# Non-response bias assessment in logistics survey research: Use fewer tests?


## Abstract

**Purpose –** The current research considers the concepts of *individual* and *complete* statistical power used for multiple testing and shows their relevance for determining the number of statistical tests to perform when assessing non-response bias.

**Methodology/approach –** A statistical power analysis of 55 survey-based research papers published in three prestigious logistics journals (*International Journal of Physical Distribution and Logistics Management, Journal of Business Logistics, Transportation Journal*) over the last decade was conducted.

**Findings –** Results show that some of the low complete power levels encountered could have been avoided if fewer tests had been used in the assessment of non-response bias.

**Originality/value of paper –** The research offers important recommendations to scholars engaged in survey research as they assess the effects of non-respondents on research findings. By following the recommended strategies for testing non-response bias, researchers can improve the statistical power of their findings.



**Keywords –** Non-response bias, statistical power analysis, survey research methods

**Paper type –** Research paper

# Non-response bias assessment in logistics survey research: Use fewer tests?

## 1. Introduction

Survey research is favored among many researchers looking to gain input from managers working in the field. As logistics scholars, we have been challenged to further engage with practitioners to observe the "real world" and "calibrate our research agendas against business needs and challenges," (Waller et al., 2012, p. 76). The input from the respondents is used to test hypotheses and build theory to help us understand the factors that lead businesses to succeed. One of the primary goals outlined by Mentzer (2008) for scholars in our field is to maximize the generalizability of the research. This means that we need to ensure that our research samples sufficiently represent the population of interest. Researchers engaged in survey administration are faced with a significant challenge in this area as participants are becoming harder to find. Over the past several years, response rates to survey requests have declined (Griffis et al., 2003; Larson, 2005). As such, the researchers are left wondering how well the respondents represent the non-respondents and if there are underlying reasons for the non-response of invited participants.

Recently, De Beuckelaer and Wagner (2012) examined research published in leading supply chain journals and addressed the issue of small sample survey research, indicating that researchers risk losing statistical power in their findings when using such samples. The authors also suggested that researchers should "test for the possibility of nonresponse bias, and reflect on reasons as to why the sample is so small," (De Beuckelaer and Wagner, 2012, p. 629). Our research expands on these suggestions by specifically addressing the issues of non-response bias testing and statistical power.

Non-response bias can be described as the result of people who respond to a survey being different from sampled individuals who did not respond, in a way relevant to the study (Dillman, 2007). When respondents differ from non-respondents, statistics (e.g., regression and path coefficients) based on responses alone often do not validly depict the population investigated and may result in predictions which are inaccurate, unreliable and misleading (Filion, 1975; Lohr, 2001; Wagner and Kemmerling, 2010). Even, low rates of non-response can have large effects on the results of a survey. Lohr (2001, p. 256-257) reports a result from a 1969 survey in Norway, concerning the voting rate for an issue of concern, in which a less than 10% non-response rate led to a 17% overestimation of the population voting rate. It is therefore important for non-response bias to be assessed, and adjustments made for the bias if detected.

An increasing amount of attention is being paid to the issue of non-response bias in logistics survey research. Wagner and Kemmerling (2010) reported the results of a content analysis of methods used by logistics scholars to assess non-response bias in three logistics journals (*International Journal of Physical Distribution and Logistics Management, Journal of Business Logistics, Transportation Journal*). They found the following four most commonly used methods to assess non-response bias:

- Comparison of responses from early vs. late respondents (assumes that late respondents are most similar to non-respondents because their replies required more prodding and took the longest time).
- Comparison of responses from respondents vs. responses from a random sample of non-respondents obtained after a pre-cutoff date.
- Comparison of respondents vs. non-respondents on multiple characteristics (usually demographic).

- Comparison of the demographics of respondents to those of the population.

The first two methods listed above were used in over 80% of the logistics journal articles surveyed by Wagner and Kemmerling (2010) and have the following procedure in common. They often involve some form of statistical analysis, usually a t-test comparing group means, and multiple instances of these tests are performed (i.e., a test for each item of interest). A survey item of interest can be a survey question, scored on a rating scale, or a survey construct, which is a factor made up of multiple survey questions. There have been different approaches used in determining the number of survey items to use in the comparisons. For example Lambert and Harrington (1990, p.17) used 51 survey items which were identified by experts as being very important to the study, resulting in 51 Analysis of Variance (ANOVA) statistical tests comparing respondents to non-respondents. Based on sample costs considerations, Mentzer and Flint (1997, p.206) suggested that the use of five non-demographic survey questions was a sufficient number for comparing respondents to non-respondents. Results shown later in Tables 2 and 3 of our survey of the logistics literature indicate that the average number of survey items employed, when using either of the previous two methods for assessing non-response bias, is 14. Probably the most significant implication of the number of survey items, used in the statistical tests to assess non-response bias, is its effect on the statistical power of the tests.

In a study where multiple statistical tests have been used to jointly assess non-response bias, all of the tests need to result in non-significance in order to provide strong support that non-response bias is not an issue in the study. The number of tests performed affects the probability that all of the tests will jointly, correctly, detect a difference between group means (Senn and Bretz, 2007). Westfall et al. (1999) describe this probability as *complete power* and labeled the power for a single test in isolation of other tests as *individual power*. The complete

power of tests is most applicable to the situation whereby multiple tests are used to assess a single issue (e.g., drug efficacy) and all the tests may fail to detect a difference between group means. This is the situation when non-response bias is being assessed with multiple statistical tests. In this study we show that the consideration of complete power has implications for the two commonly used techniques for assessing non-response bias in logistics research. Specifically, we will (i) present the concept of complete statistical power and discuss the relevance for tests comparing group means, (ii) present a summary of statistical power calculations of tests for non-response bias conducted in 55 articles published in the logistics literature within the last decade, and (iii) discuss the implications of power analysis for the determination of the number of tests performed, when assessing non-response bias in logistics research.

## 2 Literature Review

### *2.1 Statistical Power Analysis for Multiple Tests Comparing Two Means*

Statistical inference is the process of drawing conclusions from data that are subject to random variation (Upton and Cook, 2008). Statistical power analysis exploits the relationships among four variables involved in statistical inference: sample size ($N$), significance criterion ($\alpha$), population effect size ($ES$), and statistical power (Cohen, 1988; Cohen, 1992). Cohen (1992) provides a table of $ES$ indexes for eight popular statistical tests including t and F tests which were the most used statistical tests in our survey of non-response bias assessment, in the logistics literature. To convey the meaning of any given $ES$ index, Cohen (1992) proposed the operational conventions of small, medium and large $ES$ indexes. He noted that the medium $ES$ index represents 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. While

$ES$ can be estimated using sample means, standard deviations and/or proportions (Cohen, 1988), our survey of the logistics literature revealed that such sample estimates were not provided for the items used to assess non-response bias. Therefore, all power calculations performed in this study were carried out using the conventional medium $ES$ index as specified in Cohen (1992). Verma and Goodale (1995) found medium effect sizes for a majority of articles published in *Decision Sciences* and the *Journal of Operations Management*. We could not find a study that had investigated effect sizes for the three logistics journals examined in this study. Nevertheless, the Verma and Goodale (1995) study does provide support for our use of medium effect sizes in this study.

In comparing the population means ($\mu$) for two groups, we are formally testing the null hypothesis: $H_0$:$\mu_1 = \mu_2$; with an alternative hypothesis: $H_a$:$\mu_1 \neq \mu_2$. The improbability of $H_0$ is assessed by measuring the observed statistical difference between the sample means using the following test statistic ($t$):

$$t = \frac{\bar{\bar{x}}_1 - \bar{\bar{x}}_2}{\hat{\sigma}\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}, \qquad (1)$$

where $\bar{\bar{x}}_j$ is the sample mean for group $j$ ($j$=1 or 2), $n_j$ is the sample size for group $j$, and $\hat{\sigma}$ is the pooled sample standard deviation. Since the use of $t$ in (1) results in two rejection regions when used to test $H_a$: $\mu_1 \neq \mu_2$ , it is usually more straightforward to use the statistic $F = t^2$, which results in an F-test (i.e., one-way ANOVA) with a single rejection region.

Typically, when assessing non-response bias for surveys the responses of each group (i.e., early vs. late respondents or respondents vs. non-respondents) on multiple survey items are compared. This results in the statistical testing of a number of null hypotheses $H_{0,1}$, $H_{0,2}$,…….,$H_{0,k}$, ($k$=number of tests performed), for which we have a number of alternative hypotheses $H_{a,1}$, $H_{a,2}$,…….,$H_{a,\,k}$. This could be done with a set of t-test statistics $t_1, t_2, \ldots\ldots, t_k$ , of the form

represented in (1), or using a set of F-test statistics $F_1, F_2, \ldots \ldots, F_k$, where $F_i = t_i^2$ $(i = 1,2, \ldots, k)$. If each test is conducted at an $\alpha$ significance level, then the probability of falsely rejecting at least one of the null hypotheses and therefore committing a type 1 error is greater than $\alpha$. In order to cap the maximum risk of falsely rejecting any of the null hypotheses and thereby allowing significance levels for single and multiple tests to be directly comparable, a multiplicity adjustment is required. Several methods for multiplicity adjustments have been proposed. The reader is referred to Miller (1981) and Hsu (1996) for a detailed treatment of these adjustments. In this study we will be using the popular Bonferroni adjustment for controlling type 1 error. With the Bonferroni method, each of the $k$ tests is conducted at an $\alpha/k$ significance level and this caps the maximum probability of falsely rejecting any of the $k$ null hypotheses at $\alpha$ (i.e., the family error rate). Given the family error rate $(\alpha)$ and $ES$, the *complete* power for testing multiple null hypotheses, is the probability of jointly rejecting all of the false null hypotheses (Westfall et al., 1999). Complete power is particularly relevant to the assessment of non-response bias, since the consideration of the complete power of tests is most applicable to the situation whereby all the hypothesis tests may fail to reject the null hypothesis. In order to have strong support that non-response bias is indeed not an issue, all the statistical tests should result in the correct non rejection of their respective null hypothesis. Failure to correctly reject any of the null hypotheses, as a result of low statistical power, would lead to the incorrect conclusion that non-response bias may not be a concern. Previous applications of complete power have occurred in the pharmaceutical research area (Senn and Bretz, 2007; Dmitrienko et al., 2009) where joint statistical tests yielding costly non rejections have been known to occur. In order to determine the complete power, the joint distribution of the test statistics is required. We illustrate this using the set $F_1, F_2, \ldots \ldots, F_k$ of F-test statistics. The researcher must decide if the tests are consistent with the joint sampling

distribution of the $F$ statistics, if the null hypothesis is true. When the null hypothesis is true and each of the $k$ test statistics marginally follow a central F-distribution with 1 and $n_1 + n_2 - 2$ degrees of freedom then the set $F_1, F_2, \ldots \ldots, F_k$ follows a $k$-variate F-distribution with 1 and $k(n_1 + n_2 - 2)$ degrees of freedom and correlations $\rho_{ij}$ $(i, j = 1, \ldots., k; i \neq j)$ between each pair of the $k$ population measurements on which the test statistics are based upon. In practice, the population correlations $\rho_{ij}$ can be estimated by the sample correlations $\hat{\rho}_{ij}$ between the $k$ survey items being used to assess non-response bias. Given the $k$ test statistics with the Bonferroni adjustment applied to the critical values, individual and complete power are represented by the following equations:

$$\text{Individual power} = P\left(F_i > F_\alpha \,|H_{a,i}\ is\ True\right),$$
$$\text{Complete power} = \left[\cap_{i=1}^{k}\left(F_i > F_{\alpha/k}|H_{a,i}\ is\ True\right)\right], \quad (2)$$

where ∩ is the set operator for "and"; | is the operator for "given". For instance, with two independent test statistics each with a Bonferroni adjusted individual power of 0.80, the complete power would be 0.80 x 0.80 = 0.64. In general, when the test statistics are correlated, an exact solution for the probability $P\left[\cap_{i=1}^{k}\left(F_i > F_{\alpha/k}|H_{a,i}\ is\ True\right)\right]$ requires multivariate integration of the non-central F (or t) distribution, and therefore an exact solution is analytically intractable. However, approximations of the integration have been developed by Genz and Bretz (2009) for the **R** 2.6.1 programming language that was used to estimate complete power in this study. Complete power can also be estimated in SAS by using the proc MULTEST statement (Westfall et al., 1999).

### *2.1.1 Relation between Correlation Structure, Number of Tests and Complete Power*

In determining complete power, the correlations between the pairs of measurements play a key role. In the simple case when the correlations between each of the measures are identical (i.e.,

$\rho_{ij} = \rho$ for all $i \neq j$), the relationship between correlation, the number of tests, and complete power is shown in Figure 1 below.

*[Take in Figure 1 Here]*

In explaining Figure 1, we will be using the "weak" (±0.10 to 0.29), "moderate" (±0.30 to 0.49) and "strong" (±0.50 to 1.00) classifications for correlation as suggested by Cohen (1988). We will also be using the "low" (< 0.60), "medium" (≥ 0.60 and < 0.80) and "high" (≥ 0.80) classifications for power levels as suggested by Verma and Goodale (1995).

Figure 1 shows the complete power calculated as a function of the correlation coefficient for 3, 5, 10 and 20 t-tests where the power for a single test is 0.80. It clearly shows that the complete power is at its highest when there is a strong positive correlation between the measures used in the tests. This is because if the measures are highly correlated with each other, then one test correctly detecting a difference between group means would mean that the others are likely to correctly detect a difference as well. This increases the chance that all tests will, jointly, correctly detect a difference (i.e., the complete power of the tests). This indicates that in addition to a high individual power level for each test, the strength of the pair wise correlation between the tests also needs to be taken into account when assessing the complete power of the tests.

Another factor that affects complete power is the number of tests performed.

Figure 1 shows that the complete power of the tests decreases as the number of t-tests performed increases. These results show that even at strong positive correlations, performing a large number of tests will result in a complete power for the tests which is substantially smaller than the individual power of each test.

Figure 1 only shows the results when the common correlation is positive. When the common correlation between measures is negative this induces a "non-positive definite" correlation structure which is inconsistent with a multivariate t-distribution (Genz and Bretz, 2009). Therefore using our **R** code to estimate complete power with common negative correlation structures resulted in error warnings. We did not encounter any error warnings when estimating complete power with the actual correlation structures that we found in our survey of the logistics literature.

The implications of the insights gained from Figure 1 for assessing non-response bias in logistics research, is discussed in the next sections.

## 3. Methodology

### *3.1 Statistical Power Analysis of Tests Used for Assessing Non-response Bias in IJPDLM, JBL, and TJ*

A survey of 55 research articles published within the last decade in the *International Journal of Physical Distribution and Logistics Management* (vols. 30 to 42), *Journal of Business Logistics* (vols. 22 to 32), and *Transportation Journal* (vols. 40 to 51) was undertaken. Similar to other studies which have surveyed the logistics literature (e.g., Larson, 2005; Wagner and Kemmerling, 2010), we based our choice on evaluations which focus on the academic prestige, impact and readership of the logistics journals (Gibson and Hanna, 2003; Menachof et al., 2009). Also, these three journals are known for publishing empirical research in the logistics area (Spens and Kovacs, 2006; Wagner and Kemmerling, 2010).

Our study is based on articles published between 2000 and the first half of 2012, and to be included in the analysis the research article must have: i) used statistical tests to assess non-response bias by comparing early vs. late respondents or respondents vs. non-respondents, ii) used

t or F tests to assess non-response bias, iii) specified the sample size for both groups used for assessing non-response bias, iv) specified the number of tests performed, and v) provided estimates of pair wise correlations of survey items used for assessing non-response bias. The fourth and fifth criteria enabled us to obtain estimates of complete power for the tests used in the research articles, but these two criteria turned out to be the most restrictive in limiting the number of articles that could be included in our analysis. We therefore relaxed these criteria for individual power calculations and only computed complete power for those articles which had met the last two criteria. The initial dataset consisted of articles in the ProQuest and Emerald Insight online databases. The following keyword/phrases were used to identify articles that conformed to the four criteria mentioned previously: "bias", "nonresponse", "nonrespondents", "respondent", "survey", "structural equation model", "Armstrong and Overton", "Lambert and Harrington". These keywords were also used by Wagner and Kemmerling (2010) to identify articles which had dealt with non-response bias in the three logistics journals. However, they did not evaluate the statistical power of the tests used in the articles that they found. Likewise, studies investigating non-response bias in areas such as production and operations management (Malhotra and Grover, 1998), information systems (King and He, 2005), management (Werner et al., 2007), and marketing (Collier and Bienstock, 2007), did not evaluate the statistical power of the tests used to assess non-response bias in the articles that were surveyed.

We found a total of 110 articles in which non-response bias had been assessed by comparing two groups using a statistical test (e.g., t, F, chi-squared); however, only 85 of these articles used either t or F tests. Of these 85 articles, 55 met the first three criteria. Of these 55 articles, 38 met the first four criteria and only 16 of the 38 met all five criteria. The majority of the

articles provided $p$-values without specifying an $\alpha$ level for the tests used in assessing non-response bias, so we assumed an $\alpha = 0.05$ level for our power calculations.

We used **R** (version 2.6.1) code to estimate both individual and complete power. The individual power calculations, at a medium effect size, only required knowledge of the sample sizes for the two groups being compared and the $\alpha$ level used. Calculations of complete power were only made when information about the number of tests performed and estimates of pair wise correlations, of survey items used for assessing non-response bias, could be found in the article. Table 1 presents the individual and complete power values for articles in each journal when early vs. late respondents data was used, with t or F tests, to assess non-response bias. Table 1 shows that when this method was used with t or F tests for detecting non-response bias, 21 of the 40 articles (52.50%) did not have high ($\geq 0.80$) individual power levels. If we consider medium and high ($\geq 0.60$) power levels to be acceptable (Verma and Goodale, 1995, p.148) then 34 of the 40 articles (approximately 85%) in Table 1 had acceptable individual power. However, only four of the seven articles for which we were able to obtain complete power levels for in Table 1, had an acceptable level of complete power, indicating that non-response bias test results may be misleading.

*[Take in Table 1 here]*

Table 2 presents the individual and complete power levels for articles in each journal, when respondent vs. non-respondent data was used with t or F tests to assess non-response bias. Table 2 shows that when this method was used to assess non-response bias all of the tests were at the medium and high ($\geq 0.60$) levels of individual power. This could be due to the fact that most of the studies using this method referenced Lambert and Harrington (1990), who advocated that the selection of sample sizes should be based on statistical power considerations. However, both of

the articles for which we were able to obtain complete power levels for in Table 2, had low levels of complete power. The median (mean) complete power level in Tables 1 and 2 is 0.52 (0.62). This value suggests that, on average, there is approximately a forty eight (thirty eight) percent chance that the t or F tests, used to assess non-response bias in any of the articles that we surveyed, would indicate that there is **not** a significant difference between the two groups being compared on each of the items considered, even when there are real differences.

*[Take in Table 2 here]*

As mentioned, the average (median) number of tests used for all the articles in Tables 1 and 2 is fourteen (ten). In the next section we investigate the number of tests that would have resulted in medium and high ($\geq 0.60$) levels of complete power in the articles that we surveyed.

## 4. Analysis and Results

For researchers making complete power considerations, a "critical value" for the number of tests required to achieve a particular level of complete power for all t and F tests would be desirable. However, such a table is infeasible since complete power also depends on the correlation structure of the tests, which are typically unique for each study in which non-response bias is assessed. We therefore narrowed our scope to providing cut-off values for the number of tests that would have lead to acceptable ($> 0.60$) values of complete power in the journal articles that we surveyed. In each of the 9 articles for which we were able to calculate complete power levels for, the number of t or F tests used to assess non-response bias corresponded to the number of survey items used in these tests. Based on the insights from Figure 1, we realized that given the individual power levels of these articles the cut-off values for the number of tests was likely to fall in the 2 to 5 range for weak to moderate strength correlations. Therefore, for each of these articles we

created all possible sets of 2, 3, 4, and 5 survey items, out of the total number of survey items that were actually used to assess non-response bias in the article. For example, thirteen survey items corresponding to 13 t-tests were used in article #23 to assess non-response bias. If two t-tests had been used instead, then the number of all possible sets of 2 survey items that can be selected from the 13 items (without replacement) is 78. Therefore for article #23 we calculated the complete power values for each of the 78 sets of 2 survey items, and recorded the minimum and maximum of these values in addition to the percentage of the 78 that resulted in medium to high (> 0.60) values of complete power. We repeated these calculations for all possible sets of 2, 3, 4, and 5 survey items used to assess non-response bias in each of the 9 articles. The results of the analysis are shown in Table 3.

*[Take in Table 3 here]*

The average correlation between all survey items used to assess non-response bias in each of the articles in Table 3, ranged from 0.18 to 0.75 (i.e., weak to strong positive correlations on average). Given the correlation structures for the survey items used to assess non-response bias in these articles, Table 3 provides the following insights for logistics scholars.

Table 3 shows that for medium levels of individual power ranging from 0.60 to 0.70 with average correlations of 0.21 (i.e., weak), the use of 2 tests to assess non-response bias can result in medium (0.60 to 0.70) levels of complete power. However, for article #23 there was a small chance of this occurring since only 14% of all possible sets of two tests yielded power levels above 0.60, when the individual power was less than 0.70. There was an even smaller chance (i.e., 1%) of this occurring when 3 tests were used for the same article. For medium levels of power ranging from 0.71 to 0.80, and average correlations ranging from 0.18 to 0.69, the results in Table 3 show that the use of only two tests to assess non-response bias is likely to result in medium (0.60 to 0.80) levels of complete power. The lowest chance of this occurring was 67% with article #9, all other

articles with individual power levels ranging from 0.71 to 0.80 had a 100% chance that 2 tests would result in medium complete power levels. When 3 tests were used instead, 100% of all possible sets of three tests resulted in medium levels of complete power for article #3 and #53 with average correlations of 0.27 and 0.69, respectively. However, 0% of all possible sets of three tests resulted in medium levels of complete power for article #9 with an average correlation of 0.18.

Table 3 also shows that for high individual power levels ranging from 0.80 to 1.00 with average correlations ranging from 0.56 to 0.75, the use of 2 to 4 tests always resulted in medium levels of complete power.

In summary, the results from our analysis in Table 3 suggest that, given the correlation structures found in the logistics articles that we surveyed, logistics scholars wanting to ensure acceptable levels of complete power for t or F tests used to assess non-response bias should try and achieve individual power levels of the tests which are greater than 0.70. With individual power levels between 0.71 and 0.80, logistics researchers using only two (t or F) tests to assess non-response bias are likely to have medium levels of complete power. With high individual power levels ranging from 0.80 to 0.90, coupled with high average correlations between measurement items, logistics researchers using 2 to 4 tests to assess non-response bias are likely to have medium levels of power. With very high individual power levels (> 0.90), logistics researchers can achieve medium to high levels of complete power using more than four tests; however, complete power levels were highest for all of the articles in Table 3 when two, t or F tests, were jointly used to assess non-response bias.

## 5. Discussion

The results of our study show that the individual power of t and F tests used to assess non-response bias in *IJPDLM, JBL* and *TJ* ranged from a low of 0.22 to a high of 1.00. The complete

power for which we were able to compute, for some of these tests, ranged from a low of 0.18 to a high of 1.00, with almost half being less than 0.60. The low levels of complete power that we found could have been avoided if fewer survey items had been used when assessing non-response bias with t or F tests. Specifically, our analysis of the range of minimum and maximum complete power values shown in Table 3 revealed that the use of two t or F tests, to jointly assess non-response bias, always resulted in a higher level of complete power, over all articles and combinations of individual power levels and correlation structures that we encountered, than when 4 or more tests were used. The use of only two tests is likely to result in the highest level of complete power achievable, for a given correlation structure and individual power level of each test. Based on power considerations, we therefore recommend that two randomly selected survey items is a reasonable number for logistics researchers to use when assessing non-response bias with t or F tests.

Our analysis also showed that for a fixed number of tests, individual power levels greater than 0.70 for each t or F test being used to assess non-response bias, made it more likely that medium (0.60 to 0.80) levels of complete power would be achieved. This indicates that prescriptions for increasing the individual power of each test (e.g., adequate sample sizes for non-respondents) can also lead to an increase in the complete power of the tests.

In practice the number of variables needed to effectively assess non-response bias is dependent on the nature of the study and the research model. The focal set of survey items to use in the assessment of non-response bias is often based on theoretical grounds, expert opinion, and/or sampling costs. However, if the number of survey items in the focal set is large, resulting in the potential for insufficient statistical power, then a trade-off could be decided upon by the researcher (e.g., non-statistical assessment of the complete focal set versus statistical assessment of a small

subset). Consideration of individual and complete statistical power values should help the researcher in better assessing this trade-off.

According to Cohen (1992), the power of individual tests is a concept that has been around since 1928. Relative to individual power, the concept of complete power is not as well-known and is most applicable to the situation whereby multiple tests are used to assess a single issue and all the tests may fail to detect a difference between the groups being compared. Methods for estimating complete power have become available within the last decade in a handful of software programs (e.g. SAS and **R**). Evaluations of exact complete power levels of statistical tests are an additional burden and may be infeasible for logistics researchers who are unable to use SAS or **R** for their analyses. If two t or F tests are used by the researcher to assess non-response bias then a quick estimate of the complete power associated with using the two tests can be obtained from Table 4.

*[Take in Table 4 here]*

If the researcher is unable to estimate the complete power of the tests, then the researcher can improve the rigor in the assessment of non-response bias by: i) performing an individual power analysis, ii) specifying the number of tests used to assess non-response bias and iii) providing estimates of the correlations between pairs of survey items used in the assessment of non-response bias. Such information can be used by peers, who are able to compute complete power levels, to assess the tests for non-response bias performed in those research articles.

Lastly, our survey of the literature showed that none of the articles considered multiplicity adjustments (e.g., Bonferroni) when using multiple t and F tests to jointly assess non-response bias. When a multiplicity adjustment is not applied then the maximum risk of falsely rejecting any of the null hypotheses is higher than the $\alpha$ used for each test. This would result in complete power

considerations which are at higher significance levels than they should be, leading to erroneous conclusions. Several methods for multiplicity adjustments can be found in the statistics literature (Miller, 1981; Hsu, 1996), with each multiplicity adjustment having power implications. When the Bonferroni adjustment is applied then Table 4 can be used to estimate the complete power for two t or F tests.

*5.1 Adjusting Survey Results for Non-Response Bias*

When the assessment of non-response bias indicates that there is a risk for bias, it may be possible to use modeling methods to make predictions about the non-respondents. One such method is to take a representative subsample of the non-respondents and use that subsample to make inferences about the other non-respondents. This is the rationale behind the technique advocated by Hansen and Hurwitz (1946), in which survey results are weighted according to the proportion of the initial and subsample respondents in the total sample. Weights that can be used to adjust survey results for non-response bias include: weighting class adjustments (Holt and Elliot, 1991; Lin and Schaeffer, 1995), poststratification weights and raking adjustments (Oh and Scheuren, 1983). Each weighting method has an assumed model underlying its use. If weighting adjustments are made, then the researcher should always state the assumed non-response model and give evidence to justify it. A reference for more information on weighting methods, used to adjust for non-response, and the assumptions underlying them is Oh and Scheuren (1983) and Lohr (2001).

Weighting methods do not make use of any relationships between the variables of interest in a survey and the non-response. Parametric models for non-response in which a model is developed for the complete data and components are added to the model to account for the proposed non-response mechanism, do make use of the possible relationship between non-

response and variables of interest. The extrapolation method of Armstrong and Overton (1979) and the selection bias adjustment by Heckman (1979) are two examples of parametric methods. Stasny (1991) discusses several parametric methods that can be used to adjust for non-response. Parametric methods for non-response adjustments often require (i) a thorough knowledge of mathematical statistics, (ii) a powerful computer and (iii) knowledge of numerical methods for optimization (Lohr, 2001).

## 6. Conclusion

In this research we consider the concepts of *individual* and *complete* statistical power, used for multiple testing, and show their relevance for determining the number of statistical tests to perform when assessing non-response bias in logistics research. Specifically, our results showed that some of the low complete power levels encountered in our analysis of the logistics journals could have been avoided if fewer tests had been used in the assessment of non-response bias. To further improve the statistical power of non-response bias assessment, scholars should consider methods for increasing the sample of non-respondents (using only a few items) as our research shows that this can improve both individual and complete power levels.

**References**


Armstrong, J.S. and Overton, T.S. (1977), "Estimating nonresponse bias in mail surveys", *Journal of Marketing Research,* Vol. 14 No. 3, pp. 396-402.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), Erlbaum, Hillsdale, NJ.

Cohen, J. (1992), "A power primer", *Psychological Bulletin,* Vol. 112 No. 1, pp. 155-159.

Collier, J.E. and Bienstock C.C. (2007), "An analysis of how nonresponse error is assessed in academic marketing research", *Marketing Theory,* Vol. 7 No. 2, pp. 163-183.

De Beuckelaer, A. and Wagner, S.M. (2012), "Small sample surveys: increasing rigor in supply chain management research", *International Journal of Physical Distribution & Logistics Management*, Vol. 42 No. 7, pp. 615-639.

Dmitrienko, A., Tamhane, C.A., and Bretz, F. (2009), *Multiple Testing Problems in Pharmaceutical Statistics,* Chapman and Hall/CRC, Boca Raton, FL.

Dillman, D. (2007), *Mail and Internet Surveys: The Tailored Design Method 2007 Update with New Internet, Visual, and Mixed-Mode Guide,* John Wiley & Sons, Hoboken, NJ.

Filion, F.L. (1975), "Estimating bias due to nonresponse in mail surveys", *Public Opinion Quarterly,* Vol. 39 No. 4, pp. 482-491.

Gibson, B.J. and Hanna, J.B. (2003), "Periodical usefulness: the U.S. logistics educator perspective", *Journal of Business Logistics*, Vol. 24 No. 1, pp. 221-240.

Genz, A. and Bretz, F. (2009), *Computation of Multivariate Normal and t Probabilities*, Springer, Heidelberg, Germany.

Griffis, S.E., Goldsby, T.J., and Cooper, M. (2003), "Web-based and mail surveys: a comparison of response, data, and cost", *Journal of Business Logistics,* Vol. 24 No. 2, pp. 237-258.

Hansen, M.H. and Hurwitz, N.W. (1946), "The problem of nonresponse in sample surveys", *Journal of the American Statistical Association,* Vol. 41 No. 236, pp. 517-529.

Heckman, J.J. (1979), "Sample selection bias as a specification error", *Econometrica,* Vol. 47 No. 2, pp. 153-161.

Holt, D. and Elliot, D. (1991), "Methods of weighting for unit non-response", *The Statistician,* Vol. 40 No. 3, pp. 333-342.

Hsu J.C. (1996), *Multiple Comparisons: Theory and Methods,* Champan and Hall/CRC, London, England.

King, W.R. and He, J. (2005), "External validity in IS survey research", *Communications of the Association for Information Systems,* Vol. 16 No. 45, pp. 880-894.

Lambert, D.M. and Harrington, T.C. (1990), "Measuring nonresponse bias in customer service mail surveys", *Journal of Business Logistics,* Vol. 11 No. 2 pp. 5-25.

Larson, P.D. (2005), "A note on mail surveys and response rates in logistics research", *Journal of Business Logistics,* Vol. 26 No. 2, pp. 211-222.

Lin, I.F., and Schaeffer, N.C. (1995), "Using survey participants to estimate the impact of nonparticipation", *Public Opinion Quarterly,* Vol. 59 No.2, pp. 236-258.

Lohr, S. (2001), *Sampling: Design and Analysis,* Duxbury Press, Boston, MA.

Malhotra, K. and Grover, V. (1998), "An assessment of survey research in POM: from constructs to theory", *Journal of Operations Management,* Vol. 16 No. 4, pp. 407-425.

Menachof, D.A.,Gibson, B.J, Hanna, J.B., and Whiteing, A.E. (2009), "An analysis of the value of supply chain management periodicals", *International Journal of Physical Distribution and Logistics Management,* Vol. 39 No. 2, pp. 145-166.

Mentzer, J.T. (2008), "Rigor versus relevance: why would we choose only one?", *Journal of Supply Chain Management,* Vol. 44 No. 2, pp. 72-77.

Mentzer, J.T. and Flint, D.J. (1997), "Validity in logistics research", *Journal of Business Logistics,* Vol. 18 No. 1, pp. 199-216.

Miller, R.G. (1981), *Simultaneous Statistical Inference* (2nd Edition), Springer, New York, NY.

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

Senn, S. and Bretz, F. (2007), "Power and sample size when multiple endpoints are considered", *Pharmaceutical Statistics,* Vol. 6 No. 3, pp. 161-170.

Spens, K.M. and Kovács, G. (2006), "A content analysis of research approaches in logistics research", *International Journal of Physical Distribution and Logistics Management,* Vol. 36 No. 5, pp. 374-390.

Stasny, E.A. (1991), "Hierarchical models for the probabilities of a survey classification and nonresponse", *Journal of the American Statistical Association,* Vol. 86 No. 414, pp. 296-303.

Upton, J.G. and Cook, I. (2008), *A Dictionary of Statistics,* Oxford University Press, London, England.

Verma R. and Goodale J.C . (1995), "Statistical power in operations management research", *Journal of Operations Management,* Vol. 13 No. 2, pp. 139-152.

Wagner, S.M. and Kemmerling. R. (2010), "Handling nonresponse in logistics research", *Journal of Business Logistics,* Vol. 31 No. 2, pp. 357-381.

Waller, M.A., Fawcett, S.E., and Van Hoek, R. (2012), "Thought leaders and thoughtful leaders: advancing logistics and supply chain management", *Journal of Business Logistics,* Vol. 33 No. 2, pp. 75-77.

Werner, S., Praxedes, M. and Kim, H. (2007), "The reporting of nonresponse analyses in survey research", *Organizational Research Methods,* Vol. 10 No. 2 pp. 287-295.

Westfall P.H, Tobias R.D, Rom .D, Wolfinger R.D, Hochberg Y. (1999), *Multiple Comparisons and Multiple Tests Using the SAS System*, SAS, Cary, NC.

**Table I.** Individual and complete power levels for articles in which a late vs. early respondents comparison has been made with t or F tests to assess non-response bias

| # | Date | Journal | Number of Tests (k) | Sample size | | Individual Power | Complete Power |
|---|---|---|---|---|---|---|---|
| | | | | Early Respondents | Late Respondents | | |
| 1 | 2012 | IJPDLM | 5 | 82 | 80 | 0.8854 | 0.7929 |
| 2 | 2011 | IJPDLM | 21 | 252 | 73 | 0.9633 | No correlations |
| 3 | 2011 | IJPDLM | 6 | 100 | 45 | 0.7901 | 0.5151 |
| 4 | 2010 | IJPDLM | 13 | 77 | 37 | 0.6978 | No correlations |
| 5 | 2009 | IJPDLM | - | 64 | 64 | 0.8015 | No k |
| 6 | 2008 | IJPDLM | - | 30 | 30 | 0.4779 | No k |
| 7 | 2007 | IJPDLM | 17 | 26 | 8 | 0.2244 | No correlations |
| 8 | 2007 | IJPDLM | 16 | 111 | 41 | 0.7759 | No correlations |
| 9 | 2005 | IJPDLM | 6 | 101 | 34 | 0.7066 | 0.3393 |
| 10 | 2005 | IJPDLM | 21 | 229 | 76 | 0.9645 | No correlations |
| 11 | 2005 | IJPDLM | 25 | 51 | 25 | 0.5245 | No correlations |
| 12 | 2004 | IJPDLM | 21 | 121 | 31 | 0.6943 | No correlations |
| 13 | 2003 | IJPDLM | 6 | 364 | 58 | 0.9416 | No correlations |
| 14 | 2001 | IJPDLM | - | 53 | 18 | 0.4394 | No k |
| 15 | 2001 | IJPDLM | - | 19 | 22 | 0.3438 | No k |
| 16 | 2001 | IJPDLM | 7 | 107 | 141 | 0.9729 | No correlations |
| 17 | 2010 | JBL | 17 | 50 | 50 | 0.6969 | No correlations |
| 18 | 2009 | JBL | 6 | 100 | 100 | 0.9404 | 0.8721 |
| 19 | 2009 | JBL | - | 145 | 45 | 0.8302 | No k |
| 20 | 2008 | JBL | 11 | 100 | 100 | 0.9404 | No correlations |
| 21 | 2008 | JBL | 8 | 258 | 40 | 0.8347 | No correlations |
| 22 | 2007 | JBL | 7 | 518 | 173 | 0.9999 | 0.9998 |
| 23 | 2006 | JBL | 13 | 121 | 31 | 0.6943 | 0.1773 |
| 24 | 2004 | JBL | 8 | 117 | 40 | 0.7743 | No correlations |
| 25 | 2004 | JBL | 28 | 230 | 76 | 0.9646 | No correlations |
| 26 | 2003 | JBL | 11 | 49 | 58 | 0.7235 | No correlations |
| 27 | 2002 | JBL | 13 | 156 | 52 | 0.8745 | No correlations |
| 28 | 2001 | JBL | 31 | 229 | 76 | 0.9645 | No correlations |
| 29 | 2001 | JBL | - | 74 | 24 | 0.5586 | No k |
| 30 | 2000 | JBL | - | 229 | 76 | 0.9645 | No k |

**Table I.** continued

| # | Date | Journal | Number of Tests (k) | Sample size | | Individual Power | Complete Power |
|---|---|---|---|---|---|---|---|
| | | | | Early Respondents | Late Respondents | | |
| 31 | 2009 | TJ | 30 | 58 | 37 | 0.6524 | No correlations |
| 32 | 2007 | TJ | 41 | 92 | 45 | 0.7791 | No correlations |
| 33 | 2006 | TJ | 20 | 60 | 42 | 0.6919 | No correlations |
| 34 | 2006 | TJ | 30 | 51 | 41 | 0.6548 | No correlations |
| 35 | 2005 | TJ | - | 97 | 98 | 0.9349 | No k |
| 36 | 2003 | TJ | 9 | 308 | 257 | 1.0000 | No correlations |
| 37 | 2002 | TJ | - | 61 | 65 | 0.7948 | No k |
| 38 | 2002 | TJ | - | 364 | 58 | 0.9416 | No k |
| 39 | 2001 | TJ | 5 | 230 | 76 | 0.9646 | 0.9422 |
| 40 | 2000 | TJ | 65 | 43 | 46 | 0.6447 | No correlations |

Note: IJPDLM= International Journal of Physical Distribution and Logistics Management; JBL=Journal of Business Logistics; TJ = Transportation Journal. The Bonferroni adjustment was applied for all calculations of complete power. The Bonferroni adjustment was not applied to the calculations of individual power listed in the table. The individual power for each test was assessed at the 5% (unadjusted) significance level at a medium effect size.

**Table II.** Individual and complete power levels for articles in which a respondents vs. non-respondents comparison has been made with t or F tests to assess non-response bias

| # | Date | Journal | Number of Tests (k) | Sample size | | Individual Power | Complete Power |
|---|---|---|---|---|---|---|---|
| | | | | Respondents | Non-respondents | | |
| 2 | 2011 | IJPDLM | - | 325 | 30 | 0.7432 | No k |
| 41 | 2011 | IJPDLM | - | 226 | 250 | 0.9997 | No k |
| 42 | 2010 | IJPDLM | 3 | 124 | 30 | 0.6851 | 0.4517 |
| 43 | 2009 | IJPDLM | 10 | 304 | 30 | 0.7406 | No correlations |
| 44 | 2005 | IJPDLM | - | 201 | 20 | 0.5648 | No k |
| 45 | 2004 | IJPDLM | 10 | 152 | 28 | 0.6766 | No correlations |
| 46 | 2004 | IJPDLM | 6 | 143 | 33 | 0.7306 | No correlations |
| 47 | 2010 | JBL | 5 | 336 | 32 | 0.7690 | No correlations |
| 48 | 2010 | JBL | 5 | 389 | 30 | 0.7494 | No correlations |
| 49 | 2010 | JBL | 8 | 254 | 300 | 1.0000 | No correlations |
| 50 | 2009 | JBL | 10 | 304 | 30 | 0.7406 | No correlations |
| 51 | 2007 | JBL | 6 | 296 | 34 | 0.7862 | No correlations |
| 52 | 2006 | JBL | 10 | 152 | 300 | 0.9989 | No correlations |
| 53 | 2006 | JBL | 5 | 322 | 31 | 0.7554 | 0.4453 |
| 54 | 2004 | JBL | 5 | 302 | 30 | 0.7403 | No correlations |
| 55 | 2004 | JBL | 2 | 142 | 530 | 0.9997 | No correlations |

Note: IJPDLM= International Journal of Physical Distribution and Logistics Management; JBL=Journal of Business Logistics; TJ = Transportation Journal. The Bonferroni adjustment was applied for all calculations of complete power. The Bonferroni adjustment was not applied to the calculations of individual power listed in the table. The individual power for each test was assessed at the 5% (unadjusted) significance level at a medium effect size.

**Table III.** Range of complete power values of all possible sets of 2, 3, 4 and 5 survey items

| Article # | Actual # of survey items used (k) | Complete power at k | Avg. correlation | Individual power | | Number of survey items used in the analysis | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 2 | 3 | 4 | 5 |
| 1 | 5 | 0.7929 | 0.56 | 0.8854 | [min, max] | [0.8556, 0.8712] | [0.8278, 0.8464] | [0.8069, 0.8227] | |
| | | | | | Percent with power >0.60 | 100% | 100% | 100% | N/A |
| 3 | 6 | 0.5151 | 0. 27 | 0.7901 | [min, max] | [0.7096, 0.7330] | [0.6453, 0.6669] | [0.5921, 0.6138] | [0.5507, 0.5626] |
| | | | | | Percent with power >0.60 | 100% | 100% | 67% | 0% |
| 9 | 6 | 0.3393 | 0. 18 | 0.7066 | [min, max] | [0.5859, 0.6298] | [0.4891, 0.5389] | [0.4256, 0.4647] | [0.3746, 0.4050] |
| | | | | | Percent with power >0.60 | 67% | 0% | 0% | 0% |
| 18 | 6 | 0.8721 | 0.58 | 0.9404 | [min, max] | [0.9222, 0.9434] | [0.9014, 0.9338] | [0.8863, 0.9146] | [0.8772, 0.8951] |
| | | | | | Percent with power >0.60 | 100% | 100% | 100% | 100% |
| 22 | 7 | 0.9998 | 0.36 | 0.9999 | [min, max] | [0.9999, 1.0000] | [0.9998, 1.0000] | [0.9998, 0.9999] | [0.9998, 0.9998] |
| | | | | | Percent with power >0.60 | 100% | 100% | 100% | 100% |

**Table III.** continued

| Article # | Actual # of survey items used (k) | Complete power at k | Avg. correlation | Individual power | | Number of survey items used in the analysis | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | 2 | 3 | 4 | 5 |
| 23 | 13 | 0.1773 | 0.21 | 0.6943 | [min, max] | [0.5668, 0.6795] | [0.4553, 0.6448] | [0.3706, 0.5354] | [0.3170, 0.4771] |
| | | | | | Percent with power >0.60 | 14% | 1% | 0% | 0% |
| 39 | 5 | 0.9422 | 0.75 | 0.9646 | [min, max] | [0.9602, 0.9638] | [0.9544, 0.9587] | [0.9492, 0.9534] | [0.9447, 0.9484] |
| | | | | | Percent with power >0.60 | 100% | 100% | 100% | 100% |
| 42 | 3 | 0.4517 | 0.21 | 0.6851 | [min, max] | [0.6108, 0.6297] | N/A | N/A | N/A |
| | | | | | Percent with power >0.60 | 100% | N/A | N/A | N/A |
| 53 | 5 | 0.4453 | 0.69 | 0.7554 | [min, max] | [0.7025, 0.7432] | [0.6642, 0.7105] | [0.6432, 0.6781] | N/A |
| | | | | | Percent with power >0.60 | 100% | 100% | 100% | N/A |

Notes: "[min,max]" = minimum and maximum complete power values; "Percent with power > 0.60"= percent of all possible sets of the given number of survey items which had medium to high (>0.6) values of complete power ; "Avg. Correlation"=average of all the pair wise correlations between survey items used for assessing non-response bias in the numbered journal article.

**Table IV.** Complete power values for two t or F-tests jointly used to assess differences between two group means

| | Individual Power | | | |
|---|---|---|---|---|
| **Correlation** | **0.70** | **0.75** | **0.80** | **0.90** |
| **-1** | 0.40 | 0.51 | 0.60 | 0.80 |
| **-0.9** | 0.40 | 0.51 | 0.60 | 0.80 |
| **-0.8** | 0.41 | 0.51 | 0.60 | 0.80 |
| **-0.7** | 0.42 | 0.51 | 0.61 | 0.80 |
| **-0.6** | 0.43 | 0.52 | 0.61 | 0.80 |
| **-0.5** | 0.44 | 0.53 | 0.61 | 0.80 |
| **-0.4** | 0.45 | 0.53 | 0.62 | 0.81 |
| **-0.3** | 0.46 | 0.54 | 0.62 | 0.81 |
| **-0.2** | 0.47 | 0.55 | 0.63 | 0.81 |
| **-0.1** | 0.48 | 0.56 | 0.64 | 0.81 |
| **0** | 0.49 | 0.57 | 0.64 | 0.81 |
| **0.1** | 0.51 | 0.58 | 0.65 | 0.82 |
| **0.2** | 0.52 | 0.59 | 0.66 | 0.82 |
| **0.3** | 0.53 | 0.60 | 0.67 | 0.83 |
| **0.4** | 0.54 | 0.61 | 0.68 | 0.83 |
| **0.5** | 0.56 | 0.63 | 0.69 | 0.84 |
| **0.6** | 0.57 | 0.64 | 0.70 | 0.84 |
| **0.7** | 0.59 | 0.66 | 0.72 | 0.85 |
| **0.8** | 0.61 | 0.67 | 0.73 | 0.86 |
| **0.9** | 0.64 | 0.70 | 0.75 | 0.87 |
| **1** | 0.70 | 0.75 | 0.80 | 0.90 |

Note: Table 4 provides the complete power values when two t or F tests, with common individual power levels ranging from 0.70 to 0.90, are jointly used for comparing two group means. The complete power values are computed at a medium effect size, for correlations ranging from -1 to +1, with the Bonferroni adjustment applied and (unadjusted) significance level $\alpha$=0.05. Power calculations for large or small effect sizes are available upon request from the authors. The correlations refer to pair wise correlations between the two measures used in the two tests. The individual power values listed in Table 4 are the unadjusted power values.

**Figure 1.** Complete power for multiple t-tests, with common correlations between measures, where the Bonferroni adjustment has been applied and the power for each individual test at the 5% (unadjusted) significance level is 0.8.