Technical Note: Recommendations for Assessing Unit Non-response Bias in Dyadic Focused Empirical Supply Chain Management Research

Abstract

The last decade has seen an increase in empirical Supply Chain Management (SCM) research with dyadic data. Such data structures can further complicate the assessment of non-response bias, which plays a key role in establishing the credibility of research results. A survey of 75 research papers with dyadic data, published in five empirically-focused SCM academic journals, over the last decade, reveals a lack of agreement on methods used in the assessment for potential *unit* non-response bias. Of the various statistical tests found, only the Multivariate Analysis of Variance (MANOVA) approach allows for a single statistical test to be utilized in assessing for potential unit non-response bias via incorporation of the design structure of the dyadic data. We investigate the use of an *effect size* confidence interval coverage, of a MANOVA, to detect a meaningful difference between respondents and non-respondents correctly. Our results show that with dyadic data, such meaningful differences can be detected with significantly smaller sample size requirements than traditional approaches such as *t*-tests or ANOVA. Recommendations are provided for setting up and executing a MANOVA to assess for potential *unit* non-response bias with dyadic data.

Keywords: Dyadic data, unit non-response bias, effect size analysis, survey research methods, supplier-customer relationships

1. Introduction

In a business-to-business context, dyadic relationships between buying and selling firms are crucial. Over the past 20 years, there has been an increase in Supply Chain Management (SCM) research in which both the supplier's and buyer's view in a buyer-supplier relationship is of interest (i.e., dyad). Data collected from individual units, which are then paired in a meaningful way (e.g., suppliers and buyers; carriers and shippers), in an empirical study, is known as *dyadic data*. Results from a carefully conducted empirical dyadic study can be used to make statements about a population, which are based on information that is obtained from a sample of that population. However, before a researcher can generalize the findings of such a study, the quality of information based on the data collection must be assessed. A key issue that can compromise the quality of such empirical studies is non-response bias.

The assessment of non-response bias in empirical studies plays an important role in establishing the credibility of research results. Even when a clear *target population* has not been identified for a survey, the use of a sampling frame (i.e., list of units to sample), with some units not responding, could still lead to non-response bias concerns (Lohr, 2010: 257). There are two types of non-response. *Item* non-response refers to the absence of answers to specific questionnaire items, whereas *unit* non-response refers to the complete absence of responses from a sampled unit (Lohr, 2010; Wagner & Kemmerling, 2010). Item non-response is typically treated as a case of missing data with the associated techniques (e.g., imputation) for handling missing data applied to such non-response occurrences (Collier & Bienstock, 2007; Lohr, 2010). In this study, we are interested in *unit* non-response bias. Unit non-response is considered to pose a much greater threat to survey research than item non-response, due to the bias typically being much larger than that of item non-response (Yan & Curtin, 2010).

There are key differences between item and unit non-response, both in terms of the underlying causes and potential treatments (Lohr, 2010; Yan & Curtin, 2010). For example, assessments for potential item non-response bias (e.g., missing data bias) can be made without the need to survey additional units in the sampling frame who did not respond to the initial survey (Little & Rubin, 2002). Generally, assessments made for potential unit non-response bias require additional units to be sampled, via a follow-up telephone call or face-to-face interview, to collect information on some, but not necessarily all, key items on the survey (Lohr, 2010; Thompson & Washington, 2013). Likewise, if there is a concern of potential bias, adjustments for unit non-response bias are made at the unit level (i.e., responses for all items are adjusted by typically using sample weights), whereas adjustments for item non-responses are made at the item level (i.e., units with complete responses on all items would not receive any adjustments). Making adjustments for potential unit non-response bias, however, is not a trivial task (e.g., see Thompson & Washington, 2013 for a case-study example) and may result in further biased estimates if improper adjustment procedures are employed (Lohr, 2010; Thompson & Washington, 2013). It is, therefore, important for researchers to *first* assess how well the responses of those in-sample represent those of the out-of-sample respondents prior to deciding to perform an adjustment for potential unit non-response bias.

For decades, in supply chain management research and other allied areas, the approach to assessing for potential unit non-response bias has been statistical methods that involve the comparison of responses from in-sample versus out-of-sample respondents, on multiple characteristics, which are relevant to the study (Clottey & Benton, 2013; Collier & Bienstock, 2007; Lambert & Harrington, 1990; Wagner & Kemmerling, 2010; Werner *et al.*, 2007). If the responses of the two groups differ by very much, then it is good evidence that unit non-response cannot be ignored, and therefore adjustments to the data may be required to accommodate for potential unit non-response bias (Wagner & Kemmerling, 2010; Whitehead *et al*., 1993). Usually, multiple statistical tests (e.g., *t*-tests or ANOVA) are performed, and unit non-response bias is considered not an issue if *all* the tests result in non-significance. For non-dyadic data, Clottey and Benton (2013) noted that having multiple tests leads to the classical problem in statistics in which the more inferences are made, the more likely that erroneous inference has occurred. An argument can be made that dyadic data can further complicate non-response bias assessment due to the introduction of additional correlations (e.g., between the responses of members from the same dyad) in the analysis, along with an increase in the number of statistical tests performed (e.g., when separate tests are utilized for each member of the dyad). Clottey and Benton (2013: 801) recommend that researchers adopt methods, to assess for potential non-response bias, which minimizes the number of separate statistical tests needed to be performed. Our research expands on this suggestion by considering the use of a Multivariate Analysis of Variance (MANOVA) approach that allows for a single statistical test to be utilized for the assessment with dyadic data.

Due to the need to sample additional units when assessing and making adjustments for potential unit non-response bias, researchers are faced with a significant challenge in this area as survey participants are becoming harder to find. This is all the more acute in the context of a dyadic study where the time needed to recruit/sample both members of a dyad along with budgeting issues, such as incentives for both members of the dyad and follow-up expenses, can be challenging (Quinn *et al*., 2010). Therefore, an approach that is used to assess/adjust for potential unit non-response bias while also minimizing the number of out-of-sample respondents required would be of substantial benefit to researchers who work with dyadic data.

By choosing an appropriate statistical method that makes better use of the dyadic data structure in the assessment for potential non-response bias, a researcher may be able to perform such an assessment without the need for multiple testing or the ‘large’ number of non-respondents required with the traditional approaches. Thus, the objective of this study is to provide consistent guidelines for conducting a statistical test, to assess for potential unit non-response bias with dyadic data structures. Before addressing this objective, we first had to gain insights into the approaches currently used with SCM research to assess for unit non-response bias with dyadic data.

2. Current Methods Used to Assess for Potential Unit Non-Response Bias with Dyadic Data

The *SCM Journal List*™ is an annual ranking of universities’ supply chain management research output, based on the leading supply chain management journals. According to the list, the top *empirically* focused SCM journals are: Journal of Operations Management (*JOM*), Decision Sciences (*DS*), Journal of Business Logistics (*JBL*), and the Journal of Supply Chain Management (*JSCM*). A foundational goal of SCM is the integration of supply (e.g., logistics and operations) and demand (e.g., marketing), as defined by the Council of Supply Chain Management Professionals (CSCMP). Thus, empirical SCM work is also found in leading empirical marketing journals. We, therefore, added the Journal of Marketing (*JM*) and the Journal of Marketing Research (*JMR*) to our list of journals for consideration. The Journal of Marketing, the Journal of Marketing Research, and the Journal of the Academy of Marketing Science are among the top empirically focused marketing journals regarding their impact on the discipline (Collier & Bienstock, 2007; Hult *et al*., 1997; Zinkhan, 2003). Of these three journals, we found that the *Journal of Marketing* was the main publication outlet for survey-based research involving dyadic data.

2.1. Research method

The context for this study is survey-based research involving a dyad in the analysis. We conducted a survey of research articles with dyadic data over the past decade in *JOM* (vols. 25 to 52), *DS* (vols. 38 to 48), *JBL* (vols. 28 to 38), *JSCM* (vols. 43 to 53), and *JM (*vols. 71 to 81). The dataset consisted of articles in the Science Direct, American Marketing Association, and Wiley Online databases. The last decade has seen calls (see Flynn *et al*., 2018) for the use of multiple respondents (e.g., dyads) in empirical supply chain research. While such research has been prevalent on the marketing side for several decades, it is a recent trend in other SCM areas (Simpson *et al*., 2015). Thus, our decision to examine empirical articles published in the recent decade. The "dyadic data" keyword was used to identify articles that may have involved the use of dyadic data in the analysis. This resulted in a total of 344 articles across the five journals. We searched each of the articles to ensure that: (i) analysis had been performed with dyadic data, and ii) that unit non-response bias had been assessed. Of the initial 344 articles, a total of 75 met this criterion. From these 75 total articles, if provided, we recorded the statistical method used to assess for unit non-response bias, the literature reference cited in support of the statistical method, and the average and the median number of survey items and total sample sizes used in the statistical tests. The results are shown in Table 1 below.

Table 1: Unit Non-response Bias Assessment in Studies Involving the Analysis of Dyadic Data in *Decision Sciences* (DS), *Journal of Business Logistics* (JBL), *Journal of Operations Management* (JOM), *Journal of Marketing* (JM), and *Journal of Supply Chain Management* (JSCM); 2007- 2017

Journal	(a) No. of articles with key-word dyadic data	(b) No. of articles from (a) with unit non-response bias assessment involving dyadic data	(c) Average total sample sizes reported in articles from (b)	(d) Statistical methods used to assess unit non-response bias in articles from (b)*	(e) Cited references for methods used to assess unit non-response bias in articles from (b)
DS	47	16	240.5	Chi-squared only (1) *F*-test/ANOVA (3) *t*-tests only (4) *t*-test & Chi-squared (2)	Armstrong & Overton (7) Wagner & Kemmerling (1) Groves *et al*. (1)
JBL	64	6	195.7	*t*-test only (1) *t*-test & ANOVA (1) *t*-test & MANOVA (3) Wilcoxon rank test (1)	Armstrong & Overton (3) Lambert & Harrington (2) Mentzer & Flint (2) Wagner & Kemmerling (1)
JM	59	20	357.4	*t*-test only (15) *t*-test & MANOVA (1) *t*-test & Chi-squared (2) Chi-squared only (1)	Armstrong & Overton (6)
JOM	85	20	244.1	Chi-squared only (1) MANOVA (3) *t*-tests only (14) *t*-test & ANOVA (1)	Armstrong & Overton (14) Lambert & Harrington (3) Mentzer & Flint (2) Wagner & Kemmerling (1)
JSCM	89	13	284.1	ANOVA & Chi-squared (1) *t*-tests only (10) *t*-test & ANOVA (1) Mann-Whitney test (1)	Armstrong & Overton (8) Lambert & Harrington (3)

**Note*: Not all articles in (b) reported the type of statistical method or cited a reference used in the assessment of unit non-response bias. Therefore, the totals of the values in (d) and (e) do not necessarily add up to the numbers in (b).*

The results in Table 1 show the variety of statistical methods that have been utilized in assessing for unit non-response bias in SCM research involving the analysis of dyadic data. *T*-tests, either individually or combined with other tests, were employed in the majority of studies, with Armstrong and Overton (1977) being the main cited reference in support. The *JOM* and *JM* jointly had the most studies with analyses of survey-based dyadic data. Also, the average and the median number of tests used for such assessments were highest for *JSCM*, with studies in *JOM*, *JM* and *DS* using a minimum of four tests (e.g., separate tests on two survey items for buyer and supplier responses). None of the articles found in *JBL* listed the number of survey items used in their tests. The average total sample size was highest for studies in *JM*, with studies in *JB*L having the smallest total sample sizes on average. The total sample size employed, averaged over all five journals, was approximately 240.

Collectively, the results in Table 1 indicate that the *t*-test method of comparing late to early respondents, for assessing unit non-response bias, was the most commonly used in articles from each of the journals. Critics of this approach note that it is based on a tenuous assumption that late respondents are more like non-respondents (Hulland *et al*., 2018; Lohr, 2010). Others note that this approach is easy to apply and requires an exact recording of when responses are received (Wagner & Kemmerling, 2010). It is easy to see that there is no clear agreement on the appropriate statistical method to be used in the assessment of potential unit non-response bias with dyadic data. Furthermore, in Table 1, the prevalent use of four or more tests in the assessment makes it likely that one or more of the tests results in an erroneous conclusion about potential non-response bias (Clottey & Benton, 2013: 801). We believe that consideration of effect sizes, statistical power, and sample size requirements may help in providing a predominant approach to assessing for unit non-response bias in the presence of dyadic data.

We start with an illustration of the possible correlations involved in unit non-response bias assessment, with dyadic data, and then show how such correlation structures may help to detect a meaningful difference between respondents, and non-respondents, with small sample sizes. Specifically, the impact of the number of survey items on the ability (i.e., effect size confidence interval coverage) of MANOVA tests, to detect meaningful differences, is investigated. We conclude with a discussion and implications for practice when assessing for potential unit non-response bias with dyadic data.

3. Example of Correlations Involved with Unit Non-response Bias Assessment with Dyadic Data

Table 2 provides sample data for responses by 24 participants on two survey questionnaire items. Twelve supplier and twelve buyer respondents make up the 24 and are categorized in the table using indicators (i.e., value 1 if the respondent belongs to a particular party, a value of 0 otherwise). The first six respondents of each party (i.e., buyer or supplier) are respondents of the initial survey, while the last six respondents did not respond to the initial survey (i.e., responded only after the researcher had made follow-up contact once data collection for the initial survey was closed). These classifications are depicted using a color scheme and cell labels (i.e., A1: A24 and B1: B24).

Table 2: Sample Dyadic-Data Collected on Two Continuous Response Variables

Response Variables (survey questionnaire items)		Group (buyers, suppliers)	
#1	#2	Buyer	Supplier
35.74 (=A1)	50.20 (=B1)	1	0
48.97 (=A2)	50.73 (=B2)	1	0
48.18 (=A3)	42.14 (=B3)	1	0
54.05 (=A4)	42.15 (=B4)	1	0

49.46 (=A5)	49.06 (=B5)	1	0
44.22 (=A6)	52.59 (=B6)	1	0
40.95 (=A7)	49.94 (=B7)	1	0
58.15 (=A8)	48.29 (=B8)	1	0
53.24 (=A9)	49.98 (=B9)	1	0
54.88 (=A10)	56.86 (=B10)	1	0
59.65 (=A11)	49.12 (=B11)	1	0
49.13 (=A12)	60.82 (=B12)	1	0
23.18 (=A13)	60.10 (=B13)	0	1
44.09 (=A14)	39.73 (=B14)	0	1
39.18 (=A15)	54.63 (=B15)	0	1
53.32 (=A16)	22.05 (=B16)	0	1
63.81 (=A17)	61.28 (=B17)	0	1
34.92 (=A18)	61.38 (=B18)	0	1
31.18 (=A19)	56.39 (=B19)	0	1
52.09 (=A20)	56.32 (=B20)	0	1
47.18 (=A21)	54.30 (=B21)	0	1
61.32 (=A22)	64.81 (=B22)	0	1
71.81 (=A23)	51.05 (=B23)	0	1
42.92 (=A24)	70.33 (=B24)	0	1

🟦 Respondents to the initial survey

🟧 Non-respondents to the initial survey

The traditional approach to assessing for potential unit non-response bias is to compare the average responses from the non-respondent group to that of the respondent group, for each response variable. If the responses of the two groups differ by very much, that is good evidence that unit non-response cannot be ignored (Wagner & Kemmerling, 2010: 361). Thus, from Table 2, the data to be used for assessment based on response variable #1 would be the values in cells $\{A1 - A6;\ A13 - A18\}$ compared to the values in cells $\{A7 - A12; A19 - A24\}$, respectively. Likewise, the assessment based on response variable #2 would involve the values in cells $\{B1 - B6;\ B13 - B18\}$ compared to the values in cells $\{B7 - B12; B19 - B24\}$, respectively. Clottey and Benton (2013) argue that the responses in cells $\{A1 - A6;\ A13 - A18\}$ are correlated with those in cells $\{B1 - B6;\ B13 - B18\}$ for respondents, and likewise values in cells $\{A7 - A12; A19 - A24\}$ are correlated with values in cells $\{B7 - B12; B19 - B24\}$ for non-respondents. Thus, these correlations could be accounted for in a statistical test (e.g., multivariate *t*-test or Hotelling *T* squared test), meaning that only one test would need to be performed in the assessment instead of two. The single test would have a higher statistical power to correctly detect meaningful differences, in the responses of the initial non-respondent and respondent groups, than the *complete power* of the two *t*-tests (or ANOVAs) applied individually to the cells with the 'A' or 'B' labels. The probability of a statistical test to correctly detect a difference between groups when one exists is the *individual power* of the test. The probability of a set of statistical tests to *jointly* detect a difference between groups correctly, when such differences exist, is known as the complete power of the set of tests (Clottey & Benton, 2013; Westfall *et al.*, 1999). In the case of the dyadic data depicted in Table 2, responses for different survey items for respondents belonging to the same party (i.e., buyer or supplier) would result in four correlations, such as the correlations between the appropriate values in $\{A1 - A6;\ A13 - A18\}$ versus $\{B1 - B6;\ B13 - B18\}$ for respondents, and $\{A7 - A12; A19 - A24\}$ versus $\{B7 - B12; B19 - B24\}$ for non-respondents. Also, responses between buyers and suppliers may be correlated with each other, which would result in an additional four correlations, thus, yielding a total of eight correlations. This would mean that two multivariate *t*-tests (one each for buyers and suppliers) would be needed to assess for non-response bias in this case (if the researcher does not want to violate the required independence assumption for the test by combining buyer and supplier data). This would result in a complete power for the two tests that are lower than the individual power of each test (see Table S2 of Clottey & Benton, 2013).

Table 3 lists the sources of these eight correlations along with the mean values from Table 2 for each set, and the correlation coefficient between each set of values.

Table 3: Sources of Response Correlations for Dyadic Data in Table 2 When assessing for Unit Non-response Bias via the Comparison of Respondents to Non-respondents.

Source	#	Cell set #1 [Mean][*]	Cell set #2 [Mean][*]	Correlation between set #1 and set #2 values [†]
Respondents: Within party (i.e., buyers or suppliers)	1	$\{A1 - A6\}$ [46.8]	$\{B1 - B6\}$ [47.8]	-0.55
	2	$\{A13 - A18\}$ [43.1]	$\{B13 - B18\}$ [49.9]	-0.33
Non-respondents: Within party (i.e., buyers or suppliers)	3	$\{A7 - A12\}$ [52.7]	$\{B7 - B12\}$ [52.5]	-0.22
	4	$\{A19 - A24\}$ [51.1]	$\{B19 - B24\}$ [58.9]	-0.25
Respondents: Between party (i.e., buyers & suppliers)	5	$\{A1 - A6\}$ [46.8]	$\{A13 - A18\}$ [52.7]	0.83
	6	$\{B1 - B6\}$ [47.8]	$\{B13 - B18\}$ [49.9]	0.54
Non-respondents: Between party (i.e., buyers & suppliers)	7	$\{A7 - A12\}$ [52.7]	$\{A19 - A24\}$ [51.1]	0.90
	8	$\{B7 - B12\}$ [52.5]	$\{B19 - B24\}$ [58.9]	0.96

**Note: The values within the square brackets are the averages obtained from Table 2 for the respective cell.*

[†] *Note: Pearson correlation coefficient between the respective cell values in Table 2.*

It is clear that the comparison of respondents versus initial non-respondents when assessing for potential non-response bias, with all the dyadic data in Table 2, results in a complex design. If multiple separate tests are utilized in the assessment (e.g., when the design structure is not considered), then a low complete power will result. This makes it likely that one or more of the tests resulted in an erroneous conclusion about potential non-response bias (Clottey & Benton, 2013: 802). This likelihood is increased further when, due to small sample sizes (as is the case for the data in Table 2), the separate tests have low individual statistical power to start with. In such a case, consideration of the dyadic data structure in a single MANOVA test could result in the correct detection of a difference even with such a small sample size. The following numerical example illustrates this.

3.1. Example of the effect of Table 3 correlations on unit non-response bias assessment

We will illustrate how the use of a one-way ANOVA versus a MANOVA could result in different conclusions when applied to the data in Table 2. It should be noted that the square of the test statistic used in the popular *t*-test approach for assessing unit non-response bias is equal to the *F* statistic used for assessment with a one-way ANOVA (e.g., the Lambert & Harrington, 1990, approach). Thus, the *t*-test and ANOVA approaches are equivalent when used to assess for potential non-response bias on the same dataset. Given this equivalency, the one-way ANOVA applied to data from either the buyer or supplier, on either survey item #1 or #2, boils down to splitting up the variance as follows:

$$V_T = V_t + V_e$$

where, $V_j, j \in \{t, e, T\}$ are the Treatment, Error, and Total variance values, respectively. The "Treatment" factor represents whether a response is from an initial non-respondent or from a respondent to the survey, which is the main effect of interest. For the data in Table 2, the *p*-values for the one-way ANOVAs on survey item #1 are 0.15 and 0.35 for the buyer and supplier groups, respectively. For survey item #2, they are 0.12 and 0.24, respectively. Collectively, this would suggest that there are no significant differences in the average values for initial non-respondents to that of the respondents. A look at Table 3 shows that the difference in mean values for respondents versus non-respondents on survey item #1 in the buyer group is 5.9 (= 80-76.5) while that for the supplier group is 8.0 (=43.1-51.1). For survey item #2, the buyers have a mean difference of 4.7 (=47.8 -52.5) and suppliers have a mean difference of 9.0 (=49.9 - 58.9). These mean differences result in four standardized ANOVA effect size (ES) indexes that are each greater than 0.4. Cohen (1988: 26) noted that a medium ES index represented an effect likely to be visible to the naked eye of a careful observer and had been found, in effect size surveys, to approximate the average size of observed effects in various fields. One such survey was conducted by Verma and Goodale (1995) in which they found medium and large effect sizes for most of the articles published in the *DS* and *JOM*. A more recent survey is the one by Helmuth *et al*. (2015) in which they found medium effect sizes for articles published in the *DS*, *JOM*, *JBL*, and *JSCM* from 2002-2012. Hence our choice of a medium population effect size representing a meaningful difference. An ES index of 0.4 would, therefore, be categorized as large using the convention by Cohen (1988: 26). Thus, the mean differences in Table 3 would be non-trivial for SCM research even though the ANOVA results suggest otherwise. Clearly, the four one-way ANOVAs do not consider any of the correlations specified in Table 3. A two-way MANOVA would take such correlations into account.

For the two-way MANOVA, there are two dependent variables, which are survey items #1 and #2, along with a Party and Treatment factor. The 'Party' factor captures the effect (i.e., at two levels, corresponding to buyer and supplier) of the dyad, with the main effect of interest being differences in the responses of the initial non-respondents versus respondents. Table 4 below shows the data in Table 2, re-arranged for analysis using a two-way MANOVA approach.

Table 4: Table 2 Re-arranged for Analysis as a Two-way MANOVA

		Party Responding			
		Buyer		Supplier	
		Response		Response	
		Item #1	Item #2	Item #1	Item #2
Treatment	Respondents	35.74	50.20	23.18	60.10
		48.97	50.73	44.09	39.73
		48.18	42.14	39.18	54.63
		54.05	42.15	53.32	22.05

		49.46	49.06	63.81	61.28
		44.22	52.59	34.92	61.38
	Non-respondents	40.95	49.94	31.18	56.39
		58.15	48.29	52.09	56.32
		53.24	49.98	47.18	54.30
		54.88	56.86	61.32	64.81
		59.65	49.12	71.81	51.05
		49.13	60.82	42.92	70.33

Using a two-way MANOVA with the data in Table 4, the total variance is split as follows:

$$\boldsymbol{V}_T = \boldsymbol{V}_t + \boldsymbol{V}_p + \boldsymbol{V}_{t*p} + \boldsymbol{V}_e$$

where $\boldsymbol{V}_i\ ; i \in \{t, p, t * p, e, T\}$ is an appropriate (2x2) covariance *matrix* for Treatment, Party, Interaction, Error, and Total effects, respectively. The off-diagonal elements in each of the $\boldsymbol{V}_i$, $i \in \{t, p, t * p, e\}$ collectively capture all the correlations described in Table 3. For the data in Table 4, the *p*-value for the Treatment factor of the two-way MANOVA is 0.04 (i.e., <0.05). This would suggest that there is a significant difference in the average values for initial non-respondents compared to that of respondents.

This example serves to illustrate that consideration of the correlations (i.e., between variables and those resulting from the dyadic structure of the data) in the assessment of differences in the responses of initial non-respondents versus respondents, can improve the ability of a statistical test to detect a meaningful difference correctly. The use of MANOVA allowed for all the correlations listed in Table 3 to be considered in the assessment via the rearrangement of the data, as depicted in Table 4, to represent the dyadic design better. This resulted in the MANOVA being able to detect a difference in respondent and initial non-respondent groups, even with a small sample size (e.g., *n*=24), whereas the traditional approach using ANOVA was unable to detect any difference. The main advantages of using MANOVA, in this case, is that: (1) it can be used to detect differences between groups with a single test and possibly at smaller sample sizes than with a *t*-test or ANOVA, and (2) different design structures can be accommodated with relative ease in the analysis; for example, higher-level data structures would correspond to additional levels in the Party factor. This would eliminate the need to perform multiple tests for unit non-response bias assessment with such data structures. There is no way of incorporating the dyadic data design structure (e.g., as shown in Table 4), and thus only performing a single test, with the *t*-test approach attributed to Armstrong and Overton (1977), the *F*-test/ANOVA approach of Lambert and Harrington (1990), or the complete-power multivariate *t*- test approach of Clottey and Benton (2013). Likewise, for the other tests identified in Table 1, such as the Chi-squared, Mann-Whitney, and Wilcoxon rank-sum tests.

The previous example illustrates that a MANOVA could be used to detect meaningful differences in the responses of initial non-respondents versus respondents even when the sample size is small. However, it also begs the question of what is considered a *small* sample size. Statistical power analysis can help to shed light on this issue.

3.2. Sample size implications

Table 5 below shows the minimum sample size required to achieve a given level of statistical power, at a medium effect size, for an ANOVA and two-way MANOVA test for the type of data shown in Table 2. The results were obtained using the program G*Power 3.1 (Faul *et al*., 2007), which is freely available online.

Table 5: Minimum Sample Sizes Required to Detect a Medium ES with the Data in Table 2, as Recommended by the G*Power Program, based on Statistical Power Considerations

Test	Minimum sample size required per test	Individual Power	Complete/Joint Power
ANOVA	128	0.8	<0.80
	210	0.95	<0.95

MANOVA	43	0.8	na
	65	0.95	na

The ANOVA results in Table 5 were obtained by supplying a medium ES index of 0.25 (Cohen, 1988; 355), a significance level (α) value of 0.05 and a power value of 0.8 and 0.95, to G*Power 3.1 to determine the required sample size. According to Verma and Goodale (1995), 0.8 and 0.95 values are strong and very strong levels of statistical power for SCM research. Thus, the four ANOVAs used to assess for potential non-response bias, with the type of data found in Table 2, would need a sample size of at least 128 for buyers and 128 for suppliers, with each test, to have a strong individual power of correctly detecting a medium population effect size. Furthermore, the ANOVA sample sizes in Table 5 assume that the data is balanced (i.e., an equal number of respondents and initial non-respondents) implying that the researcher would need at least 64 non-respondents (i.e., half of 128) for each test in order to achieve a strong individual power. An unbalanced ANOVA would require an even larger total sample size. For example, a total sample size of at least 356 (i.e., 178 buyers and 178 suppliers) would be required to achieve a strong individual power level with the Mentzer and Flint (1997: 206) recommendation of contacting 30 non-respondents when assessing for potential *unit* non-response bias.

The complete power of four tests is always less than the individual power of each test (Clottey & Benton, 2013: 802). In the case where the correlations between buyer and supplier responses are zero (e.g., the four tests are independent of each other), then the complete power at an individual power level of 0.8 is 0.41 (=0.8^4), while it is 0.81 (=0.95^4) at the 0.95 level. Put another way, a complete power value less than 0.8 means that there is less than an 80% chance that all four ANOVA tests would correctly detect a medium effect size difference, even if each test has an 80% chance of doing so. Thus, a study would require a total sample size of 420 (i.e., 210 buyer and 210 supplier responses) to achieve at least a strong (i.e., >0.80) level of complete power when assessing for potential unit non-response bias using the traditional (*t*-test or ANOVA) approach, with the type of data shown in Table 2. Of the 75 dyadic data articles surveyed in Table 1, only 24 (32%) of the articles reported a total sample size larger than 420. Forty-nine percent of the articles reported total sample sizes of less than 256. The *DS* and *JM* had the joint highest percent (i.e., 75%) of articles reporting total sample sizes larger than 256. None of the six articles from *JBL*, in Table 1, had total sample sizes larger than 256.

The MANOVA results in Table 5 were obtained by supplying a medium (eta-squared) ES index of 0.15 (Cohen, 1988: 478), a significance level (α) value of 0.05, a power value of 0.8 and 0.95, with values of 2 for groups, predictors and response variables, to G*Power 3.1 to determine the required sample size for a MANOVA. The required sample size, to achieve a comparable level of individual power as the ANOVA, is substantially less with the MANOVA. Also, the MANOVA results in only one test; therefore, the multiple testing and complete power issue do not apply to it. Thus, to achieve a strong level of power with MANOVA, the researcher would only need to contact at least 11 buyers and their suppliers (assuming a balanced MANOVA), which is significantly less than the 64 that would be required with the traditional (i.e., ANOVA) approach.

Cohen (1988: 479) notes that a population covariance or correlation matrix should be provided to determine the sample size required for detecting a population effect size in a MANOVA. However, G*Power 3.1 did not require such a matrix to be provided for the sample size recommendations shown in Table 5. The G*Power 3.1 documentation (Faul *et al*., 2007: 183) did not directly address the covariance matrix issue but did point out that the power analysis for the MANOVA was based on the algorithm by O'Brien and Shieh (1999). We found that

O’Brien and Shieh (1999: 11) utilized an identity matrix (i.e., ones on the diagonal and zeros on the off-diagonals) to evaluate their algorithm and no evidence of the use of any other covariance matrices. This could mean that non-zero correlations (e.g., as shown in Table 3) may affect the ability of the MANOVA to correctly detect a medium effect size at the recommended minimum sample sizes given in Table 5. Indeed, the sample size of 24, for the example data in Table 4, is a lot less than the amount recommended for a MANOVA power value of 0.8. The MANOVA employed in that example was able to detect an effect correctly. The correlations listed in Table 3 likely played a part in the detection by the MANOVA, even with such a small sample size. Being able to obtain sample size recommendations for a MANOVA without having to provide a population covariance/correlation matrix is not undesirable. An SCM researcher will find it difficult to come up with a population matrix to use in the assessment of potential non-response bias since these types of details are frequently not included in published articles (e.g., none were found in any of the published articles that we surveyed in Table 1). Even if such covariance matrices could be found, a different population matrix would need to be used for different types, and number, of response variables utilized in assessing for potential non-response bias with MANOVA. This can complicate the process of sample size determination for a MANOVA. While the correlations do have an impact on the ability of the MANOVA to correctly detect a difference between respondents and non-respondents, as revealed by the results of our analysis on simulated data (available on request), we were unable to find systematic correlation patterns (e.g., a critical mass percentage of negative or positive correlation values) that researchers could use as a guide for determining when a sample size smaller than that shown in Table 5 could result in sufficient statistical power to assess for potential unit non-response bias. Our initial simulation analysis, however, led us to investigate another factor-the number of survey items to include in the MANOVA-which could likewise affect the ability of the MANOVA to detect a meaningful difference between respondents and non-respondents correctly. We were able to establish potential cut-offs for this factor, as shown in our analysis in Section 4 below.

4. Analysis: Effect of the Number of Response Variables

In addition to factors such as sample size, effect size and significance level which can affect statistical power in both ANOVA and MANOVA (Cohen, 1988), the statistical power of a MANOVA can also be affected by the number of response variables (Huberty & Morris,1989: 307). Table 6 below shows that increasing the number of response variables utilized in a MANOVA requires an increase in the sample size to maintain the current level of statistical power.

Table 6: Number of Response Variables and Minimum Sample Sizes Required for a Two-way MANOVA to Detect a Medium ES (as recommended by the G*Power 3.1 program based on statistical power considerations)

Power	# of variables	Minimum sample size required
0.8	2	43
	3	49
	4	55
	5	59
	6	64
	7	68
	8	71
0.95	2	65
	3	73
	4	80
	5	86
	6	92
	7	97
	8	102

We look at the possible effects of adding response variables while maintaining the existing pairwise correlation structure. In general, the more response variables there are in a one-way MANOVA, the higher the chance of correctly detecting a meaningful difference between the treatments of interest if one does exist (Tabachnick & Fidell, 2007: 372). As mentioned earlier, we chose a medium effect size as representing a meaningful difference in our analysis since previous surveys (e.g., Helmuth *et al*., 2015) had found medium effect sizes for most articles published in the *DS, JOM, JBL,* and *JSCM*.

4.1. Using effect-size confidence intervals to detect a meaningful difference

Rejection of the null hypothesis of a MANOVA would indicate that a significant effect has been detected, but does not provide any insight into how *large* the detected effect could actually be. In order to gain insights about the size of the effect detected, and also make inference about what the effect size would be in a census of all respondents and non-respondents (i.e., the population effect size), we have to construct effect size confidence intervals. Thus, the ability of a MANOVA to detect a medium effect size could be assessed by creating a bootstrap effect-size confidence interval (Banjanovic & Osborne, 2016; Kirby & Gerlanc, 2013) with sample effect size estimates obtained from the MANOVA. Sample **R** code for obtaining an effect-size bootstrap confidence interval, for the type of data in Table 4, is available upon request. The researcher can then examine the interval to see which population effect sizes (small, medium, or large) are contained in the interval. If the MANOVA has enough statistical power to detect a medium effect size but a small effect size is contained in the resulting confidence interval, then it means that a researcher cannot rule out the possibility that the detected difference is not meaningful (Fan, 2001: 282). Thus, the examination of such conditions is an indication of the ability of a MANOVA to detect a meaningful (i.e., medium effect size) difference in the responses of respondents versus initial non-respondents. Conversely, a MANOVA performed with insufficient statistical power (e.g., <0.8) would likely fail to detect any difference between respondents and initial non-respondents, due to a large *p*-value, while also resulting in wide effect-size confidence intervals which contain small, medium and large effects (Fan, 2001: 282). Such a situation should be a signal to the researcher that the MANOVA performed had insufficient power (likely due to the small sample size), to detect a meaningful difference, and therefore more effort should be put towards increasing statistical power. We detail these conclusions in a flowchart in Appendix A. Thus, the examination of effect size confidence intervals, in the assessment for potential *unit* non-response bias, can yield a more detailed and insightful analysis than the traditional approach of just reporting a significant *p*-value when comparing the responses of respondents to initial non-respondents.

A confidence interval that contains both a small and medium population ES would suggest that one cannot rule out that the detected difference, in respondents and initial non-respondents, may *not* be of practical significance. In such a situation, a larger sample size than that suggested in Table 5 would result in tighter confidence intervals. Such narrow intervals are more likely not to contain a small effect size. However, with previously mentioned difficulties in obtaining larger sample sizes in dyadic studies, we wanted to see if there were solutions other than increasing the sample size available to a researcher in such situations. Increasing the number of response variables included in the MANOVA may be one such solution.

What we wanted to know is if increasing the number of response variables would also increase the ability to detect a *meaningful* difference. Specifically, whether having more response variables could tighten the resulting intervals. To gain insights, we ran a simulation analysis (details available upon request) in which we generated multivariate normal datasets with means set so that the difference between the respondent and non-respondent groups represented a medium (eta-squared) ES index of 0.15. A MANOVA was then performed, based on the layout in Table 4, for 100 of such generated datasets and the percentage of confidence intervals that contained both a small and medium ES value was recorded. To investigate the effect of increasing the number of response variables included in the MANOVA, a pair of additional response variables were added with similar pairwise correlation structures as the existing variables, and the MANOVA was re-run. By adding pairs of variables in this fashion, we were able to maintain the proportion of negative/positive correlations (i.e., correlation structure) in the resulting correlation matrix, thereby ensuring that the results of the analysis could only be attributed to increases in the number of response variables, and not due to different correlation structures or differing sample sizes. We considered 2, 4, 6, and 8 response variables. We generated 100 datasets for each response variable scenario, and recorded the percentage of confidence intervals that contained a small effect size. The results are shown in Figure 1 below.

Figure 1: Impact of Increasing the Number of Response Variables on the Ability of a MANOVA to Detect a Medium Effect Size Difference

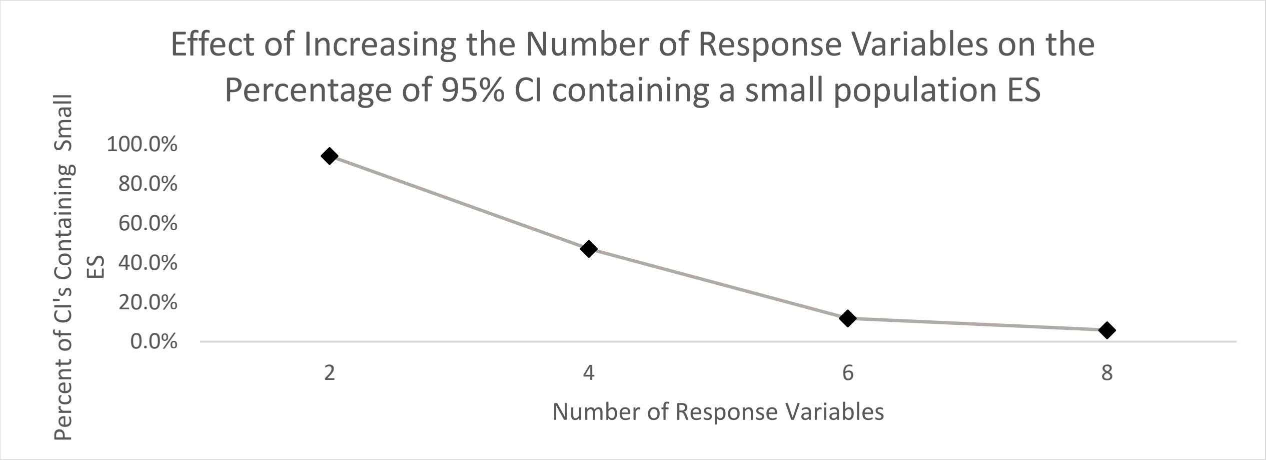


Figure 1 shows that as more pairs of response variables with similar pairwise correlation structures as existing pairs were included in the two-way MANOVA, the percent of confidence intervals containing small effect sizes reduced at a decreasing rate. Delving deeper into this result, we found that the center of the confidence intervals averaged out at 0.09, with two response variables, and continued to increase with the inclusion of more response variables culminating in a value of 0.14 in the eight-response variable scenario. Clearly, 0.14 is closer to the population effect size of 0.15 than the center of the confidence intervals created in the scenarios with less than eight response variables. We also found that the half-width of the confidence intervals averaged out at 0.09 with two response variables but was reduced to 0.08 with eight response variables. This indicates that the inclusion of more response variables, with similar pairwise correlation structures as the existing variables, improved both the accuracy and precision of the effect size confidence intervals, *while* the *total sample size was maintained at the same level* in each of the scenarios.

5. Discussion

A survey of 75 published studies involving the analysis of dyadic data in the *DS*, *JOM*, *JSCM*, *JBL*, and *JM* revealed that various statistical tests, including the *t*-test, ANOVA, Chi-squared and MANOVA have been utilized in assessing for potential unit non-response bias. Of these statistical tests, only a two-way MANOVA allows for a single test to be employed in assessing for potential *unit* non-response bias. Performing a particular statistical test, multiple times, to assess whether there is a meaningful difference in the responses of initial non-respondent versus respondent groups in the population, is an instance of the classical multiple testing problem (Clottey & Benton, 2013: 801). That is, as more inferences are made, then it is more likely that erroneous inferences will occur. The traditional *t*-test or ANOVA approach results in a minimum of two tests (e.g., one for buyers and another for suppliers) if applied to dyadic data. Results from our survey of the literature indicate that a minimum of four such tests (e.g., one each for two survey items from buyer and supplier responses) were employed, often with an insufficient sample size for the tests to have an acceptable level (i.e., >0.8) of statistical power. The collection of dyadic data is more challenging than standard data collection due to the time needed to recruit/sample both members of a dyad along with budgeting issues. This can result in low response rates (Quinn *et al.*, 2010) and difficulties in achieving the sample size required to detect a meaningful difference in respondents and initial non-respondents, for a traditional *unit* non-response bias assessment. By using the two-way MANOVA method proposed in this study, an SCM researcher can perform such an assessment without the need for multiple testing or the large sample sizes required (see Table 5) with the traditional approaches. If the test suggests that the responses of the two groups do not differ by very much, that is good evidence that *unit* non-response bias is unlikely to be an issue (Wagner & Kemmerling, 2010; Whitehead *et al.*, 1993).

We provided a numerical example where the use of the popular *t*-test/ANOVA approach to assessing for potential non-response bias with a small sample size (*n*=24) yielded a different conclusion to that when a two-way MANOVA was used to perform the same assessment. Our analyses of simulated data shows that MANOVAs with less than half the sample size required for an ANOVA to have a 0.95 statistical power of detecting a meaningful (i.e., medium effect size) difference were *always* able to detect a difference correctly (e.g., *p*-values less than 0.05), whereas the multiple ANOVAs used to perform the same assessment were not. In Table 5, we provide sample size recommendations, from a popular freely available online software - G*Power 3.1- for a two-way MANOVA to have a statistical power of at least 0.8 or 0.95 to detect a medium population effect size. However, G*Power 3.1 did not require a population correlation matrix to be provided for the sample size recommendations shown in Table 5. Not having to provide a population correlation matrix to make sample size decisions is desirable. However, it can also mean that the recommended sample sizes in Table 5 result in an ambiguous detection (e.g., a significant *p*-value, but with a confidence interval which contains a small population effect-size) of a medium population effect size, when correlations in the dyadic data are considered. While all *p*-values in our analyses were significant (i.e., <0.05), several of the effect size confidence intervals contained a small population (eta-squared) effect size value. A confidence interval that contains a small population effect size would suggest that the detected difference between respondents and initial non-respondents may *not* be meaningful. This is a significant finding since it emphasizes the importance of providing confidence intervals of effect sizes along with the *p*-value in an assessment for potential *unit* non-response bias. This was not done in any of the 75 articles listed in Table 1. We have included a step-by-step flow chart in Appendix A to guide researchers on how this can be done in practice.

If a test has enough statistical power to detect a meaningful effect size, then a non-significant *p*-value test result, coupled with an effect-size confidence interval that only contains a small population effect size, provides a much stronger argument for there being no practical or statistical differences in the responses of initial non-respondents versus respondents. Likewise, if a detected difference is not of practical importance (i.e., is consistent with a small effect size), then *unit* non-response may not be an issue even though the statistical test was significant (see Appendix A). Calls for reporting of effect-size intervals along with *p*-values are being made in settings unrelated to the assessment of non-response bias (e.g., Fan, 2001; Schwab *et al*., 2011). Our results suggest that this call is likewise important in the assessment of potential *unit* non-response bias. Annotated **R** code (available upon request) can be used by researchers to obtain such 95% bootstrap confidence intervals, of eta-squared effect sizes, when a two-way MANOVA is performed. Effect size confidence intervals can also be obtained with commercial software such as STATA and SPSS (see Lakens, 2014). However, we did not come across specific examples of the use of such commercial software to create effect-size confidence intervals for a two-way MANOVA.

MANOVA was reported as being used to assess for *unit* non-response bias in seven of the 75 articles, as shown in Table 1. However, none of the seven articles discussed statistical power considerations for the sample sizes and the number of response variables used in their MANOVA test. Also, it is not clear from the details provided in the seven articles whether the MANOVA approach utilized in those studies was adequate in assessing for potential *unit* non-response bias. In this study, we show how dyadic data can be set up to be analyzed as a two-way MANOVA with a ‘Party’ (e.g., buyer, supplier) factor and the ‘Treatment’ factor being respondents versus initial non-respondents. As noted earlier, the two-way MANOVA allows for the correlations between dyad member responses to be accounted for in an assessment for potential *unit* non-response bias, meaning that only one test needs to be performed. Also, the sample size required for enough statistical power (e.g., >0.8) to detect a medium effect size is significantly smaller for the two-way MANOVA than for the other approaches identified in Table 1. Appendix A provides step-by-step instructions for setting up and utilizing the two-way MANOVA, with effect size confidence intervals, to perform the assessment.

In addition to sample size, the statistical power of a MANOVA used to assess for potential non-response bias is affected by the number of response items included, as shown in Table 6. Results of our analysis on simulated data indicate that-provided the sample size is at least as large as those listed in Table 6, to achieve a 0.8 or 0.95 level of statistical power, including additional pairs of variables, with similar pairwise correlation structures as existing variables, improved the precision and accuracy of the effect size confidence intervals (see Figure 1). This is an intriguing and important finding since it suggests that the inclusion of more survey items, in the assessment of potential non-response bias, is desirable in the MANOVA approach. The prevalent advice given to SCM researchers for the number of survey items to include in an assessment for potential *unit* non-response bias is two key items, in order to maximize complete power (Clottey & Grawe, 2014), or five which is considered a reasonable number of key items (Mentzer & Flint, 1997). The results of our study suggest that with dyadic data structures, the number of key items to use could be expanded past five due to effect size considerations.

Our numerical example illustrates that a MANOVA with a significant outcome could have more accurate and precise effect size confidence intervals (for a given sample size) with the inclusion of more response variables in the analysis. It is becoming increasingly difficult to increase survey sample sizes in standard SCM research due to respondent fatigue (Larson, 2005). This is even more challenging in dyadic studies due to the time needed to recruit/sample both members of a dyad along with a higher chance of encountering budgeting issues (Quinn *et al.*, 2010). It should, therefore, be welcome news that factors other than an increase in sample size could be invested into the dyadic study to improve the ability of a MANOVA to detect a meaningful difference in respondents and initial non-respondents. The researcher is still responsible for ensuring that the survey items included in the assessment, due to effect size considerations, are meaningful to the study and/or to the detection of potential unit non-response bias (Hulland *et al*., 2018; Mentzer & Flint, 1997; Thompson & Washington, 2013). The researcher also must ensure that the minimum sample size requirements suggested in Table 6 are not violated with the addition of more variables, since then the MANOVA may have insufficient statistical power to detect a medium effect size. The more survey items that are included, the larger the sample size needs to be to achieve enough statistical power (e.g., >0.8) for detecting a meaningful effect size. Our results (see Appendix A) suggest that the researcher should include as many key items as their sample size will allow (see Table 6) when assessing for potential *unit* non-response bias with the proposed two-way MANOVA approach.

5.1 Methods for accommodating potential *unit* non-response bias

Adjustments for unit non-response bias are made at the *unit* level. If the proposed MANOVA assessment indicates that there is a risk for bias (see Appendix A), it may be possible to use modeling methods to make predictions about the non-respondents. One such approach is to *extrapolate* what the final "wave" of responses on a survey item would have been, based on responses from previous waves. The Mean Absolute Percent Error (MAPE) of the predictions could then be used as a measure of the amount of non-response bias present, with responses on *all* survey items adjusted accordingly by the MAPEs. Details of such an approach can be found in Armstrong and Overton (1977: 6-8). This approach is based on the critiqued assumption (e.g., Hulland *et al*., 2018; Lohr, 2010) that respondents in later waves are representative of non-respondents. Execution of this approach with a group of initial non-respondents, instead of a late-wave group, would mitigate the need for such an assumption although it would also necessitate including all survey items in the follow-up contact of non-respondents. This, in turn, would mean that a large number of non-respondents would need to be contacted (e.g., see Appendix A and Table 6) in order to apply this approach.

A different approach is to take a sub-sample of non-respondents, after the initial survey with a reduced number of survey items, and use that sub-sample to make inference about the other non-respondents. Hansen and Hurwitz (1946) suggested that survey results could be weighted according to the proportion of initial and sub-sample respondents in the total sample. Weighting adjustments are further complicated with dyadic data than with standard data for three reasons. First, the sampling design for the collection of dyadic data is more complex, and therefore, so are the resulting weighting adjustments. Second, the weights can occur at different analytical levels within the dyadic dataset. As an example, a dataset of buyer-supplier dyads could have three analytical levels corresponding to (1) buyers, (2) suppliers, and (3) the dyads representing the buyer-supplier relationship. Last, relevant auxiliary information required to create the weights may come in different forms for the various units and levels (Lohr, 2010: 266). Various weights have been proposed to adjust survey results for suspected unit non-response bias. These include: weighting class (Holt and Elliot, 1991; Lin & Schaeffer, 1995), post-stratification, and raking adjustments (Lohr, 2010: 266-273). If weighting adjustments are made, then the researcher should state the assumed non-response model and give evidence to justify it (Lohr, 2010: 272). Oh and Scheuren (1983) and Lohr (2010: 266-273) provide details and assumptions underlying weighting methods. Making weighting adjustments for potential *unit* non-response bias, however, is not a trivial task (e.g., see Thompson & Washington, 2013, for a case-study example) and may result in further biased estimates if improper adjustment procedures are employed (Lohr, 2010; Thompson & Washington, 2013). It is therefore important for researchers to *first* assess how well the responses of those in-sample represent those of the out-of-sample respondents, prior to deciding to perform an adjustment for potential unit non-response bias.

7. Conclusion

Over the last decade, there has been an increase in Supply Chain Management (SCM) research in which multiple views (e.g., buyer-supplier, shipper-carrier) in a relationship, such as dyads, are used in an analysis. Dyadic data introduce additional correlations relative to a standard study, which can not only complicate the focal analysis but also affect the ability (e.g., effect size confidence interval coverage) of a statistical test to detect a meaningful difference in respondent and initial non-respondent groups. If the responses of the two groups differ greatly, that is good evidence that *unit* non-response cannot be ignored and therefore adjustments to the data may be required to accommodate for potential non-response bias (Wagner & Kemmerling, 2010, Whitehead *et al*., 1993).

The current study provides practical guidelines (see Appendix A for a summary) that enable SCM scholars, using dyadic data, to perform a rigorous statistical test with enough power (i.e., >0.8) to assess for meaningful differences in responses from initial non-respondents. We show how dyadic data can be set up to be analyzed as a two-way MANOVA with 'Party' as one factor (e.g., buyer, supplier) and a 'Treatment' factor being the respondents versus initial non-respondents group. By accounting for correlations, inherent in dyadic data, with the use of MANOVA in the assessment of potential *unit* non-response bias, only one statistical test needs to be performed instead of the currently popular approach requiring the execution of multiple *t*-tests or ANOVAs. This reduces the chance that one or more of such multiple tests would result in erroneous inference (Clottey & Benton, 2013: 801). Also, the MANOVA can be performed with significantly smaller sample size requirements than traditional approaches, such as *t*-tests or ANOVA.

We provide a numerical example where the use of the popular *t*-test/ANOVA approach to assessing for potential *unit* non-response bias, yielded a different conclusion to that when a two-way MANOVA was used to perform the same assessment. Analyses of simulated data show that a MANOVA with less than half the sample size required for an ANOVA to have a high statistical power (i.e., >0.8) of detecting a meaningful (i.e., medium effect size) difference was *always* able to detect a difference correctly (e.g., *p*-values less than 0.05), whereas the multiple ANOVAs used to perform the same assessment were not. The reduced sample size means that fewer initial non-respondents would need to be contacted in a follow-up in order to elicit the data required to perform the test. While all *p*-values in our analyses were significant (i.e., <0.05), several of the resulting effect size confidence intervals contained both a small and medium population (eta-squared) effect size value and thus could not be used to conclude unambiguously that the detected difference had practical significance. The results of our analyses demonstrate that in such situations, the inclusion of more response variables in the MANOVA could result in more precise and accurate effect size confidence intervals. Such confidence intervals are less likely to contain a small effect size value, for a given sample size, when there is a medium effect size in the population. Our results also indicate that a researcher should include as many key items as their sample size and number of non-respondents to be contacted, will allow (see Appendix A and Table 6) when assessing for potential *unit* non-response bias with the proposed two-way MANOVA approach. This is somewhat contrary to the advice given to SCM researchers, in previous studies (Clottey & Grawe, 2014; Mentzer & Flint, 1997), to perform *unit* non-response bias assessment on only two or five key survey items.

Scholars using the MANOVA approach proposed in this study to assess for potential unit non-response bias, with dyadic data, still need to perform the standard eight validation checks (Tabachnick & Fidell, 2007: 371-435) to ensure that an honest MANOVA has been performed. These standard checks, combined with the application of MANOVA as proposed herein and summarized in a flowchart (see Appendix A), have the potential to significantly improve the rigor applied to the assessment of *unit* non-response bias in SCM research with dyadic data. Additional analyses, such as the examination of response rates at each analytical level of the dyadic dataset, or comparison of the response sample to a target population or sampling frame, coupled with our proposed approach, may be needed to yield a more complete view of the potential for *unit* non-response bias with dyadic data.

References

Armstrong, J. S. and T.S. Overton.1977. Estimating nonresponse bias in mail surveys. *Journal of Marketing Research*, 14, 396-402. https://doi.org/10.2307/3150783

Banjanovic, E.S., J.W. Osborne. 2016. Confidence intervals for effect sizes: Applying bootstrap resampling. *Practical Assessment, Research & Evaluation*, 21(5), 1-20.

Clottey, T., and W. C. Benton. 2013. Guidelines for improving the power values of statistical tests for nonresponse bias assessment in OM research. *Decision Sciences* 44, 797–812. https://doi.org/10.1111/deci.12030

Clottey, T. and S. Grawe. 2014. Non-response bias assessment in logistics survey research: use fewer tests? *International Journal of Physical Distribution & Logistics Management*, 44(5), 412-426. https://doi.org/10.1108/IJPDLM-10-2012-0314

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences*. (2nd Ed). Hillsdale, NJ: Erlbaum.

Collier, J. E. and C.C. Bienstock. 2007. An analysis of how nonresponse error is assessed in academic marketing research. *Marketing Theory*, 7(2), 163-183. https://doi.org/10.1177/ 1470593107076865

Faul F., E. Erdfelder, A.G Lang, and A. Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavioral Research Methods,* 39, 175–191. https://doi.org/10.3758/BF03193146

Fan, X. 2001. Statistical significance and effect size in education research: Two sides of a coin. *Journal of Educational Research, 94*(5), 275-282. https://doi.org/10.1080/00220670109598763

Flynn, B., M. Pagell, and B. Fugate. 2018. Survey research design in supply chain management: The need for evolution in our expectations. *Journal of Supply Chain Management* 54(1): 1– 15.

Groves, R. M., F. J. J. Fowler, M.P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*, 2nd edition. Wiley-Interscience: Hoboken, NJ.

Hansen, M.H. and N.W. Hurwitz. 1946. The problem of nonresponse in sample surveys. *Journal of the American Statistical Association,* 41 (236), 517-529.

Helmuth, C. A., C.W. Craighead, B.L. Connelly, D.Y. Collier, and J.B. Hanna. 2015. Supply chain management research: Key elements of study design and statistical testing. *Journal of Operations Management*, 36, 178-186. doi:10.1016/ j.jom.2014.12.001

Holt, D. and D. Elliot. 1991. Methods of weighting for unit non-response. *The Statistician,* 40 (3). 333-342.

Huberty, C.J. and J.D. Morris. 1989. Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin*, 105 (3), 302-308. https://doi.org/10.1037/0033-2909.105.2.302

Hulland, J., H. Baumgartner, and K. M. Smith. 2018. Marketing survey research best practices: evidence and recommendations from a review of JAMS articles. *Journal of the Academy of Marketing Science*, 46(1), 92-108. https://doi.org/10.1007/s11747-017-0532-y

Hult, G.T.M., W.T. Neese, and R.E. Bashaw. 1997. Faculty perceptions of marketing journals, *Journal of Marketing Education*, 19(1), 37–52. https://doi.org/10.1177/027347539701900105

Kirby, K. N., and D. Gerlanc. 2013. BootES: An R package for bootstrap confidence intervals on effect sizes. *Behavior Research Methods*. 45(4), 905–927. doi:10.3758/s13428-013-0330-5

Lakens, D. 2014. *Calculating confidence intervals for Cohen's d and eta-squared using SPSS, R, and Stata*. http://daniellakens.blogspot.com/2014/06/calculating-confidence-intervals-for.html (accessed May 18th, 2018).

Lambert, D.M. and T.C. Harrington. 1990. Measuring nonresponse bias in mail surveys. *Journal of Business Logistics*, 11, 5 – 25.

Larson, P.D. 2005. A note on mail surveys and response rates in logistics research. *Journal of Business Logistics*, 26 (2), 211-222. https://doi.org/10.1002/j.2158-1592.2005.tb00212.x

Lin, I.F. and N.C. Schaeffer. 1995. Using survey participants to estimate the impact of Nonparticipation. *Public Opinion Quarterly,* 59 (2), 236-258.

Little, R.J.A. and D.B. Rubin. 2002.*Statistical Analysis with Missing Data*, 2nd Ed. Wiley: NY.

Lohr, S. L. 2010. Nonresponse, in: *Sampling: Design and Analysis.* 2nd Ed. Brooks/Cole: Boston.

Mentzer, J.T. and D.J. Flint. 1997. Validity in logistics research. *Journal of Business Logistics*, 18 (1), 199–216.

O'Brien, R. G, and G. Shieh. 1999. *Pragmatic, unifying algorithm gives power probabilities for common F tests of the multivariate general linear hypothesis*. Available at www.bio.ri.ccf.org/UnifyPow.

Oh, H.L. and F.J. Scheuren. 1983. Weighting adjustment for unit nonresponse, in: Madow, W.G., I. Olkin, and D.B. Rubin, *Incomplete Data in Sample Surveys,* Academic Press, New York, NY, 143-184.

Quinn C., S.B. Dunbar, P.C. Clark, and O.L. Strickland. 2010. Challenges and strategies of dyad research: cardiovascular examples. *Applied Nursing Research*, 23, 15-20. doi: 10.1016/ j.apnr.2008.10.001

Schwab, A., E. Abrahamson, F. Fidler, and W.H. Starbuck. 2011. Researchers should make thoughtful assessments instead of null-hypothesis significance tests. *Organization Science*, 22(4), 1105–1120.

Simpson, D., J. Meredith, K. Boyer, D. Dilts, L.M. Ellram, and G.K. Leong. 2015. Professional, research, and publishing trends in operations and supply chain management. *Journal of Supply Chain Management*, 51: 87–100. https://doi.org/10.1111/jscm.12078

Tabachnick, B.G. and L.S. Fidell. 2007. Multivariate analysis of variance and covariance, in: *Using multivariate statistics (5th ed.)*. Allyn & Bacon/Pearson, Boston. 371-435.

Thompson, K.J. and K.T. Washington. 2013. Challenges in the treatment of unit nonresponse for selected business surveys: a case study. *Survey Methods: Insights from the Field*. Retrieved from: http://surveyinsights.org/?p=2991

Verma, R. and J.C. Goodale. 1995. Statistical power in operations management research. *Journal of Operations Management* 13 (2), 139-152. https://doi.org/10.1016/0272-6963(95)00020-S

Wagner, S.M. and R. Kemmerling. 2010. Handling nonresponse in logistics research. J*ournal of Business Logistics*, 31(2), 357–381. https://doi.org/10.1002/j.2158-1592.2010.tb00156.x

Westfall P.H, R.D. Tobias, D. Rom, R.D. Wolfinger, and Y. Hochberg.1999. *Multiple comparisons and multiple tests using the SAS system*, Cary, NC: SAS.

Werner, S., M. Praxedes, and H.-G. Kim. 2007. The reporting of nonresponse analyses in survey research. *Organizational Research Methods*, 10(2), 287-295. https://doi.org/10.1177 /1094428106292892

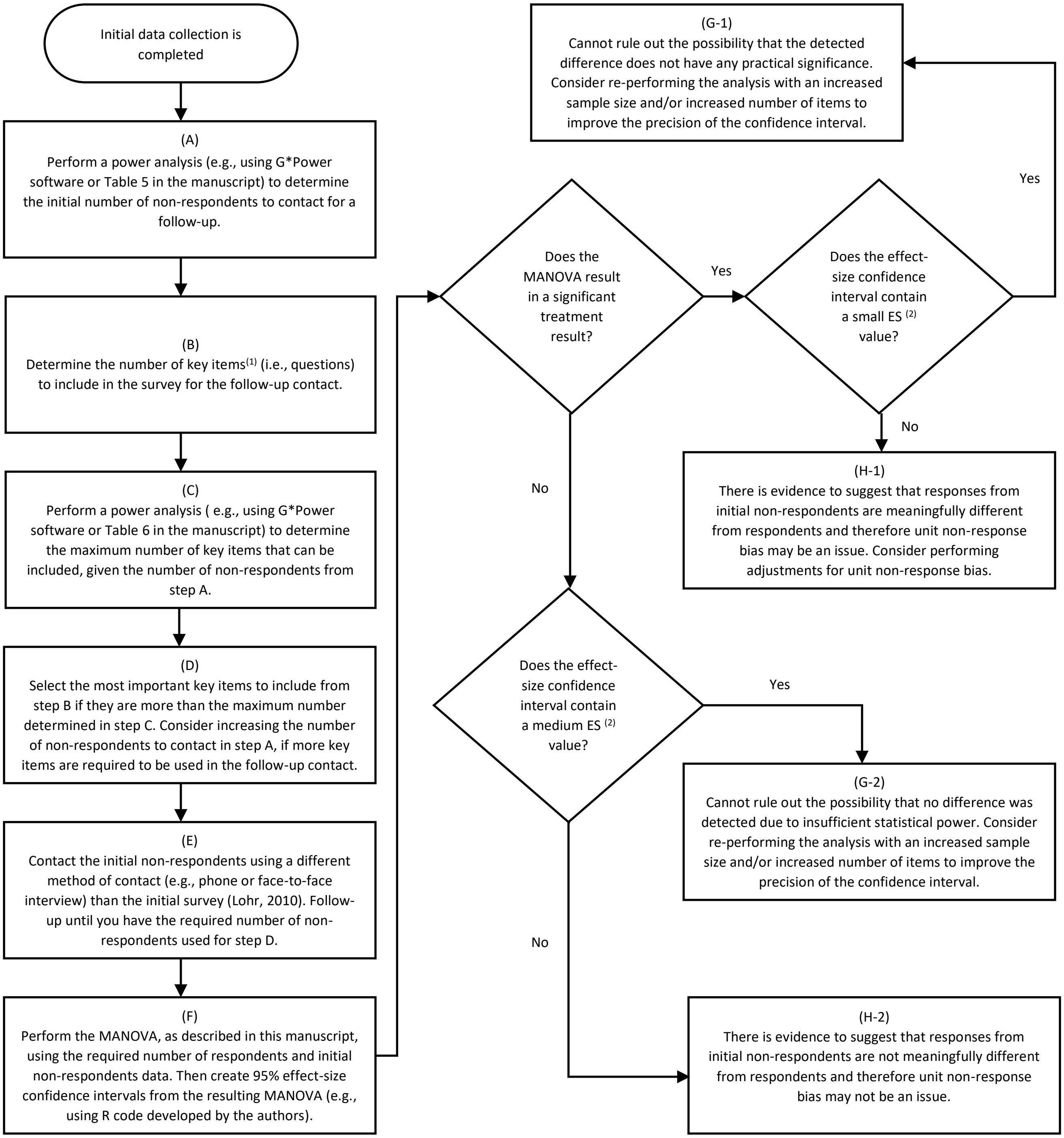
Whitehead, J. C., P. A. Groothuis and G. C. Blomquist. 1993. Testing for non-response and sample selection bias in contingent valuation. *Economic Letters* 41, 215–220. https://doi.org/ 10.1016/0165-1765(93) 90200-V

Yan, T. and R. Curtin. 2010. The relation between unit nonresponse and item nonresponse: A response continuum perspective. *International Journal of Public Opinion Research*, 22, 535- 551. doi:10.1 093/ijpor/edq

Zinkhan, G.M. 2003. A look to the future of JAMS: Three years out, thirty years out … *Journal of the Academy of Marketing Science*, 31(3), 225–228. https://doi.org/10.1177 /0092070303031003001

Appendix A

Simplified Procedure for Performing a MANOVA, with Effect Size Confidence Intervals, to Assess for Potential Unit Non-response Bias with Dyadic Data


Initial data collection is completed
(A) Perform a power analysis (e.g., using G*Power software or Table 5 in the manuscript) to determine the initial number of non-respondents to contact for a follow-up.
(B) Determine the number of key items(1) (i.e., questions) to include in the survey for the follow-up contact.
(C) Perform a power analysis (e.g., using G*Power software or Table 6 in the manuscript) to determine the maximum number of key items that can be included, given the number of non-respondents from step A.
(D) Select the most important key items to include from step B if they are more than the maximum number determined in step C. Consider increasing the number of non-respondents to contact in step A, if more key items are required to be used in the follow-up contact.
(E) Contact the initial non-respondents using a different method of contact (e.g., phone or face-to-face interview) than the initial survey (Lohr, 2010). Follow-up until you have the required number of non-respondents used for step D.
(F) Perform the MANOVA, as described in this manuscript, using the required number of respondents and initial non-respondents data. Then create 95% effect-size confidence intervals from the resulting MANOVA (e.g., using R code developed by the authors).
Does the MANOVA result in a significant treatment result?
Yes
No
Does the effect-size confidence interval contain a small ES (2) value?
Yes
No
(G-1) Cannot rule out the possibility that the detected difference does not have any practical significance. Consider re-performing the analysis with an increased sample size and/or increased number of items to improve the precision of the confidence interval.
(H-1) There is evidence to suggest that responses from initial non-respondents are meaningfully different from respondents and therefore unit non-response bias may be an issue. Consider performing adjustments for unit non-response bias.
Does the effect-size confidence interval contain a medium ES (2) value?
Yes
No
(G-2) Cannot rule out the possibility that no difference was detected due to insufficient statistical power. Consider re-performing the analysis with an increased sample size and/or increased number of items to improve the precision of the confidence interval.
(H-2) There is evidence to suggest that responses from initial non-respondents are not meaningfully different from respondents and therefore unit non-response bias may not be an issue.

[1] Some choices for key items include: 1) items which are key to the outcome of the study (Lohr, 2010), 2) items specific to the survey which are known to possibly affect the propensity of a unit to participate in the survey (Groves et al., 2009), and 3) general items (e.g., company size) which have a track record in related surveys of potentially affecting response propensities (Thompson & Washington, 2013)

[2] ES values are based on Cohen (1988: p. 478) indices.