

## CONSTRUCT MEASUREMENT IN STRATEGIC MANAGEMENT RESEARCH: ILLUSION OR REALITY?

BRIAN K. BOYD,<sup>1\*</sup> STEVE GOVE<sup>2</sup> and MICHAEL A. HITT<sup>3</sup>

<sup>1</sup> W. P. Carey School of Business, Arizona State University, Tempe, Arizona, U.S.A.

<sup>2</sup> Management/Marketing Department, University of Dayton, Dayton, Ohio, U.S.A.

<sup>3</sup> Mays Business School, Texas A&M University, College Station, Texas, U.S.A.

*Strategic management research has been characterized as placing less emphasis on construct measurement than other management subfields. In this work, we document the state of the art of measurement in strategic management research, and discuss the implications for interpreting the results of research in this field. To assess the breadth of measurement issues in the discipline, we conducted a content analysis of empirical strategic management articles published in leading journals in the period of 1998–2000. We found that few studies discuss reliability and validity issues, and empirical research in the field commonly relies on single-indicator measures. Additionally, studies rarely address the problems of attenuation due to measurement error. We close with a discussion of the implications for future research and for interpreting prior work in strategic management. Copyright © 2004 John Wiley & Sons, Ltd.*

Book 7 of *The Republic* presents what is probably the most widely known parable of Socrates: the shadows on the cave wall. As he said to Glaucon:

Behold! human beings living in an underground den, which has a mouth open toward the light and reaching all along the den; here they have been since childhood, and have their legs and necks chained so that they cannot move, and can only see before them, being prevented by the chains from turning round their heads. Above and behind them is a fire blazing at a distance, and between the fire and the prisoners there is a raised way; and you will see, if you look, a low wall built along the way, like the screen which marionette-players have in front of them, over which they show the puppets.

I see.

And do you see, I said, men passing along the wall carrying all sorts of vessels, and statues and figures of animals made of wood and stone and various materials, which appear over the wall? Some of them are talking, others silent.

You have shown me a strange image, and they are strange prisoners.

Like ourselves, I replied; and they see only their own shadows, or the shadows of one another, which the fire throws on the opposite wall of the cave. (Jowett, 1999: 209)

As the prisoners converse, and without any other vantage point, they naturally perceive the shadows to be reality—they assign names and try to explain the various flickering shapes which appear on the wall. Echoes of sound are attributed to the shadows. Consider what happens if a prisoner were to become unchained, and look into the light. Once his eyes adjusted, he would realize that ‘what he saw before was an illusion’ (Jowett, 1999: 210).

Keywords: measurement; research design; Type II error; Type I error

\*Correspondence to: Brian K. Boyd, W. P. Carey School of Business, Arizona State University, Tempe, AZ 85287-4006, U.S.A. E-mail: brianboyd@asu.edu

However, if he tried to explain the true nature of reality, he would likely be ridiculed by his peers.

What relevance does Socrates' allegory have for strategic management researchers? We propose that the cave is a metaphor for one of the most serious threats to strategic management research: poor construct measurement. While the implications of measurement error are well known, they are typically ignored in a majority of studies on strategic management topics. So, similar to the freed prisoner, many academic researchers ignore or are unaware that their measures often do not fully or accurately capture the constructs of interest. Our purpose is to highlight the extent and consequences of measurement error in strategic management research.

We begin with a brief overview of research design and methodology issues in strategic management. Next, we explore several topics in more detail, including statistical power, sample size, and measurement. Finally, we examine the potential for similar problems in other strategic management areas. We assess the 'state of the art' of measurement in strategic management research with a review and critique of 196 empirical strategic management articles published in a recent three-year period.

## RESEARCH DESIGN ISSUES IN STRATEGIC MANAGEMENT

### Background

Strategic management is generally acknowledged to be one of the younger subdisciplines within the broader management domain. Such emergent areas are typically characterized by debate, and challenges to existing paradigms (Kuhn, 1996). While the latter are often couched as theoretical discussions, empirical work plays a critical role in confirming, or challenging, a particular perspective. Contributing to this advancement of the field, there has been a small research stream that critiques empirical research in strategic management. This stream includes both narrative (Hitt, Boyd, and Li, 2004; Hitt, Gimeno, and Hoskisson, 1998; Venkatraman and Grant, 1986) and quantitative reviews. Examples of the latter are summarized in Table 1. Regardless of the topic, these reviews have been consistently critical of the rigor of strategic management research. However, one critical

dimension of research design—construct measurement—is not covered by this pool of studies.

Construct measurement is particularly relevant to strategic management research, as the variables of interest tend to be complex or unobservable (Godfrey and Hill, 1995). Paradoxically, measurement has historically been a low-priority topic for strategic management scholars (Hitt *et al.*, 1998, 2004). As a result, complex constructs have often been represented with simple measures, and with limited testing for reliability or validity (Venkatraman and Grant, 1986). Our intent is to contribute to this research stream with a critique of measurement issues in the strategic management field. We begin with a brief discussion of two related topics: statistical power and sample size, and the compounding effects of measurement error.

### Statistical power and sample size

Power represents the potential for a statistical test to produce a statistically significant outcome. Of concern with regard to power are the sample size, Type I and Type II errors, the magnitude of the effect, the test used, and the data quality (Cohen, 1987, 1992). There are two important dimensions of power: factors outside of the control of the researcher, (i.e., the true effect size), and factors that may be influenced by research and measurement design. A Type I error represents the risk of mistakenly rejecting the null hypothesis—falsely concluding that a relationship exists when it is not statistically supported. Most empirical studies control for a Type I error, with the  $p < 0.05$  level widely accepted as an appropriate threshold; alternately stated, 95 percent likelihood that a relationship is not a false-positive. A Type II error occurs when a meaningful relationship exists, but the null hypothesis is not rejected. Statistical power therefore represents the probability that a null hypothesis will be rejected for a given effect size. Cohen (1987) recommends using 0.80 as the threshold for power assessment—i.e., an 8 in 10 chance that an existing relationship will be successfully detected.

Whenever a more stringent  $p$ -level for a Type I error is used, the probability of a Type II error increases, and vice versa. Additionally, the probability of a Type I and Type II error is affected by the sample size. If a population correlation between two constructs is 0.30, and the criterion for statistical significance is  $\alpha = 0.05$ , there is only

Table 1. Methodological critiques of strategic management research

| Category and study                         | Outlet <sup>a</sup> | Journal pool reviewed <sup>b</sup><br>(# reviewed) | # Studies examined | Focus and general finding   | Potential implication                   |
|--|---------------------|--|--------------------|---|---|
| <i>Statistical power</i>                   |                     |  |                    |   |   |
| Ferguson and Ketchen (1999)                | SMJ                 | Not reported (6)                                   | 24                 | Power—sufficient power in only 8% of strategic management studies                         | Probability of Type II error            |
| Mone, Mueller, and Mauland (1996)          | P Psych             | Subset (7)   | 210                | Power—low power in management studies; especially acute for small and medium effect sizes | Probability of Type II error            |
| <i>Data analysis</i>                       |                     |  |                    |   |   |
| Bowen and Wiersema (1999)                  | SMJ                 | Not reported (1)                                   | 90                 | Cross-sectional designs—inappropriate usage of methodologies for cross-sectional data     | Inaccurate findings; flawed conclusions |
| Bergh and Holbein (1997)                   | SMJ                 | All (1)  | 203                | Longitudinal designs—Type I bias present in more than 90% of studies                      | Inaccurate findings; flawed conclusions |
| Ketchen and Shook (1996)                   | SMJ                 | All (5)  | 45                 | Cluster analysis—cluster analysis frequently utilized incorrectly                         | Inaccurate findings; flawed conclusions |
| <i>Other measurement and design issues</i> |                     |  |                    |   |   |
| Bergh and Fairbank (2002)                  | SMJ                 | All (1)  | 126                | Measurement of change—methodologies for assessing change inappropriate                    | Inaccurate findings; flawed conclusions |
| Short, Ketchen, and Palmer (2002)          | JM                  | All (5)  | 437                | Sampling—<20% of studies use random samples; analysis of generalizability uncommon        | Generalizability unknown                |
| Hubbard, Vetter, and Little (1998)         | SMJ                 | Subset (9)   | 37                 | Replications—replications uncommon; more prevalent in SMJ than AMJ or ASQ                 | Limited confirmation of findings        |

<sup>a</sup> Outlet in which the work was published. Abbreviations: SMJ, *Strategic Management Journal*; P Psych, *Personnel Psychology*; JM, *Journal of Management*.

<sup>b</sup> 'Subset' indicates that a sample of relevant articles were used; 'all' indicates that all papers meeting the study criteria/focus were included.

a 50 percent chance of successfully identifying this relationship with a sample size of  $n = 30$ . The probability of identifying the relationship increases to 70 percent if the sample size grows to 70 subjects, and the probability is over 90 percent with a sample of 100 subjects.

While both authors and reviewers generally consider a Type I error, they frequently ignore the possibility of a Type II error. Strategic management authors rarely conduct power analyses and surveys have shown that the perceived need for this analysis is low (Mone, Mueller, and Mauland, 1996). Unfortunately, the majority of studies in strategic management suffer from weak power. Strategic management studies have less than half of the recommended power levels, achieving only a 40 percent probability of rejecting the null hypothesis (Mazen, Hemmasi and Lewis, 1987). In another review, Ferguson and Ketchen (1999) concluded that only 8 percent of published studies on organizational configurations had statistical power consistent with recommended standards (e.g., Cohen, 1992). Furthermore, the statistical power in strategic management studies is substantially below that in other management subdisciplines (Mone *et al.*, 1996).

Thus, these data suggest that statistical power is an important issue in the design of scholarly research, and that research in strategic management needs improvement. In the next section we explore another potential problem: construct measurement and measurement error in strategic management, along with its implications for statistical power.

### Measurement error and attenuation

Blalock (1979) described models of social processes as consisting of three elements: (1) a theoretical language explaining causal relations between constructs; (2) an operational language for examining relationships between constructs using indicators; and (3) an integrative theory describing the causal relationships between constructs and indicators. The operational language that links certain indicators to their constructs is highly relevant to strategy research.

However, much of the research in strategic management consists of hypothesized relationships between constructs: Blalock's first element. Studies linking two unobserved constructs are prevalent in strategic management research, resulting in

'the problem of unobservables' (Godfrey and Hill, 1995). For example, a researcher may hypothesize that agency problems lead to opportunistic actions by executives. Yet, agency problems are unobservable constructs which can not be directly examined. Rather the relationship between two variables, using proxies for the respective constructs, is examined. For example, the degree of CEO ownership in the firm may be used to predict executive pay. In this case, CEO ownership is a proxy for agency problems, and executive pay is a proxy for opportunistic behavior.

If the proxies utilized perfectly represent the latent concepts without error—that is, they have a correlation of 1.00—power is unchanged. But, even a modicum of measurement error has a significant negative effect on power (Schmidt, Hunter, and Urry, 1976; Zimmerman and Williams, 1986). Still, there is a major concern because power analyses assume exact measurement of predictor and outcome variables to determine the minimum sample size needed; that is, they do not consider measurement error. As a result, sample sizes may be too small and the probability of rejecting the null hypothesis is low even when power analyses are employed (Maxwell, 1980).

Let us consider a study in which the population effect size is expected to be moderate (e.g.,  $r = 0.30$ ). When the potential for Type I error is set at the commonly accepted 5 percent level (i.e.,  $p < 0.05$ ), a sample size of 150 is needed to achieve a power level of 0.80 (Cohen, 1987). Power declines precipitously with reduced correspondence between the theoretical constructs and the operational language. If Cronbach's alpha for the independent and dependent variables are both 0.60, the observed correlation will be only about 0.10. This reduction is due entirely to measurement error. In the example of a sample size of 150, the chances of detecting the relationship decline from the accepted level of 8 in 10, to slightly more than 3 in 10.

As noted previously, prior studies describe statistical power in strategic management research in particular as unacceptably low (Mone *et al.*, 1996; Ferguson and Ketchen, 1999). Unfortunately, in assessing power, studies have not examined the effect of measurement error and may, therefore, actually underestimate the severity of the situation. Cohen (1987) concluded that the lack of reliability reduces observed effect sizes and also decreases power. Likewise, he argued that

increases in reliability improved observed effect sizes and power.

To demonstrate the consequences of measurement error, Boyd, Gove, and Hitt (2005) conducted a replication analysis of the agency–diversification research stream. Through the use of a structural model, the authors demonstrated how effect sizes diminish with the use of less precise measures. Ultimately, they concluded that the debate over Amihud and Lev's (1981) findings were largely artifacts of measurement error. Stated differently, while debate is intended to advance the discipline (Kuhn, 1996), debate that is spurred by measurement problems may actually limit the discipline's ability to advance.

## CONTENT ANALYSIS OF EMPIRICAL STUDIES

Boyd and colleagues (2005) demonstrate the consequences of measurement error in the context of a specific research stream. An important question is whether this is an isolated problem, or one endemic to a broader range of strategy research topics. To address this question, we completed a content analysis of a sample of strategy-related articles published in leading scholarly journals.

### Sample

To determine the optimal characteristics of the sample, we reviewed the design characteristics of prior methodological critiques of strategic management research, as shown in Table 1. This review suggests that the sample should have three attributes. First, a range of journals should be sampled. Second, the sample should have a multi-year time frame. Third, prior critiques have used two different approaches for selecting specific articles for analysis—some have included all articles that met the relevant criteria (e.g., Ketchen and Shook, 1996; Mone *et al.*, 1996), while others included a subset of relevant articles (e.g., Hubbard, Vetter, and Little, 1998). We chose to include all relevant studies in the interests of generalizability.

Our sample comprised the universe of empirical strategic management articles that were published in the discipline's leading scholarly outlets over a specific time period. We began with MacMillan's (1989, 1991) set of 14 primary outlets for strategy research wherein a half dozen of these

journals were ranked by an expert panel to be of 'outstanding quality': *ASQ*, *AMJ*, *AMR*, *HBR*, *MS*, and *SMJ*. We excluded *AMR* and *HBR* from our list as they generally do not publish empirical work. Therefore, strategic management articles published in the *Academy of Management Journal*, *Administrative Science Quarterly*, *Management Science*, and *Strategic Management Journal* were included in the final sample. We selected a time period of 1998–2000, as the recent studies are presumably most likely to have the highest level of methodological sophistication. Collectively, our combination of leading journals and recent timeframe should provide a 'best case' assessment of the state of construct measurement in strategy.

We reviewed each volume of the journals, selecting articles for inclusion using a two-stage approach. First, we identified articles reporting research on strategic management topics. We included all papers published in *SMJ*, as it is a discipline-specific outlet. From *AMJ*, *ASQ*, and *MS*, we selected all articles that met a liberal definition of falling into the strategic management domain. The coders who made these assignments have served as reviewers on manuscripts and have held repeat editorial board assignments on a subset of these journals. As our focus was on the use of measurement approaches, as opposed to the development of such approaches, we narrowed the pool by selecting only articles reporting empirical tests of hypotheses. We specifically excluded those relying solely on case analysis and descriptive statistics, those using meta-analyses as they are restricted in their selection of measures, and those developing measures but not testing hypotheses. For example, if an article included only the development of a scale, it was excluded from our sample because the scale was not used in a test of hypotheses as an independent statistical test. If an article developed and validated a scale and also used the scale in a hypothesis test, the article was coded as having one statistical test and is included in our sample. This screen yielded a final sample of 196 articles—a sample comparable to the prior methodological critiques listed in Table 1. A list of the sample articles is available from the authors. As a post hoc analysis, two external raters unaffiliated with the project independently assessed a random sample of 70 articles from *AMJ*, *ASQ*, and *MS*. Agreement with inclusion and exclusion decisions was consistent with our ratings in this study ( $\alpha = 0.91$ ).

A substantial number of the articles included multiple statistical tests with different independent and dependent variables, samples, sample sizes, and analyses. Therefore, each statistical test was used as the unit of analysis. We selected the most complete models presented, as multiple hierarchical-like models were common (e.g., a regression with control variables, indicators, and interaction terms tested in three separate models). We counted all tests that utilized different dependent variables and samples as unique. To avoid allocating extra weight to articles that present multiple subsample analyses, we counted subsamples only if a new dependent variable was utilized. This yielded a final sample of 625 statistical tests from the 196 articles—a sample considerably larger and more comprehensive than prior methodological critiques in the field.

### Analysis

A content analysis of each article and test was completed by an expert rater. A subset of articles was coded by a second rater with comparable results ( $\alpha = 0.96$ ). The articles were examined to evaluate the construct operationalizations employed with the intent of developing a categorization scheme. We elected to treat all variables as potential constructs. This was based on two findings from the review: differences between articles regarding what constitutes a construct and the prevalence of 'hidden' constructs masked as single variables within the studies.

We found that the definition of what constituted a construct within the strategic management literature was largely at the discretion of the researcher. Constructs are 'theoretical creations based on observations but which cannot be observed directly or indirectly' (Babbie, 1989: 109) and the basis for most strategic management theories (Godfrey and Hill, 1995). In practice, we found that many constructs in the sample were not identified as such. Organizational size, for example, appeared in our sample as a widely utilized variable and is arguably one of the most commonly used variables in strategic management research. Size was repeatedly found in the samples as an independent and a control variable. We drew on a subsample of articles and found 'size' as a proxy of available organizational resources, propensity/ability to initiate competitive action, core rigidity, and public profile among a wide

range of other constructs. These varied constructs were, however, generally operationalized using a single indicator. No attempts to examine (establish) convergent validity were reported for the association between size and other measures of the intended construct. As single indicators were the norm, reports of reliabilities and measurement error were not common.

When examined across the volume of studies, the measures of size also appeared to vary far less than the constructs' size was purported to represent. Three indicators in particular—total assets, sales, and employees—constituted over 80 percent of the size measures employed despite the range of constructs size presumably represents.

Because of the lack of specific criteria for identifying constructs in the studies and the potential for commonly used variables to represent complex constructs, we decided to treat all variables as potential constructs. The primary benefit of this approach is a lack of positive bias in the use of multiple indicators in the field for constructs. However, our analysis should be considered as a comprehensive, yet conservative, estimation of the use of construct measurement.

The measures employed in the tests were coded into one of five categories that progressively provide increased ability to assess validity, reliability, and measurement error. These categories are: single indicators, discrete items, single ratios, indexes, and scales/multiple measures.

*Single indicators*, at the nadir of methodological sophistication, provide the researcher with the least assurance that a measure is a valid and reliable proxy of a construct and no estimates of reliability, and thus error, are possible. In the context of a regression model, we coded the use of a sole variable with an accompanying beta as a single variable. For example, a regression estimate for 'sales outside of home country' was coded as a single indicator if no other variables for internationalization were included.

*Single ratios*, similar to single indicators, serve as sole indicators of a construct but they are comprised of two parts in the form of a ratio. These variables may provide an advantage over single indicators as they allow a multifaceted perspective (i.e., condition  $Y$  in relation to  $Z$ ) but they may also mask important information and do not allow for the overall association between the variables to

be examined in terms of reliability.<sup>1</sup> For example, a single ratio of the internationalization construct is 'ratio of foreign sales to total sales.' Other common single ratios include debt-to-equity, return-on-assets, and book-to-market measures (e.g., Tobin's *Q*).

*Discrete indicators* are collections of single indicators that collectively serve to indicate a construct. They are conceptually linked, but have their own beta estimates in a regression model. For example, the internationalization construct may be assessed using three separate, discrete variables (e.g., 'sales outside of home country,' 'employees outside of home country,' and 'count of products sold outside of home country'). The correlation between discrete items can be analyzed and serve as a limited assessment of reliability.<sup>2</sup>

A second category of measures, which are forms of multiple measures, allows formal assessment of reliability and thus error to be quantified. Two such approaches are indexes and scales. *Indexes* incorporate measures of one or more dimensions of a construct into a single item, commonly using a summative approach. For example, a firm's level of internationalization can be operationalized using an index calculated by summing 'foreign sales,' 'number of foreign employees,' and 'number of expatriate managers.' In a regression model, a single beta is calculated for the index. Indexes commonly utilize scale-dependent weights of each indicator that comprise the index or category weights assigned from a distribution of sampled subjects. The reliability among index components may be calculated prior to index creation to provide statistical support for the collective measure.

The final category, *scales and multiple measures*, utilizes data reduction approaches (i.e., factor analysis, principal component modeling, structural equation modeling, etc.) to explicitly assess the degree to which multiple items represent a construct and the error associated with the measure. In a regression model, a single beta is calculated

for the scale, not for the individual items. For the internationalization construct, a single 'internationalization' value may be comprised of multiple items (e.g., 'foreign sales', 'foreign employees', and 'expatriate managers') and assessing fit onto a common dimension. Assessing reliability, and thus measurement error, is inherent to such an approach.

## Results

The presentation of the results of our content analysis is focused on three areas: sample size, measurement schemes, and reliability. Highlights of these results are shown in Tables 2 and 3.

### Sample size

Other studies (Mazen *et al.*, 1987; Mone *et al.*, 1996) have documented the scope of power issues in strategy research. Consequently, a power analysis of our sample would add significant length to the paper, yet little new knowledge. Two aspects relating to statistical power—sample size and the ratio of sample size to indicators—are germane to our analysis, however.

We begin with an examination of sample size statistics. In aggregate, the mean sample size<sup>3</sup> of our pool was 2559 (S.D. = 12,909), and ranged from 20 to 158,782. The mean alone is misleading given a highly positive skew ( $S = 8.42$ ) and large kurtosis ( $K = 83.41$ ). A more accurate picture of the average sample size is garnered from an examination of the percentiles. Studies at the 50th percentile utilized a sample of  $N = 215$  with studies at the 75th, 90th, and 95th percentiles having samples of 426, 1,461 and 8699 respectively.

The number of indicators used in a study relative to the sample size also provides an indication of statistical power. The distributions for the number of independent variables and control variables along with the ratio of sample size to independent, control, and the sum of independent and control variables are highly skewed and non-normal. A statistical test at the 50th percentile was comprised of four independent variables and three control variables having a sample size to independent variable ratio of 58 to 1, a sample

<sup>1</sup> This limitation was identified by one of our reviewers.

<sup>2</sup> It is important to note indicator causality here. In most applications, indicators are seen as effects of an underlying construct. Other times, however, the indicators may drive the construct. In this context, described as causal or formative indicators, diagnostic tools such as inter-item correlations or reliabilities may not be relevant (MacCallum and Browne, 1993). From an anecdotal review of our article pool, the notion of indicator causality is discussed only occasionally. We would like to thank one of our reviewers for this important point.

<sup>3</sup> In assessing sample sizes we did not adjust for the effective reduction associated with the use of pooled cross-sectional data.

Table 2. Normality of sample size and variable distributions

|              | Sample size ( <i>N</i> ) | Independent variables | Control variables | Ratio of <i>N</i> to independent variables | Ratio of <i>N</i> to control variables | Ratio of <i>N</i> to (independent + control variables) |
|--------------|--------------------------|-----------------------|-------------------|--|--|--|
| Mean         | 2,559.22                 | 5.66                  | 8.62              | 628.70                                     | 606.69                                 | 236.12   |
| S.D.         | 12,908.56                | 5.09                  | 30.26             | 3,097.24                                   | 3,537.90                               | 1,333.82   |
| Minimum      | 20.00                    | 1.00                  | 0.00              | 2.80                                       | 1.00                                   | 1.00   |
| Maximum      | 158,782.00               | 36.00                 | 507.00            | 39,695.50                                  | 39,695.50                              | 19,847.80  |
| Skewness     | 8.42***                  | 2.29*                 | 10.47***          | 8.78***                                    | 8.91***                                | 11.73***   |
| Kurtosis     | 83.41***                 | 7.44***               | 139.33***         | 94.00***                                   | 86.08***                               | 161.41***  |
| Percentiles: |                          |                       |                   |  |  |  |
| 25th         | 98                       | 2                     | 0                 | 23   | 22                                     | 12   |
| 50th         | 215                      | 4                     | 3                 | 58   | 39                                     | 24   |
| 75th         | 426                      | 7                     | 8                 | 142  | 84                                     | 65   |
| 85th         | 825                      | 10                    | 10                | 236  | 208                                    | 112  |
| 90th         | 1,461                    | 12                    | 15                | 362  | 275                                    | 179  |
| 95th         | 8,699                    | 15                    | 18                | 1,906                                      | 1,477                                  | 597  |

Significance of skewness and kurtosis based on Z-scores with: \*  $p < 0.05$ ; \*\*  $p < 0.01$ ; \*\*\*  $p < 0.001$ .

size to control variable ratio of 39 to 1, and an overall ratio of 24 observations to independent and dependent variables. Two conclusions may be drawn from this analysis. First, while ultimately dependent on the statistical power of the measures used, ratios of subjects to indicators in the studies examined appear consistent with generally accepted norms. Second, and directly related to our emphasis on measurement, sample size insufficiencies do not appear to be a hindrance for the use of measurement schemes incorporating multiple measures. The sample sizes, on average, allow incorporation of multiple measurement approaches into the research designs without the burden of obtaining additional subjects.

#### Measurement schemes

At the nadir, the use of measures for which reliability cannot be assessed (i.e., single indicators, single ratios, and discrete items) provide the researcher and reader with the least assurance that a measure is a valid and error-less proxy of a construct. Our review of the published tests suggests that the use of such measures is a common, but not exclusive, approach in strategy research.

Results indicate that little of the published research pays attention to the problem of measurement error. Fully 70.8 percent of independent, 57.7 percent of dependent, and 92.7 percent of control variables are based on a methodology that disallows the assessment of reliability.

Of the 3388 independent variables, 1613 (47.6 percent) were single indicators. Of the 677 dependent variables, 228 (38.1 percent) were measured using single indicators. For control variables, 79.6 percent (4280 of 5376 variables) were single indicators. Across the 625 tests reviewed, 335 used at least one single indicator for an independent variable, 238 for a dependent variable, and 371 for a control variable. Thirty-four studies (5.4 percent of the sample) relied on single-item indicators *exclusively* for independent, dependent, and control variables. Single ratios also appear prevalently in the studies examined comprising the measures for 9.1 percent, 14.9 percent, and 8.1 percent of the independent, dependent, and control variables, respectively. Discrete items comprise the measures for 14.1 percent, 4.7 percent, and 5.0 percent of the respective variables.

Techniques allowing reliability and measurement error to be assessed are used in some of the strategic management studies and should be noted. A very small but laudable number of tests, only 0.32 percent of all examined in our study, utilized a full complement of multiple measures. These tests operationalized their measurement using either indexes or scales for all of the IVs, DVs, and control variables. Unfortunately, such rigor was the exception rather than the standard practice.

Indexes and scales as multiple measures constitute 14.8 percent and 11.8 percent of the independent variables used respectively, 17.3 percent



Table 3. Results of content analysis

|   | Count | IVs<br>(% of<br>variables) | % of<br>tests | Count | DVs<br>(% of<br>variables) | % of<br>tests | Count | Controls<br>(% of<br>variables) | % of<br>tests |
|---|-------|----------------------------|---------------|-------|----------------------------|---------------|-------|---------------------------------|---------------|
| Total number of variables   | 3388  |                            |               | 677   |                            |               | 5376  |                                 |               |
| Total number of tests   | 625   |                            |               | 625   |                            |               | 625   |                                 |               |
| <i>Single indicators</i>  |       |                            |               |       |                            |               |       |                                 |               |
| # of single indicators used   | 1613  | 47.6%                      |               | 258   | 38.1%                      |               | 4280  | 79.6%                           |               |
| # of tests using single indicators  | 335   |                            | 53.6%         | 238   |                            | 38.1%         | 371   |                                 | 59.4%         |
| Of the tests, # that rely solely on single Indicators as IV,<br>DV, or control      | 127   |                            | 20.3%         | 170   |                            | 27.2%         | 176   |                                 | 28.2%         |
| Of the tests, # that rely exclusively on single indicator<br>for IV, DV and control | 34    |                            | 5.4%          | 34    |                            | 5.4%          | 34    |                                 | 5.4%          |
| <i>Single ratios</i>  |       |                            |               |       |                            |               |       |                                 |               |
| # of single ratios used   | 309   | 9.1%                       |               | 101   | 14.9%                      |               | 435   | 8.1%                            |               |
| # of tests using single ratio indicators  | 119   |                            | 19.0%         | 95    |                            | 15.2%         | 146   |                                 | 23.4%         |
| Of tests, # that rely solely on single ratios as IV, DV, or<br>control              | 36    |                            | 5.8%          | 64    |                            | 10.2%         | 27    |                                 | 4.3%          |
| Of the tests, # that rely exclusively on single ratios for<br>IV, DV and control    | 0     |                            | 0.0%          | 0     |                            | 0.0%          | 0     |                                 | 0.0%          |
| <i>Discrete items</i>   |       |                            |               |       |                            |               |       |                                 |               |
| # of discrete items used  | 479   | 14.1%                      |               | 32    | 4.7%                       |               | 268   | 5.0%                            |               |
| # of tests using discrete indicators  | 96    |                            | 15.4%         | 32    |                            | 5.1%          | 36    |                                 | 5.8%          |
| Of the tests, # that rely solely on single ratios as IV,<br>DV, or control          | 36    |                            | 5.8%          | 31    |                            | 5.0%          | 15    |                                 | 2.4%          |
| Of the tests, # that rely exclusively on single ratios for<br>IV, DV and control    | 0     |                            | 0.0%          | 0     |                            | 0.0%          | 0     |                                 | 0.0%          |

(continued overleaf)

Table 3. (Continued)

|   | Count | IVs<br>(% of<br>variables) | % of<br>tests | Count | DVs<br>(% of<br>variables) | % of<br>tests | Count | Controls<br>(% of<br>variables) | % of<br>tests |
|---|-------|----------------------------|---------------|-------|----------------------------|---------------|-------|---------------------------------|---------------|
| <i>Indexes</i>  |       |                            |               |       |                            |               |       |                                 |               |
| # of indexes used   | 502   | 14.8%                      |               | 117   | 17.3%                      |               | 103   | 1.9%                            |               |
| Average reliability   | 0.86  |                            |               | 0.74  |                            |               | 0.74  |                                 |               |
| # of tests using indexes  | 162   |                            | 25.9%         | 114   |                            | 18.2%         | 64    |                                 | 10.2%         |
| Of tests using indexes, # reporting reliabilities                     | 33    |                            | 20.4%         | 4     |                            | 3.5%          | 9     |                                 | 14.1%         |
| Of the tests, # that rely solely on index as IV, DV, or control       | 47    |                            | 7.5%          | 78    |                            | 12.5%         | 0     |                                 | 0.0%          |
| Of the tests, # that rely exclusively on index for IV, DV and control | 0     |                            | 0.0%          | 0     |                            | 0.0%          | 0     |                                 | 0.0%          |
| <i>Full scales or multiple indicators</i>                             |       |                            |               |       |                            |               |       |                                 |               |
| # of scales used  | 399   | 11.8%                      |               | 133   | 19.6%                      |               | 39    | 0.7%                            |               |
| Average reliability   | 0.80  |                            |               | 0.82  |                            |               | 0.76  |                                 |               |
| Mean # of indicators  | 6     |                            |               | 6     |                            |               | 5     |                                 |               |
| # of tests using scales   | 171   |                            | 27.4%         | 100   |                            | 16.0%         | 26    |                                 | 4.2%          |
| Of tests using scales, # reporting reliabilities                      | 133   |                            | 33.3%         | 84    |                            | 63.2%         | 16    |                                 | 41.0%         |
| Of the tests, # that rely solely on scale as IV, DV, or control       | 62    |                            | 9.9%          | 79    |                            | 12.6%         | 3     |                                 | 0.5%          |
| Of the tests, # that rely exclusively on scale for IV, DV and control | 2     |                            | 0.3%          | 2     |                            | 0.3%          | 2     |                                 | 0.3%          |
| <i>None used</i>  |       |                            |               |       |                            |               |       |                                 |               |
| Not specified/detail missing  | 86    | 2.5%                       |               | 36    | 5.3%                       |               | 184   | 3.4%                            |               |
|   |       | 100.0%                     |               |       | 100.0%                     |               | 67    | 1.2%                            |               |
|   |       |                            |               |       |                            |               |       | 100.0%                          |               |

and 19.6 percent of the dependent variables, and 1.9 percent and 0.7 percent of the control variables used. Additionally, 17.4 percent of the studies rely solely on these types for independent variables and 25.1 percent for dependent variables. However, less than one half of 1 percent of the studies use control variable measures that allow reliability and measurement error to be assessed. Control variables serve the purpose of ensuring that the predictions provided by independent variables under examination are not overly inflated due to covariance with variables suggesting other explanations. If measures of control variables are not reliable proxies for their intended constructs, and our analyses suggest there is little evidence supporting reliable measurement, the true value of the explanatory variables is likely inflated.

### Reliability

While measures that allow reliability to be addressed are desirable, the reporting of reliability information is necessary to obtain the full value from this effort. Such information can be provided explicitly or implicitly in the manuscript. We find that reporting of this vital information is not universal practice.

The multiple measure approaches utilized within the research appear sound. The reported reliabilities in general are acceptable (average  $\alpha = 0.80$ , 0.82, and 0.76 for scales/multiple measures used as independent, dependent, and control variables respectively). These outcomes may, however, be an artifact of the reporting as the majority of studies do not report reliabilities. Reliability assessments were reported for only 0.5 percent of the independent and dependent variables based on indexes. Surprisingly, control variables based on indexes receive the most attention—nearly half (44 percent) of the studies using multiple indicators for control variables report reliabilities. This result indicates that the validity of the indexes is generally assumed and not subjected to statistical confirmation.

While the reporting of reliabilities for indexes is uncommon in the studies we reviewed, studies using scales did somewhat better. Of the 399 independent variables based on scales, reliability scores are reported for only 133 (33.3 percent). The reporting rate for the reliability of dependent variables measures was higher, 63.2 percent, while

the rate for measures of control variables was approximately 41 percent.

While less informative than reliabilities, correlations between indicators provide an indication of reliability. Correlation matrices were presented in fully 142 of the 196 articles (71.4 percent). Approximately one third of these (49 articles; 25 percent of the total) included all variables in the correlation matrix. This is a positive finding, but we also regularly found independent variables, dependent variables, interaction terms, squared terms, and control variables missing from the correlation matrixes. Most disturbing, fully one quarter of all articles, 50 of the 196, present no correlation matrix.

As a follow-up to Mone *et al.*'s (1996) call for greater attention on statistical power in strategy research, we examined the studies for evidence of attention to statistical power. Of the few studies that present a power analysis, none incorporated the reliability of their measures into their calculation. An understanding and appreciation of the influence of poor construct measurement on both the a priori planning of research approaches and the post hoc diagnostics of results appear to be absent.

### The typical article

Thus far we have focused largely on the statistical tests in reporting results. An analysis of the measurement approaches used in a typical article examines these findings in a context more familiar to researchers.

### Measurement approaches used

The typical strategic management article examined in our study, operationalized on the averages per 196 articles examined, includes 17 independent variables and 27 control variables.<sup>4</sup> While regression was the most common statistical tool, some studies utilized structural modeling, path analysis, and similar approaches resulting in an average of 3.5 dependent variables across the three statistical tests in the typical article. The independent variables utilized in an article, on average, consisted of 8.3 single items, 1.6 ratios, 2.4 discrete measures, 2.6 indexes, and 2 scales. This is an encouraging

<sup>4</sup> The average for control variables includes dummy variables such as industry and years.

result as the typical article published in leading strategy journals utilizes almost five independent variables (roughly 30 percent) for which reliability can be assessed. These results are shown in Table 4 and Figure 1.

For dependent variables, the results are comparable. On average, the articles in our sample utilize approximately two dependent variables for which reliability could not be assessed and only one that could. However, less than half of the studies that could report reliabilities actually did so. For control variables, the typical study relies virtually exclusively on single item approaches. Fully 97 percent of the average 27 control variables used in the average article used single items, ratios, or discrete measurement approaches.

A content analysis of a subset of articles indicated that the majority of these articles reference prior research as support for general measurement approaches.<sup>5</sup> Over half of the total variables were supported with references to prior work. The references ranged from very general support for expected relationships between variables of interest (e.g., correlations, results of hypothesis tests) to references for specific variable measures.

One quarter of the independent variables examined in the subset included references to a specific previously published source for a particular measurement approach. Approximately one third of this support was explained in terms such as 'based on' or 'a modified version' of a published measure. For dependent variables, the use of established measures was even more common. For two-thirds of the dependent variables prior work was cited to support a specific operationalization. Referencing prior work for measurement approaches that allow reliability to be assessed and for approaches that do not was roughly equally divided. Using references to prior work for support was less common for control variables, with cites for only 10 percent of the measures used.

Thus, poor measurement in the research reported within these articles oftentimes appears to be the result of relying on previously published work regardless of the quality of that measurement approach. This conclusion suggests that reviewers and journal editors share some of the responsibility, along with authors, for the persistence and pervasive reliance on poor measures.

<sup>5</sup> We would like to thank one of our reviewers for suggesting this point.

## DISCUSSION AND CONCLUSIONS

As an academic specialty, strategic management is a relatively young discipline: Depending on the metric used, the field is between two and three decades old. However, even with this youth, it plays a critical role in the study of business and management. As the field has matured, there are increasing expectations for the rigor of strategic management research. Our purpose is to extend the ongoing commentary on methodological issues by highlighting the importance of construct measurement. As demonstrated by the content analysis presented herein, there has been little emphasis placed on measurement concerns in strategic management research. Our replication study demonstrates the consequences of this inattention—including the underreporting of effects and potential for Type II errors.

Our purpose is not to criticize prior work but to identify and emphasize needs for future research designs and methodologies in strategic management. While the field has developed significantly since the *Strategic Management Journal* was founded, our results emphasize the need for better empirical research to move the field forward; we may have reached another plateau in the development of the field. For the strategic management field to develop further and to mature into a well-respected field accepted by its sister social science disciplines, significant attention should be placed on measurement in strategic management research.

Lest we seem overly critical of strategic management research, we note that similar problems are present in other fields as well. A meta-analytic review of 70 studies from various social sciences concluded: 'Measurement error, on average, accounts for most of the variance in a measure. This observation raises questions about the practice of applying statistical techniques based on the assumption that trait variance is large in relation to measurement error variance' (Cote and Buckley, 1987: 317). We strongly suspect that these concerns can be generalized to much research in the business disciplines. However, for the strategic management field to advance, it should not imitate its sister disciplines; rather it should take a leadership role in the conduct of high quality research.

We recommend that significant attention be paid to measurement in future strategic management research. To reduce measurement error, strategic

Table 4. Comparison of the use of single and multiple measurement approaches

|   | Count | Ivs<br>(% of<br>variables) | % of<br>tests | Count | DVs<br>(% of<br>variables) | % of<br>tests | Count | Controls<br>(% of<br>variables) | % of tests |
|---|-------|----------------------------|---------------|-------|----------------------------|---------------|-------|---------------------------------|------------|
| Total number of variables   | 3388  |                            |               | 677   |                            |               | 5376  |                                 |            |
| Total number of tests   | 625   |                            |               | 625   |                            |               | 625   |                                 |            |
| None used   |       |                            |               |       |                            |               | 184   | 3.4%                            |            |
| Not specified/detail missing  | 86    | 2.5%                       |               | 36    | 5.3%                       |               | 67    | 1.2%                            |            |
| <i>Measurement approach not allowing for assessment of reliability (single, ratio, and discrete measures)</i> |       |                            |               |       |                            |               |       |                                 |            |
| # of single, ratio, and discrete measures used  | 2401  | 70.8%                      |               | 391   | 57.7%                      |               | 4983  | 92.7%                           |            |
| # of tests using single, ratio, and discrete measures   | 550   |                            | 88.00%        | 365   |                            | 58.40%        | 553   |                                 | 88.48%     |
| Of the tests, # that rely solely on single, ratio, or discrete for IV, DV, or control                         | 199   |                            | 31.84%        | 265   |                            | 42.40%        | 218   |                                 | 34.88%     |
| Of the tests, # that rely exclusively on single, ratio, or discrete for IV, DV and control                    | 34    |                            | 5.44%         | 34    |                            | 5.44%         | 34    |                                 | 5.44%      |
| <i>Measurement approach allowing for assessment of reliability (indexes and scales)</i>                       |       |                            |               |       |                            |               |       |                                 |            |
| # of indexes and scales used  | 901   | 26.6%                      |               | 250   | 36.9%                      |               | 142   | 2.6%                            |            |
| Average reliability reported  | 0.83  |                            |               | 0.78  |                            |               | 0.75  |                                 |            |
| # of tests using indexes or scales  | 333   |                            |               | 214   |                            |               | 90    |                                 |            |
| # reporting reliabilities   | 166   |                            | 49.85%        | 88    |                            | 41.12%        | 25    |                                 | 27.78%     |
| Of the tests, # that rely solely on indexes or scales for IV, DV, or control                                  | 109   |                            | 17.44%        | 157   |                            | 25.12%        | 3     |                                 | 0.48%      |
| Of the tests, # that rely exclusively on indexes or scales for IV, DV and control                             | 2     |                            | 0.32%         | 2     |                            | 0.32%         | 2     |                                 | 0.32%      |

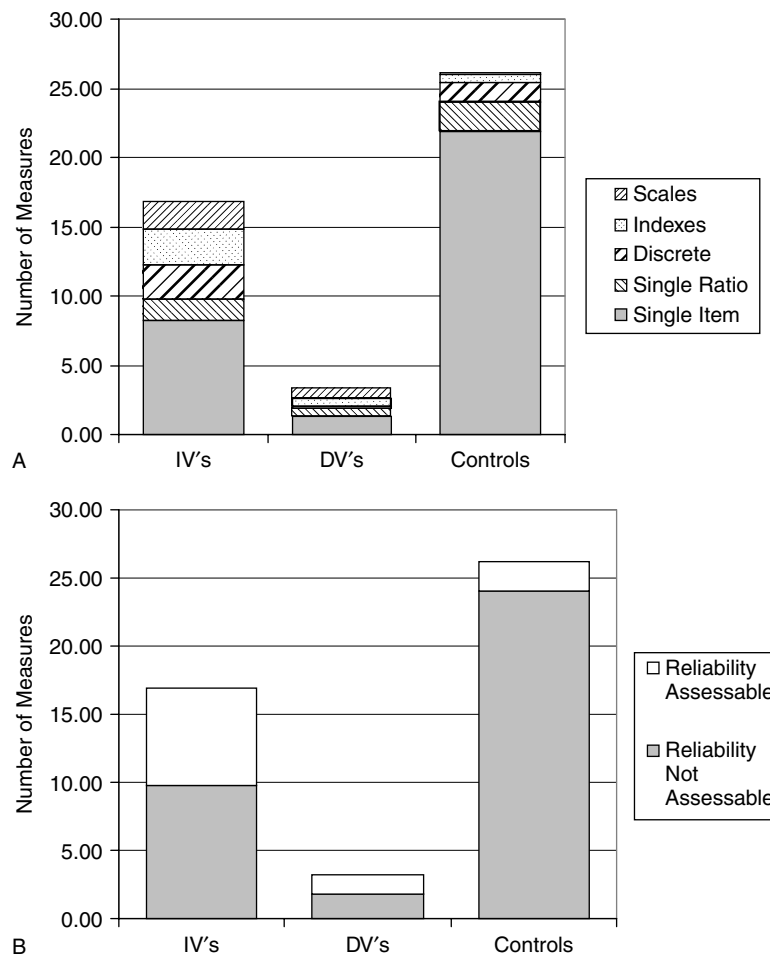


Figure 1. Article-level measurement. (A) Frequency of measurements used within typical article. (B) Ability to assess reliability within typical article

management researchers should increase their concern for the construct validity of their measures. Measurement error often can be reduced by using multiple rather than single indicators for specific constructs as suggested by the research presented herein. It is critical that steps be taken to ensure high reliability of the measures used. Thus, we strongly encourage the application of more indexes and scales in strategic management research. Reliability tests should be considered in the research design and process. Doing so will likely require a break from past and current practices in the field. As we noted, a large number of studies in strategic management rely on measures used in published research. On the one hand, such practice enhances the ability to compare the findings with previous ones. However, given the serious measurement problems identified herein, relying on past

research for measures ensures that the problems persist. Therefore, to use multiple indicators for constructs under study may require strategic management scholars to develop new measures with reliability as a major criterion.

Strategic management scholars should also display more sensitivity to the statistical power of their samples. The reduction of measurement problems may help to avoid debates as the one on agency problems and diversification strategy described by Boyd and colleagues (2005). It may also help resolve conflicting findings, thereby contributing to the resolution of important research questions. As such, it will enhance the power of strategic management research to contribute more value to the conversation involved in the practice of strategy by firms and top executives. While much of the practitioner literature, even in the

better journals, such as *Harvard Business Review* and the *Sloan Management Review*, does not rely heavily on the scholarly research, there is a closer congruence between the research in strategic management and the practitioner literature on common topics. Therefore, if research on governance, for example, is to be used to formulate improved governance practices in organizations, indeed even base laws and regulations on it, we must be assured that the results of the research are accurate. If firms are to invest heavily to develop core competencies, the research should show that having stronger core resources leads to the achievement of a competitive advantage and importantly that we can have confidence that the relationship discovered by the research is correct. Thus, such changes will enrich the field and its value added contribution to knowledge of managing business enterprises.

In point of fact, the field is unlikely to make substantial progress without such attention. For example, extend the problems of interpreting results of the research on the relationship between governance controls and product diversification noted by Boyd *et al.* (2005) to other primary research areas in strategic management. Given that diversification is one of the most researched topics in the field, other areas are likely to be less well developed. Certainly it is more difficult to do high-quality research on some content topics than on others. Some content areas may be better developed allowing more fine-grained measures to be developed and used. Additionally, it may be easier to access data on some topics than others which allows the selection of larger and higher-quality samples and the identification of multiple measures. Yet, our research suggests that the measurement problems are relatively widespread in strategic management research. Thus, we conclude that the measurement problems may be slightly easier to correct in some content areas than others.

Interestingly, we found indications that measurement problems were fairly consistent across journals regardless of their quality ranking and across scholars regardless of the quality of their training (see the Appendix). Citations were largely unrelated to the quality of the measurement used (see Appendix 1). In second-tier journals, citations were partly related to the use of better measures. While there can be several interpretations for this outcome, perhaps, the quality of the journal is used as a general proxy for the assumed quality of the research. In second-tier journals, the articles

using higher-quality measurement are more likely to be cited, whereas scholars in the field attribute high-quality research to articles published in top-tier journals (these journals are ranked in the 'A' category for a reason), and question less the measurement issues.

Doing quality research is difficult and often requires significant amounts of effort and other resources and frequently takes considerable time. Thus, the 'publish or perish' mentality that pervades many academic institutions likely has a negative effect on the conduct of quality research. Certainly, their emphasis is largely on doing quality research but this is measured by publishing the work in the highest-quality journals (based on journal rankings primarily using their acceptance and citation rates). Earlier, we noted that quality of journal does not appear to be a good proxy for the quality of the measurement used in the research. Additionally, much of the research is conducted by scholars who are younger in the field, especially by untenured assistant professors. They must work under time pressures because of the limited time for tenure decisions in most universities and they have the fewest resources on which to call for accessing samples and data. Perhaps universities should focus a higher percentage of their resources on the tenure-track untenured scholars and lengthen the time limitations for tenure decisions. However, these actions are unlikely to lessen the problem substantially as further analysis revealed few differences in the quality of measurement employed by junior and senior faculty (see Appendix 2).

To reduce measurement problems and promote more effective research designs, reviewers and editors must adopt consistent and high standards in these areas. It should be noted that *SMJ*, as the leading discipline-specific publication outlet, is the outlet for the majority of the studies in our sample. *SMJ* also leads the way in identifying design and measurement issues in the field (see Table 1). It is especially important for the gatekeepers in leading journals such as *SMJ* to continue to take such actions because, unless they do so, the quality of strategic management research is unlikely to change. Rewards (e.g., tenure, promotions, pay increases, endowed positions, teaching loads) are based on publishing in these journals; thus researchers will do what is necessary to publish in them. Promotion in the field is based on

publication and the influence of those works. However, supplemental analysis (see the Appendix) suggests that an articles' quality of measurement is largely unrelated to subsequent citation. The primary penalty for poor measurement may be the bottom-drawer solution. Studies with poor measurement that do not yield valuable results are not published, whereas studies with poor measurement that escape the gauntlet of measurement-induced Type II error pay no penalty—they are published and subsequently cited.

Alternatively, we are mindful of the 'normal science straitjacket' to which Daft and Lewin (1990) referred. We recognize that some research in new areas may require a more flexible approach and standards in order to encourage such research (e.g., empirical research on emerging markets where primary and secondary data are difficult to obtain) because of its importance. Yet, we see these situations as exceptions rather than common rules. Additionally, Daft and Lewin's (1990) intent was to promote more qualitative and alternative modes of research with the purpose of building theory and to provide richer data on which to more effectively interpret the results of large sample studies. This need continues to be of importance and therefore we reemphasize it. Our research in no way diminishes this need. In fact, high-quality qualitative research complements high-quality quantitative research. In tandem, quality research of both types can move the field forward more rapidly.

In support of the conclusions noted above, Bergh (2001) suggested that future strategic management research is likely to place greater emphasis on research designs, construct validation, and newer and more sophisticated analytical strategies. While building strong theoretical bases is highly important, measurement is at least equally important for future advances in the field of strategic management. We hope that this work serves as a catalyst to this end.

## ACKNOWLEDGEMENTS

Financial support for the first author was provided by the W. P. Carey School of Business at Arizona State University and the Hong Kong University of Science and Technology. We thank Ji Wang for her research support and Paul Sweeney for his valuable comments and suggestions. Finally, we acknowledge the insightful advice provided

by Associate Editor Will Mitchell and our two anonymous reviewers.

## REFERENCES

- Amihud Y, Lev B. 1981. Risk reduction as a managerial motive for conglomerate mergers. *Bell Journal of Economics* **12**: 605–617.
- Babbie E. 1989. *The Practice of Social Research* (5th edn). Wadsworth: Belmont, CA.
- Bergh DD. 2001. Diversification strategy research at a crossroads. In *Handbook of Strategic Management*, Hitt MA, Freeman RE, Harrison JS (eds). Blackwell: Oxford, UK.
- Bergh DD, Fairbank JF. 2002. Measuring and testing change in strategic management research. *Strategic Management Journal* **23**(4): 359–366.
- Bergh DD, Holbein GF. 1997. Assessment and redirection of longitudinal analysis: demonstration with a study of the diversification and divestment relationship. *Strategic Management Journal* **18**(7): 557–571.
- Blalock HM. 1979. *Social Statistics* (2nd edn). New York: McGraw-Hill.
- Bowen HB, Wiersema MF. 1999. Matching method to paradigm in strategy research: limitations of cross-sectional analysis and some methodological alternatives. *Strategic Management Journal* **20**(7): 625–636.
- Boyd BK, Gove S, Hitt MA. 2005. Consequences of measurement problems in strategic management research: the case of Amihud and Lev. *Strategic Management Journal* (forthcoming).
- Cohen J. 1987. *Statistical Power Analysis for the Behavioral Sciences* (2nd edn). Erlbaum: Hillsdale, NJ.
- Cohen J. 1992. A power primer. *Psychological Bulletin* **112**: 155–159.
- Cote JA, Buckley MR. 1987. Estimating trait, method, and error variance: generalizing across 70 construct validation studies. *Journal of Marketing Research* **24**: 315–318.
- Daft RL, Lewin AY. 1990. Can organization studies begin to break out of the normal science straight-jacket? An editorial essay. *Organization Science* **1**: 1–9.
- Ferguson TD, Ketchen DJ. 1999. Organizational configurations and performance: the role of statistical power in extant research. *Strategic Management Journal* **20**(4): 385–395.
- Godfrey PC, Hill CWL. 1995. The problem of unobservables in strategic management research. *Strategic Management Journal* **16**(7): 513–533.
- Hitt MA, Boyd B, Li D. 2004. The state of strategic management research and a vision of the future. In *Research Methodology in Strategy and Management*, Vol. 1, Ketchen D, Bergh D (eds). Elsevier: New York; 1–31.
- Hitt MA, Gimeno J, Hoskisson RE. 1998. Current and future research methods in strategic management. *Organizational Research Methods* **1**: 6–44.



- Hitt MA, Bierman L, Shimizu K, Kochhar R. 2001. Direct and moderating effects of human capital on strategy and performance in professional service firms: a resource-based perspective. *Academy of Management Journal* **44**: 13–28.
- Hubbard R, Vetter DE, Little EL. 1998. Replication in strategic management: scientific testing for validity, generalizability and usefulness. *Strategic Management Journal* **19**(3): 243–254.
- Jowett B. 1999. *Plato: The Republic*. Barnes & Noble: New York.
- Ketchen Jr DJ, Shook CL. 1996. The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal* **17**(6): 441–460.
- Kuhn TS. 1996. *The Structure of Scientific Revolutions* (3rd edn). University of Chicago Press: Chicago, IL.
- MacCallum RC, Browne MW. 1993. The use of causal indicators in covariance structure models: some practical issues. *Psychological Bulletin* **114**: 533–541.
- MacMillan IC. 1989. Delineating a forum for business policy scholars. *Strategic Management Journal* **10**(4): 391–395.
- MacMillan IC. 1991. The emerging forum for business policy scholars. *Strategic Management Journal* **12**(2): 161–165.
- Maxwell SE. 1980. Dependent variable reliability and determination of sample size. *Applied Psychological Measurement* **4**: 253–260.
- Mazen AM, Hemmasi M, Lewis M. 1987. Assessment of statistical power in contemporary strategy research. *Strategic Management Journal* **8**(4): 403–410.
- Mone MA, Mueller GC, Mauland W. 1996. The perceptions and usage of statistical power in applied psychology and management research. *Personnel Psychology* **49**: 103–120.
- Schmidt FL, Hunter JE, Urry VW. 1976. Statistical power in criterion-related validation studies. *Journal of Applied Psychology* **61**: 473–485.
- Short JC, Ketchen DJ, Palmer PB. 2002. The role of sampling in strategic management research on performance: a two-study analysis. *Journal of Management* **28**: 363–385.
- Venkatraman N, Grant JH. 1986. Construct measurement in organizational strategy research: a critique and proposal. *Academy of Management Review* **11**: 71–87.
- Zimmerman DW, Williams RH. 1986. Note on the reliability of experimental measures and the power of significance tests. *Psychological Bulletin* **100**: 123–124.

## APPENDIX

We conducted a series of post hoc, exploratory supplemental analyses to provide insight into possible causes and consequences of measurement error.<sup>6</sup>

<sup>6</sup> We would like to thank one of our reviewers for recommending an exploration of these issues.

We examined two aspects potentially related to the use of quality measurement: journal effects and author characteristics. We also examined the effect that measurement quality has on subsequent citation.

Exploring these factors required quantifying the quality of measurement at the article level. We calculated this via an ordinal approach. We summed the use of each measurement type (i.e., single indicators, scales, etc.) used for IVs, DVs, and controls for the article. We then calculated three overall article ratings based on the measures utilized.

First, we calculated an ordinal weighted average rating by assigning the measures with the greatest potential for assessing reliability higher values. We assigned the use of single items a value of 1, single ratios a 2, discrete items a 3, indexes a 4, and scales a 5. For example, if an article utilized 6 independent variables including 2 single items, 1 ratio, 1 index, and 1 scale, the article IV rating score was 2.16  $[(2 \times 1) + (1 \times 2) + (1 \times 4) + (1 \times 5)]/6$ . Higher values indicated more sophisticated measurement approaches with a range of 1–5 with a continuous distribution. We also rated articles based on the most sophisticated measurement utilized. This resulted in a value of 1–5 based on the single most sophisticated measure used. In the example above, the article would be coded as 5 because it included at least one scale. A third rating, a binary assessment of whether the article used any multiple measurement approaches (i.e., indexes or scales) was also calculated. Articles that used indexes or scales were coded as 1; those that did not were coded as 0.

## Journal effects

We assessed potential journal effects by developing three ANOVA models to test for differences between the measurement approaches between the four journals in the sample (i.e., *SMJ*, *AMJ*, *ASQ*, and *MS*). We utilized the ordinal weighted average rating as this is the most fine-grained of the measures, with a range of 1.00–5.00. Results suggest no differences between the measurement approaches used between the journals for IVs ( $F = 0.442$ ; sig. = 0.723), DVs ( $F = 0.965$ ; sig. = 0.411), or control variables ( $F = 1.244$ ; sig. = 0.296). None of the post hoc comparisons between specific journals were statistically significant. We cannot, however, rule out

possible effects across a broader pool of outlets. The journals included in our sample were restricted to the highest echelon of publications outlets in the field (MacMillan, 1989, 1991). Fewer differences between the quality of measurement used by research published in these journals can be expected compared to less prestigious outlets. However, these results suggest that outcomes from our overall study are not because of one or more outlier journals. Rather, the measurement approaches appear to be endemic to the field.

### Author characteristics

We examined three author characteristics: researcher skill, rank/seniority, and affiliation. We quantified the quality of author affiliation at the time of publication and the quality of author's degree-granting institution (DGI) using Gorman ratings (a similar approach was found to be valid by Hitt *et al.*, 2001).

Two variants of each of these measures were used: the average for all authors and the rating of the 'best' (e.g., most prestigious) among the authors. Higher-quality institutions were coded with higher values. The rank/seniority of the authors was roughly proxied by the years elapsed since the receipt of the terminal degree for each of the article's authors. Both an average for all authors and the greatest time elapsed since award of degree were used in the analysis.

While caution is advised in interpreting these post hoc analyses, initial results suggest a limited relationship among these variables. Correlations were low and not statistically significant, with one exception. The use of more sophisticated measurement approaches for dependent variables was significantly correlated with the rank/seniority of authors. It is important to note that the direction of this relationship is not clear—the use of higher-quality measurement approaches may alternatively be the cause or effect of rank/seniority.

### Effect of quality measurement

We examined outcomes of quality measurement by examining its relationship with citation counts. We collected citation data for all of the articles using the SSCI database. Using the three ratings of article measurement quality, we examined correlations with the cumulative count of all citations to each article as listed in SSCI. This analysis was

completed independently for IVs, DVs, and control variables.

We counted cumulative citations to articles by year for 5 years post publication. For articles published in 1998, a citation window from 1998 through most of 2003 was available (publication year to plus 5 years). For articles published in 1999 and 2000, the window was restricted to publication year plus 4 years and publication year to plus 3 years respectively.

We assessed the relationship between the quality of measurement within an article with subsequent citation counts using bivariate correlations. As this approach excludes the effect of any covariates that may dampen the relationship, we can expect any association between the variables to be apparent.

The effect of poor measurement appears largely unrelated to the value of a study within the discipline. The relationship between the quality of measurement within an article and subsequent citations to the article is not statistically significant. Correlations ranged from 0.000 to 0.176 and several were negative. Only two relationships were statistically significant at the  $p < 0.05$  level: the relationship between the most sophisticated measurement scheme used for control variables for years 0–1 and for years 0–3. While this relationship is statistically significant, it is likely spurious. To accept this finding, we must conclude that authors cite works based on the quality of control variables but not based on the quality of the independent and dependent variables. Additionally, the use of multiple measurement approaches (i.e., indexes and scales) within control variables was extremely limited: only 2.6 percent of the measures used these approaches for control variables.

While the overall relationship between measurement quality and citation appears to be insignificant, the relationship may vary based on the quality of the outlet within which the work is cited. To assess this, we separately counted citations in two of MacMillan's (1991) article pools. The 'outstanding' pool includes *SMJ*, *AMR*, *AMJ*, *ASQ*, *MS*, and *Harvard Business Review*. The 'acceptable' strategy journal pool includes *California Management Review*, *Interfaces*, *Journal of Business Strategy*, *Journal of Management*, *Journal of Management Studies*, *Long Range Planning*, *Organizational Dynamics*, and *Sloan Management Review*. We also coded citations identified in the SSCI that were in any source other than the above as 'other journals.'

Results for independent variables indicate no relationship between the quality of measurement used to assess independent variables and citation in leading journals. However, there is a statistically significant relationship between the most sophisticated measure of independent variables used within a study and subsequent citation to the article in MacMillan's (1991) 'acceptable' pool of strategy journals. No statistically significant relationships were found between measurement quality of independent variables and citation in other journals. Similar results are found for the relationships between the quality of measurement of dependent and control variables. None of the relationships are statistically significant for citations in the pool of leading or other journals, whereas they are statistically significant for the citation in other strategy journals.

Care must be taken in interpreting these two relationships because, if both are accurate, we must accept that the quality of the measurement of variables is of less importance for studies that appear in what are presumably the disciplines' highest-quality outlets than it is for those appearing in outlets of lesser quality. In a further post hoc analysis of the most cited articles, we examined a subset of articles with high citation counts in the 'acceptable' pool to identify the use of the citations in the work. In the vast majority of cases, the citations were for the results of the prior work or, when cited for a methodological purpose, they were not associated with the more sophisticated measure (i.e., index or scale) appearing in the original article. We conclude from these analyses that citation of an article is largely unrelated to the quality of the measurement used.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.