational data. Furthermore, the variables of interest, as illustrated in the questions above, are often purposeful decisions of strategic interest and when such decisions and their effects are heterogeneous and vary between firms or individuals, such self-selection can lead to a special type of endogeneity that Heckman, Urzua, and Vytlacil (2006, p. 389) call "*essential heterogeneity*": a correlation between an endogenous variable and its own

effect. Essential heterogeneity leads to difficulties in evaluating empirical models and questions about how an effect is distributed in the population.

Given the general nature of the survey-based and cross-sectional data available in MS and IOR research, it could be argued that we face what Rossi (2014, p. 655) calls "*a first order endogeneity problem.*" Said otherwise, the endogeneity issues are of such concern that they overshadow other issues like functional form or distributional assumptions because we cannot utilize, say, fixed-effects estimators to reduce the problem. This is problematic because primary survey-based data remains critical for MS and IOR researchers, as it enables them to propose and test nuanced theories and constructs at an appropriate unit of analysis. To make their theories and findings more valid and relevant, these researchers must use techniques and approaches to address this vital problem. While the awareness of these problems among MS and IOR researchers has substantially increased (see Web Appendix A for a census of MS and IOR research published in major marketing journals), the application of endogeneity-correcting techniques within MS and IOR research remains infrequent compared to other types of articles and there appears confusion about how to address endogeneity-related problems. A possible consequence is that articles in these domains get rejected in our premier journals because the researchers have either not addressed, or inadequately addressed endogeneity concerns.

Fortunately, researchers have many estimators and approaches at their disposal, but they need specific guidance regarding when and how to use them. A recent article by Germann, Ebbes and Grewal (2015) develops a useful framework to help researchers compare and choose between 1) rich data models (e.g., ordinary least squares (OLS) with control variables), 2) panel data models, 3) instrumental variable (IV) estimators, and 4) panel instrument models. However, MS and IOR researchers who use cross-sectional data are often restricted to use some kind of IV

estimator. These researchers need a framework that complements the one presented by Germann et al. (2015) and that provides guidance on implementing IV-based estimation techniques.

Our article responds to this need. First, we generate a practical *overarching framework* (see Figure 1) to help improve the understanding of endogeneity problems and how to tackle them in different situations using estimators based on IVs. Designed as a flowchart, the framework emphasizes the importance of considering potential endogeneity problems upfront at the data collection stage and developing research plans that allow for explicit consideration of endogeneity. Second, the framework enables us to provide explanations of and advice for how researchers can address six "painful" and sometimes hidden decisions related to the following key issues (highlighted in gray boxes in Figure 1): 1) Do you have an endogeneity problem? 2) What estimation technique or estimator is appropriate? 3) Given that an IV-based technique is chosen, what IVs should be chosen, and what theoretical arguments should be used to justify the IVs? 4) How should IVs be assessed empirically? 5) How should you interpret and evaluate the results? and 6) What results should you report? In our reading, most published papers spend minimal effort, if any, to explain how they make these choices.

To illustrate the use of the framework and to enable readers to follow how to address endogeneity step-by-step, in Web Appendix B we use a variable of interest to IOR researchers – formal contracting – as a *running example* and utilize publicly available cross-sectional data from a recently published IJRM article by Sande & Haugland (2015)[i] to show how we can estimate the effect of formal contracting on cost reductions and end-product enhancements in buyer-supplier relationships. In addition, Web Appendix C provides Stata source code for all the estimators described in this article.

We hope our work will find residence with multiple audiences. For readers who are relatively unfamiliar with endogeneity, we introduce them to the topic based on both equations

and graphs, give an overview of the different decisions and dilemmas involved, explain both basic and more advanced methods, and point them toward further reading. For more advanced readers, we provide an overview of the broad literature, put the various estimators in context with each other, and highlight techniques and estimators that thus far have been rarely used in marketing. For reviewers and editors, the proposed framework should function as a useful checklist for how authors should address endogeneity. The intended goal, of course remains to help generate research that is more valid and credible.

We begin in section 2 by explaining the endogeneity problem and how researchers could theoretically evaluate the threat posed by endogeneity. In section 3, we describe how to decide on the appropriate technique for addressing endogeneity by considering the nature of the endogenous variable and whether essential heterogeneity is a concern. In sections 4 through 7, we explain the remainder of the steps: choosing and justifying the IVs, empirically assessing IVs, evaluating and interpreting results, and deciding what results should be reported. Finally, in section 8, we summarize and draw conclusions.

2 DO YOU HAVE AN ENDOGENEITY PROBLEM?

Evaluating whether you have an endogeneity problem has two parts, a theoretical part and an empirical counterpart. In this section, we describe the theoretical part. Researchers also need to empirically assess whether they have an endogeneity problem. However, doing so requires that there exists an unbiased estimator, and we must evaluate the IVs before we can conclude that the IV estimator is unbiased. We will turn to the empirical assessments of endogeneity in section 6.

2.1 What is endogeneity, and why does it arise?

To theoretically evaluate whether endogeneity might be an issue in a given study, researchers must first understand what endogeneity is and how it arises. In short, endogeneity refers to situations in which an explanatory variable in a multiple regression-type setup correlates with the

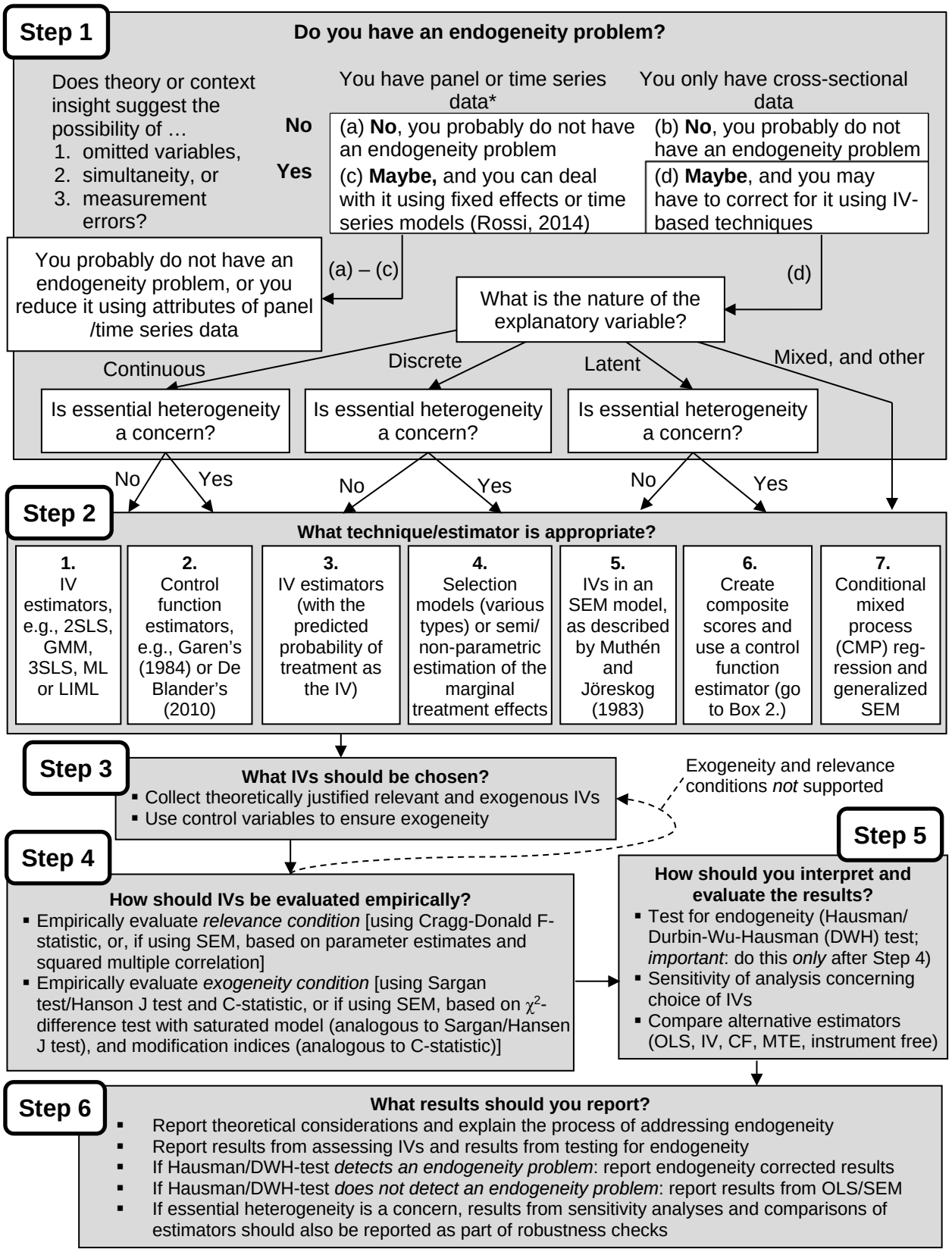


Figure 1: A framework for addressing endogeneity (*Note: panel/time series data may suffer from autocorrelation and autoregression, which are sometimes viewed as another type of endogeneity problem.)

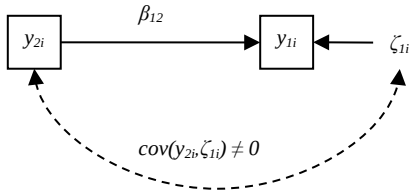
disturbance term (Wooldridge, 2010). Figure 2a) illustrates this definition using a path diagram. The squared boxes represent manifest variables, and the straight directional arrow indicates that the explanatory variable y_{2i} affects the dependent variable y_{1i} [ii]. The dotted bidirectional arrow indicates that y_{2i} correlates with the disturbance term ζ_{1i} and that the researcher has not accounted for this correlation when constructing the model. Crucially, this correlation is not a result of the effect of y_{2i} on y_{1i}. Under such conditions, the explanatory variable y_{2i} is said to be *endogenous* (Wooldridge, 2010)[iii]. Figure 2b) shows that endogeneity can also be a problem in structural equation models (SEMs) with latent variables measured through multiple indicators.

Endogeneity can arise for three reasons: omitted variables, simultaneity, and measurement errors. Figure 2c) illustrates how an *omitted variable*, u_i, can drive both the explanatory y_{2i} and the dependent variable y_{1i} and thereby lead to a correlation between them. To illustrate this, consider the case of the explanatory variable in Sande and Haugland's (2015) data – formal contracting. Despite including several control variables in their analysis, the independent variables explain only 43% of the variance in formal contracting. This means that the level of formal contracting was also chosen based on constructs not measured by the researchers which could include parties' reputation, contracting experience, technological uncertainty, agency problems within a decision-makers' organization, etc. All these variables are omitted variables. The crucial concern is that these omitted variables may also affect the performance variables, cost reductions and end-product enhancements, and hence, if they are not accounted for, we are likely to have an endogeneity problem. To see why this can occur, assume that our dependent variable y_{1i} is a function of y_{2i} and u_i in the real world, as follows:

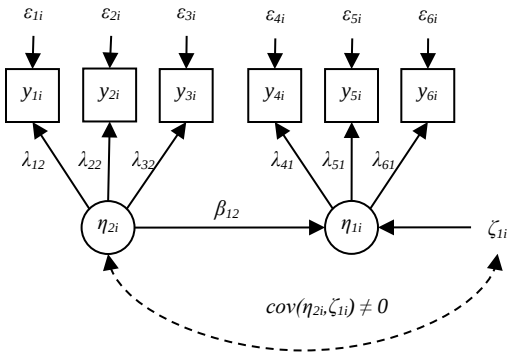
(1) $y_{1i} = \gamma_{10} + \beta_{12}\, y_{2i} + \pi_1 u_i + e_{1i}$,

where e_{1i} is a disturbance term. Assume further that y_{2i} is itself a function of u_i, as follows:

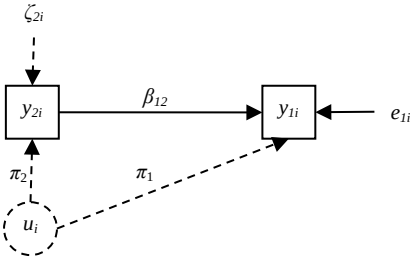
(2) $y_{2i} = \pi_0 + \pi_2\, u_i + \zeta_{2i}$,



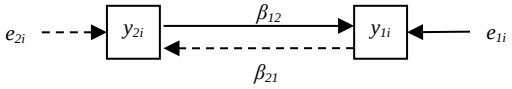
a) General definition of endogeneity, observed variables



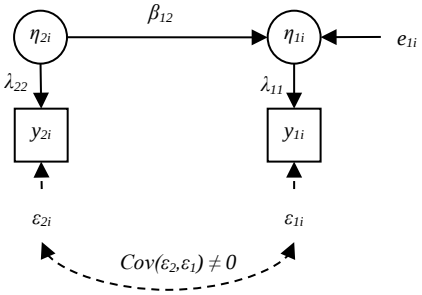
b) General definition of endogeneity, latent variables



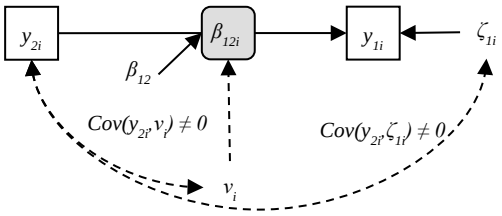
c) Omitted variables



d) Simultaneity



e) Correlated measurement errors



f) Essential heterogeneity (note that the correlation between v_i and y_{2i} reflects the effect of omitted variables u_i)

Legend:

y_i , η_i	endogenous variables
x_i, ξ_i, u_i	exogenous variables
$\mathbf{x_i}$, $\boldsymbol{\xi_i}$	vectors of several (k) exogenous variables, for example x_{11i}, x_{12i},… x_{1ki} and ξ_{11i}, ξ_{12i},… ξ_{1ki}.
ζ_i, v_i, e_i	error terms
β, δ, γ, λ	parameters
x,y	observed and modeled variables

u_i	unobserved and unmodeled variables
η_{1i}	latent variables (unobserved but modeled)
β_{12i}	Individual-specific parameter
——→	explicitly modeled directional path
- - - - - →	unmodeled directional path
←——→	explicitly modeled covariance
←- - - - - →	unmodeled covariance

Note: the intercepts are not included in the diagrams.

Figure 2: Understanding endogeneity and its different forms (note: legend is also for Figure 3).

where ζ_{2i} is a disturbance term. Finally, suppose that we do not observe u_i and therefore simply regress y_{1i} on y_{2i}. Doing so transforms Equation (1) such that $\pi_1 u_i$ moves into the error term:

(3) $y_{1i} = \gamma_{10} + \beta_{12}\, y_{2i} + \zeta_{1i}$,

where $\zeta_{1i} = (\pi_1 u_i + e_{1i})$. Because y_{2i} is a consequence of u_i and because u_i has been incorporated into ζ_{1i}, ζ_{1i} will correlate with y_{2i}, and we will have an endogeneity problem, i.e., our explanatory variable y_{2i} will correlate with the disturbance term of Equation (3) ζ_{1i}: $\text{Cov}(y_{2i}, \zeta_{1i}) = \pi_0\pi_2\text{Var}(u_i)$[iv].

Figure 2d) illustrates *simultaneity*. Two variables y_{2i} and y_{1i} mutually affect one another, and when using cross-sectional data, these effects will both be reflected in the data; thus, we have to assume that the effects occur simultaneously. In such situations, endogeneity occurs if one of the endogenous variables is treated as exogenous and OLS is applied (Wooldridge 2010). In fact, if the source of endogeneity is simultaneity, the covariance between y_{2i} and ζ_{1i} in Figure 2a) can be expressed as a function of the unmodeled effect β_{21}: $\text{Cov}(y_{2i}, \zeta_{1i}) = \beta_{21}\text{Var}(\zeta_{1i})/ (1- \beta_{12}\beta_{21})$. This simultaneity may be illustrated using Sande and Haugland's (2015) data. It is entirely possible that the level of formal contracting, in part, is a result of observing previous performance outcomes, which in cross-sectional data will be reflected in current performance levels.

Measurement errors lead to endogeneity because of two reasons. First, as discussed by Bollen (1989) and Wooldridge (2010), among others, measurement errors in the independent variables often lead to attenuation bias, i.e., they reduce the parameter estimates in the structural model. Second, and as illustrated in Figure 2e), measurement errors may correlate, for instance because of common method bias (Antonakis, Bendahan, Jacquart, & Lalive, 2010). For instance, several of the variables in the Sande and Haugland's (2015) data are multi-item Likert scales which may suffer from measurement errors arising from common method bias.

2.2 Why is endogeneity problematic?

The problem with endogeneity can be best illustrated by referring to the gold standard for

inferring causality: randomized controlled experiments. Suppose that we want to compare the use of a formal contract versus relying on an informal handshake for managing supplier relationships. Using experimental jargon, we label these two states as *treatment* (formal contract) and *control* (handshake). Then, the treatment effect on the outcome for relationship *i* is equal to the *difference* between the *potential outcomes* in the treatment *t* and control *c* states (Morgan & Winship, 2007):

(4) $\delta_i = y_i^t - y_i^c$

Ideally, we want to observe both y_i^t and y_i^c because we would then know the outcomes under both the treatment and control for each relationship. Unfortunately, we never observe both potential outcomes for each supplier relationship; we observe only the outcome for the *factual* choice, either y_i^t *or* y_i^c. As such, since we do not observe the *counterfactual* outcome, we cannot compute the outcome differences for each supplier relationship. An alternative then is to compare different supplier relationships and estimate *average effects* across these relationships (Imbens & Angrist, 1994). This average effect for a population is called the average treatment effect (ATE):

(5) $\bar{\delta} = \bar{y}^t - \bar{y}^c$,

where $\bar{y}^t$ and $\bar{y}^c$ are the average values of y_i^t and y_i^c across *all* supplier relationships in the treatment and control groups, respectively. Crucially, since we only observe factual outcomes, $\bar{\delta}$ is only an *estimate*. To estimate $\bar{\delta}$ consistently, we must assume that the counterfactual outcome for the treatment group is similar to the factual outcome for the control group, and vice versa. Random treatment assignment ensures that this assumption holds, because it ensures that the groups are, on average, similar to one another before the treatment takes place (Morgan & Winship, 2007). Randomized experiments thus provide consistent estimates.

Unfortunately, unlike randomized experiments, with observational data, treatments cannot

be randomly assigned. Therefore, the treatment and control groups are not likely to be similar and even if we could observe the counterfactual outcome for the treatment group, it would *not* be similar to the average factual outcome for the control group. In addition, with observational data, treatments are not independent of outcomes. For instance, certain relationships based on certain characteristics might choose formal contracts over a handshake agreement – thus, treatments are endogenous. The dotted correlation between y_{2i} and ζ_{1i} in Figure 2a) emphasizes this unobserved relationship, which, if unaccounted for, will confound the true effect of y_{2i} on y_{1i}, β_{12}. Hence, if we naïvely try to estimate β_{12} simply by regressing y_{1i} on y_{2i}, we will not separate the true effect β_{12} from the correlation between y_{2i} and ζ_{1i}. The resulting parameter estimate will be a mix of β_{12} and $Cov(y_{2i},\zeta_{1i})$. Indeed, simulations presented by Semadeni, Withers, and Trevis Certo (2014) suggest that even when the level of endogeneity is low, we could have coefficient estimates that are biased by as much as 100 percent, for example by shifting a truly negative coefficient to a significantly positive one.

In Figure 1, we suggest that a researcher should theoretically evaluate the endogeneity problem by (a) asking whether theory or context suggests the possibility of omitted variables, simultaneity, or measurement errors and (b) evaluating what type of data are available. If the researcher has access to time series or panel data, techniques for addressing endogeneity problems other than those described here are available and it is typically better to rely on these other methods than IVs (Rossi, 2014)[v]. However, researchers who only have access to cross-sectional (survey) data should consider the techniques described here.

2.3 The role of essential heterogeneity

Essential heterogeneity arises because the effect of decisions y_{2i} varies across individuals and because decision makers self-select and sort on components of the outcomes from their decisions y_{2i}. Decision makers are often aware of at least some of the components of the potential gains and

costs that accrue from choosing y_{2i}; based on this awareness, they will choose the level of y_{2i} that they expect will optimize outcomes. However, in any observational dataset, some components of these potential gains/costs from y_{2i} will remain unobserved—i.e., omitted variables—which may both influence the *chosen level* of y_{2i} and the *effect* of y_{2i}. Figure 2f) illustrates this phenomenon, where outcome y_{1i} is a function of the endogenous variable y_{2i} and an error term ζ_{1i}:

(6) $y_{1i} = \gamma_{10} + \beta_{12i} y_{2i} + \zeta_{1i}$.

where γ_{10} is the intercept. The subscript i in β_{12i} indicates that each individual i faces a unique effect of the endogenous explanatory variable y_{2i}. β_{12i} is given as follows:

(7) $\beta_{12i} = \beta_{12} + \nu_i$,

where β_{12} is the average effect of y_{2i} on y_{1i}, and ν_i is a random error term.

In Figure 2f), we show β_{12i} as a *correlated random coefficient*: it is random and correlates with observed and unobserved variables, including y_{2i} (Heckman & Vytlacil, 1998, p. 974). Therefore, while correlating with the error term ζ_{1i}, the explanatory variable y_{2i} may also correlate with its own effect β_{12i}. Heckman, Urzua, and Vytlacil (2006, p. 389) call this phenomenon "essential heterogeneity" whereas Luan & Sudhir (2010, p. 244) call it "slope endogeneity".

To illustrate, recall that there are several possible omitted variables that could affect formal contracting in the Sande and Haugland (2015) data that could cause endogeneity related problems. These omitted variables may in addition impact components of the gains/costs incurred by the firms from choosing higher levels of formal contracting. For example, high supplier reputation could potentially curb opportunism and as such the effect of formal contracts on outcomes would be weaker; however, since we do not measure supplier reputation, we are likely to find that the chosen level of formal contracting correlates with its own effect on outcomes.

Essential heterogeneity is of concern for three reasons. First, we may need special techniques. As explained in section 3 (and Figure 1), control function estimators are particularly

useful. Second, and as discussed in section 5, essential heterogeneity makes assessing the assumptions underlying our models more difficult. Third, and as discussed in section 6, essential heterogeneity triggers questions regarding how selection takes place, how the effect is distributed across the population, and what unobserved variables might drive the heterogeneous effects.

3 WHAT TECHNIQUE/ESTIMATOR IS APPROPRIATE?

We now turn to the task of choosing the technique to address endogeneity. We offer separate discussions for continuous, discrete, latent, and mixed endogenous variables.

3.1 Addressing endogeneity with continuous explanatory variables

3.1.1 The control variables approach

In principle, the use of control variables can mitigate the omitted variable bias. Germann et al. (2015) refer to this approach as "rich data models", which essentially suggests that the regression equation should not omit any conceivable control variable. However, this approach will only work if we are able to measure *all* the omitted variables *perfectly*, which is quite unfeasible using field or observational data. At best, the available control variables represent imperfect measures of some of the potential control variables. However, as we discuss later, regardless of these limitations, control variables can still play an important role in ensuring the exogeneity of IVs.

3.1.2 The IV approach (Box 1, Figure 1)

The IV approach is the most commonly used strategy for handling endogeneity and identifying effects of interest to us (Reiersøl, 1945; Wright, 1928). Figure 3a) illustrates a path model[vi], in which an exogenous variable x_{1i} affects an endogenous regressor y_{2i}, which in turn affects the dependent variable of interest y_{1i}. Endogeneity here is the correlation between ζ_{2i} and ζ_{1i}.

Given that our core interest is the estimation of β_{12}, could x_{1i} help us identify the correct parameter estimate for β_{12}? The answer is yes if x_{1i} satisfies *two key criteria*: (1) x_{1i} must be significantly correlated with y_{2i}, and (2) x_{1i} must be uncorrelated with the disturbance term ζ_{1i},

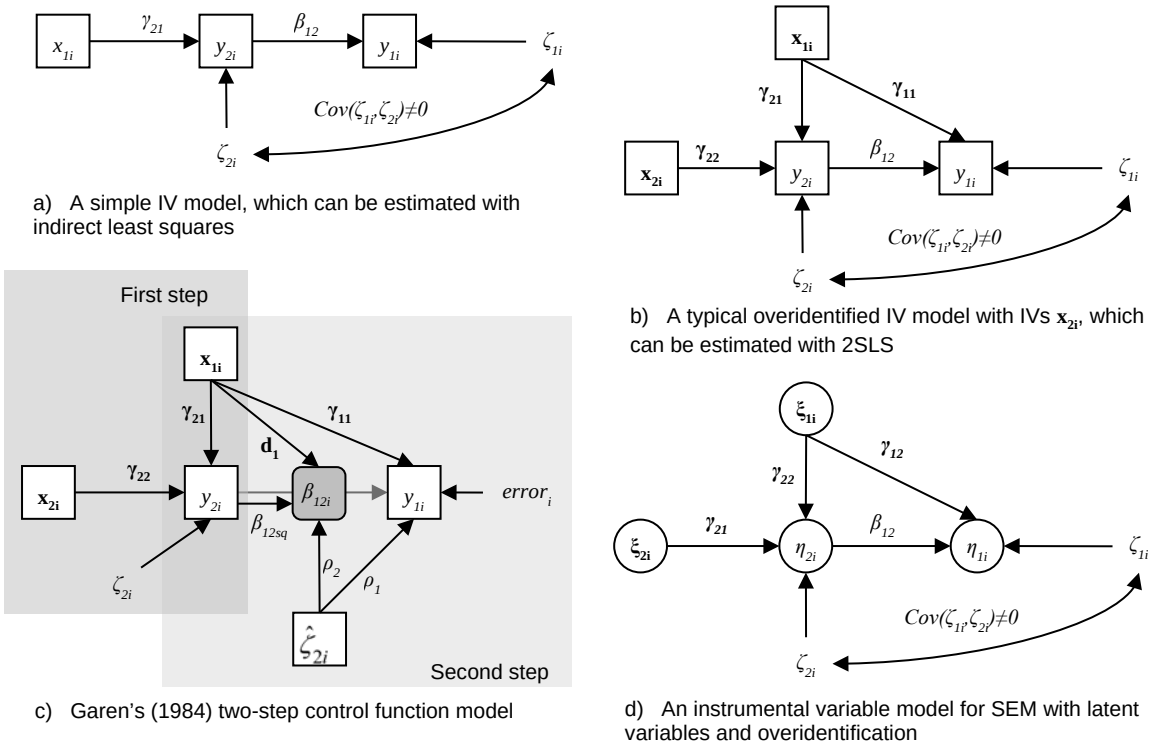


Figure 3: Graphical illustration of different types of estimators (see Figure 2 for legend).

Cov$(x_{1i}, \zeta_{1i}) = 0$. The first assumption is called the *relevance condition*, while the second one is called the *exogeneity* or *orthogonality condition*. It is also called the *exclusion restriction* because we restrict x_{1i} from having any direct relationship with y_{1i}. (Note: this condition consists of three parts that we discuss further in section 4.) If these two conditions hold, we can express the effect of y_{2i} on y_{1i} (Angrist & Pischke, 2009)[vii] as follows:

$$(8)\quad \beta_{12} = \frac{\mathrm{Cov}(x_{1i}, y_{1i})}{\mathrm{Cov}(x_{1i}, y_{2i})}$$

Cov$(x_{1i},\zeta_{2i}) = 0$ enables the identification of γ_{21}. We can also express β_{12} in terms of regression coefficients, which can be easier to understand than Equation (8)[viii]:

$$(9)\quad \beta_{12} = \frac{\beta_{12}\cdot\gamma_{21}}{\gamma_{21}} = \frac{\text{the indirect effect of } x_{1i} \text{ on } y_{1i}}{\text{the effect of } x_{1i} \text{ on } y_{2i}}$$

Assuming that the relevance and exogeneity conditions hold, we use Equation (1) to infer the

effect of y_{2i} on y_{1i}. Equations (1) and (2) are called *indirect least squares*.

3.1.2.1 The logic of the IV approach

In a population of units under investigation, we can distinguish between three hypothetical unobserved subpopulations: *compliers*—those who respond positively to the IV (i.e., take the treatment only when induced to by the IV); *always takers*—those who choose the treatment regardless of the IV; and *never-takers*—those who never choose the treatment regardless of the IV. In addition, a potential fourth subpopulation is called *defiers*—those who respond negatively to the IV (the opposite of the compliers' reaction). The IV estimation assumes that a population cannot contain both compliers and defiers (the monotonicity assumption), which means that an IV can only affect the treatment variable in one direction (Imbens & Angrist, 1994). As such, an IV estimation is solely based on the complier subpopulation, where some units take the treatment because of the high value of the IV and others do not because of the low value of the IV. If the IV satisfies the exogeneity condition, then, among the compliers, we can compare those who have taken the treatment with those who have not and assume that the average factual outcome for the treated is comparable to the average counterfactual outcome for the non-treated, and vice versa (Morgan & Winship, 2007).

Hence, the logic of IV methods follows Haavelmo's (1944, p. 14) idea of exploiting "the stream of experiments that Nature is steadily turning out from her own enormous laboratory." IVs essentially work like lab assistants who do the actual task of randomly assigning subjects to treatment and control groups. In a situation where the business environment approximates a randomized experiment, we call it a "natural experiment" (Angrist & Krueger, 2001).

3.1.2.2 IV estimates and heterogeneous effects

That IV estimates are based only on the complier subpopulation is not problematic if we expect that the effect is uniform across all units. However, the effect quite possibly differs between

individual units and subpopulations. Recognizing this potential issue, Imbens and Angrist (1994) call the IV estimate a local average treatment effect (LATE) rather than an ATE. A different IV will often provide a different estimate of the effect because those firms that comply with one IV may experience a different effect of the endogenous variable than those that comply with another IV. Later in this article, we discuss the implications of heterogeneous effects for the choice of estimators and IVs.

3.1.2.3 IV estimation with two-stage least squares (2SLS)

One of the most common IV approaches involves the use of 2SLS. The indirect least squares in Equations 8 and 9 provide the logic of this approach. Essentially, with 2SLS, we first regress y_{2i} on two row vectors of exogenous control variables $\mathbf{x_{1i}}$ and IVs $\mathbf{x_{2i}}$:

(10) $y_{2i} = \gamma_{20} + \mathbf{x_{1i}\gamma_{21}} + \mathbf{x_{2i}\gamma_{22}} + \zeta_{2i}$

where γ_{20} is the intercept; $\boldsymbol{\gamma_{21}}$ and $\boldsymbol{\gamma_{22}}$ are column vectors of coefficients; and ζ_{2i} is the error term. Next, we regress y_{1i} on the prediction of the endogenous variable $\hat{y}_{2i}$ and control variables $\mathbf{x_{1i}}$:

(11) $y_{1i} = \gamma_{10} + \beta_{12}\hat{y}_{2i} + \mathbf{x_{1i}\gamma_{11}} + error_i$

where γ_{10} is the intercept; β_{12} is the effect of y_{2i}; $\boldsymbol{\gamma_{11}}$ is column vector of coefficients; and $error_i = \left[\zeta_{1i} + \beta_{12}\left(y_{2i} - \hat{\gamma}_{20} - \mathbf{x_{1i}}\hat{\boldsymbol{\gamma}}_{\mathbf{21}} - \mathbf{x_{2i}}\hat{\boldsymbol{\gamma}}_{\mathbf{22}}\right)\right]$ (see Figure 3b). Because the error term depends on the sampling error in $\hat{\gamma}_{20}$, $\hat{\boldsymbol{\gamma}}_{\mathbf{21}}$ and $\hat{\boldsymbol{\gamma}}_{\mathbf{22}}$, we should always use specialized software (e.g., Stata) that automatically estimates both stages and obtains the correct standard errors. Web Appendix C provides the basic Stata code for implementing 2SLS and other related IV estimators.

3.1.2.4 Interaction terms and IVs

If we suspect that the effect of an endogenous variable varies across observed variables, we can examine this heterogeneity explicitly using an IV estimator. We can include interaction terms in

Equation (11) between the endogenous explanatory variable y_{2i} and observed exogenous moderators $\mathbf{x_{1i}}$. However, the interaction terms $y_{2i}\mathbf{x_{1i}}$ are endogenous, and finding suitable IVs for these terms can be challenging. One possibility is to look for IVs among the interaction terms between $\mathbf{x_{1i}}$ and $\mathbf{x_{2i}}$. An alternative is to first predict y_{2i} according to Equation (10) and to use $\hat{y}_{2i}\mathbf{x_{1i}}$ as IVs for $y_{2i}\mathbf{x_{1i}}$. In such cases, we should never regress the dependent variable y_{1i} directly on $\hat{y}_{2i}\mathbf{x_{1i}}$; we should only use $\hat{y}_{2i}\mathbf{x_{1i}}$ as IVs for $y_{2i}\mathbf{x_{1i}}$ (Wooldridge, 1997; 2010, p. 262-268). Such 2SLS estimations are easy to implement.

3.1.3 The CF approach (Box 2 in Figure 1)

As we described in section 2.3, under essential heterogeneity, unobserved heterogeneity interacts with the effect of the endogenous variable such that it correlates with its own effect. The leading approach for estimation is the CF approach (Heckman & Robb, 1985; Wooldridge, 2008).

The simplest possible CF estimator (which *does not* account for essential heterogeneity) is similar to 2SLS in most respects. The first stage is identical to that of the 2SLS, as shown in Equation (10): we regress the endogenous explanatory variable y_{2i} on two row vectors of control variables $\mathbf{x_{1i}}$ and IVs $\mathbf{x_{2i}}$. However, unlike 2SLS, in the second stage, we do not use the predicted value of y_{2i}. Instead, we regress the dependent variable y_{1i} on the explanatory variable y_{2i}, the control variables $\mathbf{x_{1i}}$, and the estimated first-stage residual $\hat{\zeta}_{2i} = y_{2i} - \hat{y}_{2i}$, which is a composite estimate of all variables not included in $\mathbf{x_{1i}}$ and $\mathbf{x_{2i}}$ that affect y_{2i} (Wooldridge, 2010):

(12) $y_{1i} = \gamma_{10} + \beta_{12}y_{2i} + \mathbf{x_{1i}\gamma_{11}} + \rho_1\hat{\zeta}_{2i} + error_i$

where γ_{10}, β_{12}, $\boldsymbol{\gamma_{11}}$, and ρ_1 are parameters, and $error_i = \hat{\zeta}_{1i} + \rho_1\left[\hat{\gamma}_{20} - \gamma_{20} + \mathbf{x_{1i}}\left(\hat{\boldsymbol{\gamma}}_{\mathbf{21}} - \boldsymbol{\gamma}_{\mathbf{21}}\right) + \mathbf{x}_{2i}\left(\hat{\boldsymbol{\gamma}}_{\mathbf{22}} - \boldsymbol{\gamma}_{\mathbf{22}}\right)\right]$. We include $\hat{\zeta}_{2i}$ in the second-stage regression to *control* for all the unobservables that lead to the endogeneity of y_{2i}. Both stages can be consistently estimated using OLS.

The estimate of β_{12} obtained in (12) will be identical to that obtained with 2SLS because the CF approach relies on the same identification conditions as the IV approach: the relevance and exogeneity of the IVs $\mathbf{x_{2i}}$ (Wooldridge, 2010). Consequently, when using a CF estimator, the exogeneity and relevance conditions holding is equally important.

The drawback of this basic model is that if $\rho_1 \neq 0$, standard errors from OLS are not valid. Because $\hat{\zeta}_{2i}$ is a generated regressor (i.e., it is calculated as $\hat{\zeta}_{2i} = y_{2i} - \hat{y}_{2i}$), the error term becomes a function of the first-stage parameter estimates ($\hat{\boldsymbol{\gamma}}_{21}$ and $\hat{\boldsymbol{\gamma}}_{22}$). Fortunately, the standard errors are easy to correct using bootstrapping techniques (Wooldridge, 2008; 2010).

Despite these drawbacks, a simple CF estimator is useful under some conditions. First, if we have relevant and exogenous instruments and the parameter estimate for ρ_1 is significant, endogeneity is a problem, and we must account for it, which is equivalent to conducting the Hausman (1978) test for endogeneity. A non-significant ρ_1 means that endogeneity is not a major problem, and we should opt to present the results from using OLS because OLS is more efficient than 2SLS. Second, ρ_1 tells us the direction of bias. For example, if ρ_1 is negative, omitted variables positively related to y_{2i} may have a negative effect on y_{1i}.

3.1.3.1 Garen's (1984) two-step model

Even though the simple CF estimator discussed above has little to offer beyond what 2SLS does, CF estimators are flexible and useful in more complex situations. Garen (1984), for instance, developed the first estimator with correlated random coefficients when the endogenous variable is continuous. To explain this estimator, we first present the *correlated random coefficient model* (Heckman & Vytlacil, 1998). The first stage is identical to Equation (10) and is restated below:

(13) $y_{2i} = \gamma_{20} + \mathbf{x_{1i}}\boldsymbol{\gamma_{21}} + \mathbf{x_{2i}}\boldsymbol{\gamma_{22}} + \zeta_{2i}$

The second stage is similar to Equation (6), except that it includes $\mathbf{x_{1i}}$:

(14) $y_{1i} = \gamma_{10} + \beta_{12i} y_{2i} + \mathbf{x_{1i}\gamma_{11}} + \zeta_{1i}$

The total heterogeneous effect of y_{2i} now depends on both observed and unobserved variables. We therefore modify Equation (7) by setting $v_i = \mathbf{x_{1i}d_1} + h_i$, where h_i is a random variable that represents unobserved heterogeneity and $\mathbf{d_1}$ a column vector of parameters:

(15) $\beta_{2i} = \beta_{12} + \mathbf{x_{1i}d_1} + h_i$

Equations (13) through (15) show the correlated random coefficient model. Estimating Equation (14) is difficult because ζ_{2i}—and hence y_{2i}—correlate with h_i and ζ_{1i}. Assuming the joint normality distribution of ζ_{1i}, ζ_{2i} and h_i, Garen's (1984) estimator accounts for these correlations by allowing the estimated error term $\hat{\zeta}_{2i}$ from the first-stage regression to moderate the effect of y_{2i} on y_{1i}, as illustrated in Figure 3c). His model thereby allows the effect β_{12i} to correlate with unobserved heterogeneity. In addition, it includes y_{2i}^2 and interaction terms between $\mathbf{x_{1i}}$ and y_{2i}:

(16) $y_{1i} = \gamma_{10} + \beta_{12} y_{2i} + \beta_{12sq} y_{2i}^2 + \mathbf{x_{1i}\gamma_{11}} + \rho_1 \hat{\zeta}_{2i} + \rho_2 \hat{\zeta}_{2i} y_{2i} + y_{2i}\mathbf{x_{1i}d_1} + error_i$

where γ_{10}, β_{12}, β_{12sq}, $\mathbf{\gamma_{11}}$, ρ_1, ρ_2, and $\mathbf{d_1}$ are parameters, and $error_i = \zeta_{1i} + \rho_1\left[\hat{\gamma}_{20} - \gamma_{20} + \mathbf{x}_{1i}\left(\hat{\boldsymbol{\gamma}}_{\mathbf{21}} - \boldsymbol{\gamma}_{\mathbf{21}}\right) + \mathbf{x}_{2i}\left(\hat{\boldsymbol{\gamma}}_{\mathbf{22}} - \boldsymbol{\gamma}_{\mathbf{22}}\right)\right] + \rho_2 y_{2i}\left[\hat{\gamma}_{20} - \gamma_{20} + \mathbf{x}_{1i}\left(\hat{\boldsymbol{\gamma}}_{\mathbf{21}} - \boldsymbol{\gamma}_{\mathbf{21}}\right) + \mathbf{x}_{2i}\left(\hat{\boldsymbol{\gamma}}_{\mathbf{22}} - \boldsymbol{\gamma}_{\mathbf{22}}\right)\right]$. The marginal effect of y_{2i} is

(17) $\delta y_{1i} / \delta y_{2i} = \beta_{12} + 2\,\beta_{12sq}\, y_{2i} + \rho_2 \hat{\zeta}_{2i} + \mathbf{dx_{1i}}$

Assuming mean-centered variables, the ATE of y_{2i} is β_{12}. We can estimate Equation (16) using OLS, but the error term depends on first-stage estimates. We must therefore use bootstrapping to correct for the generated regressor problem. Note that this model relies on the same identification conditions as IV methods. Hence, instrument relevance and exogeneity must also hold here.

Garen's (1984) estimator is useful for several reasons. One reason is that if the assumptions underlying the model are correct, it is more efficient than 2SLS. By including more terms and predicting a larger share of the variance in y_{1i}, standard errors will typically be smaller.

In addition, Garen's estimator is more informative than 2SLS in the context of heterogeneous effects. First, it can calculate the marginal effect of y_{2i} and confidence bands around this marginal effect across different values of $\hat{\zeta}_{2i}$. Doing so provides information about how the effect of the endogenous variable is distributed across the population. Second, we can use the parameter estimates of ρ_1 and ρ_2 to tell us what kinds of firms or individuals choose high and low values of y_{2i}. For example, if the derivative of Equation (16) with respect to $\hat{\zeta}_{2i}$ is greater than zero at a high value of y_{2i}, firms or individuals with unexpectedly high y_{2i} (i.e., $\hat{\zeta}_{2i} > 0$) tend to achieve higher outcomes y_{1i} with high values of y_{2i} than what those with unexpectedly low y_{2i} would achieve had they chosen a high level of y_{2i}. Third, by calculating counterfactual outcomes for different types of individuals, we can examine the presence of comparative or absolute (dis)advantages; this search is a central rationale in empirical MS research. For example, if firms with unexpectedly high y_{2i} would have earned less than others had they chosen a low level of y_{2i} and those with an unexpectedly low level of y_{2i} would have earned less than others had they chosen a high level of y_{2i}, we can conclude that the unobserved variables captured in $\hat{\zeta}_{2i}$ represent the unobserved comparative advantages of choosing a high value of y_{2i}.

However, Garen's (1984) estimator has limitations because it rests on relatively strong assumptions of the joint normality and homoscedasticity of error terms ζ_{1i}, ζ_{2i} and h_i in Equation (13)–(15), which implies that the relationships between the error terms are linear. If these assumptions are incorrect, then IV estimators, particularly those that include interaction terms between observed heterogeneity $\mathbf{x_{1i}}$ and the endogenous variable y_{2i}, are generally more robust in estimating the ATE β_{12} than Garen's (1984) estimator is because such IV estimators rest on weaker assumptions than Garen's (1984) estimator (Wooldridge, 1997; 2003).

3.1.3.2 Extensions

To relax some of the strong assumptions in the model, Garen's (1984) estimator has been extended in several ways. Card (2001) suggests adding more interaction terms to Garen's (1984) model: two-way interaction terms between the IVs $\mathbf{x_{2i}}$ and $\hat{\zeta}_{2i}$ and three-way interaction terms between y_{2i}, $\mathbf{x_{2i}}$ and $\hat{\zeta}_{2i}$. Luan and Sudhir's (2010) model also relaxes some of the stronger assumptions of Garen's (1984) model and accommodates multiple endogenous variables. Wooldridge (2015) presents several possible extensions of Garen's (1984) model, highlighting the flexibility of CF estimators: they can be applied to both linear and non-linear models and be modified for different kinds of assumptions. An example of such a model is De Blander's (2010) estimator, which accommodates multiple endogenous variables and relaxes the normality and linearity assumption of Garen's (1984) estimator. He does so by adding more quadratic and interaction terms to Garen's (1984) estimator. Below, we present a simplified version of this estimator that includes only one endogenous variable:

(18) $$y_{1i} = \gamma_{10} + \beta_{12} y_{2i} + \beta_{12sq} y_{2i}^2 + \mathbf{x_{1i}}\boldsymbol{\gamma_{11}} + \rho_1 \hat{\zeta}_{2i} + \rho_2 \hat{\zeta}_{2i} y_{2i} + y_{2i}\mathbf{x_{1i}}\mathbf{d_1}$$

$$+ y_{2i}\mathbf{x_{2i}}\mathbf{d_2} + \hat{\zeta}_{2i}\mathbf{x_{1i}}\boldsymbol{\rho_3} + \hat{\zeta}_{2i}\mathbf{x_{2i}}\boldsymbol{\rho_4} + \rho_5 \hat{e}_i + error_i$$

where the error term includes elements of the first-stage parameter estimates, and $\hat{e}_i$ is equal to $\hat{\zeta}_{2i}^2$ orthogonalized with respect to a vector containing all cross-products of the exogenous variables. This model is open to the unobserved heterogeneity having a non-linear relationship with the dependent variable because it interacts with both control variables $\mathbf{x_{1i}}$ and IVs $\mathbf{x_{2i}}$. This estimator is more efficient, has lower variance and is no less robust than Wooldridge's (2003) IV estimator. However, the cost of such efficiency is that, unlike Wooldridge's (2003) IV estimator, it assumes that the first-stage regression (Equation 13) is not mis-specified.

An additional advantage of De Blander's (2010) estimator, which can be implemented by OLS and bootstrapping, is that all the extra interaction terms provide additional information on how selection takes place and how unobserved heterogeneity is related to the dependent variable. However, the cost of these additional interaction terms is that more observations are needed across all levels of the moderating variables, which increases the requirement for a larger sample size. Web Appendices B (p. 22) and C demonstrate how to implement Garen's (1984) and De Blander's (2010) estimators in Stata.

3.2 Addressing endogeneity with discrete explanatory variables

Marketing researchers are often interested in discrete explanatory variables. Instead of using a Likert scale, Sande and Haugland (2015) could, for example, have measured formal contracting as a binary variable (formal contract=1, handshake=0). Again, consistent with the counterfactual approach described in Section 2.2, $y_{2i} = 1$ is the *treatment state* and $y_{2i} = 0$ is the *control state*.

3.2.1 The IV approach (Box 3, Figure 1)

When we can assume that there is no essential heterogeneity, we can use 2SLS as described in Equations (10) and (11). However, we can use a more efficient version of 2SLS than the usual one. Here, the first stage involves a probit estimator:

$$y_{2i}^{*} = \mathbf{x_{1i}}\boldsymbol{\gamma}_{\mathbf{21}}^{\mathbf{p}} + \mathbf{x_{2i}}\boldsymbol{\gamma}_{\mathbf{22}}^{\mathbf{p}} + \zeta_{2i} \tag{19}$$

where $\boldsymbol{\gamma}_{\mathbf{21}}^{\mathbf{p}}$ and $\boldsymbol{\gamma}_{\mathbf{22}}^{\mathbf{p}}$ are the probit slope parameters, and y_{2i}^{*} is a latent variable, which we do not observe. Instead, we observe y_{2i}, which takes on values 0 or 1 according to the following rule:

$$y_{2i} = \begin{cases} 1 & \text{if } y_{2i}^{*} > 0 \quad \Leftrightarrow \mathbf{x_{1i}}\boldsymbol{\gamma}_{\mathbf{21}}^{\mathbf{p}} + \mathbf{x_{2i}}\boldsymbol{\gamma}_{\mathbf{22}}^{\mathbf{p}} > \zeta_{2i} \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

We obtain fitted probabilities of treatment given the observed variables $\hat{P}_i\left(\mathbf{x_{1i}}, \mathbf{x_{2i}}\right) = \Phi\left(\mathbf{x_{1i}}\hat{\boldsymbol{\gamma}}_{\mathbf{21}}^{\mathbf{p}} + \mathbf{x_{2i}}\hat{\boldsymbol{\gamma}}_{\mathbf{22}}^{\mathbf{p}}\right)$ (i.e., the propensity score), where $\Phi(\bullet)$ is the cumulative density

function for the standard normal distribution. Next, we estimate Equation (11) using a 2SLS estimator, where $\hat{P}_i(\mathbf{x_{1i}}, \mathbf{x_{2i}})$ is used as an IV for y_{2i} and $\mathbf{x_{1i}}$ are control variables. It is important to recognize that when using 2SLS, we use $\hat{P}_i(\mathbf{x_{1i}}, \mathbf{x_{2i}})$ as an IV for the endogenous binary variable y_{2i} rather than as a regressor. Equation (19) does not have to be correctly specified for IV estimation to be consistent, and other link functions, such as logit, can be used. In situations when the endogenous binary variable y_{2i} is expected to interact with the control variables $\mathbf{x_{1i}}$, $\hat{P}_i(\mathbf{x_{1i}}, \mathbf{x_{2i}})\mathbf{x_{i1}}$ can be used as IVs for $y_{2i}\mathbf{x_{i1}}$ (Wooldridge, 2010). Web Appendix C demonstrates how we can implement IV estimation with a binary endogenous variable in Stata.

3.2.2 The selection model (a CF approach) (Box 4, Figure 1)

Under essential heterogeneity, the assumptions of 2SLS are not satisfied (Heckman & Robb, 1985). Therefore, a common approach is to use a selection model (Heckman, 1974; 1979; Lee, 1978) that has been applied in many MS and IOR studies. The selection model is a CF approach and is quite similar to Garen's (1984) estimator, except that the explanatory variable is discrete. Therefore, the selection model involves a probit in the first stage, i.e., Equation (19). We cannot directly estimate $\hat{\zeta}_{2i}$ from the probit. Instead, we calculate a "generalized residual" $\hat{r}_i$, which is a function of y_{2i}, $\mathbf{x_{1i}}$, and $\mathbf{x_{2i}}$ (Wooldridge, 2015, p. 428)[ix]:

(21) $$\hat{r}_i(y_{2i}=1, \mathbf{x_{1i}}, \mathbf{x_{2i}}) = \phi\left(\mathbf{x_{1i}}\hat{\boldsymbol{\gamma}}^{\mathbf{p}}_{\mathbf{21}} + \mathbf{x_{2i}}\hat{\boldsymbol{\gamma}}^{\mathbf{p}}_{\mathbf{22}}\right) \Big/ \Phi\left(\mathbf{x_{1i}}\hat{\boldsymbol{\gamma}}^{\mathbf{p}}_{\mathbf{21}} + \mathbf{x_{2i}}\hat{\boldsymbol{\gamma}}^{\mathbf{p}}_{\mathbf{22}}\right)$$

(22) $$\hat{r}_i(y_{2i}=0, \mathbf{x_{1i}}, \mathbf{x_{2i}}) = -\phi\left(\mathbf{x_{1i}}\hat{\boldsymbol{\gamma}}^{\mathbf{p}}_{\mathbf{21}} + \mathbf{x_{2i}}\hat{\boldsymbol{\gamma}}^{\mathbf{p}}_{\mathbf{22}}\right) \Big/ \left[1 - \Phi\left(\mathbf{x_{1i}}\hat{\boldsymbol{\gamma}}^{\mathbf{p}}_{\mathbf{21}} + \mathbf{x_{2i}}\hat{\boldsymbol{\gamma}}^{\mathbf{p}}_{\mathbf{22}}\right)\right]$$

where $\phi(\bullet)$ and $\Phi(\bullet)$ are the probability density and the cumulative density functions for the standard normal distribution. The ratios, $\phi(\bullet)/\Phi(\bullet)$ and $\phi(\bullet)/[1-\Phi(\bullet)]$ are called inverse Mills ratios. The second step of the Heckman two-step method can be obtained from Garen's (1984) estimator by replacing $\hat{\zeta}_{i2}$ with $\hat{r}_i$ in Equation (16) and removing the quadratic effect of y_{2i}:

(23) $y_{1i} = \gamma_{10} + \beta_{12} y_{2i} + \mathbf{x_{1i}\gamma_{11}} + \rho_1 \hat{r}_i + \rho_2 \hat{r}_i y_{2i} + y_{2i}\mathbf{x_{1i}d_1} + error_i$

The ATE of y_{2i} is then β_{12}. By rearranging, we obtain the following two equations, termed "switching regressions" because individuals or firms switch between treatment and control states:

(24) $y_{1i}^1 = k_1 + \mathbf{x_{1i}\gamma_{11}^1} + \sigma_1 \hat{r}_i \left(y_{2i} = 1, \mathbf{x_{1i}}, \mathbf{x_{2i}} \right) + error_i^1$

(25) $y_{1i}^0 = k_0 + \mathbf{x_{1i}\gamma_{11}^0} + \sigma_0 \hat{r}_i \left(y_{2i} = 0, \mathbf{x_{1i}}, \mathbf{x_{2i}} \right) + error_i^0$

where y_{1i}^1 and y_{1i}^0 are outcomes for the treatment and control group; $k_0 = \gamma_{10}$ and $k_1 = \gamma_{10} + \beta_{12}$ are intercepts; $\boldsymbol{\gamma_{11}^0} = \boldsymbol{\gamma_{11}}$ and $\boldsymbol{\gamma_{11}^1} = \boldsymbol{\gamma_{11}} + \mathbf{d_1}$ are the slope parameters for $\mathbf{x_{1i}}$; and $\sigma_1 = \rho_1 + \rho_2$ and $\sigma_0 = \rho_1$ are the slope parameters for the inverse Mills ratios. The ATE is $\beta_{12} = k_1 - k_0$.

We can estimate equations (19), (24) and (25) manually as a two-step procedure using any statistical software with bootstrapping, a probit estimator, and the possibility of calculating inverse Mills ratios. As with Equation (18), we must correct the standard errors to account for the generated regressors. Alternative estimation procedures are also available (see Web Appendix C).

We can use the output from estimating Equations (24) and (25) to better understand the effects of the endogenous explanatory variable and how it is selected. First, we can use the results to predict the expected outcomes for a random observation in the sample. Second, we can predict the expected factual and counterfactual outcomes for both the treatment and control groups. Finally, analogous to the interpretation of ρ_1 and ρ_2 for Garen's (1984) estimator, we can interpret the estimates of σ_1 and σ_0 to substantively discuss positive and negative selection into particularistic strategy choices as well as the comparative and absolute advantages. Maddala (1983) and Hamilton and Nickerson (2003) provide in-depth discussions of these issues.

3.2.3 Marginal treatment effect (MTE) estimation (Box 4, Figure 1)

A drawback of the selection model is that (similar to Garen's (1984) estimator), it assumes the joint normality of the error terms of the first- and second-stage equations[x]. Recent years have

seen the development of new methods for estimating the MTE. Some of these methods accommodate less restrictive assumptions and facilitate the calculation of different types of treatment effects. Heckman and Vytlacil (2005) define the MTE as the expected effect of treatment conditional on observed characteristics $\mathbf{x_{1i}}$ and unobservables $U_{2i} = F_{\zeta_{2i}}(\zeta_{2i})$, where $F_{\zeta_{2i}}$ is the cumulative distribution function of ζ_{2i} so that U_{2i} is the propensity to not be treated:

$$(26)\quad \text{MTE} = E\left(y_{1i}^{1} - y_{1i}^{0} \middle| \mathbf{x_{1i}} = \mathbf{x_1}, U_{2i} = u_2\right) = E\left(\beta_{12i} \middle| \mathbf{x_{1i}} = \mathbf{x_1}, U_{2i} = u_2\right)$$

where $\mathbf{x_1}$ is a vector of the values of $\mathbf{x_{1i}}$, and u_2 is the value of U_{2i} for individual *i*. In other words, the MTE is defined at particular values of $\mathbf{x_{1i}}$ and U_{2i}: MTE is the treatment effect for individuals with observed characteristics $\mathbf{x_{1i}} = \mathbf{x_1}$ and who are at the u_2^{th} quantile in the distribution of ζ_{2i} (Cornelissen et al. 2016).

When we evaluate the MTE at the point where the propensity to not be treated, U_{2i}, is equal to the propensity to be treated (i.e., the propensity score $\hat{P}_i(\mathbf{x_{1i}}, \mathbf{x_{2i}})$), the MTE is the effect of treatment for those actors that are indifferent to treatment (i.e., $\mathbf{x_{1i}}\boldsymbol{\gamma}_{\mathbf{21}}^{\mathbf{p}} + \mathbf{x_{2i}}\boldsymbol{\gamma}_{\mathbf{22}}^{\mathbf{p}} = \zeta_{2i}$) (Heckman, Urzua, and Vytlacil, 2006). These marginal individuals are important because they are often the ones who policymakers want to "treat" (Björklund and Moffitt, 1987). For example, a company might be interested in implementing policies that motivate employees or customers to behave in a certain way when interacting with customers. The MTE of this behavior on, for example, customer satisfaction or repeat purchases should be informative in such cases.

MTE estimation is important for at least three reasons. First, Heckman, Urzua, and Vytlacil (2006) show that all treatment effects, including the ATE, the treatment effect on the treated, the treatment effect on the untreated, and the LATE, can be constructed as the weighted

averages of the MTE by integrating over U_{2i}. Second, we can estimate confidence bands around the MTE across the range of U_{2i} to understand how selection occurs and how the effect is distributed among the population. Third, the MTE can be used to construct estimators that answer policy-relevant questions, even if we do not have IVs that identify an interesting complier subpopulation (e.g., Carneiro, Heckman, and Vytlacil, 2011).

The MTE can be estimated in several ways. For instance, Equation (23), which is also called the fully parametric normal model, can be used. In that case, the MTE is equal to

(27) $$\text{MTE} = \beta_{12} + \mathbf{x_{1i}}\mathbf{d_1} + \rho_2 \Phi^{-1}(U_{2i})$$

where Φ^{-1} is the inverse of the standard normal cumulative density function, and $\Phi^{-1}(U_{2i})$ is equal to the residual in Equation (19).

Other methods relax some of the restrictive assumptions of the fully parametric normal model, including nonparametric local instrumental variable (LIV) estimation with minimal assumptions, semiparametric and parametric polynomial models (Brave & Walstrom, 2014; Cornelissen et al., 2016). The LIV estimator is the derivative of the conditional expectation of y_{2i} with respect to the propensity score (Heckman and Vytlacil, 2005). Nonparametric LIV estimation can recover the MTE pointwise, which—loosely speaking—means that we obtain a treatment effect for each combination of the values in $\mathbf{x_{1i}} = \mathbf{x_1}$ and U_{2i} (Cornelissen et al. 2016).

Unfortunately, the nonparametric LIV approach requires large amounts of data that provide empirical support of the propensity score conditional on $\mathbf{x_{1i}}$, which is difficult to achieve in practice. Therefore, researchers estimating MTEs use models that employ additional assumptions (Cornelissen et al. 2016). An example of such a model is the parametric polynomial model, which relies on stronger assumptions than both nonparametric LIV and semiparametric

polynomial models. However, it is less restrictive than the fully parametric normal model because it relaxes the assumption of the joint normality of the error terms. Parametric polynomial models involve regressing the performance variable onto $\mathbf{x_{1i}}$, the propensity score $\hat{P}_i(\mathbf{x_{1i}}, \mathbf{x_{2i}})$ and its interaction terms with $\mathbf{x_{1i}}$, and polynomials of the propensity score. Jointly significant linear and higher-order polynomial expansion terms of the propensity score indicate selection on unobservables (Brave & Walstrom, 2014, Heckman, Schmierer, and Urzua, 2010).

Cornelissen et al. (2016) provide a review of the MTE literature. Illustrative empirical applications of MTE estimation include Aakvik, Heckman, and Vytlacil (2005), Carneiro, Heckman, and Vytlacil (2011), and Brinch, Mogstad, and Wiswall (2017). In Web Appendix C, we describe how to implement selection models (the manual procedure, in addition to built-in and user-written commands) and commands for estimating the MTE using the fully parametric normal model and parametric polynomial models.

3.3 Addressing endogeneity with latent explanatory variables

3.3.1 IVs in SEM (Box 5, Figure 1)

The use of multi-item survey measures has traditionally been prominent in marketing. Muthén and Jöreskog (1983) suggest how IVs can be used in SEM to tackle endogeneity. We illustrate this use in Figure 3d), where $\boldsymbol{\xi_{1i}}$ and $\boldsymbol{\xi_{2i}}$ are vectors of exogenous latent variables and η_{1i} and η_{2i} are endogenous latent variables. The conditions for identifying the effect of interest β_{12} are the same as those for other IV estimators: the latent IVs $\boldsymbol{\xi_{2i}}$ should be significantly related to η_{2i}; the effects of $\boldsymbol{\xi_{2i}}$ on η_{1i} should be completely mediated by η_{2i}; and $\boldsymbol{\xi_{2i}}$ must be as good as randomly assigned (i.e., uncorrelated with ζ_{1i} and ζ_{2i}). The main differences between Figure 3d) and 2SLS are that the variables of interest are latent and measured by multiple measures in SEM and that the covariance between the disturbance terms ζ_{1i} and ζ_{2i} is explicitly estimated as part of the

model in SEM (using, e.g., maximum likelihood). By estimating the covariance between the error terms, we control for omitted variables and other sources of endogeneity. With a few exceptions, this approach has not been used much in MS and IOR research.

Notably, the co-variances among the disturbance terms of endogenous variables in an SEM model should not be confused with correlations among measurement errors. SEM users have long been skeptical about post hoc opening up for correlated error terms to improve model fit, particularly across different constructs, unless there are substantial or theoretical reasons for doing so. This skepticism stems from its potential capitalization on random sample specific characteristics (Cole, Ciesla, & Steiger, 2007). However, co-variances among the disturbance terms of endogenous variables are not a misspecification or a way of capitalizing on chances to improve model fit. Instead, they are part of a conscious strategy to identify the theoretical effect of interest. In Web Appendices B (p. 18 and 26) and C, we demonstrate how to control for endogeneity in an SEM model using the `sem` command in Stata (StataCorp, 2017).

3.3.2 Accounting for essential heterogeneity in SEM (Box 6, Figure 1)

To our knowledge, standard SEM software packages do not allow for interactions between latent variables and residuals, which could account for essential heterogeneity. One option then is to create (latent variable) scores, treat the scores as continuous variables, and use a CF estimator.

3.4 Mixes of different types of endogenous variables (Box 7, Figure 1)

Sometimes, we are interested in more complex models that include several different types of endogenous and dependent variables. Two Stata commands can be useful in such situations: conditional mixed-process (`cmp`) regression and generalized structural equation modeling (`gsem`). Both can be used to control for endogeneity using IVs in more complex models.

The user-written `cmp` command is based on the classical linear model and the normal

distribution, and its primary strength is that it can estimate a recursive set of regression equations that mix different models for the dependent and endogenous variables (classic linear regression, truncated regression, censored (tobit) regression, probit, ordered probit, interval regression, and multinomial probit). `cmp` relies upon simulated maximum likelihood for estimation of the equations and is written as a seemingly unrelated regression estimator, which enables the error terms of the different equations to correlate enabling us to account for endogeneity. We refer to Roodman (2011) for further details about the `cmp` command.

Stata's built-in `gsem` command fits structural equation models with generalized linear response variables. Outcome variables can be continuous, binary, count, categorical, and ordinal with many different distribution families and link functions. `gsem` can thus be used to fit a variety of models that also include latent variables and enable us to control for endogeneity by allowing correlations between error terms. For example, StataCorp (2017, p. 451) shows how we can include Heckman selection models in more complex SEMs. We refer to Stata's reference manual for further details about `gsem` (StataCorp, 2017).

4 WHAT IVS SHOULD BE CHOSEN?

All the techniques described in this article rely on the exogeneity and relevance assumptions. For example, Figure 3a)–d) show the exogeneity condition by restricting the IVs from having paths directly linking them with the dependent variables. These restrictions represent non-trivial assumptions and are no less important to the estimation than the paths that are actually modeled (i.e., not restricted to zero). In fact, simulations conducted by Semadeni et al. (2014) suggest that even low levels of endogeneity among the IVs could increase reported parameter estimates by as much as 1000 percent. Hence, to trust the results, it is important that we have very good reasons to believe that the exogeneity condition holds. The IVs must also be relevant, i.e., explain

sufficient unique variance for each endogenous regressor. Weak IVs increase the finite sample bias of 2SLS in the direction of the OLS estimate, and this bias worsens if we include more IVs, the sample size is small, the IVs are less exogenous, or the correlation between the error terms of the first- and second-stage regressions is stronger (e.g., Hahn and Hausmann, 2003). In addition, weak IVs undermine the validity of asymptotic standard errors (e.g., Nelson & Startz, 1990). Therefore, if we have weak IVs, we cannot trust the t-statistics and confidence intervals.

As we will describe in section 5, we should assess these assumptions empirically, but the test procedures described there also rely on assumptions. Therefore, any empirical model will rely on certain untestable assumptions and the credibility of the empirical findings thus ultimately relies on theoretical and contextual arguments. Germann et al. (2015, p. 10) rightly point out that "researchers explore the meaning of the various models' identifying assumptions in light of their context and then determine the appropriate specifications."

MS and IOR researchers face two key questions before choosing their IVs: How can we find IVs and theoretically justify that the exogeneity and relevance conditions should hold? And, can control variables help us ensure that the exogeneity condition holds? This is where the role of theory becomes important, and we discuss this role below.

4.1 Finding relevant IVs

Relevant IVs are variables for which one or more mechanisms link the IV with the endogenous explanatory variable. In general, identifying relevant IVs in a context is dependent on how endogeneity arises in the setting. For instance, if endogeneity is likely to arise out of common method bias (e.g., social desirability), we should collect IVs from other data sources that are not contaminated by the same bias but that are still closely related to the endogenous explanatory variable of interest. If simultaneity is the problem, a solution may involve using data where the IVs are gathered at an earlier point in time than the endogenous explanatory variable. If the

endogenous explanatory variable is self-selected and essential heterogeneity is a concern, we should consider how selection occurs and find variables affecting the marginal costs and benefits of selecting high or low values of the endogenous explanatory variable. For example, in Web Appendix B (p. 8), we identify potential IVs by looking for variables in the Sande & Haugland's (2015) data that affect the marginal costs of writing contracts and the marginal benefits of formal contracting. Notice also how we explain the mechanisms through which we believe the IVs affect formal contracting.

Usually, the relevance of an IV increases as the IV nears the unit of analysis (e.g., firm, dyad, individual). Table 1 illustrates different IVs in MS and IOR research, and they typically become less relevant the further removed they are from the unit of analysis.

4.2 Finding exogenous IVs

The exogeneity condition consists of three different parts, as illustrated in Figure 4. The first part is that no omitted variables (OVs) affect both the IV and the dependent variable. The second part is that no omitted mechanisms (OM1s) transmit an effect of the dependent variable on the IV, i.e., there is no reverse causality. The third part of the exogeneity condition is that no omitted mechanisms (OM2s) should lead to a direct effect of the IV on the dependent variable.

A step toward ensuring that the first two parts of the exogeneity condition hold is to look for IVs among the variables determined outside our model or the unit of analysis. Heckman (2000) calls variables not set or caused by the variables in the model "external"; they should reduce the pool of omitted variables and are less likely to suffer from reverse causality (OM1s).

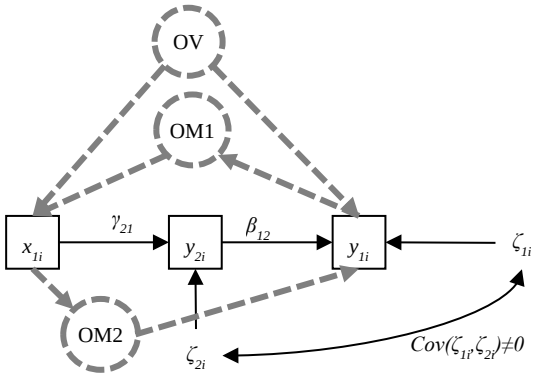


Figure 4: The three parts of the exogeneity condition.

Table 1 illustrates how the potential IVs are

Type of IV	Examples
Variables clearly determined outside the unit of analysis	▪ Physical environment, e.g., weather data ▪ Biological data, e.g., genetic markers ▪ Variables describing the institutional environment, such as laws and regulations
Variables describing phenomena outside the unit of analysis that may still be affected by the unit of analysis	▪ Variables describing immediate organizational environment, e.g., competitors and network ▪ Variables describing higher-level organizational, e.g., corporate policies or structures
Variables describing the unit of analysis	▪ Sticky variables that change less frequently than the endogenous variables, e.g., the transaction size as an IV for governance choice ▪ Lagged values of variables closely related to the endogenous variable ▪ Lagged values of the endogenous variable, industry averages of the endogenous variable

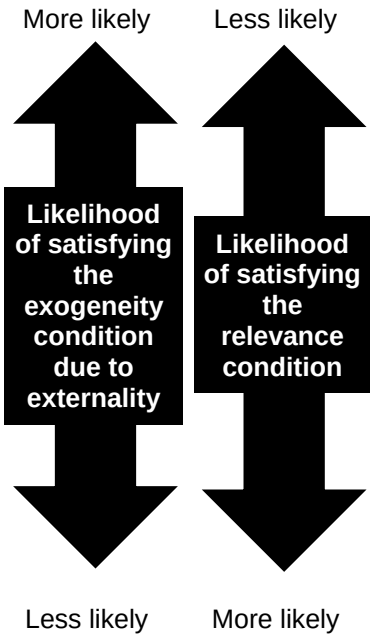


Table 1: Potential IVs in MS and IOR research.

more likely to satisfy the exogeneity condition if they describe an attribute of the environment external to the unit of analysis rather than the unit of analysis itself. For example, the external market environment of a firm is often unlikely to be affected by an individual firm or its relationships with other firms. Labor economists have drawn IVs from the institutional environment surrounding decision makers (e.g., policies applying only to people with certain birthdates). Data from the natural environment (e.g., weather data) or biological data about managers (e.g., genetic markers) could also be used as IVs (Antonakis, et al., 2010). In general, when using survey data, it would be beneficial to combine these data with secondary data because IVs drawn from secondary data will not suffer from endogeneity due to common method bias.

If adequate IVs outside the focal unit of analysis are either difficult to find or irrelevant, we may turn to variables that describe phenomena closer to our unit of analysis and even those that are part of it but that can still be considered somewhat external to our model. At least two

approaches are possible. First, if the interest lies in lower-level organizational phenomena, we could use variables that tap into higher-level organizational phenomena as IVs. This is because higher-level organizational variables not only are stickier but are also likely to represent decisions made based on a range of factors that may not be directly related to outcomes at a particular individual, dyad, team, or business unit level. For example, in Web Appendix B (p. 11), we argue that headquarters' influence over purchasing is external to the unit of analysis (dyad) because it is defined at a higher organizational level (the firm level). Few omitted variables (OVs) are likely to affect both formal contracting and headquarters' influence, and formal contracting in a given supplier relationship is unlikely to affect the purchasing function (i.e., there are no OM1s).

Second, we can exploit the timing of events and the stickiness of certain variables. For instance, some decision variables are stickier and change much more slowly than the endogenous explanatory variable. Then, decision makers are likely to choose the level of the endogenous explanatory variable with respect to the stickier decision variable—not the other way around. For example, in Web Appendix B (p. 10), we describe how annual purchasing value and relationship complexity should satisfy the exogeneity conditions—even though they are both attributes of the dyad—in part because they are stickier than formal contracting. A related approach that we could have used (if we had the data) is to use as IVs lagged values of variables closely related to formal contracting (e.g., relationship complexity) or the lagged value of formal contracting itself (e.g., formal contracting a few years earlier). However, compared with truly external IVs, these types of IVs do have a higher risk of endogeneity due to OM1s.

Unfortunately, although external variables are more likely to be exogenous, externality is no guarantee that omitted variables (OVs) do not affect both the IV and the dependent variable. To prevent this possibility, we may identify potential omitted variables related to the dependent variable and then evaluate the likelihood of these omitted variables affecting the otherwise

external IVs. In Web Appendix B (p. 10), we use this approach to evaluate theoretically whether possibly omitted variables likely affect both the IVs and the dependent variable, cost reduction.

Unfortunately, even if we have strong arguments that no omitted variables affect both the IV and the dependent variable, externality still cannot guarantee exogeneity because the third part of the exogeneity condition may be violated, i.e., that the IV has direct effects on the dependent variable (i.e., OM2). Even "perfect instruments," such as lottery numbers and randomization in field experiments, may not satisfy this assumption (Deaton, 2010). Therefore, we must develop arguments based on theory and contextual insights to justify this assumption. Two types of arguments are possible. One approach is to identify the possible consequences of the IV, other than the endogenous explanatory variable, and to then evaluate the likelihood that these other consequences affect the dependent variable. Alternatively, we identify possible antecedents of the dependent variable, other than our endogenous explanatory variable, and evaluate the likelihood that the IV affects these variables. In both cases, we identify potential omitted mechanisms (OM2s) linking the IV and the dependent variable. In Web Appendix B (p. 10), we use these approaches to evaluate the likelihood that problem-solving processes function as a mechanism (OM2) between relationship complexity and cost reduction.

Three warnings are warranted at this point. First, IVs based on endogenous variables, such as lagged endogenous variables and industry averages, may be inadequate. They will only work if they predict the exogenous (and not the endogenous) variation in the endogenous explanatory variable. For example, IVs based on industry averages will not work if the industry reflects omitted variables associated with both the endogenous explanatory variable and the dependent variable. In addition, IVs based on industry averages prevent us from using industry fixed effects to account for unobserved heterogeneity across industries (Larcker and Rusticus, 2010). Second, as highlighted in Table 1, we face a fundamental trade-off when using IVs: as variables become

more external (and exogenous), they also tend to become less relevant (Stock, 2010). Third, as we will explain in section 5, researchers should strive to use IVs that rely on different theoretical mechanisms because this practice increases the likelihood of discovering a lack of exogeneity.

4.3 Re-introducing the role of control variables

Sometimes finding IVs that satisfy the exogeneity condition is difficult. Figure 4 hints at a potential solution to this problem: we can introduce control variables that measure the omitted variables or mechanisms that undermine the exogeneity of the IVs and thereby break the links between the IVs and the dependent variable, ensuring that the exogeneity condition holds. Suppose that we cannot initially provide a good rationale for $Cov(x_{1i},\zeta_1) = 0$ in Figure 4. In that case, if we can identify and include control variables that proxy for or are indicators of OV, OM1, or OM2 in the model, we can ensure exogeneity (Angrist & Pischke, 2009; Stock, 2010)[xi]. In Web Appendix B (p. 10) we use this approach to argue that that the IVs are exogenous.

However, we should be aware that parameter estimates for the control variables are not necessarily informative. If we use control variables as proxies for unmeasured variables that are thought to affect both the IV and the dependent variable, the existence of these unmeasured variables may bias the coefficients for the control variables. Thus, control variables may contribute to consistent estimates of the effects that are of primary interest to us, but their own parameter estimates are not necessarily informative (Stock, 2010).

Unfortunately, endogenous control variables (i.e., they are caused by omitted variables correlated with the dependent variable) may increase rather than decrease bias in IV estimation because they can introduce new omitted variables that link the IV and the dependent variable. To avoid this problem, we can use IVs for the endogenous control variable as well (Frölich, 2008), but doing so will add complexity and increase the number of assumptions underlying the model.

5 HOW SHOULD IVS BE EVALUATED EMPIRICALLY?

5.1 Continuous and discrete endogenous explanatory variables

Researchers should always assess the exogeneity and relevance assumptions empirically and report the results of these assessments, along with results of estimating the first-stage regressions [i.e., Equation (10)]. Several tests are available to assess the *exogeneity condition*, the most prominent being the Sargan (1958) statistic, Hansen's (1982) J-statistic, and the C-statistic (Baum, Schaffer, & Stillman, 2003; Hayashi, 2000). These tests are all are based on testing the overidentifying restrictions on the estimated model. Hence, they require that the system of equations be overidentified, meaning that we have more IVs than endogenous regressors. The Sargan and Hansen statistics test for any IV's failure to satisfy the exogeneity condition. The difference between these two tests is that the former assumes homoscedasticity while the latter accommodates heteroscedasticity. They can both be viewed as analogous to the Lagrange multiplier or score tests (Baum et al., 2003); they follow a χ^2 distribution and are thus analogous to $\Delta\chi^2$ tests in SEM (a likelihood ratio test)[xii]. A significant p-value implies that at least one of the IVs is invalid. In contrast, the C-statistic enables us to test *individual* overidentification restrictions (just like the modification indices in SEM) and requires at least two more IVs than endogenous explanatory variables. A significant p-value for the C-statistic for a given IV (or subset of IVs) means that it does not satisfy the exogeneity assumption (Baum et al., 2003).

These tests suffer from two major limitations. First, exogeneity tests assume that at least one of the IVs is truly exogenous. The tests will be biased if none of the IVs is in fact exogenous (Murray, 2006). Therefore, if none of the IVs is exogenous, the tests may erroneously conclude that all the IVs are exogenous. The risk of such erroneous conclusions is greater if the different IVs all rely on the same theoretical explanation because if this explanation is wrong in the sense that links between the IVs and the dependent variable actually exist, then empirical assessments

of the exogeneity condition may fail to detect this lack of exogeneity. An implication of this limitation is that researchers should strive to find IVs that rely on different theoretical explanations because it is then more likely that at least one of them will hold (Murray, 2006).

Second, when effects are heterogeneous and when different IVs identify different effects, there is a risk that the opposite may happen: they may reject the null hypothesis that the IVs are valid when each of the IVs are in fact valid because, at the most fundamental level, tests of overidentification restrictions simply test whether the different IVs identify the same parameter. Therefore, with heterogeneous effects, such tests are not necessarily relevant (Parente & Santos Silva, 2012), and we may still judge models to be acceptable even when the IVs do not pass exogeneity tests. A finding that the effect indeed depends on unobserved variables (e.g., using a CF estimator) may support this judgment. However, in such situations, we have no evidence of exogeneity except for theory-based arguments, and we should be particularly careful when choosing IVs and justifying their exogeneity.

For the *relevance condition* to hold, the IVs must have a strong relationship with the endogenous regressor. The most common test of instrument relevance is the first-stage F-statistic form of the Cragg-Donald Wald F-statistic proposed by Stock and colleagues (Stock, Wright, and Yogo, 2002; Stock and Yogo, 2005), which tests for bias resulting from such weak IVs. Relative bias and maximal size are two variants of this test. The former is based on assessing the bias of the IV estimator relative to the bias of OLS: the IV is considered weak if the maximum bias of the IV estimator is greater than a certain threshold (e.g., 10% of the OLS bias). The latter is more conservative and is based on how the Wald test performs: the IVs are considered weak if they lead to a rejection rate of the null hypothesis that is larger than a certain threshold (e.g., 15%) when the true rejection rate of the test is 5% (Stock, Wright, and Yogo, 2002). If the IVs are too weak, we should use Moreira's (2003) conditional likelihood ratio (CLR) estimator, which is

fully robust to weak IVs, or limited information likelihood (LIML). LIML and Fuller's (1977) modified LIML are not fully robust to weak IVs but are better than 2SLS. Stock and Yogo (2005) provide critical F-values for LIML and Fuller's LIML. An extensive econometric literature covers weak instruments. We refer to Bascle (2008), Cameron and Trivedi (2009), and Larcker and Rusticus (2010) for accessible reviews of this literature.

In Web Appendix B (p. 12), we use the tests described above to assess the IVs in Sande and Haugland's (2015) data. Web Appendix C provides an overview of relevant Stata commands.

5.2 Latent endogenous explanatory variables in SEM

Assessing the exogeneity and relevance conditions when the endogenous explanatory variable is latent is somewhat different from cases in which we have manifest variables because we must use SEM software. Regarding the exogeneity condition, the $\Delta\chi^2$-square statistic of maximum likelihood estimation in SEM is analogous to the Sargan-Hansen test in the sense that they test the overidentifying restrictions of the model. Using IVs in SEM as illustrated in Figure 3d) implies that we place restrictions on the model. Therefore, as per Anderson and Gerbing's (1988) advice for SEM, we should first estimate a confirmatory measurement model in which all constructs freely correlate with each one another. Next, we compare this saturated model with the overidentified structural IV model using a $\Delta\chi^2$ square test[xiii] (a likelihood ratio test). If the models are significantly different, one or more of the IVs do not satisfy the exogeneity condition. Analogous to the C-statistic, we should also consider the modification indices (a Lagrange multiplier test) for the paths that are restricted to zero (i.e., the paths from individual IVs in ξ_{2i} to η_{1i} in Figure 3f) to determine whether any of these parameters are significant at the 10% level. If so, the findings imply that the restriction does not hold and that the IV in question does not satisfy the exogeneity condition. Note that the $\Delta\chi^2$ square test and the modification indices are the only relevant statistics for judging overidentified structural models. Approximate fit indices, such

as the RMSEA or the CFI, are inadequate because they will not detect invalid IVs.

Regarding the relevance condition, a statistic identical to the F-statistic that tests for bias resulting from weak IVs (Stock and Yogo 2005) does not exist in SEM models. Instead, we should report a joint χ^2 square difference test of the consequences of restricting γ_{22} to zero, which would be analogous to the F-statistic. In addition, we should report the partial squared multiple correlations coefficient for the IVs (i.e., the share of variance explained by IVs). In Web Appendix B (p. 16), we demonstrate these procedures using Sande and Haugland's (2015) data.

6 HOW SHOULD YOU INTERPRET AND EVALUATE THE RESULTS?

An important consequence of endogeneity and essential heterogeneity is that researchers should carefully evaluate the meaning and relevance of the effects that they estimate. Parameter estimates from OLS and IV estimators typically provide summary treatments of effects by aggregating treatment effects across parts of the population. Before gathering data and estimating models, researchers should therefore consult the theory and consider what effects they are interested in estimating and why and for whom these effects are likely to be positive, negative, or not significant. To interpret and evaluate the results, we generally recommend researchers to compare the results obtained using different estimators and to conduct sensitivity analyses.

At this step in the research process, we have (hopefully) established that our IVs are relevant and exogenous, which means that we have a consistent estimator. We must therefore *compare IV estimates* (consistent, but not efficient) *with OLS estimates* (efficient, but perhaps not consistent). This is a different way of conducting the previously mentioned Hausman (1978) *test for endogeneity*, and this test should not be conducted until after a researcher has established that the IVs are valid. (In Web Appendix C, we provide source code for conducting the *Durbin-Wu-Hausman* version of this test in Stata.) If there is no significant difference, we should rely on the more-efficient OLS estimate and interpret it as an ATE. However, significant differences

between IV and OLS estimates suggest that we should account for endogeneity and rely on the less-efficient but consistent IV estimate. In this case, interpretation requires more care.

Given that IV estimators identify the effect for the complier subpopulation, can we generalize the IV estimate (LATE) to the whole population and interpret it as an ATE? If we have strong reason to assume that the effect of interest is homogenous, such generalizations are unproblematic as long as the IVs satisfy the assumptions of relevance and exogeneity. If the researchers have access to several IVs, they can combine all the IVs into a single 2SLS model and create a weighted average of the LATEs (Angrist & Pischke, 2009). Using several IVs reduces the sources of variation in the choice of treatment and thus increases the likelihood of arriving at an estimate that is close to the ATE (Bascle, 2008).

If we instead have a case of essential heterogeneity, the situation becomes more complex. In such cases, different IVs may identify the effect for different complier subpopulations, and if the LATEs that the IVs identify are very different, a weighted average of the LATEs can be difficult to interpret (Morgan & Winship, 2007). Researchers should then *conduct sensitivity analyses* by estimating individual LATEs for each IV and judge how sensitive the estimates are to changes in the choice of IVs. If the parameter estimates that each IV identifies are widely different, some of the IVs are not exogenous (and it should have been detected at Step 4 in Figure 1), or the effect of interest is indeed heterogeneous and the IVs identify different effects. If researchers suspect that the latter is the case, they should compare the pattern of parameter estimates across the different IVs with the pattern that they would expect to observe based on theory. In other words, researchers should carefully consider the subpopulation that each IV identifies and the effect that they would expect for each subpopulation.

Alternatively, researchers can use IVs that identify interesting complier subpopulations. This approach is useful if the IV defines a policy meant to induce individuals or firms to adopt a

particular treatment (Morgan & Winship, 2007). Basically, the specific theory that we want to test in the specific context should direct our choice of IVs. The choice of IVs forces us to be clear on what theory we are testing. For example, if we are studying the performance consequences of using formal contracts, we might use as IVs company-wide policies regarding hiring or training that motivate formal contracting among procurement professionals. Alternatively, in cooperation with a company, we could randomize invitations to training sessions in writing formal contracts, which may lead us closer to the experimental ideal. Managerially, the question then is as follows: what are the performance consequences of changing the contract among those suppliers that are induced by the policy to change the contract? The LATE will provide us with an answer to this question, but we may not be able to generalize outside the targeted complier subpopulation.

However, if theory suggests that the effect of interest is heterogeneous, why should researchers limit themselves to estimating the LATE or the ATE? The information revealed through the LATE may sometimes be trivial or of little managerial interest. Instead, heterogeneity should prompt research questions, such as: Why does the effect differ between different firms or individuals? How is the effect distributed across the population? What firms or individuals select into treatment versus no treatment? By using estimators that account for essential heterogeneity, researchers may gain more insight into these types of questions. Such techniques can also detect essential heterogeneity and thus indicate whether researchers should rely on exogeneity tests to determine whether the exogeneity condition holds. Hence, we recommend that researchers *compare IV estimates with estimates based on using CF estimators or estimators of the MTE*. In Web Appendix B (p. 19) we provide an example of the type of comparisons discussed above, comparing parameter estimates of the effect of formal contracting on cost reductions across several different estimators. Using a simplified version of De Blander's (2010) estimator (p. 29) we also examine an alternative performance variable, end-product enhancements, and find that

the effect of formal contracting on buyer's end-product enhancements depends on unobserved heterogeneity, and the results indicate positive selection: when the formal contract is less detailed and explicit than predicted by our first-stage estimates, the effect on end-product enhancements is insignificant. However, if the formal contract is more detailed and specific than predicted, the effect on end-product enhancements is significant and positive. These results suggest that firms possess private information about the effect of formal contracting on end-product enhancements and act accordingly when choosing the level of formal contracting.

Finally, if we lack IVs, or if we want an additional point of comparison, we may also turn to what are sometimes called instrument-free estimators (Park & Gupta, 2012). These estimators rely on various restrictions and properties of the data to identify the effect of interest even in the absence of observed IVs. For example, the latent instrumental variable method (Ebbes, et al., 2005) assumes that the variance in the endogenous explanatory variable can be separated into exogenous and endogenous parts and that there exists a latent discrete instrument that, along with certain distributional assumptions, enables us to identify the effect of interest. We refer to Ebbes, Wedel, and Böckenholt (2009) for a review of these techniques.

7 WHAT RESULTS SHOULD YOU REPORT?

To increase the credibility of their research, MS and IOR researchers should provide better descriptions of how they address endogeneity, and they must justify the decisions involved, from Steps 1 to 5 in Figure 1. First, proper reporting should describe possible sources of endogeneity and essential heterogeneity or, alternatively, why endogeneity is *not* a first-order concern.

Second, in case endogeneity *is* a first-order concern, researchers must justify their choice of estimators. The framework presented in Figure 1 should be helpful for this choice.

Third, researchers must theoretically justify the relevance and exogeneity of the IVs. In particular, it is important to evaluate the threats of omitted variables and mechanisms (OM1s and

OM2s) illustrated in Figure 4. In cases where the researchers only have access to a single IV or when they suspect heterogeneous effects, the theoretical justifications are particularly important because they are the only basis for trusting the empirical model and results.

Fourth, researchers should always report empirical assessments of the IVs: tests of instrument relevance in addition to overall overidentification tests, such as the Sargan test, and tests of individual restrictions, such as the C-statistic.

Finally, researchers should provide careful interpretation and evaluation of the results that they obtain. They must report tests of endogeneity based on valid IVs, and in case endogeneity is *not* a problem, they should report results obtained using standard OLS or SEM. If endogeneity *is* detected, researchers should report results obtained using IV-based estimators. If the IVs are weak, results obtained using IV-estimators that are robust to weak IVs should be reported (e.g., Moreira's CLR) rather than others (e.g., 2SLS). If essential heterogeneity is a concern, researchers should evaluate whether the IV-based estimate can be generalized to the entire population or for whom the IV estimate is valid. Proper evaluations may require the reporting of sensitivity analyses, results from the use of CF estimators, and/or results from other estimators of the MTE. If researchers lack relevant IVs, they may turn to instrument-free estimators.

8 CONCLUSIONS

In this article, we provide a broad overview of the topic of endogeneity, especially in the context of empirical research in MS and IOR research that uses observational/field data, and we provide suggestions for addressing endogeneity. Although endogeneity is of increasing concern, we feel that many MS and IOR researchers, especially those who use primary survey-based data, either believe it is a non-issue or apply it in a haphazard manner. However, this stance needs to change because the inferences that we can draw from our data and the guidance that we provide to managers critically depends on whether we have identified the true underlying effects.

Our goal is to provide a pedagogical, overarching and practical tool/guide that brings the reader up to date in terms of understanding endogeneity; to help MS and IOR researchers choose the appropriate technique; and to appropriately implement (justify and assess assumptions), interpret, and report the results obtained via their chosen technique(s). To this end, we describe why endogeneity is a problem and provide relatively non-technical explanations of some of the most important techniques available for tacking endogeneity in MS and IOR research. Although highly relevant, some of these techniques have not been used much in marketing (e.g., the CF methods), and some of them have not been used at all (MTE estimation).

In addition to developing a novel framework for understanding and tackling endogeneity, we make three important arguments with implications for future MS and IOR research. First, we emphasize the importance of theoretically justifying the relevance and exclusion restrictions. Current research in marketing does not put much effort in doing so. To this end, we provide specific advice for how to find and justify relevant and exogenous IVs. We suggest sources of relevant IVs (Table 1), and we provide a detailed explanation of what the three different parts of the exclusion restriction really mean, what it takes to avoid violating this restriction, and how control variables may help us ensure the exogeneity of the IVs.

Second, relatively few articles in marketing empirically assess whether the relevance and exogeneity conditions underlying the use of IVs hold. Not conducting such analyses increases the risk of biased parameter estimates. However, the tests available for assessing the relevance and exogeneity of the IVs suffer from several weaknesses, and we describe how researchers can mitigate those weaknesses. Ultimately, the solution to these weaknesses is to recognize that—given the limitations of the available data—all estimators rely on assumptions; to evaluate what set of assumptions is most realistic; and to compare the results from the different models. In our view, this approach to empirical MS and IOR research is more fruitful than the common practice

of relying on saturated OLS or SEM models that disregard endogeneity concerns.

Third, as opposed to many previous treatments of endogeneity in marketing, we emphasize the roles of heterogeneity and essential heterogeneity. The variables of interest in many MS and IOR studies are self-selected based on observing components of the gain. The complexity of the context typically renders it impossible for researchers to identify and measure all possible components of the gain, which leads to essential heterogeneity. Facing essential heterogeneity, MS and IOR researchers should consider what their parameter estimates mean and for whom the estimated effects are valid. We suggest several estimators that give more insight into how effects are distributed in the population and how firms and individuals make choices. However, choosing between the estimators is often difficult because they rely on different assumptions. One approach then is to present and compare results from different estimators.

Finally, we want to urge MS and IOR researchers that endogeneity is not just a pesky methodological or empirical issue; rather, endogeneity issues have a strong theoretical foundation and addressing, and tackling, endogeneity is very likely to ply valid. Said otherwise, if endogeneity is not accounted for or if the assumptions underlying the techniques that we use to tackle endogeneity are invalid, we are likely to get biased parameter estimates that give us flawed conclusions about the underlying theoretical relationships, and consequently make us provide poor and perhaps even harmful advice to practicing managers. This article aims to help MS and IOR researchers avoid these pitfalls.

9 REFERENCES

Aakvik, A., Heckman, J. J., & Vytlacil, E. J. (2005). Estimating treatment effects for discrete outcomes when responses to treatment vary: An application to Norwegian vocational rehabilitation programs. *Journal of Econometrics, 125*(1–2), 15-51.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*(3), 411-423.

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives, 15*(4), 69-85.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics - an empiricists companion*. Princeton, NJ: Princeton University Press.

Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, *21*(6), 1086-1120.

Bascle, G. (2008). Controlling for endogeneity with instrumental variables in strategic management research. *Strategic Organization*, *6*(3), 285-327.

Baum, C. F., Schaffer, M. E., & Stillman, S. (2003). Instrumental variables and GMM: Estimation and testing. *Stata Journal, 3*(1), 1-31.

Björklund, A., & Moffitt, R. (1987). The estimation of wage gains and welfare gains in self-selection models. *The Review of Economics and Statistics, 69*(1), 42-49.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Brave, S., & Walstrum, T. (2014). Estimating marginal treatment effects using parametric and semiparametric methods. *Stata Journal*, *14*(1), 191-217.

Brinch, C. N., Mogstad, M., & Wiswall, M. (2017). Beyond LATE with a Discrete Instrument. *Journal of Political Economy*, 125(4), 985–1039.

Cameron, A. C., & Trivedi, P. K. (2009). *Microeconometrics using stata*. College Station, TX: Stata Press.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica, 69*(5), 1127-1160.

Carneiro, P., Heckman, J. J., & Vytlacil, E. J. (2011). Estimating marginal returns to education. *The American Economic Review, 101*(6), 2754-2781.

Cole, D. A., Ciesla, J. A., & Steiger, J. H. (2007). The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological Methods*, *12*(4), 381-398.

Cornelissen, T., Dustmann, C., Raute, A., & Schönberg, U. (2016). From LATE to MTE: Alternative methods for the evaluation of policy interventions. *Labour Economics*, *41*, 47-60.

Deaton, A. (2010). Instruments, Randomization, and Learning about Development. *Journal of*

Economic Literature, 48(2), 424–455.

De Blander, R. (2010). A simple estimator for the correlated random coefficient model. *Economics Letters, 106*(3), 158-161.

Ebbes, P., Wedel, M., & Böckenholt, U. (2009). Frugal IV alternatives to identify the parameter for an endogenous regressor. *Journal of Applied Econometrics*, 24(3), 446–468.

Ebbes, P., Wedel, M., Böckenholt, U., & Steerneman, T. (2005). Solving and testing for regressor-error (in) dependence when no instrumental variables are available: With new evidence for the effect of education on income. *Quantitative Marketing and Economics*, 3(4), 365–392.

Frölich, M. (2008). Parametric and nonparametric regression in the presence of endogenous control variables. *International Statistical Review*, 76(2), 214–227.

Fuller, W. A. (1977). Some properties of a modification of the limited information estimator. *Econometrica*, 45(4), 939–953

Garen, J. (1984). The returns to schooling: A selectivity bias approach with a continuous choice variable. *Econometrica, 52*(5), 1199-1218.

Germann, F., Ebbes, P., & Grewal, R. (2015). The chief marketing officer matters! *Journal of Marketing, 79*(3), 1-22.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica, 12* (Supplement, July), iii-115.

Hahn, J., & Hausman, J. (2003). Weak instruments: Diagnosis and cures in empirical econometrics. *The American Economic Review, 93*(2), 118-125.

Hamilton, B. H., & Nickerson, J. A. (2003). Correcting for endogeneity in strategic management research. *Strategic Organization, 1*(1), 51-78.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica, 50*(4), pp. 1029-1054.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica, 46*(6), 1251-1271.

Hayashi, F. (2000). *Econometrics*. Princeton, N.J.: Princeton University Press.

Heckman, J. J. (1974). Shadow prices, market wages, and labor supply. *Econometrica, 42*(4), 679-694.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153-161.

Heckman, J. J., & Robb, R. J. (1985). Alternative methods for evaluating the impact of interventions. In J. J. Heckman, & B. S. Singer (Eds.), *Longitudinal analysis of labor market data* (pp. 158-246). Cambridge, UK: Cambridge University Press.

Heckman, J. J., Urzua, S., & Vytlacil, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *Review of Economics and Statistics, 88*(3), 389-432.

Heckman, J. J., & Vytlacil, E. (1998). Instrumental variables methods for the correlated random coefficient model: Estimating the average rate of return to schooling when the return is correlated with schooling. *The Journal of Human Resources, 33*(4), 974-987.

Heckman, J. J., Schmierer, D., & Urzua, S. (2010). Testing the correlated random coefficient model. *Journal of Econometrics, 158*(2), 177-203.

Heckman, J. J., & Vytlacil, E. (2005). Structural equations, treatment effects, and econometric policy evaluation. *Econometrica, 73*(3), 669-738.

Heckman, J. L. (2000). Causal parameters and policy analysis in economics: A twentieth century retrospective. *The Quarterly Journal of Economics, 115*(1), 45-97.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica, 62*(2), 467-475.

Larcker, D. F., & Rusticus, T. O. (2010). On the use of instrumental variables in accounting research. *Journal of Accounting and Economics, 49*(3), 186-205.

Lee, L.-F. (1978). Unionism and wage rates: A simultaneous equations model with qualitative and limited dependent variables. *International Economic Review, 19*(2), 415-433.

Luan, Y. J., & Sudhir, K. (2010). Forecasting marketing-mix responsiveness for new products. *Journal of Marketing Research, 47*(3), 444-457.

Maddala, G. S. (1983). *Limited dependent and qualitative variables in econometrics*. Cambridge, MA.: Cambridge University Press.

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica, 71*(4), 1027–1048.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference*. New York: Cambridge University Press.

Murray, M. P. (2006). Avoiding invalid instruments and coping with weak instruments. *Journal of Economic Perspectives, 20*(4), 111-132.

Muthén, B., & Jöreskog, K. G. (1983). Selectivity problems in quasi-experimental studies. *Evaluation Review, 7*(2), 139-174.

Nelson, C. R., & Startz, R. (1990). The distribution of the instrumental variables estimator and its t-ratio when the instrument is a poor one. *Journal of Business, 63*(1), 125-140.

Parente, P. M. D. C., & Santos Silva, J. M. C. (2012). A cautionary note on tests of overidentifying restrictions. *Economics Letters, 115*(2), 314-317.

Park, S., & Gupta, S. (2012). Handling endogenous regressors by joint estimation using copulas. *Marketing Science*, 31(4), 567–586.

Reiersøl, O. (1945). Confluence analysis by means of instrumental sets of variables. *Arkiv För Matematik, Astronomi Och Fysik., 32a*(4), 1-119.

Rindfleisch, A., Malter, A. J., Ganesan, S., & Moorman, C. (2008). Cross-sectional versus longitudinal survey research: Concepts, findings and guidelines. *Journal of Marketing Reasearch, 45*(3), 261-279.

Roodman, D. (2011). Fitting fully observed recursive mixed-process models with cmp. *The Stata Journal*, 11(2), 159–206

Rossi, P. E. (2014). Even the rich can make themselves poor: A critical examination of IV methods in marketing applications. *Marketing Science, 33*(5), 655-672.

Sande, J. B., & Haugland, S. A. (2015). Strategic performance effects of misaligned formal contracting: The mediating role of relational contracting. *International Journal of Research in Marketing*, 32(2), 187–194

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica, 26*(3), pp. 393-415.

Semadeni, M., Withers, M. C., & Trevis Certo, S. (2014). The perils of endogeneity and instrumental variables in strategy research: Understanding through simulations. *Strategic Management Journal, 35*(7), 1070–1079.

StataCorp. (2017). Stata structural equation modeling reference manual. In *Stata: Release 15. Statistical Software* (pp. 1 – 659). College Station, Texas: Stata Press.

Stock, J. H. (2010). Comment on Angrist and Pischke: The other transformation in econometric practice: Robust tools for inference. *Journal of Economic Perspectives, 24*(2), 83-94.

Stock, J. H., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In D. W. K. Andrews, & J. H. Stock (Eds.), *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg* (pp. 80-108). Cambridge, New York: Cambridge University Press.

Stock, J. H., Wright, J. H., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic*

Statistics, 20(4), 518-529.

Wooldridge, J. M. (1997). On two stage least squares estimation of the average treatment effect in a random coefficient model. *Economics Letters, 56*(2), 129-133.

Wooldridge, J. M. (2003). Further results on instrumental variables estimation of average treatment effects in the correlated random coefficient model. *Economics Letters, 79*(2), 185-191.

Wooldridge, J. M. (2008). Instrumental variable estimation of the average treatment effect in the correlated random coefficient model. In D. L. Millimet, J. A. Smith & E. J. Vytlacil (Eds.), *Modelling and evaluating treatment effects in econometrics* (pp. 93-116). Amsterdam: Elsevier.

Wooldridge, J. M. (2010). *Econometric analysis of cross-section and panel data* (2nd ed.). Cambridge, MA.: The MIT Press.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources, 50*(2), 420-445.

Wright, P. G. (1928). *The tariff on animal and vegetable oils*. New York: The Macmillan Company.

FOOTNOTES

[i] Sande & Hauglands (2015) provide theoretical background, variable definitions, and a complete description of the data.

[ii] *i* denotes observation *i*. Note also that we denote all exogenous variables as *x* variables and all endogenous variables as *y* variables. y_{2i} is therefore an *endogenous explanatory variable—not* an exogenous explanatory variable.

[iii] Note that this definition of an endogenous variable is slightly different from what is common in structural equation modeling (SEM), where an endogenous variable is typically any variable determined within the context of a model (Bollen 1989).

[iv] Assuming that Cov(u_i, ζ_{2i}) = 0, Cov(u_i,e_{1i})=0, and Cov(ζ_{2i},e_{1i})=0.

[v] Note that time series and panel data are not a complete panacea. Even if a fixed effects panel data estimator controls for sources of endogeneity that do not change over time, endogeneity may still pose a problem. In such cases, one may use techniques that combine panel data estimators with IVs. Stata implements such procedures through the `xtivreg` and `xtivreg2` procedures (Cameron & Travedi, 2009). We consider such procedures to be beyond the scope of this article. In addition, panel and time series data may introduce new problems not present in cross-sectional data, such as autocorrelation and autoregression, which are sometimes viewed as endogeneity problems in their own right (Semadeni, Withers, and Trevis Certo, 2014).

[vi] Note that path analysis and the IV solution to endogeneity share some common intellectual roots. In fact, Sewall Wright, a biometrician and the inventor of path analysis, was the son of economist Phillip G. Wright, who is credited with the first ever use of instrumental variables (Angrist & Krueger, 2001).

[vii] The reason is that Cov(x_{1i},y_{1i}) = β_{12}Cov(x_{1i},y_{2i}) + Cov(x_{1i},ζ_{1i}) and Cov(x_{1i},ζ_{1i}) = 0.

[viii] To see why, note that $Cov(x_{1i}, y_{1i})/Cov(x_{1i}, y_{2i}) = [\beta_{12} Cov(x_{1i}, y_{2i}) + Cov(x_{1i}, \zeta_{1i})]/Cov(x_{1i}, y_{2i}) =$ $[\beta_{12}\gamma_{21}Var(x_{1i}) + \beta_{12}Cov(x_{1i}, \zeta_{2i}) + Cov(x_{1i}, \zeta_{1i})]/[\gamma_{12} Var(x_{1i}) + Cov(x_{1i}, \zeta_{2i})]$. If both $Cov(x_{1i}, \zeta_{2i}) = 0$ and $Cov(x_{1i}, \zeta_{1i}) = 0$, then $\beta_{12} = Cov(x_{1i}, y_{1i})/Cov(x_{1i}, y_{2i}) = \beta_{12}\gamma_{21}/\gamma_{21} = [\text{the indirect effect of } x_{1i} \text{ on } y_{1i}]/[\text{the effect of } x_{1i} \text{ on } y_{2i}]$.

[ix] We can also express Equations (21)-(22) using a single equation (where $\bullet = \mathbf{x_{1i}}\hat{\boldsymbol{\gamma}}^{\mathbf{p}}_{\mathbf{21}} + \mathbf{x_{2i}}\hat{\boldsymbol{\gamma}}^{\mathbf{p}}_{\mathbf{22}}$):

$$\hat{r}_i(y_{2i}, \mathbf{x_{1i}}, \mathbf{x_{2i}}) = y_{2i}\,\phi(\bullet)/\Phi(\bullet) - (1 - y_{2i})\,\phi(\bullet)/[1 - \Phi(\bullet)]$$

[x] A more precise description of this problem in the context of binary endogenous variables might be useful. First, reformulate Equation (24)–(25) so that the potential performance outcomes are defined as $y^1_{1i} = k_1 + \mathbf{x_{1i}}\boldsymbol{\gamma}^{\mathbf{1}}_{\mathbf{11}} + U^1_{1i}$ (when y_{2i}=1) and $y^0_{1i} = k_0 + \mathbf{x_{1i}}\boldsymbol{\gamma}^0_{11} + U^0_{1i}$ (when y_{2i}=0), where $k_0 = \gamma_{10}$ and $k_1 = \gamma_{10} + \beta_{12}$ are intercepts, and $\boldsymbol{\gamma}^{\mathbf{0}}_{\mathbf{11}} = \boldsymbol{\gamma}_{\mathbf{11}}$ and $\boldsymbol{\gamma}^{\mathbf{1}}_{\mathbf{11}} = \boldsymbol{\gamma}_{\mathbf{11}} + \mathbf{d_1}$ are the slope parameters for $\mathbf{x_1}$. $U^1_{1i} = h_i y_{2i} + \zeta_{1i}$ and $U^0_{1i} = \zeta_{1i}$ are unobservables when y_{2i} = 1 (treatment state) and y_{2i} = 0 (control state), respectively. Second, note that just as Garen's (1984) model assumes the joint normality of the error terms ζ_{1i}, ζ_{2i} and h_i, the selection model (Equation (19), (24)–(25)) assumes the joint normality of ζ_{1i}, U^1_{1i}, and U^0_{1i}.

[xi] Proxy variables are antecedents to the unobserved variable, whereas indicators are caused by the unobserved variable.

[xii] The Sargan and Hansen J-tests are analogous to the Lagrange multiplier or score tests (Baum et al., 2003) and to the $\Delta\chi^2$ test in structural equation modeling (SEM). For instance, compared to the model in Figure 3c, the one in Figure 3d is overidentified (these models could easily be estimated in SEM) because the endogenous regressor y_{2i} is predicted by several IVs $\mathbf{x_{2i}}$ that are restricted to have no direct relations with the dependent variable y_{1i}. The Sargan and Hansen tests assess whether these restrictions hold as a whole, as in the case of the $\Delta\chi^2$ test in SEM. The C-statistic assesses whether they hold individually, as in the case of modification indices in SEM.

[xiii] Notice that all $\Delta\chi^2$ square tests comparing two models assume that the base model (in our case, the measurement model) is correctly specified.

WEB APPENDIX A: A CENSUS OF MARKETING JOURNAL PUBLISHING

The purpose of this web appendix is to provide an overview of our census and substantiate the claims we make in the introduction of the article concerning endogeneity concerns and method trends among marketing journals that publish survey-based MS and IOR research.

Table WA1 below lists journals that we examine, along with some key statistics on each journal. We consider the years 2010-2016, a period in which these journals, according to Google Scholar, published 8375 articles.

We rely on Google Scholar for three reasons: (1) search results vary between search engines; (2) alternative search engines often miss articles that Google Scholar finds, and in some of them it is difficult to conduct nested Boolean searches; and (3) limiting the search to a single search engine ensures a uniform method across journals. When using Harzing's Publish or Perish software (Harzing, 2016), the searches can be performed relatively efficiently.

A disadvantage of Google Scholar is that it sometimes includes too many articles in any given search, or it includes the same article twice. However, the Publish or Perish software helps weed out double registrations.

Table WA1: Overview of journals examined

Rank in AJG 2015[1]	2016 JIF w.o. self-cites[2]	2016 AIS[2]	2016 SJR index[3]	Used by UTD T100 BS RR[4]?	Journal	# of articles	% of articles that mention endogeneity			
							All articles		Only survey-based MS/IOR	
							2010-2016	Only 2016	2010-2016	Only 2016
4*	4.635	3.100	5.947	Yes	Journal of Marketing (JM)	336	24.7 %	44.2%	17.0%	37.5%
4*	3.439	3.225	6.319	Yes	Journal of Marketing Research (JMR)	517	18.8 %	16.9%	9.2%	0.0%
4*	1.877	2.391	4.261	Yes	Marketing Science (MSC)	452	30.1 %	32.7%	33.3%	33.3%
4	5.487	2.218	3.997	No	Journal of the Academy of Marketing Science (JAMS)	354	10.2 %	22.2%	7.4%	19.0%
4	1.641	1.272	1.674	No	International Journal of Research in Marketing (IJRM)	344	19.8 %	27.0%	13.1%	0.0%
3	1.250	2.042	2.299	No	Quantitative Marketing and Economics (QME)	91	30.8 %	36.4%	50.0%	N.A.
4	3.506	1.352	2.556	No	Journal of Retailing (JR)	288	17.5 %	26.8%	11.1%	0.0%
3	1.704	0.905	1.160	No	Marketing Letters (ML)	322	8.1 %	4.8%	4.9%	7.1%
3	2.125	0.704	2.332	No	Journal of International Marketing (JIM)	140	7.1 %	25.0%	5.8%	16.7%
3[5]	2.322	0.628	1.815	No	Journal of Business Research (JBR)	2506	4.4 %	4.9%	3.0%	4.5%
3	2.096	0.645	1.830	No	Industrial Marketing Management (IMM)	985	0.9 %	0.7%	2.3%	2.3%
3	1.443	0.532	0.933	No	International Marketing Review (IMR)	221	5.9 %	13.9%	12.8%	29.4%
3	1.172	0.436	1.003	No	European Journal of Marketing (EJM)	655	3.1 %	7.5%	2.0%	8.3%
3	(N.A.)	(N.A.)	0.843	No	Journal of Marketing Management (JMM)	591	0.7 %	1.2%	0.6%	0.0%
2	0.909	0.262	0.828	No	Journal of Business and Industrial marketing (JBIM)	445	0.9 %	2.4%	1.1%	2.3%
2	0.875	0.210	0.792	No	Journal of Business-to-Business Marketing (JBBM).	128	0.0 %	0.0%	0.0%	0.0%

[1] Chartered Association of Business Schools (CABS) Academic Journal Guide (AJG) 2015, https://charteredabs.org/academic-journal-guide-2015/

[2] Journal Impact Factor (JIF) and Article Influence Score (AIS), which are based on the Thomson Reuters database. Available from Thompson Reuters' Journal Citation Reports: https://jcr.incites.thomsonreuters.com/

[3] Scimago Journal Rank (SJR), which is based on the Scopus database: http://www.scimagojr.com

[4] The UTD Top 100 Business School Research Rankings™ (UTD T100 BSRR): http://jindal.utdallas.edu/the-utd-top-100-business-school-research-rankings/index.php

[5] Journal of Business Research is not listed in the AJG 2015 as a marketing journal. However, this journal publishes many marketing articles and has separate associate editors on buyer behavior, marketing, business-to-business research, advertising and marketing communication, service research, sales research, and retailing.

Methods and analysis

As part of the census, we identify different kinds of articles as follows:

Survey-based MS and IOR articles:	Articles that mention the terms "survey", "questionnaire", "factor analysis", or "structural equation modeling" and NOT "consumer behavior" (2482 articles).
Non-survey and/or non-MS/IOR-articles	Articles not belonging to the category "survey-based MS or IOR articles," such as editorials, conceptual papers, qualitative papers, quantitative papers based on secondary data, and survey-research and experimental research dealing with consumer behavior (5893 articles)
Articles that mention "endogeneity"	Articles that mention the word "endogeneity" (115 articles among survey-based MS and IOR articles, and 580 articles among non-survey non-MS/IOR articles)
Articles that theoretically justify or empirically evaluate instrumental variables (IVs)	Articles that mention the words "exogeneity," "Hansen J test," "Hansen test," "Sargan" OR "ivreg2," "instrument relevance," "Cragg-Donald F-test," or "weak instrument" (13 articles among survey-based MS and IOR articles, and 137 articles among non-survey non-MS/IOR articles)

In addition, we choose to split the journals into three different groups:

Group 1: JM, JMR, MSC;
Group 2: JAMS, IJRM, QME and JR; and
Group 3: ML, JIM, JBR, IMM, IMR, EJM, JMM, JBIM, and JBBM.

In 2016, these groups can be characterized as follows:

Group 1: Used by The UTD Top 100 Business School Research Rankings™ and rated at level 4* on the CABS Academic Journal Guide in 2015.
Group 2: Article Influence Score higher than 1 and rated at level 4 on the CABS Academic Journal Guide in 2015.
Group 3: Article Influence score lower than 1, and they are all ranked at level 3 or lower on the CABS Academic Journal Guide in 2015.

Assessing the overall trend in endogeneity

Figure WA1 below shows that, overall, there is a positive trend of increased endogeniety concern among all the marketing journals we examined.

However, the frequency with which endogeneity is mentioned among survey-based MS and IOR articles is lower than that among other types of articles. This is somewhat surprising, given that "all other articles" includes articles that we normally do not expect to address endogeneity, such as editorials, experimental studies, theory development and qualitative studies.

These findings supports two claims in the introduction of the article:

- *"the awareness of these problems* [i.e., endogeneity problems] *among MS and IOR researchers has substantially increased" and*
- *"the application of endogeneity-correcting techniques within MS and IOR research remains infrequent compared to other types of articles, and there appears to be confusion about how to address endogeneity-related problems."*

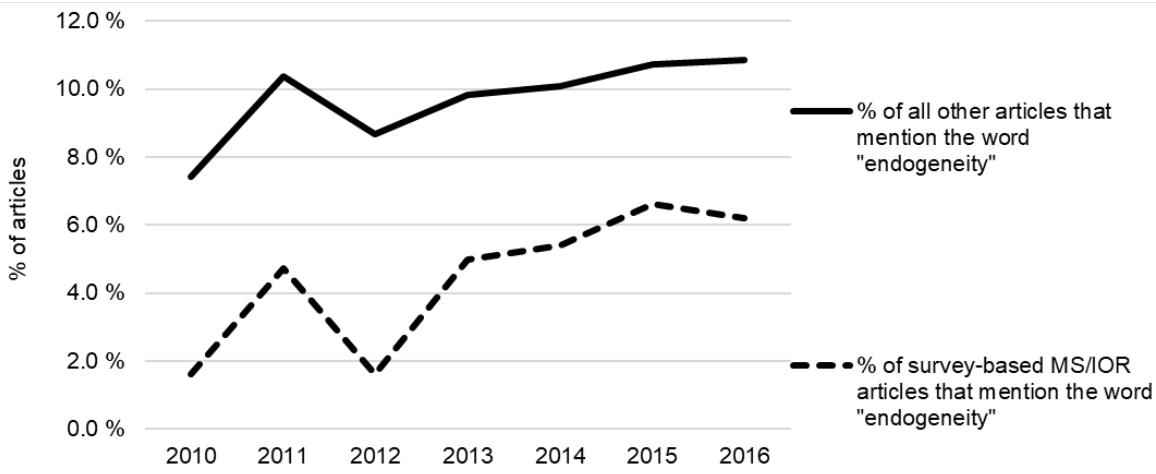


Figure WA1: Share of articles showing concern with endogeneity.

Exploring the heterogeneity between journals in endogeneity concern

The statistics in Table WA1 indicate that there are differences between journals. In the following, we explore how the trends in endogeneity concern differ between journals.

Figure WA2 below shows that among survey-based IOR-studies, there are clear differences between the three groups. Whereas 16% of the survey-based MS and IOR studies published in Group 1 journals mention endogeneity, only 10% of the Group 2 journals and 2.8% of the Group 3 journals mention endogeniety.

However, the share of articles that mention endogeneity has generally increased among all groups of journals (despite some ups and downs). *Our claim that "the awareness of these problems* [i.e., endogeneity problems] *among MS and IOR researchers has substantially increased" is therefore supported in each of these groups when only considering survey-based MS and IOR articles.*

Figure WA3 below illustrates the same trends for the same journals but for all other kinds of articles, i.e., non-survey and/or non-MS/IOR articles. As evident here as well, we see that the frequency with which the word endogeneity is mentioned varies by journal group: 26%, 19%, and 3.5% of the non-survey and/or non-MS/IOR articles published in Group 1, 2, and 3 journals, respectively, mention endogeneity

However, the trends are generally upwards in all three groups of journals. *Our claim that "the*

awareness of these problems [i.e., endogeneity problems] *among MS and IOR researchers has substantially increased" is therefore supported in each of these groups of journals when only considering non-survey and/or non-MS/IOR articles.*

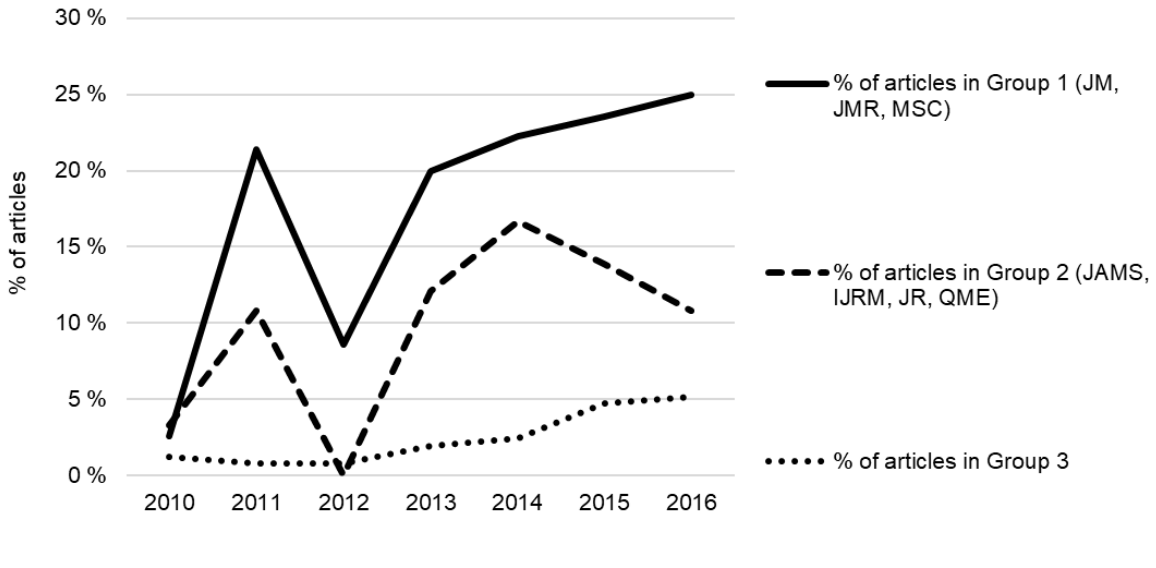


Figure WA2: Share of survey-based MS and IOR articles showing concern with endogeneity, across journal groups

Figures WA2 and WA3 provide further support for the conclusion drawn based on Figure WA1 that survey-based MS and IOR articles are less concerned with endogeniety than other types of articles. For each of the three categories of journals, the share of articles that mention endogeneity is lower among survey-based MS and IOR articles than among other types of articles: 16% vs 26 in Group 1, 10% vs 19% in Group 2, and 2.8% vs 3.5% in Group 3. *These findings support our claim that "the application of endogeneity-correcting techniques within MS and IOR research remains infrequent compared to other types of articles, and there appears to be confusion about how to address endogeneity-related problems" for each of the three groups of journals.*

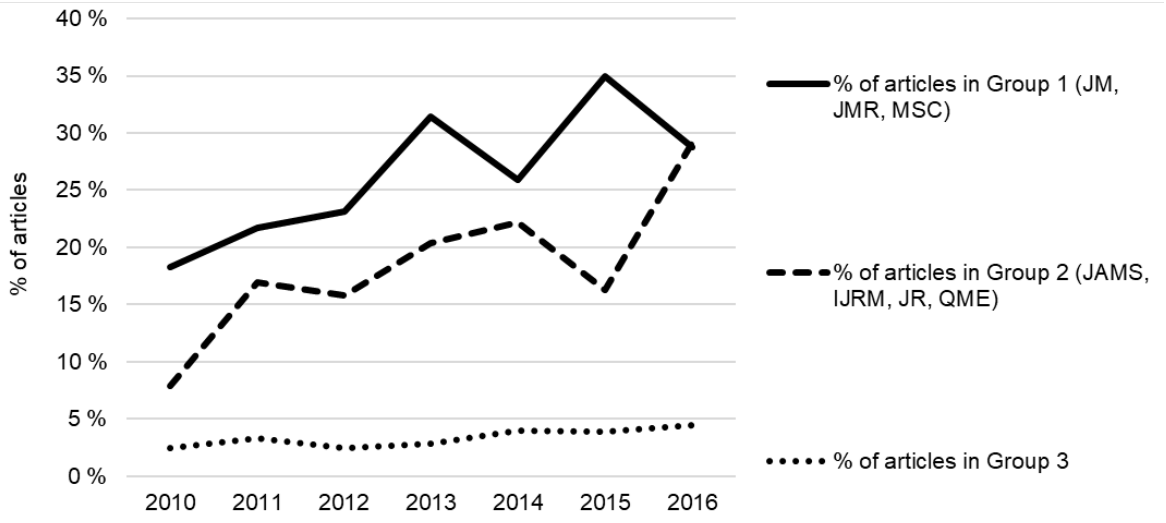


Figure WA3: Share of other non-survey and/or non-MS/IOR articles showing concern with endogeniety, across jornal groups

Do authors theoretically justify and empirically evaluate IVs?

To obtain a better impression of method practice in marketing, we also perform a search aimed at revealing the extent to which authors attempt to theoretically justify and empirically evaluate IVs. Figure WA4 illustrates the trends in terms of how frequently articles mention various keywords, indicating that such justifications and evaluations have been performed among those that also mention endogenity (see page 2 of this web appendix for the keywords).

As evident, the share of articles that mention the key-words that indicate attempts to justify and/or evaluate the IVs is overall quite low, particularly among survey-based MS and IOR articles. On average, among survey-based MS and IOR articles, only 11% mention one or more of the key-words we are looking for among those that already show concern for endogeneity. Among non-survey and/or non-MS/IOR articles, this number is higher: 21%.

This finding provides further support for our claim that "the application of endogeneity-correcting techniques within MS and IOR research remains infrequent compared to other types of articles, and there appears to be confusion about how to address endogeneity-related problems".

In addition, Figure WA4 specifically supports our claim in the article that "there appears to be confusion about how to address endogeneity-related problems" for two reasons. First, if no researchers performing survey-based MS and IOR research and being concerned with endogeneity were confused about how to address endogeneity, we should not have observed differences between survey-based MS and IOR articles and other types of articles in Figure WA4. Second, if there was no confusion about how to address endogeniety, the numbers in Figure WA4 would probably have been higher, particularly for survey-based MS and IOR research because IV-based methods are often the only way to address endogeneity in such research.

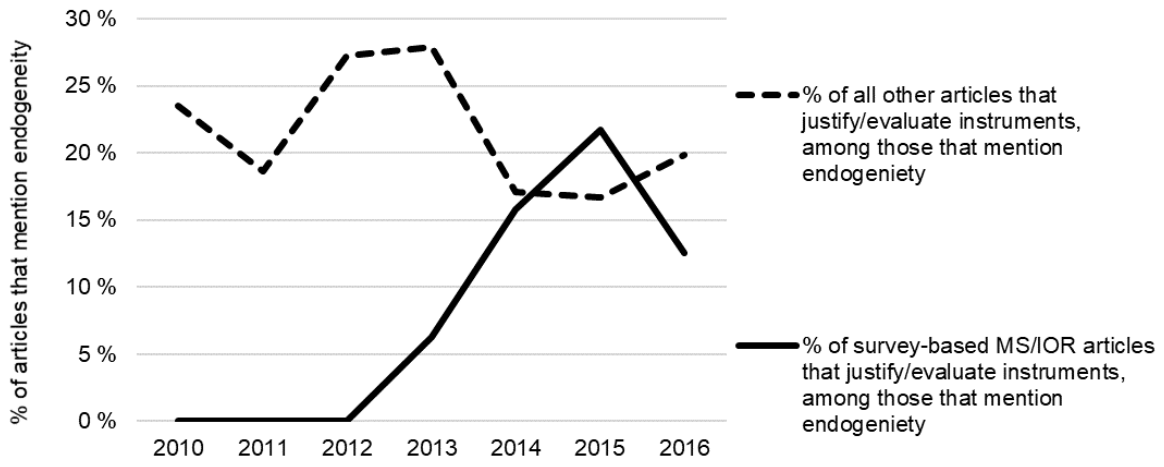


Figure WA4: Share of articles that seem to theoretically justify and/or empirically evaluate IVs, among those articles that mention endogeneity

Decline in the share of survey-based MS and IOR articles in top journals

WA5 illustrates the trends in the share of survey-based MS and IOR articles for each of the three groups of journals. As is evident, the share of survey-based MS and IOR research remained more or less stable from 2010 to 2014 for Groups 1 and 2, but in 2014, this share started to decline. Whereas survey-based MS and IOR articles accounted for approximately 20% of the articles in Groups 1 and 2 from 2010 through 2014, in 2015 and 2016, these shares dropped to 16% and 14%, respectively. In absolute numbers, survey-based MS and IOR articles in Groups 1 and 2 declined from a total of 69 and 79 articles in 2010 and 2011, respectively, to 53 articles in both 2015 and 2016.

This drop accounts for much of the relative increase we have seen in the share of articles in Groups 1 and 2 that mention endogeneity, because this number has remained fairly stable and low throughout the period, i.e., between 2 (in 2010) and 12 (in 2014) articles per year, as shown in the gray area in Figure WA5.

From Figure WA5, it is also worth noting that the share of survey-based MS and IOR articles increases as we go from Group 1 via Group 2 to Group 3, which suggests that the top journals are more sceptical towards such research.

These findings thus provide support to our claims in the article that a "A possible consequence is that articles in these domains get rejected in our premier journals because the researchers have either not addressed or have inadequately addressed endogeneity concerns."

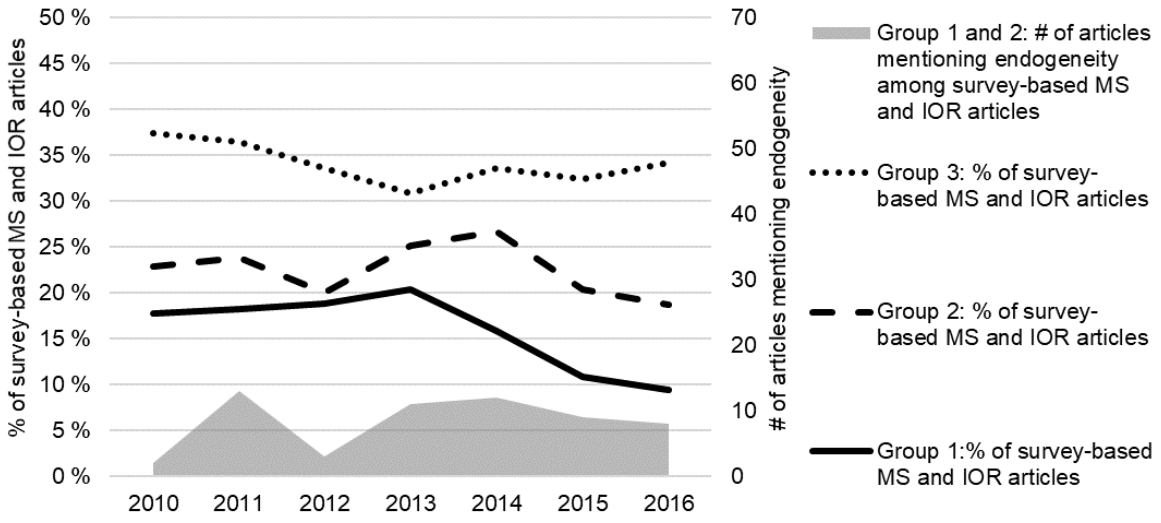


Figure WA5: Trends in the share of survey-based MS and IOR articles across journal groups

WEB APPENDIX B: APPLYING AND DEMONSTRATING THE FRAMEWORK

The purpose of this web appendix is to apply and demonstrate the proposed framework described in Figure 1 as well as some of the estimators described in the article.

To apply the framework, we will utilize the data from a recently published IJRM article by Sande & Haugland (2015) on formal contracting. We use this article for two reasons:

(1) we can refer to this article for most issues except the analyses performed here (e.g., theory, definitions and explanations of variables, explanations of measures, context description), and
(2) the authors have made the data and their source code available for anyone to download and use at http://www.runmycode.org/doidata/10.1016/j.ijresmar.2015.02.002.

Given that the purpose of this web appendix is to demonstrate and apply the framework and some of the estimators in our article, *our goal for the analyses in this web appendix is to estimate the effect of formal contracting on cost reductions and to interpret this parameter estimate. The substantive question of interest is therefore whether using more detailed formal contracts in a particular buyer-supplier relationship leads to cost reductions for the buyer.*

In addition, towards the end of this web appendix, *we compare the results of using De Blander's (2010) estimator for the effect of formal contracting on cost reductions with the effect of formal contracting on end-product enhancements* to demonstrate the role of heterogeneity.

We choose not to propose any particular hypothesis concerning the effect of formal contracting in this web appendix, beyond noting that using more detailed and explicit contracts can both increase costs (due to higher costs of writing contracts) and reduce costs (due to higher ex post costs of, e.g., renegotiations, errors, lower quality). We refer to Mooi & Ghosh (2010) for a discussion and specific empirical results concerning these effects.

(Note that in contrast to Sande & Haugland, we are not interested in estimating the effects of *misaligned* formal contracting, and we are not interested in the role of relational contracting. Sande & Haugland build on research in the strategy literature that measures misalignment as the deviation between observed and predicted levels of formal contracting as well as econometric literature on the correlated random coefficient model to give an interpretation of the parameter estimate for the absolute value of the first-stage residual.)

Throughout this web appendix, we will refer to a Stata do-file (code-for_web-appendix_B.do) accompanying this web appendix, where we include all the Stata code for conducting the analyses here. Readers will therefore be able to replicate our analyses after downloading the dataset from the above link. In addition, in this web appendix, we will include several tables of results, but not all. Therefore, readers who want the full details from the analyses should download the data and use the do-file to generate the results.

Here is a table of contents for the rest of this document:

Preparations

Before applying the framework, we need to prepare the data, and we want to conduct a simple OLS regression to see what ordinary least squares tell us about the association between formal contracting and cost reduction outcomes while controlling for observed variables.

Preparing the dataset

In this web appendix, we will not provide edited tables of results. Instead, we will include output as it is displayed in Stata. To make the interpretation of the output easy, Box WB1 below provides an overview of the variables and their names (see also section Preparing the data: A in the do-file). Most of the variables used in the study are measured using multiple-item Likert scales (indicated by (L) in Box WB1 below), while others are based on objective descriptions (e.g., annual purchasing value).

Box WB1: Overview of variables and their names

```
* We use the following names for each of the variables in our syntax:

* cro        = Cost reductions (L)
* eeo        = End-product enhancements(L)
* formrole   = Detailed role specification (dimension of formal contracting) (L)
* formadp    = Detailed contingency planning (dimension of formal contracting) (L)
* formcon    = Formal contracting (L)
* relnorm    = Relational contracting (L)
* bsa        = Buyer specific assets (L)
* ssa        = Supplier specific assets (L)
* unc        = Environmental uncertainty (L)
* perfamb    = Performance ambiguity (L)
* complex    = Relationship complexity (L)
* hqinflu    = Headquarters influence (L)
* knsim      = Knowledge similarity (L)
* lnval      = Natural logarithm of annual purchasing value
* lnempl     = Natural logarithm of number of employees (firm size)
* lnintproc  = Natural logarithm of 1 + share of internal procurement
* bexp       = Purchasing manager sales and marketing experience (L)
* sexp       = Supplier representative sales and marketing experience (L)
* c_process  = Dummy for processing firms
* c_trade    = Dummy for reselling firms
* c_construc = Dummy for construction firms
```

Note from the do-file (see section Preparing the data: C) that we also create and use quadratic and interaction terms for buyer and supplier asset specificity (i.e., `bsasq=bsa*bsa`, `ssasq=ssa*ssa`, and `bsassa=bsa*ssa`).

Second, to use most of the techniques described in this article, we must use scores for each of the latent variables. We follow Sande & Haugland in using loading-weighted mean scores for each latent variable. Sande & Haugland also considered the use of simple mean scores. However, when weighting the items on their loadings in the confirmatory factor model, they observed that the correlation matrix between the loading-weighted scores is more similar to the correlation matrix from the confirmatory factor model than the correlation matrix obtained from simple mean scores. We present the code for generating the loading-weighted scores in the section Preparing the data, B and C in the do-file.

Estimating OLS regression

We estimate an OLS regression (see section Preparing the data: D in the do-file) to obtain a parameter estimate for the association between formal contracting and cost reduction after controlling for all other observed variables. As evident from Box WB2 below, formal contracting

(formcon) has a relatively small (b=0.07) and weakly significant (p=0.088) relationship with cost reductions.

If we interpret this result as reflecting an effect of formal contracting on cost reductions, we would conclude that formal contracting has little consequence for realizing cost reductions. This conclusion rests on the assumption that there is no endogeneity, and if there is endogeneity, the control variables effectively control for it.

We now turn to applying the framework, and the first question we ask is: Do we have an endogeneity problem?

Box WB2: Stata output from estimating OLS regression

```
      Source |       SS           df       MS      Number of obs   =       305
-------------+----------------------------------   F(16, 288)      =     10.95
       Model |  175.180329        16  10.9487706   Prob > F        =    0.0000
    Residual |   287.99203       288  .999972326   R-squared       =    0.3782
-------------+----------------------------------   Adj R-squared   =    0.3437
       Total |  463.172359       304  1.52359329   Root MSE        =    .99999

------------------------------------------------------------------------------
         cro |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     formcon |    .069488    .040645     1.71   0.088    -.0105109    .1494869
      md_bsa |   .2209995   .0735364     3.01   0.003     .0762626    .3657365
      md_ssa |   .3115533    .068111     4.57   0.000     .1774949    .4456117
       bsasq |  -.0164646   .0371751    -0.44   0.658     -.089634    .0567047
       ssasq |   .0663493   .0402627     1.65   0.100    -.0128972    .1455959
      bsassa |  -.0672726   .0527343    -1.28   0.203     -.171066    .0365209
         unc |   .1630084   .0555275     2.94   0.004     .0537173    .2722995
     perfamb |  -.1352719   .0459749    -2.94   0.004    -.2257614   -.0447825
      lnempl |  -.1504017    .049172    -3.06   0.002    -.2471837   -.0536197
       knsim |   .1008271   .0488181     2.07   0.040     .0047416    .1969125
   lnintproc |  -.3853949   .1196454    -3.22   0.001    -.6208852   -.1499046
  c_construc |  -.1374284   .1870815    -0.73   0.463    -.5056488     .230792
     c_trade |   .0935846   .1933125     0.48   0.629    -.2868998    .4740691
   c_process |   .1570329   .1977344     0.79   0.428    -.2321548    .5462206
        bexp |   -.049268    .042597    -1.16   0.248     -.133109     .034573
        sexp |   .1660323   .0476259     3.49   0.001     .0722932    .2597713
       _cons |    3.00335   .3926294     7.65   0.000     2.230563    3.776137
```

Step 1: Do you have an endogeneity problem?

The framework in the article suggests that we should consider four questions to assess theoretically if we might have an endogeneity problem. Note that in evaluating the possibility of omitted variables, simultaneity and measurement errors, we also speculate about the possible direction of bias. We consider each of these questions below:

What kind of data do you have?	Sande & Haugland's data are cross-sectional. Hence, we cannot, for example, utilize a fixed effects panel data model to reduce or eliminate endogeneity problems.
Does theory or contextual insight	The explanatory variable of interest is formal contracting, which is measured using a set of Likert scale items describing "the detail with which explicit contract terms specify the agreement and

suggest the possibility of omitted variables?

formalize the parties' roles and contingency plans" (Sande & Haugland, 2015, p.188). The level of formal contracting in a particular buyer-supplier relationship can in principle be determined by many factors, including the variables we observe in Sande & Haugland's dataset.

However, Sande & Haugland lack a number of potential variables that might also be taken into account when firms decide on the level of formal contracting, such as supplier's reputation, buyer's previous experience in using formal contracts, personalities, knowledge hazards, and technological uncertainty. Indeed, Sande & Haugland's web appendix shows that the independent variables explain only 43.3% of the variance in formal contracting. In other words, 57.7% of the variance in formal contracting must be explained by variables unobserved by us.

Some of these unobserved variables might in turn be related to cost reduction outcomes, thereby representing a threat of creating omitted variable bias if not accounted for. For example, a supplier's or buyer's bad reputation or the presence of knowledge appropriation hazards might have direct negative effects on cost reductions. If these omitted variables also motivate the parties to use more detailed formal contracts, their omission might lead to the parameter estimate in Box WB2 being *downward biased*. Other variables, such as experience in writing contracts, might bias this parameter estimate *upwards*, for example, if experience in writing contracts increases formal contracting while having a direct positive relationship with cost reduction.

In sum, we are likely to have omitted variable bias. The direction of the potential bias in Box WB2 is uncertain.

Does theory or context insight suggest the possibility of simultaneity?

Firms may sometimes respond to poor realized levels of cost reductions by increasing or decreasing the level of formal contracting. Hence, formal contracting and cost-reductions may affect each other. Unfortunately, our cross-sectional data give us only a snapshot of this process and require us to make the assumption of equilibrium, which means that the effects have already taken place and that the system is in a steady state. Therefore, we must also assume that the variables affect each other simultaneously.

Simultaneity might therefore bias the parameter estimate we are interested in. For example, if firms in general respond to poor cost reduction outcomes by increasing the level of detail in the contracts, it means that cost reduction outcomes have a negative effect on formal contracting. If we do not account for this reverse

negative effect, the OLS estimate in Box WB2 will be *biased downwards*.

Does theory or contextual insight suggest the possibility of measurement errors?

Sande & Haugland rely for the most part on variables measured by multiple-item Likert scales. The use of multiple measures reduces the degree of measurement errors and should reduce the degree of attenuation of the correlations between the variables that result from measurement errors. However, because Likert scales are perceptual, they are vulnerable to common method variance, namely, that part of the variance in the measures can be attributed to the method used to measure the variables. Common method bias is essentially a form of omitted variables bias in which unobserved measurement errors correlate across the different latent variables, as illustrated in Figure 2 in the article.

There are two reasons to believe that common method bias is not a major concern in Sande and Haugland's data. First, most of the variables concern concrete topics, such as asset specificity, head-quarter influence, relationship complexity, formal contracting, and cost reductions. Psychological biases, such as social desirability bias, mood states, and implicit theories, should be less likely to affect responses on such variables than, for example, less-concrete variables such norms, trust, or personality variables. Second, Sande & Haugland estimate a measurement model in which they add an additional latent variable that is allowed to affect all perceptual scale items, and this variable explains only 3.2% of the variance in the perceptual scale items.

However, the test used by Sande & Haugland is a fairly weak test, because it lumps all possible sources of common bias into one single factor of common variance. There is a wide variety of sources of common method variance, and they may not affect all the variables equally. It is therefore possible that if they had had measures of, for example, social desirability response, it would have explained more of the variance in the perceptual scale items than 3.2%. Therefore, although common method variance does not seem to be of major concern here, we should not completely rule out the possibility that the parameter estimate in Box WB2 could be biased by common method variance as well. If there is a common method bias, it is difficult to speculate about the direction of the bias, both because several of the control variables will be affected by the same biases and because there are many different sources of common method variance.

Based on the discussion above, our conclusion must be that we may have endogeneity problems. It is also difficult to say precisely in what direction the bias might be, although most of the arguments in the previous paragraphs indicate that it might be biased downwards. Given that we do not have panel data, it is impossible to use, for example, a fixed effects estimator, which could

have handled some types of omitted variables, simultaneity, and some sources of common method variance. We must therefore conclude that endogeneity is a *first-order problem,* and we must rely on methods based on using instrumental variables (IVs) to both handle and test for endogeneity.

At Stage 1, our framework suggests that two additional questions should be asked, and we consider them below.

What is the nature of the explanatory variable?	Sande & Haugland operationalize formal contracting as a second-order variable consisting of two dimensions, detailed role specification and detailed contingency planning, each of which are measured using multiple-item Likert scales. Hence, these variables can be used in two ways: ▪ Create a composite score ▪ Model formal contracting as a latent variable within a structural equation modeling (SEM) model For the purpose of this web appendix we will do both and will thereby demonstrate the estimators in the article for both types of variables.
Is essential heterogeneity a concern?	As described above, when considering the possibility of omitted variable bias, formal contracting is self-selected. In addition, it is reasonable to assume that the effect of formal contracting on cost reductions is heterogeneous and varies from relationship to relationship. For example, firms with little experience in writing and using detailed formal contracts are likely to face higher costs in writing and using detailed formal contracts, and if they choose to take on these costs, their lack of experience is also likely to reduce the effectiveness of using detailed formal contracts to reduce other types of costs as well. For example, detailed but poorly designed formal contracts are unlikely to be helpful. As a consequence of unobserved, relationship-specific heterogeneous effects of formal contracting, firms will sort on the gains from formal contracting, and we are likely to have the problem of essential heterogeneity.

Step 2: What technique/estimator is appropriate?

Given that formal contracting and several other variables in the dataset are measured using multiple-item Likert scales, using a model based on SEM that treats these variables as latent variables seems appropriate. We will therefore present results based on using SEM and IVs in SEM.

However, given that essential heterogeneity is a concern, it would also be useful to try techniques that more directly address how the effect of formal contracting could be heterogeneous and that actors sort on the gains from formal contracting. Hence, we should also try techniques like those developed by Garen (1984) and De Blander (2010).

In this web appendix, we also use the following techniques: (1) 2S2L and (2) instrumental variable estimation using Moreira's CLR, LIML and Fuller's LIML. These techniques enable us to make better assessments of instrument relevance compared to when using SEM, and they enable us to better account for weak IVs.

The results from using these techniques will be presented in Step 5, which concerns how to interpret the results.

Step 3: What IVs should be chosen?

Sande and Haugland provide three potential IVs: relationship complexity, headquarter influence, and annual purchasing value.

In the following, we try to theoretically justify why these variables should satisfy both the relevance and exogeneity criteria. The important point for us here is to demonstrate the *structure of the argumentation*. We do as we recommend in the article: To theoretically justify the relevance of the IVs, we describe the mechanisms linking the IV and the endogenous explanatory variable. To theoretically justify the exogeneity condition, we utilize Figure 4 in the article and differentiate among omitted variables (OVs), omitted mechanisms that lead to an effect of the dependent variable on the IV (OM1s), and omitted mechanisms that lead to a direct effect of the IV on the dependent variable (OM2s).

Because the important point is to demonstrate the structure of the argumentation, we provide fairly extensive arguments here to show how such arguments should be made. We further choose not to make numerous references to the theoretical and empirical literature on formal contracting literature here, although it is certainly possible and in other contexts desirable. However, given the purpose of this discussion, it would create unnecessary clutter.

It is also possible to disagree with our arguments. They are based on our own understanding of the theory, the context, and previous empirical results. This also means that the theoretical basis for the model is to some extent subjective, which makes empirical assessments of the relevance and exogeneity conditions important.

<u>Theoretical justification that IVs are relevant:</u>

The three variables are likely to be relevant because we can posit various types of mechanisms that should link them with formal contracting. Relationship complexity and annual purchasing value both describe attributes of the relationship between the parties (see Table 1 in the article), which means that they measure variables at the same unit of analysis as formal contracting, which should make them particularly relevant. Headquarter influence is defined at a higher organizational level and is possibly less relevant.

In the following, we first define the three IVs and then describe various theoretical mechanisms that link the IVs with formal contracting, such that a change in the IVs should lead to a change in formal contracting.

Relationship complexity	*Definition:* Relationship complexity is the extent to which the parties constitute a system with many different parts that interact to a high degree. It reflects the number of issues or contingencies that the parties can potentially specify in a contract. *Mechanisms linking it with formal contracting:* The more complex a relationship is, the easier it is for the parties to add additional issues to the contract, which decreases the marginal costs of writing more detailed formal contracts. In other words, it is easy to write a long and detailed contract simply because there is a lot to write about. Writing more detailed formal contracts has benefits as well, when relationships are complex, because more complex relationships require more coordination and detailed formal contracts that specify the parties' roles and how they should respond to unexpected changes, might help the parties coordinate better by creating common ground. *In sum, relationship complexity should motivate more detailed formal contracts both because relationship complexity reduces the marginal costs of writing more detailed formal contracts and because it increases the marginal benefits of formal contracting in terms of lower coordination and communication costs.*
Annual purchasing value	*Definition:* Annual purchasing value is a proxy for transaction frequency and is measured as the natural logarithm of the total amount of purchases per year in Norwegian kroners (NOK). *Mechanism linking it with formal contracting:* High frequency allows firms to utilize more specialized governance structures, such as detailed formal contracts. *In other words, a high annual purchasing value decreases the marginal costs of detailed formal contracting.*
Headquarter influence	*Definition:* Headquarter influence is the extent to which a company's/chain's headquarters influences purchasing decisions in the business unit. *Mechanisms linking it with formal contracting:* Headquarter influence indicates that some employees are likely to be specialized in the purchasing function. Such specialized employees are likely to have greater experience and access to contract templates and previous contracts, which should make it

easier to add more detail the contracts. Hence, headquarter influence should reduce the costs of writing more detailed and explicit contracts.

In addition, firms with a more centralized purchasing function are more likely to use formal contracts as part of the "operating mode" and to provide a paper trail documenting their procurement, i.e., it is more likely that there is an internal norm of writing more detailed formal contracts. Purchasing managers would face personal costs of breaking this norm.

In sum, headquarter influence motivates more detailed formal contracts both because it reduces the marginal costs of writing more detailed formal contracts and because it increases purchasing managers' personal costs of breaking internal rules and norms of writing detailed formal contracts.

Theoretical justification that IVs are exogenous:

In the following, we evaluate each of the IVs in terms of the likelihood that each of the three parts of the exogeneity condition are violated. Note that the control variables are important for justifying the exogeneity of the IVs.

Relationship complexity	OV: After controlling for the observed variables in the model, it is difficult to conceive of any omitted variables that that might drive both relationship complexity and cost reductions. OM1: After controlling for the observed variables in the model, it is difficult to conceive of any omitted mechanisms through which cost reductions might affect relationship complexity. OM2: In isolation, relationship complexity might directly affect cost reductions through other mechanisms than formal contracting. For example, relationship complexity might initiate complex problem solving processes that eventually lead to higher or lower cost reductions. However, in controlling for asset specificity, environmental uncertainty, and performance ambiguity, we control for several of the potential mechanisms through which such effects might take place. For example, complex problem solving will likely be related to both relationship-specific assets and formal contracting to govern the problem solving process. We therefore do not expect relationship complexity to have a direct effect on cost reductions.
Annual purchasing value	OV: We might imagine omitted variables that drive both annual purchasing value and cost reductions, and this may be the greatest weakness of this IV. One possible example could be supplier reputation or supplier capabilities, which might both reduce costs for the buyer and motivate the buyer to purchase more from a

	given supplier. However, we control for this danger to some extent by including supplier representative sales/marketing experience and asset specificity as control variables. In addition, given that Sande & Haugland measure this variable in NOK, it is unlikely to suffer from common method variance.
	OM1: We might imagine that cost reductions directly motivate higher purchasing value. However, by controlling for asset specificity and firm size, we mitigate this source of bias. In general, annual purchasing value is stickier than formal contracting, so it is more likely that annual purchasing value affects formal contracting than the reverse.
	OM2: In isolation, annual purchasing value might have direct effects on cost reductions through other mechanisms than formal contracting. However, by controlling for asset specificity and firm size, we control for some of the most important mechanisms through which such effects might take place.
	In addition, it should be mentioned that annual purchasing value has previously been used as an IV for formal governance forms in several studies (e.g., Gulati & Nickerson, 2008; Poppo & Zenger, 2002), which lends credibility to annual purchasing value as an exogenous IV.
Headquarter influence	OV: We find it difficult to imagine what omitted variables could affect both headquarter influence and cost reductions without going through formal contracting or any of the other control variables in the model.
	OM1: Headquarter influence on procurement is an attribute of the buyer firm rather than the buyer-supplier relationship itself. As such, it is defined at a higher organizational level than the endogenous variables and can be considered *external* to the unit of analysis. According to Table 1 in the article, this should increase the likelihood that this variable is exogenous. It is unlikely that cost reductions in a given buyer-supplier relationship should affect the degree of headquarter influence over purchasing in the buyer firm.
	OM2: Headquarter influence could have direct effects on cost reductions, but we control for several variables that capture some of the mechanisms through which these effects might take place, including asset specificity and buyer firm size.

An additional benefit of these variables for identification purposes is that the potential IVs are all somewhat different and are likely to work through different mechanisms, affecting both the

marginal benefits (relationship complexity) and the costs (all IVs) of formal contracting as well as purchasing managers' personal costs of breaking the norms of writing detailed contracts (headquarter influence). This should increase our ability to discover failure in order to satisfy the exogeneity condition when performing empirical assessments of exogeneity.

In sum, we have several reasons to believe that these potential IVs can satisfy the assumptions of both relevance and exogeneity. The weakness of the IV method is that these assumptions may not hold, even though we have reason to believe they do.

Note also the importance of the control variables in ensuring exogeneity. However, some of the control variables could also cause endogeneity problems. Buyer- and supplier-specific assets might themselves be endogenous variables, and, as discussed in the article, endogenous control variables can sometimes exacerbate rather than reduce endogeneity bias by introducing new omitted variables to the model. However, it is difficult to conceive of what specific omitted variables might be related to asset specificity and cost reductions and not formal contracting.

Given that we cannot be certain that our assumption that the IVs satisfy the relevance and exogeneity conditions, we must also conduct an empirical assessment of the IVs.

Step 4: How should IVs be evaluated empirically?

In Step 3, we argued that we have reason to believe that the potential IVs can satisfy the requirements for both relevance and exogeneity. In the following, we undertake an empirical examination of this assumption, first using the `ivreg2`-command in Stata 15 and then using SEM in Stata using the `sem`-command. Note that we could use Stata's built-in `ivregress`-command for many of the same functions as `ivreg2`. However, we prefer to use `ivreg2`, because it offers some statistics not available in the built-in command (in particular the C-statistic).

Empirically assessing instrument relevance and exogeneity using `ivreg2`

Checking the relevance and exogeneity of the IVs is relatively straightforward using 2SLS in the `ivreg2`-command (see Step 4: A.1. in the do-file).

Formal contracting (`formcon`) is specified as a function of all the explanatory variables. However, cost reductions (`cro`) are not directly affected by relationship complexity (`complex`), headquarter influence (`hqinflu`) or annual purchasing value (`lnval`), which only affect formal contracting. `first` will prompt Stata to display the first-stage regression, and `orthog(complex)` specifies that we want to use the C-statistic so we can assess the exogeneity of the relationship complexity.

The results are displayed in Box WB3 below. The first part of the output displays results from estimating the first-stage regression. Here, we can see that several of the explanatory variables have significant effects on formal contracting, particularly buyer- and supplier-specific assets and their quadratics and interaction effects (`bsa`, `ssa`, `bsasq`, `ssasq`, `bsassa`), firm size (`lnempl`), the three dummy variables for sub-industries, and the three IVs. In total, the first stage explains 43.3% of the variance in formal contracting.

Box WB3: Stata output from estimating 2SLS using ivreg2v

```
First-stage regressions
-----------------------

First-stage regression of formcon:

OLS estimation
--------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =      305
                                                      F( 18,   286) =    12.15
                                                      Prob > F      =   0.0000
Total (centered) SS     =  982.6887852                Centered R2   =   0.4333
Total (uncentered) SS   =   5091.96444                Uncentered R2 =   0.8906
Residual SS             =  556.9176611                Root MSE      =    1.395

------------------------------------------------------------------------------
     formcon |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      md_bsa |   .2454875   .1023036     2.40   0.017      .044124     .446851
      md_ssa |   .0627932   .0963087     0.65   0.515    -.1267706     .252357
       bsasq |  -.1067756   .0516558    -2.07   0.040    -.2084494    -.0051018
       ssasq |  -.0588699   .0561012    -1.05   0.295    -.1692935     .0515537
      bsassa |   .1560605   .0735601     2.12   0.035     .0112726     .3008484
         unc |  -.1218452   .0773346    -1.58   0.116    -.2740624      .030372
     perfamb |   .0331575   .0652618     0.51   0.612    -.0952968     .1616117
      lnempl |   .2184095   .0764481     2.86   0.005     .0679373     .3688817
       knsim |   .0192442   .0686014     0.28   0.779    -.1157834     .1542719
   lnintproc |   .1812763   .1687369     1.07   0.284    -.1508473        .5134
  c_construc |    1.08803   .2535486     4.29   0.000     .5889718     1.587088
     c_trade |   1.098646   .2738397     4.01   0.000      .559649     1.637643
   c_process |    1.28477   .2670754     4.81   0.000      .759087     1.810452
        bexp |  -.0108129   .0599388    -0.18   0.857    -.1287901     .1071643
        sexp |  -.0291743   .0665377    -0.44   0.661      -.16014     .1017915
     complex |     .17755   .0761917     2.33   0.020     .0275824     .3275176
     hqinflu |   .1392554   .0563426     2.47   0.014     .0283568     .2501541
       lnval |   .2510981   .0784567     3.20   0.002     .0966723     .4055239
       _cons |    1.25784   .5563606     2.26   0.025      .162759     2.352921
------------------------------------------------------------------------------
Collinearities detected among instruments: 1 instrument(s) dropped
Included instruments: bsa ssa bsasq ssasq bsassa unc perfamb lnempl knsim
                      lnintproc c_construc c_trade c_process bexp sexp complex
                      hqinflu lnval
------------------------------------------------------------------------------
Partial R-squared of excluded instruments:   0.0799
Test of excluded instruments:
  F(  3,   286) =     8.28
  Prob > F      =   0.0000



Summary results for first-stage regressions
-------------------------------------------

Variable    | Shea Partial R2 |   Partial R2    |  F(  3,   286)    P-value
formcon     |     0.0799      |     0.0799      |        8.28       0.0000

Underidentification tests
Ho: matrix of reduced form coefficients has rank=K1-1 (underidentified)
Ha: matrix has rank=K1 (identified)
Anderson canon. corr. N*CCEV LM statistic   Chi-sq(3)=24.38    P-val=0.0000
```

```
Cragg-Donald N*CDEV Wald statistic          Chi-sq(3)=26.50    P-val=0.0000

Weak identification test
Ho: equation is weakly identified
Cragg-Donald Wald F-statistic                        8.28
See main output for Cragg-Donald weak id test critical values

Weak-instrument-robust inference
Tests of joint significance of endogenous regressors B1 in main equation
Ho: B1=0 and overidentifying restrictions are valid
Anderson-Rubin Wald test     F(3,286)= 3.23      P-val=0.0229
Anderson-Rubin Wald test     Chi-sq(3)=10.33     P-val=0.0160
Stock-Wright LM S statistic  Chi-sq(3)=9.99      P-val=0.0187

Number of observations              N  =        305
Number of regressors                K  =         17
Number of instruments               L  =         19
Number of excluded instruments      L1 =          3

IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =      305
                                                      F( 16,   288) =     9.04
                                                      Prob > F      =   0.0000
Total (centered) SS     =  463.1723592                Centered R2   =   0.2235
Total (uncentered) SS   =  4959.101471                Uncentered R2 =   0.9275
Residual SS             =  359.6518278                Root MSE      =    1.086

------------------------------------------------------------------------------
         cro |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     formcon |   .4135612    .156109     2.65   0.008     .1075932    .7195293
      md_bsa |   .1120059     .09288     1.21   0.228    -.0700356    .2940474
      md_ssa |   .2700803   .0761333     3.55   0.000     .1208617    .4192989
       bsasq |   .0213728   .0435984     0.49   0.624    -.0640785    .1068241
       ssasq |    .087913   .0447178     1.97   0.049     .0002676    .1755584
      bsassa |  -.1303423   .0635033    -2.05   0.040    -.2548065    -.005878
         unc |   .2089363   .0635247     3.29   0.001     .0844301    .3334425
     perfamb |  -.1667023   .0517649    -3.22   0.001    -.2681596   -.0652449
      lnempl |  -.2849436   .0792438    -3.60   0.000    -.4402585   -.1296287
       knsim |   .0916887   .0531614     1.72   0.085    -.0125057    .1958832
   lnintproc |  -.4864315   .1371641    -3.55   0.000    -.7552682   -.2175949
  c_construc |  -.5572759   .2732349    -2.04   0.041    -1.092806   -.0217453
     c_trade |  -.3871973   .2963887    -1.31   0.191    -.9681086    .1937139
   c_process |  -.3156091   .2973474    -1.06   0.289    -.8983992    .2671811
        bexp |  -.0586063   .0464351    -1.26   0.207    -.1496174    .0324048
        sexp |   .1740545   .0518356     3.36   0.001     .0724586    .2756504
       _cons |   1.343083   .4233663     3.17   0.002     .5133006    2.172866
------------------------------------------------------------------------------
Underidentification test (Anderson canon. corr. LM statistic):          24.381
                                                   Chi-sq(3) P-val =    0.0000
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):                8.283
Stock-Yogo weak ID test critical values:  5% maximal IV relative bias    13.91
                                         10% maximal IV relative bias     9.08
                                         20% maximal IV relative bias     6.46
                                         30% maximal IV relative bias     5.39
                                         10% maximal IV size             22.30
                                         15% maximal IV size             12.83
```

```
                                        20% maximal IV size              9.54
                                        25% maximal IV size              7.80
Source: Stock-Yogo (2005).  Reproduced by permission.
------------------------------------------------------------------------------
Sargan statistic (overidentification test of all instruments):           1.060
                                                    Chi-sq(2) P-val =    0.5885
-orthog- option:
Sargan statistic (eqn. excluding suspect orthogonality conditions):      0.024
                                                    Chi-sq(1) P-val =    0.8772
C statistic (exogeneity/orthogonality of suspect instruments):           1.037
                                                    Chi-sq(1) P-val =    0.3086
Instruments tested:   complex
------------------------------------------------------------------------------
Collinearities detected among instruments: 1 instrument(s) dropped
Instrumented:         formcon
Included instruments: md_bsa md_ssa bsasq ssasq bsassa unc perfamb lnempl knsim
                      lnintproc c_construc c_trade c_process bexp sexp
Excluded instruments: complex hqinflu lnval
------------------------------------------------------------------------------
```

Are the IVs relevant?

In the second part of the output, we see the results from the weak identification test and reports on the Cragg-Donald Wald F-statistic, which is 8.28. This is a little too low, being lower than the cut-off for 10% maximal IV relative bias, which is 9.08. Hence, we can conclude that within the model estimated, the IVs are weak.

Another problem also occurs. `Ivreg2` reports: "`Collinearities detected among instruments: 1 instrument(s) dropped`". However, apparently, none of the regressors have in fact been removed, and we do not get this message when using Stata's built-in `ivregress`-command. Related to this issue, it may be that the test for heteroscedasticity (`ivhettest`) that one can conduct after using `ivreg2` does not yield any results (we get an error message). As we explain below, this problem vanishes when we remove the sub-industry dummy variables from the model.

We address the weakness of the IVs in two ways. First, the model used above includes rather many control variables. In Step 4: A.2. in the do-file, we therefore first test the joint significance of the dummy variables that indicate processing industry, construction industry and resellers. These variables have strong significant effects on formal contracting in the first stage, but jointly they have no significant effect on cost reductions ($\chi^2(3)=4.25$, p-value=0.24). After this test, we remove the three dummy variables and re-run the estimation. Without the three dummy variables, we find that the IVs are now relevant when measured against the 10% maximal relative IV bias, because the Cragg-Donald F-statistic is 11.811, which is higher than 9.08. In addition, we find that the `ivhettest`-command now yields results and that heteroscedasticity is not a problem.

Second, we re-estimate the above model using three alternative estimators, Moreira's Conditional Likelihood Ratio (CLR) approach, LIML and Fuller's LIML (see Step 4: A.3. in the do-file). These estimators are less sensitive to the finite sample bias associated with weak IVs. To use Moreira's CLR, we use the `condivreg`-command. To use LIML or Fuller's LIML, we add the following commands to the `ivreg2`–command: `liml` or `fuller(4)`. Moreira's CLR is regarded as the preferred estimator when IVs are weak (see Bascle, 2008, for an overview of the arguments). `condivreg` does not report results from assessing the relevance of the IVs, but the

latter two cases do. The results (not displayed here) show that the Cragg-Donald F-statistic remains 8.283, but it is measured against lower critical test values. In the LIML case, the critical test value is 6.46 (for 10% maximal LIML size). Using Fuller's LIML, the critical test value for 10% maximal Fuller relative bias is 7.90. (We refer to Bascle (2008) for an accessible review of these different estimators and explanations of how they handle weak IVs.)

Based on all these results, we conclude that the IVs are sufficiently relevant for further analyses.

Are the IVs exogenous?

As evident from Box WB3, the Sargan test is insignificant (χ^2 (2)=1.060, p-val=0.5885), which indicates that overall, the instruments are exogenous, or more precisely, the IVs identify the same effect. The C-statistic for relationship complexity is also insignificant (χ^2(1)=1.037, p-value=0.3086). If we want to conduct similar tests for headquarter influence and annual purchasing value, we must replace `orthog(complex)` with `orthog(hqinflu)` or `orthog(lnval)` (see Step 4: A.1. in the do-file). By doing this, we find that the C-statistic for both headquarter influence and annual purchasing value are also insignificant (χ^2 (1)=0.204, p-value=0.6513 and χ^2 (1)=0.315, p-value=0.5747, respectively). When re-estimating the model using LIML and Fuller's LIML, we obtain similar results (see Step 4: A.4. in the do-file).

However, it is theoretically possible that some or all of the IVs are in fact not exogenous and that we fail to detect the lack of exogeneity. The reason is that the overidentification tests simply test whether the different IVs identify the same effect. We should be able to detect failure to satisfy the exogeneity condition if at least one of the IVs is truly exogenous. If this assumption does not hold, there is a risk that the different IVs will accidentally identify the same parameter, even if it is the wrong one. This is one of the weaknesses of the IV approach, and we simply have to assume that at least one IV is exogenous.

The assumption that at least one of the IVs is truly exogenous is untestable, but we have two reasons to believe that it holds:

1. Theory suggests that the IVs are exogenous, and the chance that *all* of the IVs are in fact not exogenous seems small.
2. The IVs rely on different theoretical mechanisms, and they do not correlate very much. These arguments strengthen the credibility of the assumption of exogeneity.

<u>Empirically assessing instrument exogeneity and relevance in SEM</u>

In the previous section, we examined the relevance and exogeneity of the IVs using the `ivreg2`-command in Stata. This command requires us to create composite scores for each latent variable in the model. However, given that most of the variables in the model are measured using multiple-item Likert scales, a better approach might be to use SEM software that explicitly models the variables in our model as latent constructs reflected in measurement items. In the following, we use Stata 15 and its `sem`-command to show how one can assess the relevance and exogeneity of the IVs in a SEM model.

Estimating a measurement model

The first step of the assessment is to estimate a measurement model, and Step 4: B.1. in the do-file provides source code for doing this. Before estimating this model, we create scores for each of the dimensions of formal contracting, and in the measurement model we treat each of the two scores as items in a two-item construct. In this measurement model we also eliminate the sub-industry dummies. These simplifications reduce the number of parameters that must be estimated (thereby making it easier for the model to converge), but they do not hinder us in demonstrating how to account for endogeneity in a SEM model.

The results from estimating this model show that the approximate model-to-data fit is acceptable (χ^2 (543) = 863.43, p-value = 0.0000, RMSEA=0.044, CFI=0.938, SRMR=0.044).

The important point with this model is that it is saturated and it has the same degrees of freedom as a model in which cost reduction outcomes are regressed onto the other variables in the model.

Assessing the relevance condition in a SEM model

We start by estimating a structural model in which (see Step 4: B.2. in the do-file)

1. formal contracting is posited as a mediator between the exogenous variables and cost reductions,
2. all exogenous variables (including the three IVs) are allowed to affect both formal contracting and cost reductions, and
3. we do not account for the endogeneity of formal contracting (i.e., we restrict the correlation between the error terms to zero).

This model has a χ^2(df)= 863.43(543), the same as for the measurement model. We also find the following parameter estimates, z-values and p-values:

- Relationship complexity to formal contracting `(CPLEX ->FORMCON)`: b=0.262, z-value=1.99, p-value=0.047
- Headquarter influence to formal contracting `(HQINFLU -> FORMCON)`: b=0.214, z-value=3.66, p-value=0.000
- Annual purchasing value to formal contracting `(lnval ->FORMCON)`: b=0.320, z-value=3.78, p-value=0.000

From these parameter estimates, we can conclude that individually, the IVs have strong and significant relationships with formal contracting.

Next, we delete three paths from this model: `(CPLEX ->FORMCON)`, `(HQINFLU -> FORMCON)`, and `(lnval ->FORMCON)`, i.e., we restrict the IVs from affecting formal contracting. This model has a χ^2(df)= 896.98 (546). In other words, fit is significantly worsened by these restrictions: $\Delta\chi^2$(df)= 33.55(3). The CHIDIST-formula in Microsoft Excel shows that the p-value for this test is 0.000, which supports joint significance of the IVs.

Finally, after estimating each of the two models above, we run the post-estimation command `estat eqgof`, which provides us with equation-level goodness-of-fit statistics. The results show that when we remove the three IVs as predictors of formal contracting, we reduce the

Bentler-Raykov squared multiple correlation (and the R^2) from 40% to 32%. This result is nearly identical to the partial R^2 reported by `ivreg2`.

In sum, the IVs have significant and strong effects on formal contracting, but they do not explain a major portion of the variance in formal contracting.

Assessing instrument exogeneity in an overidentified SEM model

We now estimate a structural model where we account for the endogeneity of formal contracting. This model is overidentified and nested within the measurement model. Compared to the measurement model, this model has (see Step 4: B.3.1. in the do-file):

- paths from all the exogenous variables to formal contracting (e.g., `CPLEX -> FORMCON`);
- paths from almost all the exogenous variables to cost reduction outcomes (e.g., `KNSIM -> CRO`), except for the three IVs (`CPLEX`, `HQINFLU`, and `lnval`) that have no direct path to cost reduction outcomes;
- no free covariances between the exogenous variables and the two endogenous variables (`FORMCON` and `CRO`);
- a path from formal contracting to cost reduction outcomes (`FORMCON -> CRO`); and
- a free covariance between cost reduction outcomes and formal contracting (`e.CRO*e.FORMCON`) to account for possible endogeneity.

The model has a χ^2(df)= 865.19(545). In other words, compared to the saturated measurement model, we have a $\Delta\chi^2$(df)=1.76(2). The CHIDIST-formula in Microsoft Excel shows that the p-value for this test is 0.41. This test is analogous to a Sargan test for exogeneity, as it shows that the overidentifying restrictions do not significantly reduce the fit of the model.

However, we should also assess the individual overidentifying restrictions, just as with the C-statistic. One way to do this is to use the modification indices (a Lagrange multiplier test). The benefit of these indices is that no additional models have to be estimated for comparison. Using the `sem`-command in Stata, we can normally obtain the modification indices in Stata by the command `estat mindices`.

Using this procedure, we find that the modification index is 1.74 for relationship complexity, 0.46 for headquarter influence, and 0.14 for annual purchasing value. The corresponding p-values for these tests are, respectively, 0.19, 0.50 and 0.71. In other words, the modification indices suggest that individually the overidentifying restrictions hold (we obtain similar results when estimating the same model in Lisrel 8.80).

An alternative way to test the overidentifying restrictions is to open each overidentifying restriction individually and perform several individual $\Delta\chi^2$-tests. As opposed to the modification indices (a Lagrange multiplier test), this is a likelihood ratio test. It requires us to estimate several additional models, and in each model, we open up a path from one of the IVs to the dependent variable cost reduction outcomes. This test is asymptotically equivalent to modification indices, and, as can be seen from the results below, it yields in our case results that are nearly identical to the modification indices. In each of the models below, we add one path to the `sem`-command in Step 4: B.3.1. (see Step 4: B.3.3 in the do-file):

1. `(CPLEX -> CRO)`: χ^2(df)= 863.43(544). Hence, compared to the overidentified IV model with three IVs, we have a $\Delta\chi^2$(df)= 1.75(1). This test has a p-value of 0.18.
2. `(HQINFLU -> CRO)`: χ^2(df)= 864.72(544). Hence, compared to the overidentified IV model with three IVs, we have a $\Delta\chi^2$(df)= 0.47(1). This test has a p-value of 0.49.
3. `(lnval -> CRO)`: χ^2(df)= 865.05 (544). Hence, compared to the overidentified IV model with three IVs, we have a $\Delta\chi^2$(df)= 0.14(1). This test has a p-value of 0.71.

Note that an assumption underlying the chi-square difference test between two nested SEM models is that the base-line model must be correctly specified. If it is incorrectly specified, there is a danger that changing the structural model will lead to changes in the loadings and other parameters that really should be identical across the two nested models. To check that this is not the case, we compared all the parameters across the models and find that most of them do not change and that none of them change more than by 0.014. In other words, the model is fairly stable, which indicates that the baseline model is correctly specified. Additional evidence suggesting that the model is stable is that the results from estimating a SEM with latent variables are quite similar to those we obtain when using composite scores and IV estimation in `ivreg2`.

Step 5: How should you interpret and evaluate the results?

In step 5, we suggest that researchers should conduct sensitivity analysis concerning choice of IV, that they should compare alternative estimators, and that they should test for endogeneity.

In the following, we will consider how to do this in two different situations. First, we will assume that formal contracting is a *continuous* variable and use the composite score that we have used earlier. Next, we will assume that formal contracting is a *latent* variable and consider the results we obtain when using a SEM model.

Formal contracting as a continuous variable

Results from using IV estimators

Box WB3 above shows the results from a 2SLS estimation of the effect of formal contracting on cost reductions. However, when assessing the IVs, we find that they are too weak for 2SLS and that we should use techniques that are less sensitive to weak IVs. We therefore use the Stata code displayed in Step 5: A.1. of the do-file. The list below provides the different estimates for each of the different estimators we use:

- 2SLS: b=0.414, st.error=0.156, z-value=2.65, p-value=0.008
- 2SLS fewer control variables: b=0.318, st.error=0.119, z-value=2.68, p-value=0.007
- Moreira's CLR: b=0.414, st.error=0.161, z-value=2.57, p-value=0.011
- LIML: b=0.429, st.error=0.161, z-value=2.67, p-value=0.008
- Fuller's LIML: b=0.373, st.error=0.144, z-value=2.59, p-value=0.010

As evident, there is little difference between the results, except that Moreira's CLR and Fuller's LIML (as expected) have higher p-values due to somewhat larger standard errors compared to the parameter estimates. From these results, we can conclude that, as a whole, the effect of formal contracting on cost reduction is positive.

Testing for endogeneity

The result that formal contracting has a significant and positive effect on cost reductions is in contrast to the result presented in Box WB2, where the effect is small and only weakly significant. By testing for the difference between the IV results and the results from OLS, we perform a Durbin-Wu-Hausman test of the endogeneity of formal contracting. Using the `ivreg2`-command in Stata, we can simply add `endog(formcon)`. Using the code displayed in Step 5: A.2. in the do-file, we find that we must reject the hypothesis that formal contracting can be treated as exogenous (in both LIML and Fuller's LIML, we obtain: $\chi 2(df)=6.594$, p-value=0.0102). In other words, these results suggest that we have an endogeneity problem, that the OLS estimate is biased downward, and that we should correct for it.

Assessing heterogeneity using IV estimators

However, as indicated at the beginning of this document, the effect of formal contracting may be heterogeneous. As a first step in examining this issue, we estimate three just-identified 2SLS models. In each of these models, we remove two of the IVs in the `ivreg2`-command, so that we use only one IV to estimate the effect in each of the models (see Step 5: A.3.1. in the do-file). We find that the three IVs in isolation identify the following effects:

- Relationship complexity: 0.644, p-value=0.043
- Annual purchasing value: 0.348, p-value=0.091
- Headquarter influence: 0.291, p-value=0.309

These parameters are all substantially higher than the results from using OLS. There are some differences between them, but under Step 4, we found that the IVs passed the Sargan test and they had satisfactory C-statistics, which suggests that the above parameters are not significantly different from each other.

Another way to examine heterogeneity is to test whether the effect of formal contracting varies with observed variables. A general way to do this is to estimate the interaction effects between formal contracting and other exogenous variables. Indeed, transaction cost theory suggests that the effect of governance on performance should vary with transaction attributes. We might therefore be interested in estimating these interaction effects from a theoretical viewpoint as well. Unfortunately, it is difficult to find good IVs for all the different IVs. In the following, we therefore only examine the interaction effects between formal contracting and four of the exogenous variables in the model: buyer-specific assets (`bsa`), supplier-specific assets (`ssa`), environmental uncertainty (`unc`), and performance ambiguity (`perfamb`). Step 5: A.3.2. in the do-file provides the source code for estimating the interaction effects using the `ivregress` command in Stata 15.

The procedure in Step 5: A.3.2 consists of five steps. In the first step, we generate the endogenous interaction variables. In the second step, we predict formal contracting and create interaction terms between the predicted level of formal contracting and the four transaction attributes. The interaction terms between the predicted level of formal contracting and the four transaction attributes will be used as IVs for the interaction terms between the actual level of formal contracting and the four transaction attributes. In the third step, we perform a LIML estimation using the `ivreg2`-command in Stata (we use LIML because it is less sensitive to

weak IVs). Note that `md_formcon`, `fconXbsa`, `fconXssa`, `fconXunc`, and `fconXperfamb` are endogenous variables, and we use `complex`, `hqinflu`, `lnval`, `pfconXbsa`, `pfconXssa`, `pfconXunc`, and `pfconXperfamb` as IVs to identify their effects. In the fourth step, we perform a test of the joint significance of the various endogenous parameters. Finally, in the fifth step, we perform a joint test of the endogeneity of the interaction terms.

Box WB4, below, shows some of the results from estimating the effect of formal contracting and its interaction terms with buyer-specific assets, supplier-specific assets, environmental uncertainty and performance ambiguity. As is evident, the main effect of formal contracting is nearly identical to that found earlier (0.423) and is highly significant (p-value: 0.009). The interaction terms are not significant, however. The test of the joint significance of these interaction terms does not reject the null hypothesis (χ^2(df)=3.68(4), p-value=0.45). We also perform a joint test of the endogeneity of the interaction terms (by specifying the option `endog(fconXbsa fconXssa fconXunc fconXperfamb)`). It turns out that we have an endogeneity problem (χ^2(df)=12.611(4), p-value=0.013).

These results suggest that there is little or no heterogeneity in the effect of formal contracting across the observed transaction attributes when the change in formal contracting occurs due to a change in relationship complexity, headquarter influence or annual purchasing value.

Box WB4: Output from estimating the effect of formal contracting and its interaction terms

```
------------------------------------------------------------------------------
         cro |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  md_formcon |   .4245134   .1634636     2.60   0.009     .1041306    .7448961
    fconXbsa |  -.0862247   .0787655    -1.09   0.274    -.2406022    .0681528
    fconXssa |   .0707482   .0769094     0.92   0.358    -.0799915     .221488
    fconXunc |   .0084847   .0565907     0.15   0.881    -.1024311    .1194004
fconXperfamb |   .0454936   .0450467     1.01   0.313    -.0427963    .1337835
```

A potential problem with the above model, however, is that we have many IVs, and we may suffer from multiple weak IVs that may not be exogenous. The Sargan statistic is satisfactory (p-value=0.63), and none of the C-statistics have significant p-values. Although each of the first-stage regressions are significant, the Cragg-Donald F-statistic is quite low (3.37), and no cut-off-criteria exist for cases with five endogenous variables. We therefore also estimate models where each of these interactions enter the model in isolation (see Step 5: A.3.3 in the do-file for Stata code). In each of these cases, the Cragg-Donald F-statistic is higher than the cut-off criterion for 10% maximal LIML size when there are two endogenous variables and 4 IVs (4.72), which indicates that the IVs have sufficient relevance. The Sargan and C-statistics are also insignificant for each regression. In each of these regressions, we also test for endogeneity by using the `endog`-option. We find that the interaction term between formal contracting and buyer-specific assets suffers from endogeneity (χ^2(df)=9.545(1), p-value=0.00), but none of the others do. Concerning the parameter estimates for the interactions, the results are similar to the joint test: none of the interaction terms are significant.

We also try an alternative approach to using the predicted level of formal contracting to create IVs for the interaction terms: we create interaction terms between the IVs and the moderating variables (e.g., we use the interaction term between performance ambiguity and relationship complexity (`peraXcplex`) as an IV for the interaction term between performance ambiguity and formal contracting (`fconXperfamb`); see Step 5: A.3.4 and Step 5: A.3.5 in the do-file for the Stata code). This creates many more IVs (three per endogenous variable). With more IVs, we generally predict more of the variation in the endogenous variables compared to when we use just one IV constructed based on predicting formal contracting. However, due to the higher number of IVs, the Cragg-Donald F-statistic is lower (1.332). Furthermore, although the Sargan statistic is insignificant (p-value=0.62), two of the interaction terms—that between environmental uncertainty and headquarter influence (`uncXhqinflu`) and that between performance ambiguity and annual purchasing value (`peraXlnval`)—have weakly significant C-statistics (i.e., higher p-value than 0.05 but lower p-value than 0.1). This finding indicates that the exogeneity condition does not hold for these two interaction terms and they should not be used as IVs. We therefore remove them from the model. After doing so, none of the C-statistics are significant. In addition, we estimate simpler models with only a single endogenous interaction term in each model. In each model, the Cragg-Donald F-statistic is higher than the cut-off criterion for 10% maximal LIML size when there are two endogenous variables and five or six IVs, which indicates that the IVs have sufficient relevance. However, the results do not change: regardless of what model we estimate, none of the interaction terms are significant. Also, we find that none of the interaction terms, jointly or in isolation, suffer from endogeneity.

Assessing heterogeneity using control function estimators

In the previous section, we examined the extent to which formal contracting has a heterogeneous effect on cost reductions across observed variables. We do not detect significant heterogeneity.

In the following, we examine whether formal contracting might have heterogeneous effects across both observed and unobserved variables, using control function techniques. We first use a simplified version of Garen's (1984) estimator. Step 5 A.4.1. in the do-file displays the Stata code for implementing this estimator. Several things can be noted:

- We first generate interaction terms between formal contracting and several of the observed exogenous variables in our model. However, we do not create interaction terms between formal contracting and all the exogenous variables in the model, which this estimator originally prescribes. The reason for not including all the interaction terms is twofold: 1) the model would become very large, 2) and when trying this, we find that none of them are significant. Therefore, we only include those exogenous variables that measure transaction attributes known from previous research to have governance implications.
- We use the bootstrap to ensure correct standard errors.
- Garen's (1984) estimator starts by regressing formal contracting onto the exogenous variables. Next, we predict the first-stage residual, `zhat`, and create an interaction term between `zhat` and formal contracting. In the final stage, we regress cost reductions onto formal contracting, control variables, interaction terms between formal contracting and the transaction attributes, the square of formal contracting, and the first-stage residual and its interaction with formal contracting.

- The post-estimation command asks for bias-corrected bootstrap confidence intervals: `estat bootstrap, bc`.

The results from using Garen's (1984) estimator are presented in Box WB5 below (these results are not substantially different from the normal-based confidence intervals). As before, we find a significant and positive effect of formal contracting. We also find that few of the interaction effects are significantly different from zero, except the one between formal contracting and performance ambiguity. The interaction term between formal contracting and buyer-specific assets is not significant but is positive and has a relatively low normal-based p-value (0.14). Moreover, we find that the first-stage residual has a negative association with cost reductions, which indicates that we have an endogeneity problem. However, the interaction effect between formal contracting and the first-stage residual is not significant, which indicates that there is no essential heterogeneity and that cost reductions are not a major reason for self-selecting into higher/lower levels of formal contracting.

In sum, these results indicate that when formal contracting increases, the effect of formal contracting is positive and that the effect increases with performance ambiguity. However, it should be noted that the interaction term between formal contracting and performance ambiguity is not corrected for endogeneity in a similar way as in the IV estimator. We also find that there is an endogeneity problem, but the problem does not seem to be that firms self-select based on expected cost reductions. Rather, there seem to be absolute disadvantages related to formal contracting. Those firms that choose a higher level of formal contracting for reasons other than the variables included in the first-stage regression generally have lower cost reductions. This is an absolute disadvantage *associated* with formal contracting.

Box WB5: Results from using Garen's (1984) estimator (with 95% asymmetric bias-corrected bootstrap confidence intervals)

	Observed		Bootstrap			
	Coef.	Bias	Std. Err.	[95% Conf.	Interval]	
md_formcon	.39384804	-.0068997	.15163476	.1641503	.7758238	(BC)
fcXfc	-.02047123	-.0046782	.03873233	-.0931246	.0629257	(BC)
md_bsa	.10423348	-.0116127	.11110347	-.109529	.3343315	(BC)
md_ssa	.27884051	.0092317	.08549572	.0949028	.4341185	(BC)
bsasq	.00563337	-.003747	.05520225	-.0993161	.1193507	(BC)
ssasq	.08453113	-.0064239	.0460821	-.0020849	.1838147	(BC)
bsassa	-.12237158	.0053176	.08530414	-.2943842	.038064	(BC)
md_unc	.20186034	-.0027099	.06822576	.0758641	.3491657	(BC)
md_perfamb	-.15921241	.0087527	.05977701	-.2968571	-.0547392	(BC)
md_lnempl	-.28980836	.0006219	.08260147	-.5035042	-.1642886	(BC)
md_knsim	.08220024	.0030753	.06152417	-.0367741	.1971217	(BC)
md_lnintproc	-.47075961	.0051286	.1528306	-.816524	-.2007596	(BC)
md_c_const~c	-.50256307	.0125625	.27329951	-1.18422	-.0347821	(BC)
md_c_trade	-.33499624	.0127307	.31882926	-1.15115	.1721069	(BC)
md_c_process	-.26169366	.0114574	.29579411	-1.067097	.2272916	(BC)
md_bexp	-.05512187	-.0000627	.05109494	-.1494542	.0336768	(BC)
md_sexp	.17297876	-.0017114	.06101038	.0594215	.3018633	(BC)
fcXbsa	.0671044	.0028239	.04601742	-.0205552	.1537722	(BC)
fcXssa	-.03939512	.0007098	.03594963	-.1071579	.0294906	(BC)
fcXunc	-.00739283	.0020762	.03245464	-.070643	.0544172	(BC)
fcXperfamb	.04484673	.0013074	.02439538	.0011754	.0957703	(BC)
zhat	-.33485839	.0089342	.16137053	-.7454978	-.0903585	(BC)
formconzhat	-.01203462	.003728	.04377482	-.0999675	.0669502	(BC)

```
       _cons |   3.8455534   .0144317   .10165309    3.637351   4.025746  (BC)
```

De Blander's (2010) estimator relaxes some of the strong assumptions underlying Garen's (1984) estimator. In the following, we try to implement the De Blander estimator. Please see Step 5 A.4.1. in the do-file for the Stata source code.

However, this estimator involves adding many more terms than what Garen's (1984) estimator has, and doing so reduces the degrees of freedom and introduces high degrees of multicollinearity. Trying to estimate the full estimator yields results where most of the variables are insignificant and many of the bootstrap replications fail to generate any results. Therefore, we opt to estimate a more limited model. Compared to Garen's (1984) model, we add the following:

- Interaction terms between the first-stage residual and the four transaction attributes (buyer-specific assets, supplier-specific assets, environmental uncertainty and performance ambiguity). Doing so helps us account for the possible endogeneity of the interaction terms between formal contracting and transaction attributes.
- Interaction terms between the IVs and formal contracting and between the IVs and the first-stage residual to account for how the effect of formal contracting may depend on the IVs.
- The square of the first-stage residual orthogonalized with respect to all other exogenous variables, their squares and their cross-products, like in De Blander's original model (using the `orthog`-command in Stata).

We specify the use of 5000 bootstrap replications. It should be noted that in 149 of these 5000 replications, there are no results, because one or more parameters could not be estimated. The problem is likely that in some of the replications, Stata is incapable of performing the `orthog`-command, because when we eliminate the orthogonalized square of the first-stage residual from the model, we do not have this problem. The parameter estimates we present in Box WB6 are therefore based on the remaining 4851 successful bootstrap replications.

Box WB6 shows the results from estimating this model with bias-corrected bootstrap confidence intervals (these results are not substantially different from the normal-based confidence intervals), and we can observe the following:

- The main effect of formal contracting is significant and positive.
- The square of formal contracting has no significant effect.
- None of the interaction terms between formal contracting and the observed transaction attributes are significant.
- Similarly to the results from using Garen's (1984) estimator, the first-stage residual has a significant negative relationship with cost reductions, suggesting that endogeneity should be a concern.
- The interaction term between formal contracting and the first-stage residual is insignificant, which indicates that essential heterogeneity and self-selection should not be a major concern.
- Most of the interaction terms between the first-stage residual and the observed transaction attributes are insignificant, except for the term between buyer-specific assets and the first-stage residual. This finding means that the interaction term between formal contracting

and buyer-specific assets suffers from an endogeneity problem, and the test we have conducted here is essentially similar to the one we performed earlier when using interaction terms in `ivreg2` and testing for the endogeneity of this interaction term. This finding is difficult to interpret. Although firms that choose a higher level of formal contracting for other reasons than the variables included in the first-stage regression generally have lower cost reductions, this is not the case if they have made high relationship-specific investments. Thus, for those firms with high levels of buyer-specific assets, there are no absolute disadvantages associated with formal contracting. For some firms, there may even be absolute advantages associated with formal contracting.

- The interaction term between formal contracting and buyer-specific assets (fcXbsa) turns from positive when using Garen's (1984) estimator to negative when using De Blander's (2010) estimator.

Box WB6: Output from using De Blander's (2010) estimator to estimate the effect of formal contracting on cost reductions (with 95% asymmetric bias-corrected bootstrap confidence intervals)

	Observed		Bootstrap			
	Coef.	Bias	Std. Err.	[95% Conf.	Interval]	
md_formcon	.40505555	-.0195111	.18960251	.1299614	.9390723	(BC)
fcXfc	-.0692217	-.0005464	.06662077	-.2022434	.058199	(BC)
md_bsa	.07309021	.0021745	.1221772	-.1815668	.3006525	(BC)
md_ssa	.31258187	-.0038397	.0898753	.125565	.4821348	(BC)
bsasq	.05338531	-.0176813	.06067685	-.0443825	.1911579	(BC)
ssasq	.0817045	-.0093783	.04969921	-.0048457	.188616	(BC)
bsassa	-.11515541	.0163691	.08588973	-.3026997	.0342491	(BC)
md_unc	.17918235	.0052854	.06962393	.0433959	.3155622	(BC)
md_perfamb	-.15194585	.009591	.06298999	-.3027193	-.0457508	(BC)
md_lnempl	-.30681695	.0086932	.09084542	-.5418206	-.165983	(BC)
md_knsim	.09459209	-.0058808	.0645145	-.0236426	.2273255	(BC)
md_lnintproc	-.42301431	.0043403	.16310093	-.7742348	-.1274406	(BC)
md_c_const~c	-.49897852	.0353765	.31079443	-1.327642	-.0295246	(BC)
md_c_trade	-.35994137	.0377697	.33532786	-1.26756	.1566404	(BC)
md_c_process	-.21453982	.0214518	.34944458	-1.181844	.3141055	(BC)
md_bexp	-.05768115	.0033535	.05238807	-.1655134	.039174	(BC)
md_sexp	.16592543	-.0034807	.06153341	.0519154	.2927498	(BC)
fcXbsa	-.08643156	.0358605	.08236763	-.2739454	.0452062	(BC)
fcXssa	-.01125478	-.0140843	.067993	-.1297126	.1488097	(BC)
fcXunc	.01185764	.0063864	.05074414	-.086843	.1106137	(BC)
fcXperfamb	.06130039	-.0036826	.04461532	-.0219246	.1531058	(BC)
fcXcomplex	.02590895	-.0024076	.04606218	-.0595632	.1231849	(BC)
fcXhqinflu	.03394737	.0008238	.03161101	-.0281497	.0952022	(BC)
fcXlnval	.02069581	-.0007377	.04891841	-.0745917	.1204939	(BC)
zhat	-.3451195	.0204289	.19957935	-.8777171	-.0375193	(BC)
newvar190	.02840412	-.032141	.08201798	-.0822648	.334538	(BC)
formconzhat	.0043372	.0082242	.07020725	-.1388522	.1342825	(BC)
bsazhat	.26172007	-.042601	.09500324	.1200792	.4777275	(BC)
ssazhat	-.03934027	.0160605	.08300433	-.2260408	.1090732	(BC)
unczhat	-.06203881	-.0078426	.0725287	-.2080273	.0747459	(BC)
perfambzhat	-.05194291	.0116875	.06136258	-.1840498	.057901	(BC)
complexzhat	-.01248201	-.0019219	.06690453	-.1401781	.1190991	(BC)
hqinfluzhat	.00210807	-.0086591	.04200567	-.0735008	.0878226	(BC)
lnvalzhat	.04480824	-.00479	.05037154	-.0572304	.1428652	(BC)
_cons	3.9238117	-.0055099	.11072683	3.711139	4.146208	(BC)

In conclusion, so far in this web appendix, we have utilized variation in formal contracting that arises due to changes in relationship complexity, headquarter influence and annual purchasing value, which we argue can be assumed to be exogenous, and we have estimated the effect of formal contracting for those firms that change the degree of formal contracting due to changes in these IVs. This effect on cost reductions does not seem to change much, regardless of whether observed or unobserved variables are used as moderators. In other words, the results suggest that essential heterogeneity should not be a major concern when estimating the effect of formal contracting on cost reductions using Sande and Haugland's data. We can rely on the IV estimators to test the effect of formal contracting on cost reduction outcomes, which in all the models is positive and significant.

Formal contracting as a latent variable

Results from using IV estimators in SEM

When estimating the effect of formal contracting on cost reductions using the Stata code exhibited Step 4: B.3.1 in the do-file (see also Step 5: B.1.1), we find that the effect of formal contracting on cost reductions is positive and significantly different from zero (b=0.236, s.e.=0.092, z=2.58, p-value=0.010).

From this result, we can conclude that as a whole, the effect of formal contracting on cost reduction is positive, when increases in formal contracting occur due to higher relationship complexity, greater centralization of the purchasing function, and higher annual purchasing value.

Testing for endogeneity

In contrast to the SEM model with IVs, in a SEM model where formal contracting is treated as an exogenous variable, like in the OLS model presented in Box WB2, formal contracting does not have a significant effect on cost reductions (b=0.040, s.e.=0.034, z=1.16, p-value=0.248). Hence, there is reason to suspect that we may have an endogeneity problem.

The output from estimating the SEM model exhibited in Step 5: B. 1. of the do-file provides a test for endogeneity in the form of the significance of the correlation between the residuals for formal contracting and cost reductions. This correlation is negative and significant at nearly the 5% level (b=-0.356, s.e.=0.182, z=1.95, p-value=0.051).

Another way to test for endogeneity is to conduct a likelihood ratio test by restricting the correlation between the two residuals to zero. Removing `e.CRO*e.FORMCON` from the `sem`-command in Step 5: B.1. (so that we obtain the code displayed in Step 5: B.2.) increases the χ^2(df) from 865.19(545) to 869.81(546), which implies a $\Delta\ \chi^2$(df) =4.628(1) and a p-value of 0.032, which is a significant drop in model-to-data fit. We must therefore reject the hypothesis that formal contracting can be treated as exogenous.

Assessing heterogeneity

In principle, it may be possible to assess heterogeneity in SEM by including interaction terms between the latent variables (i.e., between formal contracting and other latent and observed variables). However, it would require many interaction terms, and the model would become very

complex. It is also, to our knowledge, impossible to include interactions between residuals and latent variables in SEM. Hence, we refer to the results using `ivreg2` and the control function estimators for assessments of heterogeneity.

Summary and discussion of how to interpret and evaluate the results

In the preceding sections, we started out in Step 1 by outlining the possible reasons why we might expect to have an endogeneity problem. Given the nature of the data and the kind of variable we are examining, we saw several reasons why we may have an endogeneity problem.

In Step 2, we considered what kind of estimator to use. Because formal contracting is measured using a multiple-item Likert scale, it seems reasonable to try to use both SEM modeling and IV models that treat formal contracting as a continuous variable.

In Step 3, we identified some potential IVs and argued that we have reason to believe that they are both relevant and exogenous. However, the arguments are not watertight, and we should perform an empirical assessment.

In Step 4, we evaluated the IVs empirically. The IVs are slightly too weak for 2SLS. Therefore, we use alternative estimators that are more robust to weak IVs. Using these estimators, the IVs are sufficiently relevant.

In Step 5, we first test the effect of formal contracting on cost reductions, using several different estimators. All of them report a positive and significant effect. Next, we test for endogeneity, and in all the tests, we reject the hypothesis that formal contracting can be treated as an exogenous variable. We further find that the interaction term between formal contracting and buyer-specific assets suffers from endogeneity.

Regardless of the estimation technique, none of the exogeneity tests suggest that the IVs do not satisfy the exogeneity condition. However, the tests for exogeneity rely on the untestable assumption that at least one of the IVs is truly exogenous. It is useful to compare this untestable assumption with the assumption when using OLS (without accounting for endogeneity) that there is no simultaneity, no omitted variables and no measurement error. Given the theoretical arguments that we probably have an endogeneity problem and that the Durbin-Wu-Hausman test rejects the hypothesis that formal contracting can be treated as an exogenous variable, it is probably better to rely on results from an IV-based model than to rely on results from OLS or SEM models that do not account for endogeneity. The IV-based model rests on narrower and more specific assumptions that we have reasons to believe are more realistic compared to the more open assumptions underlying the OLS model.

Also in Step 5, we assess the degree to which the effect of formal contracting is heterogeneous and suffers from essential heterogeneity. We find little evidence of that. The IV models provide no evidence of heterogeneity, nor does De Blander's estimator. Garen's estimator suggests that performance ambiguity moderates the effect of formal contracting, but De Blander's estimator does not. Neither Garen's nor De Blander's estimators find evidence that the effect of formal contracting varies across different levels of the first-stage residual. In other words, it does not seem that unobserved variables moderate the effect of formal contracting either and that firms select into higher or lower degrees of formal contracting based on anticipated effects in terms of cost reductions (these results are consistent with those found by Sande & Haugland). Also, the

different IVs are quite different from each other, yet they identify parameters that are fairly similar. Hence, we have reason to believe that the effect of formal contracting on cost reductions is fairly homogenous. This also means that the effect we estimate can probably be generalized beyond the complier subgroup, i.e., the group of firms that increase formal contracting because of increases in the IVs.

One surprising finding when using De Blander's estimator is that there seem to be absolute disadvantages associated with formal contracting when buyer-specific assets are small, whereas these disadvantages disappear when buyer-specific assets are high. This is an issue for further theorizing. A starting point for such theorizing could be our arguments for the different sources of endogeneity, for example, if the absolute disadvantages associated with formal contracting arise because firms respond to low levels of formal contracting.

However, a major limitation of using De Blander's (2010) model on these data is that even though we choose a restricted version of this model, the model is still large, with many different interaction terms. An even simpler and more restricted model may be more appropriate and easier to interpret, such as the one used by Sande and Haugland (2015).

Step 6: What should we report?

Our framework suggests in general that we should report results from assessing IVs and results from testing for endogeneity. If the Durbin-Wu-Hausman test detects an endogeneity problem, we should report endogeneity-corrected results. In addition, given that formal contracting is measured using multiple-item Likert scales and we can therefore treat this variable as both a continuous and a latent variable, we should report results from treating it in both ways. `Ivreg2` gives special possibilities for assessing instrument relevance and accounting for weak IVs that the `sem`-command does not. The `sem`-command explicitly accounts for how many of the variables in the data are measured using multiple-item scales.

Given that the purpose here is to estimate the effect of formal contracting on cost reductions, robustness checks might report the results from using the control function estimators. However, given that these techniques in this case suggest that the effect is fairly homogenous, extensive reports on these estimations should not be required. However, the results from these estimations are useful, because they have implications for the generalizability of the main results from the IV models.

Comparison with end-product enhancements as dependent variable

Finally, in this part, we examine the effect of formal contracting on end-product enhancements. This effect is more complex than the effect of formal contracting on cost reductions, and we can more readily see the potential benefits of accounting for essential heterogeneity.

First, it is useful to perform the same analyses as earlier for cost reductions (we do not report Stata code for these analyses; readers can easily write this code by replacing cro (cost reductions) with eeo (end-product enhancements) in the code for cost reductions). These analyses will reveal the following:

- The effect of formal contracting on end-product enhancements, according to the OLS regression, is close to zero (b=0.019, p-value=0.712).

- The instrumental variables work equally well for end-product enhancements as for cost reductions.
- The effect of formal contracting on end-product enhancements is weakly significant when using IV estimation (e.g., when using Moreira's CLR, we obtain: b=0.345, p-value=0.071).

When using Garen's estimator (see Different dependent variable: end-product enhancement: A in the do-file), the results are quite similar to IV estimation, but with a few differences:

- the main effect of formal contracting is slightly stronger (b=0.417, normal-based p-value=0.041, bias-corrected bootstrap confidence interval: [0.050; 0.857])
- there is a significant negative quadratic effect (b=0.097, normal-based p-value=0.048, bias-corrected bootstrap confidence interval: [-0.197; -0.002])
- there is a significant interaction effect between formal contracting and buyer-specific assets (b=0.153, normal-based p-value=0.002, bias-corrected bootstrap confidence interval: [0.060; 0.254])
- the first-stage residual, zhat, has a significant negative relationship with end-product enhancements (b= -0.410, normal-based p-value=0.063, bias-corrected bootstrap confidence interval: [-0.876; -0.013])
- there is a non-significant but positive interaction between formal contracting and the first-stage residual, zhat, in the effect on end-product enhancements (b=0.063, normal-based p-value=0.270, bias-corrected bootstrap confidence interval: [-0.047; 0.176])

However, Garen's estimator is known to rely on stronger assumptions than IV estimation. We can therefore also use De Blander's estimator (see the part of the do-file called Different dependent variable: end-product enhancement: B.). Box WB7 displays the results from using this estimator. We find several differences from and similarities to the previous estimators:

- Differences:
    - The average effect of formal contracting is somewhat higher than when using Garen's estimator (b=0.516, normal-based p-value=0.037, bias-corrected bootstrap confidence interval: [0.130; 1.151])
    - There is a strong significant negative quadratic effect compared with Garen's estimator (b=-0.194, normal-based p-value=0.024, bias-corrected bootstrap confidence interval: [-0.432; -0.076])
    - There is a negative interaction effect between formal contracting and performance ambiguity (b=-0.101, normal-based p-value=0.064, bias-corrected bootstrap confidence interval: [-0.217; -0.007])
    - There is a significant positive interaction between the formal contracting and the first-stage residual, zhat, in the effect on end-product enhancements (b=0.139, normal-based p-value=0.137, bias-corrected bootstrap confidence interval: [0.038; 0.385]).
    - The square of the first-stage residual, orthogonalized with respect to all other observed variables, their interactions and their cross-products, has a significant negative relationship with end-product enhancements (b= -0.126, normal-based p-value=0.188, bias-corrected bootstrap confidence interval: [-0.626; -0.069])
- Similarities:

- Most of the interaction terms involving formal contracting are insignificant, with the exception of the term with performance ambiguity.
- As with Garen's estimator, the first-stage residual, zhat, has a significant negative relationship with end-product enhancements (b= -0.516, normal-based p-value=0.045, bias-corrected bootstrap confidence interval: [-1.163; -0.097])

Note that the results here are based on 4851 bootstrap replications, because 149 replications yielded no results.

Box WB7: Output from using De Blander's (2010) estimator to estimate the effect of formal contracting on end-product enhancements (with 95% asymmetric bias-corrected bootstrap confidence intervals)

```
------------------------------------------------------------------------------
             |    Observed               Bootstrap
             |       Coef.       Bias    Std. Err.  [95% Conf. Interval]
-------------+----------------------------------------------------------------
  md_formcon |   .51573694  -.0128581   .24775968    .1295518   1.151205  (BC)
       fcXfc |  -.19350164    .042503   .08596381   -.4324436  -.0757258  (BC)
      md_bsa |  -.04106978     .00312   .14754897   -.3731087   .2145392  (BC)
      md_ssa |   .18637231  -.0003214   .11389638   -.0488145    .401834  (BC)
       bsasq |   .07271666  -.0109018   .06962548   -.0509652   .2239468  (BC)
       ssasq |   .02082424   .0013961   .06472479    -.099119    .149783  (BC)
      bsassa |  -.08802124   .0173274   .10165129    -.321455   .0790221  (BC)
      md_unc |   .18773149  -.0114164   .08516882    .0402486   .3816325  (BC)
  md_perfamb |  -.08696652   .0034806   .07521185    -.252124   .0484519  (BC)
   md_lnempl |  -.31072462  -.0033951   .11238389   -.5732058  -.1260349  (BC)
    md_knsim |   .07116219  -.0021736   .08244403    -.097133   .2289371  (BC)
md_lnintproc |  -.41500244  -.0141135   .20014464    -.807538  -.0289653  (BC)
md_c_const~c |  -.68491912   -.004453   .38109192   -1.544337  -.0335766  (BC)
  md_c_trade |   .17790491   .0474085   .41912698   -.9880052   .8204784  (BC)
md_c_process |   -.1798427   .0333769   .48604193   -1.456472   .5879469  (BC)
     md_bexp |   .03515251  -.0107239   .06554472   -.0840157   .1706248  (BC)
     md_sexp |   .08703696   .0008287   .07437225   -.0636166   .2284995  (BC)
      fcXbsa |   .04567451  -.0063569   .08725628   -.1381955   .2055357  (BC)
      fcXssa |   .02866298  -.0233635   .08801737   -.1147744   .2377432  (BC)
      fcXunc |  -.01847999  -.0048636   .06415792   -.1383817   .1108731  (BC)
  fcXperfamb |  -.10061475   .0118386    .0544069   -.2165566  -.0070895  (BC)
  fcXcomplex |   .02648477  -.0064751   .06100729   -.0801984   .1665515  (BC)
  fcXhqinflu |   .03352785  -.0058194   .04323133   -.0432811   .1269033  (BC)
    fcXlnval |  -.00110578  -.0185056   .06459753   -.1066688   .1486171  (BC)
        zhat |  -.51573541   .0092142   .25708162   -1.163406   -.097363  (BC)
   newvar190 |  -.12596181   .1130049   .09575355   -.6261857  -.0691451  (BC)
 formconzhat |   .13867515  -.0746697   .09323956    .0358281    .385298  (BC)
     bsazhat |   .18143023  -.0165651   .10691591   -.0084829   .4156351  (BC)
     ssazhat |  -.13097793   .0339326   .11174698   -.3920567   .0472229  (BC)
     unczhat |  -.09291517   .0134277   .09086325   -.2982903    .067896  (BC)
 perfambzhat |   .09979906  -.0107696   .07393147   -.0350196   .2558229  (BC)
 complexzhat |   .03709187   .0086836   .08801727    -.143559   .2017662  (BC)
 hqinfluzhat |   .01782029   .0036729   .05138036   -.0896345   .1132838  (BC)
   lnvalzhat |   .15751184   .0119292    .0670194    .0072004   .2722773  (BC)
       _cons |   4.1112618    .034977   .14876092    3.761472    4.3556  (BC)
------------------------------------------------------------------------------
```

After using De Blander's estimator to examine the effect of formal contracting on end-product enhancements, we find that the effect is heterogeneous and depends on the first-stage residual, i.e., we have a case of essential heterogeneity.

We can explore this finding further by examining the marginal effect of formal contracting on end-product enhancements conditional on the value of the first-stage residual. To do so, we create a Stata program quite similar to the previous one (see the part of the do-file called Different dependent variable: end-product enhancement: B). However, we make two important changes: (1) we calculate the conditional effect of formal contracting for several different values of the zhat, the first-stage residual, and (2) we use 10 000 bootstrap replications to obtain stable bootstrap confidence intervals for the conditional effect along the entire range of zhat. As evident from this code, for each of the 10 000 bootstrap replications, we calculate the conditional effect of formal contracting for 17 different values of zhat ranging from -4 to 4.

The results from running this bootstrap program are similar to the previous results. In addition, we obtain 95% bootstrap confidence intervals for the conditional effect of formal contracting for each of the different values of zhat, the first-stage residual. We present these confidence intervals graphically in Figure WB1 on the next page.

As evident from Figure WB1, there are differences between the normal-based and the bias-corrected confidence intervals, but the main result is similar: for those firms in which the level of formal contracting is lower than predicted by the first-stage regression (i.e., the value of the first-stage residual is negative), the effect of formal contracting on end-product enhancements is not significantly different from zero. For those firms where the level of formal contracting is higher than predicted by the first-stage regression, the conditional effect is positive and significant. This finding indicates positive selection into higher levels of formal contracting, i.e., firms have private knowledge of what effects formal contracting will have on end-product enhancements, and those firms that for some reason face more positive effects of formal contracting will choose higher levels of formal contracting.

This finding is interesting because we do not make similar findings when cost reduction outcomes is the dependent variable. In other words, it means that the choice of formal contracting is, in part, made based on what effects the parties anticipate formal contracting will have on end-product enhancements (and not cost reduction outcomes).

We cannot say for certain what these results mean, because we do not know which variables the first-stage residual reflects. But we can speculate based on the theoretical differences between the two dependent variables cost reductions and end-product enhancements. In their web appendix, Sande & Haugland (2015) present similar findings when using relational contracting as a dependent variable (note that they use a slightly different empirical model from our model here), and they suggest that the parties have private knowledge of how the formal contract will support or undermine the relational contract between the parties. In other words, the first-stage residual reflects, to some extent, the degree to which the formal contract in a particular relationship will support relational contracting between the parties, which indicates that the parties consider the relational contract when writing formal contracts.

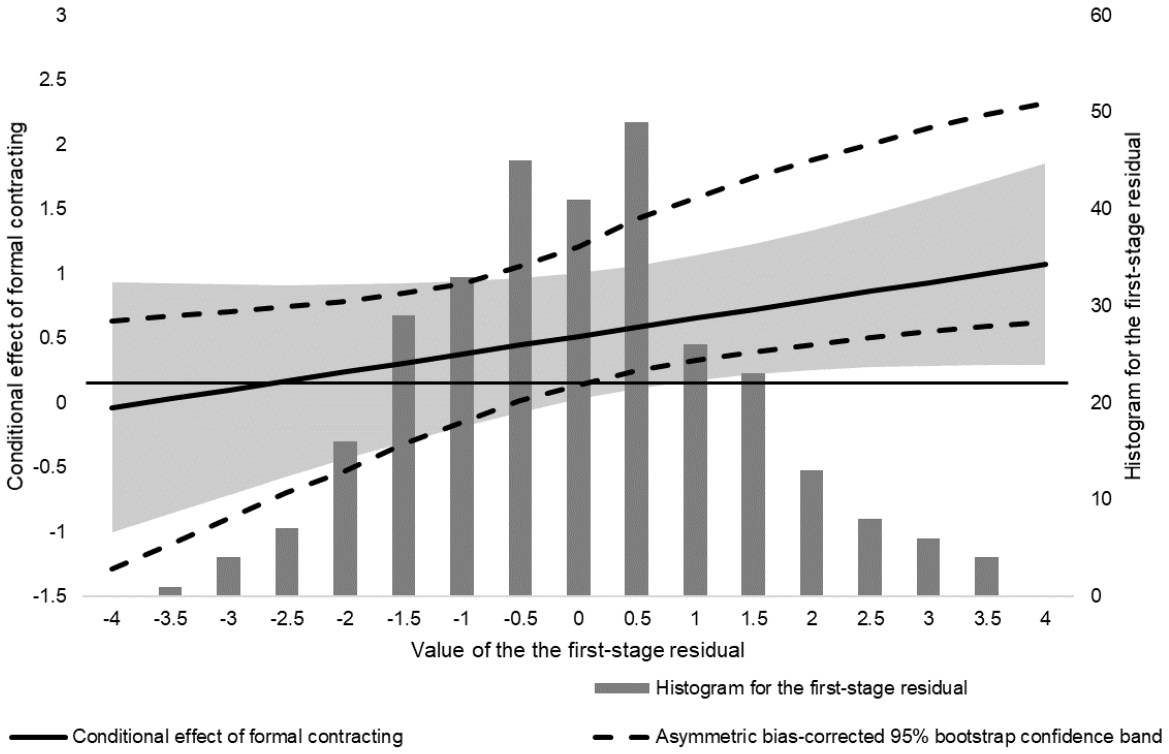


Figure WB1: The conditional effect of formal contracting on end-product enhancements as well as 95% bootstrap confidence intervals (the gray area represents the 95% normal-based bootstrap confidence band, the dotted lines represent the bounds of the 95% asymmetric bias-corrected bootstrap confidence band, and the bars illustrate the histogram for the first-stage residual)

REFERENCES IN WEB APPENDIX B

Bascle, G. (2008). Controlling for endogeneity with instrumental variables in strategic management research. *Strategic Organization, 6*(3), 285-327.

Gulati, R., & Nickerson, J. A. (2008). Interorganizational Trust, Governance Choice, and Exchange Performance. *Organization Science, 19*(5), 688–708.

Mooi, E. A, & Ghosh, M. (2010). Contract Specificity and Its Performance Implications. *Journal of Marketing*, 74(2), 105–120.

Poppo, L., & Zenger, T. R. (2002). Do formal contracts and relational governance function as substitutes or complements? *Strategic Management Journal*, *23*(8), 707–725.

Sande, J. B., & Haugland, S. A. (2015). Strategic performance effects of misaligned formal contracting: The mediating role of relational contracting. *International Journal of Research in Marketing*, 32(2), 187–194.

WEB APPENDIX C: STATA COMMANDS

This document contains Stata commands for Boxes 1 to 5 in Figure 1 in the article. Box 6 can be implemented using techniques described in Box 2. We refer to Roodman (2011) and StataCorp (2017) for details in Box 7.

Table of contents

We use the same name for the various variables and vectors of variables:

Dependent (performance) variable:	*y1*
Endogenous decision variable:	*y2*
Squared term of endogenous variable:	*y2sq*
Vector of control variables:	**x1**
The first variable in the vector of control variables	*x1v1*
The last variable in the vector of control variables	*x1vn*

Vector of instrumental variables for *y2*:	**x2**
Vector of interaction terms between *y2* and control variables:	**y2_x1**
Vector of interaction terms between endogenous variable and instrumental variables:	**y2_x2**
Vector of quadratic of control variables:	**x1sq**
Vector of cross-products of control variables	**x1cp**
Vector of quadratic of instrumental variables:	**x2sq**
Vector of cross-products of instrumental variables:	**x2cp**
Vector of cross-products between control and instrumental variables:	**x1_x2cp**

Note that not all the above variables and vectors of variables are used in all of the estimators.

Some commands are used in several estimators:

`regress`	requests a regression, followed by the dependent variable and a list of explanatory variables
`predict` *predicted_variable*, `xb`	requests the prediction of the dependent variable from the previous regression
`predict` *residual_variable*, `residual`	requests the prediction of the residual from the previous regression
`generate` *new_variable* `=` (expression)	requests the generation of a new variable, followed by an expression, such as *y2*res2*, which multiplies two variables.

Box 1: Instrumental variable estimation (equation 11)

The `ivregress` command:

The built-in IV estimator in Stata is called `ivregress` (StataCorp., 2017a). `Ivregress` can be used to implement 2SLS, LIML (limited information maximum likelihood, which is more robust against weak instruments), and GMM (generalized method of moments, which is more efficient and robust against heteroscedasticity), as follows:

`ivregress 2sls` *y1* **x1** (y2 = **x2**)
`ivregress liml` *y1* **x1** (y2 = **x2**)
`ivregress gmm` *y1* **x1** (y2 = **x2**)

To test for overidentification, use the following post-estimation command:

`estat overid`

After 2SLS, `overid` will report Sargan's (1958) and Basmann's (1960) chi-square tests. After LIML, `overid` will report Anderson and Rubin's (1950) chi-square test and Basman's F-test. After GMM, `overid` will report Hansen's (1982) J-statistic chi-square test.

Entering the option `first`, as in "`ivregress 2sls` *y1* **x1** (y2 = **x2**), `first`", instructs Stata to also report the first-stage regression. The following post-estimation command gives additional useful statistics on the relevance of the IVs:

```
estat firststage
```

It provides the Cragg-Donald Wald F-statistic and the relevant critical values for testing 2SLS relative bias and 2SLS or LIML maximal size.

Another useful post-estimation command is
`estat endogenous`

This command instructs Stata to test whether the endogenous variable(s) is/are actually exogenous. After 2SLS, Stata will report Durbin's (1954) and Wu-Hausman's (Wu 1974, Hausman 1978) statistics. After GMM, Stata will report the C-statistic (difference-in-Sargan). This post-estimation command is not available after LIML.

The `ivreg2` command:

The benefit of using `ivregress` is that it is a built-in command in Stata, supported by Stata. However, the user-written command `ivreg2`, by Baum, Schaffer, and Stillman (2003, 2007), has several additional features. `Ivreg`2 is implemented as follows:

`ivreg2` *y1* (*y2* = **x2**) **x1**, `first orthog(`name_of_variable_to_assess`) endog(`*y2*`)`

As with `ivregress`, in `ivreg2`, the option `first` instructs Stata to display the first-stage regression. One particular benefit of using `ivreg2` is that the option `orthog` enables the researcher to assess the exogeneity of individual IVs or subsets of IVs. After the `orthog`-option, Stata will report the C-statistic for the IV or subset of IVs in question as well as the chi-square statistic for the entire model after removing the IVs under evaluation. The `endog`-option instructs Stata to report tests of whether the endogenous variable is in fact exogenous.

Moreover, `ivreg2` implements several different estimation procedures, including LIML, two-step GMM, and Fuller's (1977) modified LIML, as follows:

```
ivreg2 y1 (y2 = x2) x1, liml
ivreg2 y1 (y2 = x2) x1, gmm2s
ivreg2 y1 (y2 = x2) x1, fuller(4)
```

Fuller's (1977) modified LIML requires the researcher to supply an unknown parameter, usually 1 or 4. See Bascle (2008) for more details.

Finally, another benefit of `ivreg2` is that it allows us to run the following postestimation command:

```
ivhettest
```

`ivhettest` performs Pagan and Hall's (1983) test of heteroscedasticity for IV estimation.

The `condivreg` command:

If the IVs are weak, we should use other estimators than 2SLS, such as LIML or Fuller's (1977) modified LIML, which are both partly robust to weak IVs, or Moreira's (2003) conditional likelihood ratio (CLR) estimator, which is *fully* robust to weak IVs. Moreira's CLR is implemented through the user-written command `condivreg` (Moreira & Poi, 2003):

```
condivreg y1 (y2 = x2) x1
```

If the IVs are weak, researchers should always compare their results from other estimators with those from Moreira's CLR, because it draws correct inferences regardless of the strength of the IVs (Bascle, 2008).

Interaction terms in 2SLS using the `ivregress`- and `ivreg2` commands

The following syntax tests a model where *y2* interacts with all the n variables in the vector of control variables **x1**. We use the interaction terms between *y2hat* and **x1** as instruments for **y2_x1**. An alternative is to use various quadratic and interaction terms involving **x1** and **x2** as instruments for **y2_x1**.

The following syntax requests the estimation of Equation 12 and the prediction of *y2*.

```
regress y2 x1 x2
predict y2hat, xb
```

The following syntax requests the generation of interaction terms between *y2hat* and all the variables in **x1**. We label the vector of these n interaction terms as **y2hat_x1**.

```
generate y2hat_x1v1 = y2hat*x1v1
…
generate y2hat_x1vn = y2hat*x1vn
```

The following syntax requests the estimation of Equation 13 enhanced with interaction terms between *y2* and all the control variables in **x1** (using either `ivregress` or `ivreg2`):

```
ivregress 2sls y1 x1 (y2 y2_x1 = x2 y2hat_x1), first
estat overid

ivreg2 y1 (y2 y2_x1 = x2 y2hat_x1) x1, first
orthog(name_of_variable_to_assess)
```

Box 2: Control function estimators

Garen's estimator (equation 16) with bootstrap:

Variables generated through the two-step procedure:
First-stage residual *res2*
Interaction term between endogenous decision variable and first-stage residual: *y2_res2*

The following syntax creates a small program called "`garen_estimator`" that returns the

estimates from using Garen's estimator.

```
program garen_estimator, eclass
version 15.0
tempname b V
tempvar res2 y2_res2
capture drop res2 y2_res2
regress y2 x1 x2
predict res2, residual
generate y2_res2 = y2*res2
regress y1 y2 y2sq x1 res2 y2_res2 y2_x1
matrix b=e(b)
matrix V=e(V)
ereturn post b V
end
```

The following syntax requests a bootstrap routine that draws 1000 subsamples with replacement and runs the "garen_estimator"-program on each of the sub-samples. Confidence intervals are based on the 1000 sets of estimates from these sub-samples. This procedure accounts for how two of the variables in the second-stage performance equation (*res2* and *y2_res2*) are generated based on the previous regression estimation.

```
bootstrap _b , reps(1000) level(95) seed(10101) nodots:
garen_estimator
```

The following syntax returns bias-corrected bootstrap confidence intervals:

```
estat bootstrap, bc
```

The following syntax returns normal-based, percentile, and bias-corrected bootstrap confidence intervals.

```
estat bootstrap, all
```

Note that a test of the joint significance of the terms that include the first-stage residual (*res2* and *y2_res2*) is a test of endogeneity.

De Blander's estimator (equation 18) with bootstrap:

Variables generated through the two-step procedure:	
First-stage residual	*res2*
First-stage residual squared	*res2sq*
Interaction term between endogenous decision variable and first-stage residual:	*y2_res2*
Vector of interaction terms between first-stage residual and control variables:	**res2_x1**
Vector of interaction terms between first-stage residual and instrumental variables:	**res2_x2**

Vector of d orthogonalized variables produced by the `orthog`-command. [The variables are named *newvar1*, *newvar2*,.., *newvard*, where the last variable (the d'th variable) is the orthogonalized squared first-stage residual.] **newvar**

The orthogonalized squared first-stage residual (the last variable in newvar. [The *d* in *newvard* refers to the number of variables in the **newvar** vector of variables. Hence, when writing up the code, the analyst must check how many variables will be included in **newvar**. If **newvar** includes, for example, 70 variables, the name of the orthogonalized squared first-stage residual will be *newvar70*.] *newvard*

The following syntax creates a small program called "`blander_estimator`" that returns the estimates from using De Blander's estimator. Next follows the bootstrap, similarly to the procedure for Garen's estimator. Note that if **newvar** includes many variables, the `orthog` command will be time consuming. Note also that it is important that *res2sq*, the squared first-stage residual, is placed after all the other terms when specifying the `orthog` command, because then *res2sq* will be orthogonalized with respect to all the other variables in **x1**, **x1sq**, **x1cp**, **x2**, **x2sq**, **x2cp**, and **x1_x2cp**.

```
program blander_estimator, eclass
version 13.1
tempname b V
tempvar res2 res2sq y2_res2 res2_x1 res2_x2 newvar
capture drop res2 res2sq y2_res2 res2_x1 res2_x2 res_res2sq newvar
regress y2 x1 x2
predict res2, residual
generate res2sq = res2*res2
generate y2_res2 = y2*res2
generate res2_x1 = res2*x1 *Note that this expression repeats for each variable in x1*
generate res2_x2 = res2*x2 *Note that this expression repeats for each variable in x2*
orthog x1 x1sq x1cp x2 x2sq x2cp x1_x2cp res2sq, generate(newvar*)
reg y1 y2 y2sq x1 res2 y2_res2 y2_x1 res2_x1 y2_x2 res2_x2 newvard
matrix b=e(b)
matrix V=e(V)
ereturn post b V
end
bootstrap _b , reps(1000) level(95) seed(10101) notdots:
blander_estimator
```

The following syntax requests only bias-corrected bootstrap confidence intervals.

```
estat bootstrap, bc
```

The following syntax requests normal-based, percentile, and bias-corrected bootstrap confidence intervals.

```
estat bootstrap, all
```

Note that a test of the joint significance of the terms that include the first-stage residual (*res2*,

y2_res2, **res2_x1**, and **res2_x2**) is a test of endogeneity.

Box 3: IV estimators for discrete variables

2SLS:

Note that in the following, *y2* is a binary variable, and the expression "`predict p, p`" generates a new variable *p*, which is the propensity score. The first stage does not have to be a probit, a logit can also be used.

```
probit y2 x1 x2
predict p, p
ivregress 2sls y1 x1 (y2 = p)
```

Other estimators available in the `etregress` command:

An alternative procedure for IV estimation in Stata when facing a binary variable is to use the `etregress` command, which is a built-in Stata command (StataCorp., 2017b). As opposed to 2SLS, `etregress` opens up for a correlation between the first- and second-stage residuals to control for the endogeneity of the binary endogenous variable. The following syntax can be used, and Stata will then use a maximum likelihood estimator:

```
etregress y1 x1, treat(y2 =x1 x2)
```

The coefficient for *y2* in the outcome function is the average treatment effect. The confidence interval for the rho-coefficient (which is the correlation between the residuals (errors) in the first-stage probit and the errors in the second-stage outcome regression) indicates whether endogeneity is a problem.

With `etregress`, other estimators can be used, including a one-step control function estimator (a GMM estimator obtained by the option `cfunction`) and a two-step control function estimator (obtained by the option `twostep`):

```
etregress y1 x1, treat(y2 =x1 x2) cfunction
etregress y1 x1, treat(y2 =x1 x2) twostep
```

2SLS with interaction terms:

```
probit y2 x1 x2
predict p, p
generate p_x1=p*x1          *Note that this expression repeats for each variable in x1*
ivregress 2sls y1 x1 (y2 y2_x1= p p_x1)
```

Box 4: Selection models

In general, a difficulty when estimating selection models is that rather many commands could potentially be used: a manual procedure, the `etregress` command, the `ivtreatreg` command, the `margte` command, and the `oheckman` command. Depending on what option is

used, they can all produce more-or-less identical results (given a binary endogenous explanatory variable). First, we present a quick overview of the commands:

manual procedure:	This is a two-step estimator that you program yourself in Stata. It will directly estimate Equations (23), (24) or (25) in the article. We use the bootstrap to correct the standard errors. A disadvantage of this procedure is that it is a bit cumbersome.
`etregress`	This is the built-in Stata command. It will estimate Equation (23) but not Equations (24) and (25). In addition to the two-step procedure, we can use maximum likelihood. In general, the `etregress` command has many options and possibilities.
`ivtreatreg`	This procedure is in many respects very similar to `etregress`, but it includes several more options that we can compare.
`margte`	`margte` does not estimate Equation (23) but Equations (24) and (25). The principal advantage of `margte` is that it can be used to estimate the marginal treatment effect.
`oheckman`	This command can produce identical estimates to the manual procedure, Equations (24) and (25). But the principal advantage of the `oheckman` command is that it can be used when the endogenous explanatory variable is discrete with multiple values.

A manual procedure to estimate equations (23), (24), and (25), Heckman two-step selection model:

In the following, we outline a manual procedure, a program we create in Stata, to estimate equation (23), i.e., the Heckman two-step selection model:

```
program selection_model, eclass
version 13.1
tempname b V
tempvar lp invmills y2_invmills
capture drop lp invmills y2_invmills
probit y2 x1 x2
predict lp, xb
generate invmills = y2*(normalden(lp)/(normal(lp)))-(1-
y2)*(normalden(lp)/(1-normal(lp)))
generate y2_invmills = y2*invmills
regress y1 y2 x1 invmills y2_invmills y2_x1
matrix b=e(b)
matrix V=e(V)
ereturn post b V
end
bootstrap _b, reps(1000) level(95) seed(10101) notdots:
```

```
selection_model
```

The following syntax requests only bias-corrected bootstrap confidence intervals.

```
estat bootstrap, bc
```

The following syntax requests normal-based, percentile, and bias-corrected bootstrap confidence intervals.

```
estat bootstrap, all
```

Note that the above program estimates equation (23) in the article, but it is easy to replace equation (23) with either equation (24) or (25), as follows:

Program for estimating Equation 24 (*y2* is equal to 1):

```
program selection_model_24, eclass
version 13.1
tempname b V
tempvar  lp invmills y2_invmills
capture drop  lp invmills y2_invmills
probit  y2  x1  x2
predict  lp, xb
generate  invmills  =  y2*(normalden(lp)/(normal(lp)))-(1-
y2)*(normalden(lp)/(1-normal(lp)))
generate  y2_invmills  =  y2*invmills
regress  y1  y2  x1  invmills  if y2==1
matrix b=e(b)
matrix V=e(V)
ereturn post b V
end
bootstrap _b, reps(1000) level(95) seed(10101) notdots:
selection_model_24
```

Program for estimating Equation 25 (*y2* is equal to 0):

```
program selection_model_25, eclass
version 13.1
tempname b V
tempvar  lp invmills y2_invmills
capture drop  lp invmills y2_invmills
probit  y2  x1  x2
predict  lp, xb
generate  invmills  =  y2*(normalden(lp)/(normal(lp)))-(1-
y2)*(normalden(lp)/(1-normal(lp)))
generate  y2_invmills  =  y2*invmills
regress  y1  y2  x1  invmills  if y2==0
matrix b=e(b)
```

```
matrix V=e(V)
ereturn post b V
end
bootstrap _b, reps(1000) level(95) seed(10101) notdots:
selection_model_25
```

Using the built-in `etregress` command to estimate selection models:

Stata's built-in `etregress` command can also be used to estimate selection models. Above (in the section concerning Box 3), we presented syntax for a constrained model, which assumes that there is no essential heterogeneity. However, `etregress` allows parameters to vary depending on the endogenous binary variable. A major benefit of the `etregress` command is that it is a built-in command supported by Stata.

The unconstrained model can be specified as follows, and it will provide the same parameter estimates as the manual procedure and the `ivtreatreg-` (with the `heckit`-option) and `margte-` (with the `bsopts`-option) commands (which we will describe shortly), because we allow for the endogenous binary variable to interact with the observed variables and the first-stage generalized residual (called "hazard" in StataCorp's (2017b) reference manual, p.58):

`etregress` *y1* i.*y2* **x1** i.*y2*#c.**x1**, `treat(`*y2* = **x1 x2**`) cfunction poutcomes`

Note that i.*y2*#c.**x1** means that we create interaction terms between *y2* and each of the control variables in **x1**. Therefore, we must repeat this term for each interaction. For example, if we have three variables in **x1** – *x11, x12,* and *x13* – we must write i.*y2*#c.*x11* i.*y2*#c.*x12* i.*y2*#c.*x13*.

The `cfunction`-option specifies that a one-step control function (performed by using the GMM) will be used to estimate the parameters, standard errors and covariance matrix. The `poutcomes`-option specifies that Stata will use a potential outcomes model with different variance and correlation parameters across the different values of the binary endogenous variable, i.e., the effect of the endogenous variable is allowed to vary depending on unobserved heterogeneity. The standard errors will not be similar to the manual procedure.

The above model can also be estimated with maximum likelihood and is specified as follows:

`etregress` *y1* i.*y2* **x1** i.*y2*#c.**x1**, `treat(`*y2* = **x1 x2**`) poutcomes`

This model will not yield the same parameter estimates. In general, if the model is correctly specified, it will be more efficient. We refer to StataCorp (2017) for further details.

Using the user-written `ivtreatreg` command to estimate equation (23) (Heckman two-step selection model):

Another alternative when estimating selection models is to use the user-written command `ivtreatreg` by Cerulli (2014). The following command will produce the same parameter estimates as the manual procedure and `etregress` with the `cfunction`- and `poutcomes`-options (but the standard errors are slightly different):

```
ivtreatreg y1 y2 x1, hetero(x1) model(heckit) iv(x2)
```

Although the output from this model is similar to the output from the manual procedure, the inverse Mills ratios enter in a slightly differently way. In the manual model, the inverse Mills ratios enter as a generalized residual that interacts with the endogenous binary variable. In `ivtreatreg`, the inverse Mills ratios enter separately (i.e., there is one parameter for each of the inverse Mills ratios).

Note that in addition to `heckit` (the Heckman two-step selection model), `ivtreatreg` gives the possibility to fit three other binary treatment models: `direct-2sls` (IV regression fit by direct two-stage least squares), `probit-2sls` (IV regression fit by probit and two-stage least squares), and `probit-ols` (IV two-step regression fit by probit and OLS). Using several of these models can be useful for robustness checks. We refer to Cerulli (2014) for further details.

Using the user-written command `margte` to estimate selection models and marginal treatment effects:

The user-written command `margte`, by Brave and Walstrum (2014), can also be used to estimate the Heckman two-step selection model, and it will produce the same parameter estimates as the manual procedure and `ivtreatreg`. In contrast to the previous ones, `margte` will produce outputs for equations (24) and (25). In addition, `margte` automatically estimates the ATE and MTE for all values of the propensity score and creates graphs that illustrate how the MTE varies with U_{2i}.

The Heckman two-step selection model is in `margte` called the 'parametric normal procedure'. It is in almost every respect identical to the manual procedure, except that (1) the sign for the effect of the inverse Mills ratio is the reverse of what we used in the manual procedure (i.e., the two inverse Mills ratios have a correlation coefficient of -1), (2) the outputs include an estimate of the average treatment effect, and (3) the outputs include a graph plotting the marginal treatment effect for different values of the propensity to not take the treatment (i.e., the probability that *y2* is 0 given **x1** and **x2**).

```
margte y1 x1 , treatment(y2 x1 x2 ) first bsopts(reps(1000))
```

The output also provides estimates of the effects of the inverse Mills ratios, and they are in almost all respects identical to those obtained with the two-step selection model with the bootstrap described above.

The following syntax requests only bias-corrected bootstrap confidence intervals.

```
estat bootstrap, bc
```

The following syntax requests normal-based, percentile, and bias-corrected bootstrap confidence intervals.

```
estat bootstrap, all
```

The following syntax is quite similar to the previous alternative but relies on stronger assumptions. The estimates are similar to those obtained with the `movestay` command, and they utilize maximum

likelihood.

```
margte y1 x1 , treatment(y2 x1 x2 ) first ml
```

Above, we describe how `margte` can be used to estimate selection models. However, the main purpose of `margte` is to estimate marginal treatment effects (MTEs). Below, we present the syntax for the parametric polynomial. In this case, we allow for logit rather than probit as the link function and a fourth-order polynomial expansion of the propensity score. Without including `link(logit)`, the link function would be a probit (probit is the default).

```
margte y1 x1 , treatment(y2 x1 x2 ) polynomial(4) link(logit)
```

Local instrumental variables and semiparametric estimation are beyond the scope of this article. However, `margte` can also be used for this purpose, as described by Brave and Walstrom (2014).

We refer to Brave and Walstrum (2014) for further details.

Using the user-written `oheckman` command to estimate ordered probit selection models

Equations (24) and (25) in the article are written for the situation when we have a binary endogenous explanatory variable (e.g., handshake vs. formal contracting). Sometimes, however, we are interested in estimating selection models where the endogenous variable is not binary but perhaps has three or more values. We could, for example, be interested in a variable that takes on the following values: 0=handshake, 1=formal contract, 2=vertical integration. In this case, we would have to create three equations. In such cases, Chiburis and Lokshin's (2007) user-written `oheckman` command can be useful.

The `oheckman` command will estimate two or more outcome equations, one for each value of the ordered discrete endogenous explanatory variable. In the first-stage, `oheckman` estimates a probit regression. `oheckman` will compute estimates using either a two-step procedure or maximum likelihood.

The full information maximum likelihood option is the default option and will be reported with the following command:

```
oheckman y1 x1, select(y2 = x1 x2 )
```

The following command will execute the two-step procedure:

```
oheckman y1 x1, select(y2 = x1 x2 ) twostep
```

Note that if we use a binary endogenous explanatory variable here, `oheckman` with the two-step option will return identical parameter estimates to the manual procedure and to what `margte` returns when we use the `bsopts`-option. Likewise, with a binary endogenous explanatory variable, `oheckman`'s default full information maximum likelihood option will produce the same results as `margte` when we use the `ml`-option.

We refer to Chiburis and Lokshin (2007) for further details on the `oheckman` command and results from Monte Carlo simulations that compare the performance of the full information maximum likelihood and the two-step estimator under different conditions.

Box 5: IVs in an SEM model, as described by Muthén and Jöreskog (1983)

We illustrate the Stata code for an SEM model below using Stata's `sem` command (StataCorp., 2017c). Suppose we have measured eight variables, each using three indicators. There is one dependent variable, Y1; we want to estimate the effect of Y2 on Y1; and Y2 is possibly endogenous. We have three control variables, X11, X12, and X13, and three IVs, X21, X22, and X23, as illustrated below:

Variables:	
Latent dependent variable:	Indicators:
Y1	y11, y12, y13
Latent endogenous explanatory variable:	Indicators:
Y2	y21, y22, y23
Latent control variables:	Indicators:
X11, X12, X13	x111 – x113, x121 – x123, x131 –x133
Latent instrumental variables:	Indicators:
X21, X22, X23	x211 – x213, x221 – x223, x231 – x233

An IV-model using the `sem` command in Stata can be set up as follows in this case:

```
sem (Y2 -> Y1,) (X11 -> Y1,) (X12 -> Y1,) (X13 -> Y1,)  ///
(X11 -> Y2,) (X12 -> Y2,) (X13 -> Y2,)                  ///
(X21 -> Y2,) (X22 -> Y2,) (X23 -> Y2,)                  ///
(Y1 -> y11,) (Y1 -> y12,) (Y1 -> y13,)                  ///
(Y2 -> y21,) (Y2 -> y22,) (Y2 -> y23,)                  ///
(X11 -> x111,) (X11 -> x112,) (X11 -> x113)             ///
(X12 -> x121,) (X12 -> x122,) (X12 -> x123)             ///
(X13 -> x131,) (X13 -> x132,) (X13 -> x133)             ///
(X21 -> x211,) (X21 -> x212,) (X21 -> x213)             ///
(X22 -> x221,) (X22 -> x222,) (X22 -> x223)             ///
(X23 -> x231,) (X23 -> x232,) (X23 -> x233),            ///
covstruct(_lexogenous, diagonal)                        ///
cov(_lexogenous*_oexogenous@0) nomeans latent(Y1 Y2 X11 ///
X12 X13 X21 X22 X23) cov(e.Y1*e.Y2 e.X11*e.X12          ///
e.X11*e.X13 e.X11*e.X21 e.X11*e.X22 e.X11*e.X23         ///
e.X12*e.X13 e.X12*e.X21 e.X12*e.X22 e.X12*e.X23 e.X13*e.X21   ///
e.X13*e.X22 e.X13*e.X23 e.X21*e.X22 e.X21*e.X23 e.X22*e.X23)  ///
nocapsulent
```

The above model is over-identified because there are multiple items per latent variable, and the structural model is overidentified compared to the measurement model. The reason is that X21, X22, and X23 only affect Y2 and not Y1. To test the exogeneity condition, we can compare the chi-square statistic for this model with that for the measurement model. We can test individual overidentifying restrictions by using the post-estimation command `estat mindices` (a score test) or by opening each of the paths from X21, X22, and X23 to Y1 and compare the change in chi-square (a likelihood

ratio test). We control for the endogeneity of Y2 by specifying a covariance between the error terms of Y1 and Y2, `e.Y1*e.Y2`. If this covariance is significant, it is an indication that we must correct for endogeneity.

REFERENCES IN WEB APPENDIX C

Anderson, T. W., & Rubin, H. (1950). The Asymptotic Properties of Estimates of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *The Annals of Mathematical Statistics*, 21(4), 570–582.

Bascle, G. (2008). Controlling for endogeneity with instrumental variables in strategic management research. *Strategic Organization*, 6(3), 285-327.

Basmann, R. L. (1960). On finite sample distributions of generalized classical linear identifiability test statistics. *Journal of the American Statistical Association*, 55(292), 650–659.

Baum, C. F., Schaffer, M. E., & Stillman, S. (2003). Instrumental variables and GMM: Estimation and testing. *The Stata Journal*, 3(1), 1–31.

Baum, C. F., Schaffer, M. E., & Stillman, S. (2007). Enhanced routines for instrumental variables/generalized method of moments estimation and testing. *The Stata Journal*, 7(4), 465–506.

Brave, S., & Walstrum, T. (2014). Estimating marginal treatment effects using parametric and semiparametric methods. *The Stata Journal*, 14(1), 191–217

Cerulli, G. (2014). ivtreatreg: A command for fitting binary treatment models with heterogeneous response to treatment and unobservable selection. *The Stata Journal*, 14(3), 453–480

Chiburis, R., & Lokshin, M. (2007). Maximum likelihood and two-step estimation of an ordered-probit selection model. *Stata Journal*, *7*(2), 167-182.

Durbin, J. (1954). Errors in Variables. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 22(1/3), 23-32

Fuller, W. A. (1977). Some properties of a modification of the limited information estimator. *Econometrica*, 45(4), 939–953

Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica*, 46(6), 1251-1271

Moreira, M. J. (2003). A conditional likelihood ratio test for structural models. *Econometrica : Journal of the Econometric Society*, *71*(4), 1027–1048.

Moreira, M. J., & Poi, B. P. (2003). Implementing tests with correct size in the simultaneous equations model. *The Stata Journal*, 3(1), 57–70

Roodman, D. (2011). Fitting fully observed recursive mixed-process models with cmp. *The Stata Journal*, *11*(2), 159–206

Sargan, J. D. (1958). The Estimation of Economic Relationships using Instrumental Variables. *Econometrica*, *26*(3), 393–415.

StataCorp. (2017a). Stata base reference manual. *In Stata: Release 15. Statistical Software* (pp. 1–

2985). College Station, Texas: Stata Press.

StataCorp. (2017b). Stata treatment effects reference manual: potential outcomes/counterfactual outcomes. *In Stata: Release 15. Statistical Software (pp. 1–327).* College Station, Texas: Stata Press.

StataCorp. (2017c). Stata structural equation modeling reference manual. *In Stata: Release 15. Statistical Software* (pp. 1 – 659). College Station, Texas: Stata Press.

Wu, D.-M. (1974). Alternative Tests of Independence between Stochastic Regressors and Disturbances: Finite Sample Results. *Econometrica*, 42(3), 529–546.