APPENDIX C

STATISTICAL METHODS USED IN OUR RESEARCH

This appendix is a brief summary of the statistical methods used in our research. It is meant to serve as a reference, not a detailed statistical text. We have included pointers to the relevant academic references where appropriate. The appendix roughly follows our path through research design and analysis.

SURVEY PREPARATION

Once we have decided on the constructs and hypotheses we want to test each year, we begin the research process by designing the survey instrument.[1]

When possible, previously validated items are used. Examples include organizational performance (Widener 2007) and noncommercial performance (Cavalluzzo and Ittner 2004). When we create our own measures, the survey instrument is developed following commonly accepted procedures adapted from Dillman (1978).

DATA COLLECTION

Armed with our research design and survey questions, we set out to collect data.

We collected data using snowball sampling, a nonprobabilistic technique. Details on why this is an appropriate technique, how we collected our sample, and strategies we used to counteract limitations of the technique are given in Chapter 15.

TESTS FOR BIAS

Once we have our data, we start by testing for bias.

- **Chi-square tests.** A test for differences. This is used to check for significant differences in variables that can only take on categorical values (for example, gender).
- **T-tests.** A test for differences. This is used to check for significant differences in variables that can take on scale values (for example, Likert values). We used this to check for differences between early and late responders.
- **Common method bias (CMB)** or **common method variance (CMV).** This involves conducting two tests:
  - **Harman's single-factor test** (Podsakoff and Dalton 1987). This checks to see if a single factor features significant loading for all items.
  - **The marker variable test** (Lindell and Whitney 2001). This checks to see if all originally significant correlations remain significant after adjusting for the

second-lowest positive correlation among the constructs.

We did not see bias between early and late responders. Common-method bias does not seem to be a problem with our samples.

TESTING FOR RELATIONSHIPS

Consistent with best practices and accepted research, we conducted our analysis in two stages (Gefen and Straub 2005). In the first step, we conduct analyses on the measures to validate and form our latent constructs (see Chapter 13). This allows us to determine which constructs can be included in the second stage of our research.

TESTS OF THE MEASUREMENT MODEL

- **Principal components analysis (PCA).** A test to help confirm convergent validity. This method is used to help explain the variance-covariance structure of a set of variables.
  - Principal components analysis was conducted with varimax rotation, with separate analyses for independent and dependent variables (Straub et al. 2004).
  - There are two types of PCA that can be done: confirmatory factor analysis (CFA) and exploratory

factor analysis (EFA). In almost all cases, we performed EFA. We chose this method because it is a stricter test used to uncover the underlying structure of the variables without imposing or suggesting a structure a priori. (One notable exception was when we used CFA to confirm the validity for transformational leadership; this was selected because the items are well-established in the literature.) Items should load on their respective constructs higher than 0.60 and should not cross-load.

- **Average variance extracted (AVE).** A test to help confirm both convergent and discriminant validity. AVE is a measure of the amount of variance that is captured by a construct in relation to the amount of variance due to measurement error.
  - AVE must be greater than 0.50 to indicate convergent validity.
  - The square root of the AVE must be greater than any cross-diagonal correlations of the constructs (when you place the square root of the AVE on the diagonal of a correlation table) to indicate divergent validity.
- **Correlation.** This test helps confirm divergent validity when correlations between constructs are below 0.85 (Brown 2006). Pearson correlations were used (see below for details).
- **Reliability**
  - **Cronbach's alpha:** A measure of internal consistency. The acceptable cutoff for CR is 0.70 (Nunnally 1978); all constructs met either this cutoff or CR (listed next).

Note that Cronbach's alpha is known to be biased against small scales (i.e., constructs with a low number of items), so both Cronbach's alpha and composite reliability were run to confirm reliability.

- **Composite reliability (CR):** A measure of internal consistency and convergent validity. The acceptable cutoff for CR is 0.70 (Chin et al. 2003); all constructs either met this cutoff or Cronbach's alpha (listed above).

All of the above tests must pass for a construct to be considered suitable for use in further analysis. We say that a construct "exhibits good psychometric properties" if this is the case, and proceed. All constructs used in our research passed these tests.

TESTS FOR RELATIONSHIPS (CORRELATION AND PREDICTION) AND CLASSIFICATION

In the second step, we take the measures that have passed the first step of measurement validation and test our hypotheses. These are the statistical tests that are used in this phase of the research. As outlined in Chapter 12, in this research design we test for inferential prediction, which means all tested hypotheses are supported by additional theories and literature. If no supporting theories exist to suggest that a predictive relationship exists, we only report correlations.

- **Correlation.** Signifies a mutual relationship or connection between two or more constructs. We use Pearson correlation in this research, which is the correlation most often used in business contexts today. Pearson correlation measures the strength of a linear relationship between two variables, called Pearson's r. It is often referred to as just correlation and takes a value between -1 and 1. If two variables have a perfect linear correlation, i.e., move together exactly, $r = 1$. If they move in exactly opposite directions, $r = -1$. If they are not correlated at all, $r = 0$.

- **Regression.** This is used to test predictive relationships. There are several kinds of regression. We used two types of linear regression in this research, as described below.
  - **Partial least squares regression (PLS).** This was used to test predictive relationships in years 2015 through 2017. PLS is a correlation-based regression method that was selected for our analysis for a few reasons (Chin 2010):
    - This method optimizes for prediction of the outcome variable. As we wanted our results to be beneficial to the practitioners in the industry, this was important to us.
    - PLS does not require assumptions of multivariate normality. Said another way, this method doesn't require that our data be normally distributed.
    - PLS is a great choice for exploratory research—and that's exactly what our research program is!
  - **Linear regression.** This was used to test predictive relationships in our 2014 research.

TESTS FOR CLASSIFICATION

These tests could be done at any time, because they don't rely on constructs.

- **Cluster analysis.** This was used to develop a data-driven classification of software delivery performance, giving us high, medium, and low performers. In cluster analysis, each measurement is put on a separate dimension, and the clustering algorithm attempts to minimize the distance between all cluster members and maximize the distance among clusters. Cluster analysis was conducted using five methods: Ward's (1963), between-groups linkage, within-groups linkage, centroid, and median. The results for cluster solutions were compared in terms of: (a) change in fusion coefficients, (b) number of individuals in each cluster (solutions including clusters with few individuals were excluded), and (c) univariate F-statistics (Ulrich and McKelvey 1990). Based on these criteria, the solution using Ward's method performed best and was selected. We used the hierarchical cluster analysis method because:
  - It has strong explanatory power (letting us understand parent-child relationships in the clusters).
  - We did not have any industry or theoretical reasons to have a predetermined number of clusters. That is, we wanted the data to determine the number of clusters we should have.
  - Our dataset was not too big. (Hierarchical clustering is

not suitable for extremely large datasets.)

- **Analysis of variance (ANOVA).** To interpret the clusters, post hoc comparisons of the means of the software delivery performance outcomes (deploy frequency, lead time, MTTR, and change fail rate) were conducted using Tukey's test. Tukey's was selected because it does not require normality; Duncan's multiple range test was also run to test for significant differences and in all cases the results were the same (Hair et al. 2006). Pairwise comparisons were done across clusters using each software delivery performance variable, and significant differences sorted the clusters into groups wherein that variable's mean value does not significantly differ across clusters within a group, but differs at a statistically significant level ($p < 0.10$ in our research) across clusters in different groups. In all years except 2016 (see Chapter 2 callout for the Surprise), high performers saw the best performance on all variables, low performers saw the worst performance on all variables, and medium performers saw the middle performance on all variables—all at statistically significant levels.

[1] We decide on our research model each year based on a review of the literature, a review of our previous research findings, and a healthy debate.