



# Rarely pure and never simple: Assessing cumulative evidence in strategic management

Strategic Organization

2014, Vol. 12(2) 142–154

© The Author(s) 2014

Reprints and permissions:

[sagepub.co.uk/journalsPermissions.nav](http://sagepub.co.uk/journalsPermissions.nav)

DOI: 10.1177/1476127014529895

[soq.sagepub.com](http://soq.sagepub.com)



**Scott L Newbert**

Villanova University, USA

**Robert J David**

McGill University, Canada

**Shin-Kap Han**

Seoul National University, Korea

## **Abstract**

As empirical evidence in strategy has accumulated, scholars have shown increasing interest in assessing the empirical record of leading theories. Two methods of assessment have figured prominently: vote counting and meta-analysis. Recently, critics have denounced the former in favor of the latter. While meta-analysis is certainly a powerful assessment tool, we argue that both vote counting and meta-analysis are characterized by certain strengths and weaknesses and that these methods should be seen as complementary means of understanding bodies of empirical evidence. We provide guidance regarding when to employ each method and how to improve the process of cumulative assessment.

## **Keywords**

Firm performance, meta-analysis, research methods, resource-based view, topics and perspectives, transaction cost economics

Strategic management is a vibrant field, with a rapidly growing body of empirical evidence. It is not surprising, therefore, that increased attention has been given to assessing the empirical status of core theories. While the narrative review has been used for decades, quantitative forms of

---

## **Corresponding author:**

Robert J David, Desautels Faculty of Management, McGill University, 1001 Sherbrooke Street West, Montreal, QC H3A 1G5, Canada.

Email: [robert.david@mcgill.ca](mailto:robert.david@mcgill.ca)

assessment have become increasingly popular among strategy scholars as a means of aggregating results across individual studies. These assessments have generally taken two forms: vote counting, which tabulates the percentage of statistical tests supportive of a theory, and meta-analysis, which calculates a global effect size for a theorized relationship.

Recent vote counts and meta-analyses in strategy have garnered significant attention. For example, using the vote-counting method, David and Han (2004) and Newbert (2007) assessed the empirical record of transaction cost economics (TCE) and the resource-based view (RBV), respectively. Both vote-count studies were followed quickly by meta-analyses: Geyskens et al. (2006) in the case of TCE, and Crook et al. (2008) in the case of the RBV. Each of these assessments revealed strengths and weaknesses in the empirical records of the respective theories, and extensive citation to these studies in subsequent research suggests that each has filled and continues to fill important needs for systematic ways of sorting out the voluminous bodies of empirical research in question.

Notwithstanding the scholarly impact of these studies, Combs et al. (2011) argue that only meta-analysis can provide the “truth” (p. 190) about a theory’s empirical standing and that the results of vote-count studies cannot be trusted. Such a conclusion has serious implications not only for scholars interested in TCE and the RBV but also for those concerned with how *any* body of empirical research can and should be assessed. Thus, in what follows, we evaluate the relative merits and shortcomings of vote counting and meta-analysis. We ultimately conclude, borrowing from Oscar Wilde (1898), that “the truth is rarely pure and never simple” (p. 15) and, as such, these forms of assessment should be seen as complementary, each with its own goals, strengths, and limitations. In turn, we provide guidance to scholars interested in assessing bodies of literature as well as to those conducting empirical studies that may ultimately be assessed.

## Choosing between vote counting and meta-analysis

Vote counting is a form of quantitative literature review that tabulates the number of statistical tests that are supportive versus non-supportive of a hypothesized relationship along various dimensions of interest (e.g. construct operationalization). The value of such assessments is their ability to identify hypotheses that have received high (or low) levels of support, those in need of more attention, measurement models that yield more (or less) support for a given hypothesis, and ways in which empirical attention and support have evolved over time. Meta-analysis, on the other hand, estimates the average effect size corresponding to a hypothesized relationship. Meta-analysis is well suited to revealing the strength of a relationship and any moderators of it, the measurement models that yield the strongest results, and how the strength of a relationship has evolved over time.

Given that meta-analysis is designed to compute global effect sizes with confidence intervals and significance levels (Borenstein et al., 2009), it is clearly superior on statistical grounds to vote counting. Indeed, meta-analytical studies can, at least in theory, conduct a variety of tests to determine whether or not the individual correlations are equal across studies, the global effect size is non-zero, and/or significant differences in effect sizes exist between groups (Hunter and Schmidt, 1990). Despite claims to the contrary (Combs et al., 2011), however, the statistical superiority of meta-analysis over vote counting does not necessarily mean that vote counting has no place in assessing a body of research. We argue that both methods are useful in this capacity and that, similar to the decision informing the appropriate method in any piece of scholarship, two fundamental issues inform their use: the research question and the data. We discuss these issues below and summarize our key points in Table 1.

**Table 1.** Choosing between vote counting and meta-analysis.

Decision criteria	Vote counting	Meta-analysis
Research question	Frequency of support for a hypothesis (surviving repeated attempts at falsification)	Strength of a hypothesized relationship (and rejection of the null hypothesis)
Size of study population ( <i>k</i> )	Small to large (during early to advanced stages of evidence accumulation)	Large (likely during later stages of evidence accumulation)
Measurement models	Low to high level of consensus on construct measurement	High level of consensus on construct measurement
Size of study samples ( <i>n</i> )	Medium to large	Small to large

### The research question

Because different analytical methods yield different information, the method a researcher chooses must be informed by the research question and not by the researcher's preference or training (Hathaway, 1995). Vote counting answers the question "how often has a hypothesis received empirical support?" whereas meta-analysis answers the question "how strong is a hypothesized relationship?" Both questions are relevant in assessing a theory. Drawing on Popper (1972), Godfrey and Hill (1995) note that a theory can be accepted as valid if the "theory survives repeated attempts to falsify it" (p. 526). Because vote counts report how often hypotheses survive empirical testing, they are well suited to facilitating the evaluation of a theory according to this metric. The statistical tests aggregated by vote counts reflect complex models with multiple independent and control variables carefully designed to isolate the hypothesized relationship between two variables in a given context; hypotheses that stand up to repeated scrutiny of this kind are more worthy of acceptance than those that do not.

Godfrey and Hill (1995) also suggest a second metric for theory evaluation: whether or not a theory's "predictions correspond to reality observed for populations of firms" (p. 530, emphasis in original). Meta-analysis is quite useful for this task. As vote counting's critics point out, "even where the voting method correctly leads to the conclusion that an effect exists, the critical question of the size of the effect is still left unanswered" (Hunter and Schmidt, 1990: 469). By estimating the strength of a hypothesized relationship, meta-analysis can help scholars understand the extent to which a theory describes the "reality" experienced by firms. Thus, both vote counting and meta-analysis can provide insight into a theory's validity, but by providing different information.

It is important to note, however, that both of Godfrey and Hill's (1995) criteria are somewhat ambiguous. To begin, how often does a theory need to survive attempts at empirical falsification to warrant acceptance? While we can certainly have increasing confidence in a hypothesis the more attempts at falsification it survives, declaring it to be true or false on the basis of simple majorities would be tenuous. Similarly, how large must an effect be to constitute a meaningful part of firms' "reality?" While Cohen's (1988) widely cited categorization of effect sizes would seem to provide a definitive answer, there is no statistical basis for what constitutes a small, medium, or large effect (Lipsey and Wilson, 2001). Given the admittedly "arbitrary" (Cohen, 1988: 12) nature of his categorization, it is not surprising that meta-analysts disagree on how to interpret effect sizes. For example, Crook et al. (2008: 1151) interpret their overall mean correlation of 0.22 (well below Cohen's (1988) "medium" effect size threshold) to connote "strong support" for the RBV, while other meta-analysts are more guarded when interpreting small effect sizes, characterizing them as "modest" (Carney et al., 2011: 446) or "weak at best" (Heugens and Lander, 2009: 72).

In sum, both support levels from vote counts and point estimates from meta-analysis provide useful information in assessing a theory. At the same time, for reasons we elaborate below, neither can independently and unambiguously offer the final word on a theory. However, because each method answers a distinct and important question, they can complement each other in the overall assessment of a theory.

### *The data*

The decision of whether to use vote counting or meta-analysis should also be driven by the nature of what is being assessed. First developed in the 1970s to assess the education literature (Glass, 1976), meta-analysis was quickly applied to other disciplines, but with substantial emphasis on fields related to psychology (e.g. industrial–organizational psychology, organizational behavior, and social psychology) (Hunter and Schmidt, 1990). What is important here is that certain features of the fields for which meta-analysis was developed are fundamentally different from the strategy field to which it has only recently been applied: namely, the size of the study population and both the measurement models and sample sizes in the primary studies.

*Size of study population.* Strategy theories are quite new compared to those in psychology and education. Neither TCE nor the RBV, for example, was sufficiently formalized to permit empirical testing until the publication of Williamson's (1975) seminal book and Barney's (1991) groundbreaking paper, respectively. Thus, scholars were already meta-analyzing research in education and psychology before any empirical tests of TCE or the RBV had even been conducted. This temporal advantage has yielded far more empirical studies in these established fields than in strategy, a reality which has important ramifications on the relative utility of vote counting and meta-analysis.

A key determinant of the significance of any statistic is the number of data points from which it derives. In meta-analyses, the sample size ( $k$ ) is generally equal to the number of studies being assessed. The fact that small samples make it harder to find statistical significance becomes problematic when trying to meta-analyze research in young fields that simply lack samples large enough to allow detection of statistically significant point estimates. Given this challenge, Combs et al. (2011) acknowledge that “meta-analysis may not provide definitive evidence about a theory's viability until a large body of evidence emerges” (p. 190). We agree; however, in contrast to Combs et al. (2011), we argue that in this interim period, vote counting is an informative way to assess what is known. Because sample sizes in vote counts are not limited by the number of studies ( $k$ ) but rather by the number of tests in those studies, what vote counting sacrifices in terms of statistical sophistication, it gains in its ability to explore many fine-grained issues. For example, Geyskens et al. (2006) found that too few studies explored the interactive effect of asset specificity and uncertainty on governance choice to meta-analyze it. Consequently, they used a vote count to assess support for this important TCE hypothesis.

The above supports Combs et al.'s (2011) contention that “as a research stream reaches maturity, the advantages of meta-analysis multiply” (p. 190). We agree here too; however, we do not believe that scholars seeking to gauge a theory's validity ought to wait until the field fully matures before they assess it. Nor do we recommend that scholars seeking to test a theory in the early stages of its life cycle proceed without an understanding of which hypotheses need the most attention and/or which theoretical approaches and measurement models have had more (or less) empirical success. It is precisely by providing such insights that vote counting can yield value. Based on the state of the literature at the time of their publication, David and Han's (2004) and Newbert's (2007) vote counts, for example, provided systematic overviews of emerging areas of inquiry, and each was

followed by meta-analyses that complemented their preliminary insights as the evidence continued to grow. Indeed, Geyskens et al. (2006) frame their meta-analysis by writing, “our meta-analysis builds on and extends [David and Han’s (2004)] study” (p. 524). Yet, even after 2 years, Geyskens et al. (2006) were unable to add precision to certain of David and Han’s (2004) insights given sample size concerns. Only 1 year after Newbert’s (2007) vote count, Crook et al. (2008) appear to also have faced sample size limitations, as suggested by the lower-than-normal bar ( $p \leq 0.10$ , one-tailed test) used to accept their “best case” hypothesis. Given the rapid growth of research in strategy, we expect that meta-analyses will eventually provide the types of rich insights for which the technique was developed; in the meantime, vote counting can provide meaningful information about the empirical record of leading theories.

**Measurement models.** A related feature that renders meta-analysis better suited to fields such as psychology and education than to strategy is the relative lack of paradigm consensus in strategy research. In the case of the former fields, scholars frequently study individual behavior in controlled, experimental settings using measures with proven reliability and validity (O’Reilly, 1991). In strategy, scholars cannot maintain the same level of control over their method. Because most firm behavior cannot unfold in a laboratory setting, scholars generally rely on archival proxies for their measures. Even when primary research is conducted in strategy, measures are often adapted in substantive ways due to vast differences in context. As a result, scholars have repeatedly reported little agreement regarding the operationalization of key constructs in strategy (Boyd et al., 2005; Venkatraman and Grant, 1986).

Unfortunately,

meta-analysis is most useful for assessing bivariate relationships wherein the constructs of interest and their measures are well defined ... There is no established way to account for the wide disparity of measures in many macro research streams, and the size of the disparity’s effect on meta-analytic results is unknown. (Combs et al., 2011: 192)

Thus, even though *statistically* superior to vote counting, whether the statistics yielded by meta-analysis are *practically* superior is dependent upon the degree of measurement consensus. Before conducting a meta-analysis in strategy, therefore, researchers “must decide whether an overall quantitative summary will be useful and substantively sound. Just throwing together disparate measures because the title of each study contains the [keyword] can be foolish, no matter how statistically elegant or precise the review” (Light and Pillemer, 1984: 99–100).

This, of course, is the “apples and oranges” criticism that has long been leveled against meta-analysis. Our intention in raising it here is not to debate what makes an apple an apple and an orange an orange, or whether or not the fact that both are fruits justifies their inclusion in the same meta-analysis, but simply to highlight an inherent paradox of meta-analysis: that the generation of statistical power is antithetical to the disaggregation of not only apples from oranges, but more importantly of *different kinds of apples* (e.g. Granny Smith, Macintosh, Red Delicious). Consider, for example, the performance construct. A review of the studies in Newbert’s (2007) vote count shows that in 55 different studies, performance was measured 57 different ways, with 32 different measures used in only one study and only three measures used in five or more. While some might argue that even 57 different types of apples are all still apples and so they warrant aggregation as in Crook et al.’s (2008) meta-analysis, research shows poor intercorrelations among various measures of performance both within and across studies (e.g. Dubofsky and Varadarajan, 1987). Because the correlation among these “Granny Smith,” “Macintosh,” and “Red Delicious” apples is weak, the way they correlate with other variables will vary considerably (Geyskens et al., 2006).

Unfortunately, when disparate measures are aggregated in meta-analyses, the results “are difficult or impossible to interpret” (Hunter and Schmidt, 1990: 481). While moderator analysis can address this limitation, the wide range of operationalizations of constructs in strategy, with each typically present in only a small number of studies, often results in inadequate statistical power for such analyses to be robust.

While this criticism may apply equally to vote counting, in practice, vote counts are not beholden to statistical power, and there is thus no incentive to either combine apples with oranges or different types of apples together. In fact, highlighting diversity of measurement can be a core contribution of vote counts, as in both David and Han (2004) and Newbert (2007). For example, Newbert (2007) disaggregated the resource–performance relationship that Crook et al. (2008) aggregated across seven individual independent–dependent variable pairs and concluded, albeit without an associated significance level, that “the level of support varies considerably with the relationship tested” (Newbert, 2007: 128). Thus, while meta-analysis can provide precise estimations of effect sizes when a high level of agreement exists with regard to construct definition and measurement, vote counting is a useful tool for revealing important theoretical and empirical distinctions in fields such as strategy where measurement precision is lacking.

*Size of study samples.* A final issue that informs the choice of assessment method concerns the size of the samples in the studies being assessed. When the size of these individual samples is small, vote counting suffers in ways that meta-analysis does not. Because studies with larger samples are more likely to achieve significance than studies with smaller sample sizes, vote counts that rely on a high percentage of studies with small samples may fail to detect an effect. Moreover, in study populations with a high percentage of small samples, detecting an effect becomes *more* difficult for vote counting as the number of studies in the assessment increases. To demonstrate this nuance, Hedges and Olkin (1985) develop models detailing how the number of studies ( $k$ ), effect size ( $\delta$ ), and study sample size ( $n$ , assumed identical across studies) predict Type II error in vote counts and conclude that

When effect sizes are moderate to small ( $\delta \leq 0.5$ ), standard vote counting frequently fails to detect the effects. More important the situation does not always improve as the number  $k$  of studies increases. For example, when  $\delta \leq 0.3$ , the probability that a standard vote count detects an effect decreases as  $k$  increases from 10 to 50. (p. 50)

Many advocates of meta-analysis have since cited Hedges and Olkin’s (1985) arguments in order to dismiss vote counting as “fundamentally” (Borenstein et al., 2009: 252) or “fatally” (Hunter and Schmidt, 1990: 471) flawed. Yet, it is important to note that this association between Type II error and the number of studies ( $k$ ) actually depends on the sample sizes ( $n$ ) in the underlying studies. Looking more closely at Hedges and Olkin’s (1985) models, the probability that a vote count of  $k = 10$  studies, each with a sample size of  $n = 10$ , and a medium effect size of  $\delta = 0.50$ , will detect an effect using the plurality criterion (e.g. that more than one-third of studies finds significant results in the hypothesized direction) is only 1.5%, and this probability decreases to 0% as  $k$  approaches 50. However, given the same effect size and initial number of studies, but where the sample size in each of those studies is  $n = 50$ , the probability that a vote count will detect an effect is 64.2%, and this success rate *increases* to 87.6% for  $k = 50$  studies. Moreover, when effect sizes are in the medium to large range ( $\delta \geq 0.60$ ), the probability that a vote count of  $k = 50$  studies, each with a sample size of  $n = 50$ , will detect an effect is 99.9%. Yet, vote counting need not be restricted to assessing studies with only medium and large effects. Vote counts of “a large number of studies” (presumably  $k \geq 50$ ) can detect effects of  $\delta = 0.26$  so long as the samples in each of those studies is  $n \geq 100$  (Hedges and Olkin, 1985: 51).

It appears then that vote counting's flaws are heavily dependent on the size of the samples in the studies being assessed. Indeed, Hedges and Olkin (1985) admit,

if the combinations of effect sizes and sample sizes for which vote counting fails (has power tending to zero) are not typical of social science research, then the theoretical failure of the procedure might not have practical consequences. Unfortunately, vote counting can fail for sample sizes and effect sizes that most commonly appear in educational and psychological research. (p. 51)

Whereas Hedges and Olkin's critique is based on models assuming studies with maximum sample sizes of 50 and 100, sample sizes in strategy research are typically much larger. For example, the average sample size of the studies assessed in David and Han (2004) and Newbert (2007) is 865, with 93% of studies having samples greater than 50 and 76% greater than 100. Thus, while meta-analysis is clearly the more appropriate method in fields where hypotheses are tested on large numbers of very small samples, criticisms of vote counting's use in assessing strategy research on the basis of Hedges and Olkin's (1985) findings seem overstated given that sample sizes in the field are generally large enough to minimize or eliminate the drastic Type II error rates that complicate vote counting in other fields.

## Improving our ability to assess bodies of empirical research

As the above discussion suggests, the first step in quantitatively assessing a body of empirical literature is to select a method based on the question of interest and the nature of the data. Yet, once the appropriate tool is selected, great care must be taken to ensure that the results obtained by it properly summarize that literature. To this end, we offer below a brief discussion of common concerns and how they might be mitigated, using the four recent assessments in strategy as illustrations. Our goal is not to provide a detailed "how to" guide for either vote counting or meta-analysis, but rather to highlight issues that pertain to assessments of research in strategy. Given that criticism of any assessment method tends to center on its ability to accurately assess support for a theory, we begin our discussion around the risks of committing Type I and Type II errors and proceed to address more broadly how scholars might improve their research practices so as to facilitate knowledge accumulation in the future.

### *Type I error*

Claiming a relationship exists when in fact it does not is a concern for both vote counting and meta-analysis. In the case of vote counting, in the absence of a sound statistical metric for determining that a hypothesis is "true" or "false," we advise assessors to avoid such conclusions even when a vote count shows a plurality of support for a hypothesized relationship. Vote counts should simply note where various hypotheses have received more or less support and not adjudicate on their validity. Indeed, the term "vote count" is a misleading label: unlike the common practice of early vote counts, the "votes" should *not* be used to declare a "winner" (e.g. validate a hypothesis) (Light and Smith, 1971), but rather indicate where support has (and has not) been found across a range of empirical contexts.

While meta-analysis' statistical apparatus allows it, at least in theory, to make more definitive statements regarding the validity of a hypothesis, it is subject to Type I error as a result of its aggregation of statistics computed from disparate samples. The greater the differences between samples across studies, the more important it is to correct for between-studies variance. Historically, meta-analyses overwhelmingly employed fixed-effects modeling, which "a priori assumes that

differences among studies are not influential and true between-studies variance is zero” such that the differences between studies “have no effect on the relationships among constructs of interest” (Erez et al., 1996: 280). Yet, the “underlying assumption that population effect size is the same in all studies is usually false” (Geyskens et al., 2009: 398). As a result, meta-analyses using fixed-effects modeling in contexts where between-studies variance is non-zero are likely to generate inaccurate effect sizes, “in which case Type I error rates are (often far) too high” (Geyskens et al., 2009: 398). In fields such as strategy, in which researchers tend to use data from very different contexts, such a concern is heightened.

In cases of study populations with highly distinct samples, we therefore echo calls for meta-analysts to employ a random-effects approach, a technique that “expressly models both between-studies and within-studies (e.g. error) variance, and correctly allows them to influence parameter estimates” (Erez et al., 1996: 282). Unfortunately, estimates using random-effects modeling “can become unstable when the number of studies meta-analysed is small” (Geyskens et al., 2009: 401), considerations which may limit its application in nascent fields, such as strategy, with relatively small study populations, thereby reinforcing the complementary value that vote counting can provide.

An important issue concerning Type I error for both vote counting and meta-analysis arises from publication bias. Because the effect sizes in unpublished studies reflect (in part) the realities of a given theory (Bushman et al., 2010), it is important that those aggregating a body of literature not restrict their sampling to the published literature only. Given evidence that correlations in published studies are greater than those in unpublished studies (Starbuck, 2006), the omission of unpublished studies in literature assessments will artificially inflate the number of positive tests (in the case of vote counting) or the population effect size (in the case of meta-analysis), thereby increasing the risk of Type I error (Lipsey and Wilson, 2001). We, therefore, repeat previous calls for researchers seeking to aggregate bodies of research to both include unpublished studies in their assessments and test for the presence of publication bias in order to ensure accurate vote counts and meta-analytically derived effect sizes (Kepes et al., 2012). We also call on scholars to pursue empirical research intended to uncover null findings in order to generate more evidence refuting previously accepted relationships in the public record. In so doing, they must ensure that their measurement models and the statistical power of their analyses are sufficiently robust (see Lane et al. (1998) for a noteworthy example). Relatedly, we call on reviewers and editors to temper their resistance to studies reporting insignificant results, which may contradict studies previously published in their journals, provided they meet the above methodological specifications.

### *Type II error*

As noted above, because studies with small samples tend not to have enough statistical power to detect effects that actually exist in the population, when vote counts rely on a high percentage of studies with very small samples, they risk concluding that a relationship does not exist when it in fact does. The most straightforward remedy for this concern is for researchers conducting empirical research to conduct and report the results of power analyses so readers can be assured that the absence of statistical significance is meaningful. Unfortunately, despite many such calls over the years, power analyses remain conspicuously absent in empirical research (Borenstein, 2000). Thus, we advise authors of vote counts to adjust for differences in the underlying studies’ sample sizes ( $n$ ).

To illustrate the impact sample size may have on vote counts, we weighted each of the tests in our own assessments by sample size and found that the weight-adjusted level of support increased from 47% to 49% for TCE and decreased from 53% to 50% for the RBV, suggesting that sample



size may indeed impact support levels in vote counts. We further advise authors of vote counts to perform robustness checks on their results based on the sample sizes of the underlying studies they assess. As Hedges and Olkin (1985) show, vote counts done on studies with sample sizes of 100 or less may fail to detect small effects. We, therefore, recomputed our vote counts with studies with sample size of  $n < 100$  excluded and found that levels of support decreased from 47% to 46% for TCE and from 53% to 52% for the RBV. In issuing these calls, we caution, however, that excluding studies for whatever reason “would discard useful data by requiring a researcher to ignore all studies with sample sizes below the minimum. This is inconsistent with an underlying goal of cumulative reviews—fully capturing the available literature in order to draw valid conclusions” (Combs et al., 2011: 181–182). Thus, weighting the data based on sample size as noted above is preferred to outright elimination. Yet, if underlying studies with small sample sizes are weighted rather than omitted, scholars should frame their findings guardedly, noting that the absence of support in such studies may not necessarily mean that no effect exists, but that more data is required (Borenstein et al., 2009).

Like vote counting, meta-analysis is also subject to Type II error, in particular when testing for the presence of moderators to a focal relationship. Subgroup analysis, the most widely used method to test for moderators in meta-analysis (Geyskens et al., 2009), involves splitting the full sample into groups and comparing their effect sizes with bivariate statistical tests. Unfortunately, because subgrouping results in a decrease in sample size ( $k$ ), the probability of failing to detect a moderator when one exists “increases drastically” when using this technique (Hunter and Schmidt, 1990: 464). For these reasons, scholars advise using weighted least squares regression to test for moderators as it provides a far more accurate estimation of moderator effects than subgroup analysis under a variety of conditions, including small sample size (Geyskens et al., 2009). For meta-analysts still wishing to conduct subgroup analysis, we advise them to choose moderators judiciously in order to increase the likelihood of accurately identifying differences that actually exist (Hunter and Schmidt, 1990). In testing for such differences, we further advise scholars to only proceed if there is a sufficient number of studies ( $k$ ) in each subgroup to allow for the detection of significant differences. If adequate sample size is still not available—a possibility in young fields such as strategy—a vote count on the subgroup, as used by Geyskens et al. (2006), can be used to complement the principal meta-analysis.

### *Beyond the vote-count/meta-analysis choice*

Because any cumulative assessment is only as good as the studies it aggregates, scholars conducting primary research also play an important role in improving cumulative assessments. To begin, we echo the concern that the “lack of replication [of empirical findings] dilutes the quality and meaning of statistical inference in social science research generally and strategic management research specifically” (Mezias and Regnier, 2007: 287). Building on Tsang and Kwan (1999), Mezias and Regnier (2007) describe six types of replication studies based on the source of data, the measurement model, and the analytic technique, but note that most replication research in strategy constitutes “extensions” of prior empirical studies, in which relationships are retested by applying different analytical or measurement models on different samples, either from the same or different populations of firms. On the one hand, such extensions can aid in theory development by providing evidence that a relationship is applicable beyond the specific context in which it was first observed and that it is robust to different analytic models (Singh et al., 2003). At the same time, however, these extensions do not allow for direct comparison of any two (or more) empirical tests. Given this state of affairs, we echo previous calls for research that replicates the data and methods used in prior studies, replication types which are to date exceedingly rare in strategy (McKinley, 2010).

Were such replications to accumulate, assessments (be they vote counts or meta-analyses) could assess support across the various types of replications (e.g., same data or same measures) to see where agreement/disagreement exists for a hypothesis. This would allow assessment of how much volatility there is in the level of support and/or the strength of a relationship based on the population/sample, the measurement model, and/or analytical technique. We recognize that replications suffer from a prestige problem (Mezias and Regnier, 2007) and that change in what is valued by journals does not happen overnight; yet, we hope that the value of replications may one day be realized, at least in part, through their use in cumulative assessments.

Finally, we highlight the important relationship between cumulative reviews and the measurement models in primary studies. As noted above, the operationalization of constructs in strategy research can vary widely. For example, Newbert (2007) found great variety in the measurement of performance, and David and Han (2004) found 27 measures of asset specificity. While some might argue that measurement variety is a strength of strategy research, vote counting and meta-analysis can identify operationalizations that enjoy more/less or stronger/weaker support, such that ultimately, scholars may conclude that some measures simply do not adequately capture underlying constructs, or that the constructs themselves need to be further refined. Cumulative reviews can thus highlight issues of theory–measurement disconnects that the authors of subsequent studies can avoid. This, of course, relates to the argument for replication above: replications of the most promising measurement models—as identified by cumulative reviews—would allow strategy scholars to raise their confidence in their core theories. And, iteratively, a focus on a smaller set of operationalizations in primary studies would facilitate the undertaking of subsequent cumulative reviews.

## Conclusion

We conclude from our discussion that assessing a body of research is no simple task and that while both vote counting and meta-analysis can be useful in this regard, neither offers a pure representation of the truth. As with all scientific inquiry, the choice of method depends, in part, on the question of interest to the researcher. For scholars interested in understanding how strong a given relationship is, meta-analysis is an ideal method. However, when researchers are interested in understanding how often a theory has survived repeated attempts at falsification, vote counting becomes the more appropriate tool. By providing answers to these distinct and important questions, each method can provide valuable and complementary information about a theory's empirical record.

Although meta-analysis has clear advantages over vote counting from a statistical perspective, its appropriateness also depends on the nature of the underlying data. In mature fields that have reached paradigm consensus, meta-analysis can provide robust estimates of effect sizes. However, in fields where the number of studies is small and/or diversity of measurement is high, meta-analysis' advantages decline. The strategy field in particular is characterized by far fewer empirical studies and far greater measurement diversity than the fields for which meta-analysis was developed, conditions which pose serious problems for the method. Because vote counting is not beholden to large samples of studies ( $k$ ), but rather large samples in those primary studies ( $n$ ), it can succeed in detecting effects so long as those samples ( $n$ ) are sufficiently large. With sample sizes in strategy research almost always above 50 and typically above 100, vote counting can accurately detect even small effects in ways it cannot in other fields. Thus, scholars need not wait until a research area fully matures in order to begin to assess it. Rather, vote counting can be used in the absence of the large body of work required for a robust meta-analysis.

In sum, vote counting and meta-analysis should not be seen as an either-or choice, but rather as complementary tools that can enable scholars to more accurately understand a theory's

empirical record. We would thus expect vote counting and meta-analysis performed on the same literature to produce results that, while providing different information, are not inconsistent, such as is the case with the two recent pairs of assessments in strategy. Consistent with Geyskens et al. (2006), David and Han (2004: 52) find that TCE was “quite successful” in some areas; the two studies also agree that certain TCE relationships have not been empirically corroborated, and that others require more attention. For the RBV, Newbert (2007) finds “low level[s] of support” (p. 137) in some areas and “overwhelming support” (p. 139) in others. Similarly, Crook et al. (2008) find that support for their meta-analytic hypotheses varies from “modest” (p. 1149) to “strong” (p. 1151).

Of course, vote counting and meta-analysis are also subject to certain limitations that may obscure an unambiguous assessment of a given theory, and readers must thus exercise judgment when interpreting the results of any cumulative assessment by considering the strengths and weaknesses of the method used. Ultimately, when it comes to assessing a body of empirical literature, “the truth is rarely pure and never simple” (Wilde, 1898: 15). Neither vote counting nor meta-analysis has a monopoly on this truth; instead, these forms of assessment should be seen as complementary techniques, each providing a distinct perspective on the extant evidence in support of a theory.

### Acknowledgements

Scott Newbert and Robert David contributed equally to this article and are joint first authors. We thank Abhirup Chakrabarti for helpful comments on an earlier version, and the co-editors of *Strategic Organization* for their constructive suggestions.

### Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

### References

- Barney, J. B. (1991) “Firm Resources and Sustained Competitive Advantage,” *Journal of Management* 17(1): 99–120.
- Borenstein, M. (2000) “The Shift from Significance Testing to Effect Size Estimation,” in A. S. Bellack and M. Hersen (eds) *Comprehensive Clinical Psychology*, vol. 3, pp. 313–49. Oxford: Pergamon.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T. and Rothstein, H. R. (2009) *Introduction to Meta-Analysis*. West Sussex: John Wiley & Sons.
- Boyd, B. K., Gove, S. and Hitt, M. A. (2005) “Consequences of Measurement Problems in Strategic Management Research: The Case of Amihud and Lev,” *Strategic Management Journal* 26(4): 367–75.
- Bushman, B. J., Rothstein, H. R. and Anderson, C. A. (2010) “Much Ado about Something: Violent Video Game Effects and a School of Red Herring: Reply to Ferguson and Kilburn,” *Psychological Bulletin* 136(2): 182–7.
- Carney, M., Gedajlovic, E. R., Heugens, P. P. M. A. R., Van Essen, M. and Van Oosterhout, J. H. (2011) “Business Group Affiliation, Performance, Context, and Strategy: A Meta-Analysis,” *Academy of Management Journal* 54(3): 437–60.
- Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd edn. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Combs, J., Ketchen, D., Crook, T. and Roth, P. (2011) “Assessing Cumulative Evidence within Macro Research: Why Meta-Analysis Should be Preferred over Vote Counting,” *Journal of Management Studies* 48(1): 178–97.
- Crook, T., Ketchen, D., Combs, J. and Todd, S. (2008) “Strategic Resources and Performance: A Meta-Analysis,” *Strategic Management Journal* 29: 1141–54.

- David, R. J. and Han, S. K. (2004) "A Systematic Assessment of the Empirical Support for Transaction Cost Economics," *Strategic Management Journal* 25: 39–58.
- Dubofsky, P. and Varadarajan, P. (1987) "Diversification and Measures of Performance: Additional Empirical Evidence," *Academy of Management Journal* 30(3): 597–608.
- Erez, A., Bloom, M. C. and Wells, M. T. (1996) "Using Random Rather than Fixed Effects Models in Meta-Analysis: Implications for Situational Specificity and Validity Generalization," *Personnel Psychology* 49: 275–306.
- Geyskens, I., Krishnan, R., Steenkamp, J. E. M. and Cunha, P. V. (2009) "A Review and Evaluation of Meta-Analysis Practices in Management Research," *Journal of Management* 35: 393–419.
- Geyskens, I., Steenkamp, J. B. and Kumar, N. (2006) "Make, Buy, or Ally: A Transaction Cost Theory Meta-Analysis," *Academy of Management Journal* 49: 519–43.
- Glass, G. (1976) "Primary, Secondary and Meta-Analysis of Research," *Educational Researcher* 5: 3–8.
- Godfrey, P. C. and Hill, C. W. L. (1995) "The Problem of Unobservables in Strategic Management Research," *Strategic Management Journal* 16: 519–33.
- Hathaway, R. S. (1995) "Assumptions Underlying Quantitative and Qualitative Research: Implications for Institutional Research," *Research in Higher Education* 36(5): 535–62.
- Hedges, L. V. and Olkin, I. (1985) *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press.
- Heugens, P. P. M. A. R. and Lander, M. W. (2009) "Structure! Agency! (and Other Quarrels): A Meta-Analysis of Institutional Theories of Organization," *Academy of Management Journal* 52(1): 61–85.
- Hunter, J. and Schmidt, F. (1990) *Methods of Meta-Analysis*. Newbury Park, CA: Sage.
- Kepes, S., Banks, G. C., McDaniel, M. and Whetzel, D. L. (2012) "Publication Bias in the Organizational Sciences," *Organizational Research Methods* 15: 624–62.
- Lane, P. J., Cannella, A. A. and Lubatkin, M. H. (1998) "Agency Problems as Antecedents to Unrelated Mergers and Diversification: Amihud and Lev Reconsidered," *Strategic Management Journal* 19: 555–78.
- Light, R. J. and Pillemer, D. B. (1984) *Summing Up: The Science of Reviewing Research*. Cambridge, MA: Harvard University Press.
- Light, R. J. and Smith, P. V. (1971) "Accumulating Evidence: Procedures for Resolving Contradictions Among Different Research Studies," *Harvard Educational Review* 41(4): 429–71.
- Lipsey, M. W. and Wilson, D. B. (2001) *Practical Meta-Analysis*. Thousand Oaks, CA: Sage.
- McKinley, W. (2010) "Organizational Theory Development: Displacement of Ends?" *Organization Studies* 31(1): 47–68.
- Mezias, S. J. and Regnier, M. O. (2007) "Walking the Walk as Well as Talking the Talk: Replication and the Normal Science Paradigm in Strategic Management Research," *Strategic Organization* 5(3): 283–96.
- Newbert, S. L. (2007) "Empirical Research on the Resource-Based View of the Firm: An Assessment and Suggestions for Future Research," *Strategic Management Journal* 28: 121–46.
- O'Reilly, C. A. (1991) "Organizational Behavior: Where We've Been, Where We're Going," *Annual Review of Psychology* 42: 427–58.
- Popper, K. R. (1972) *Objective Knowledge: An Evolutionary Approach*. Oxford: Clarendon Press.
- Singh, K., Ang, S. H. and Leong, S. M. (2003) "Increasing Replication for Knowledge Accumulation in Strategy Research," *Journal of Management* 29(4): 533–49.
- Starbuck, W. H. (2006) *The Production of Knowledge: The Challenge of Social Science Research*. New York: Oxford University Press.
- Tsang, E. W. K. and Kwan, K. M. (1999) "Replication and Theory Development in Organizational Science: A Critical Realist Perspective," *Academy of Management Review* 24(4): 759–80.
- Venkatraman, N. and Grant, J. H. (1986) "Construct Measurement in Organizational Strategy Research: A Critique and Proposal," *Academy of Management Review* 11(1): 71–87.
- Wilde, O. (1898) *The Importance of Being Earnest: A Trivial Comedy for Serious People*. London: Chiswick Press.
- Williamson, O. E. (1975) *Markets and Hierarchies*. New York: Free Press.

### Author biographies

Scott L Newbert is an Associate Professor at Villanova University. He received his PhD in strategic management and entrepreneurship from Rutgers University. His research interests include the processes by which firms create value through the entrepreneurial use of resources, the determinants of firm creation, and the socioeconomic impacts of entrepreneurial activity. His research on these and related topics has been published in the top management and entrepreneurship journals, including *Strategic Management Journal*, *Journal of Business Venturing*, *Entrepreneurship Theory and Practice*, and *Journal of Business Ethics*.

Robert J David is an Associate Professor of strategy and organization and the Cleghorn Faculty Scholar at the Desautels Faculty of Management, McGill University. He is also the Director of the Centre for Strategy Studies in Organization at McGill. He holds a PhD in organizational studies from Cornell University. His research interests include the role institutional change and entrepreneurial action in new market formation. His research has been published in *Organization Science*, *Industrial and Corporate Change*, and *Academy of Management Journal* among other leading outlets.

Shin-Kap Han is a Professor of sociology at Seoul National University. He received his PhD in Sociology from Columbia University in 1994. His areas of interest include social networks, organizations and institutions, culture and consumption, and methodology. Among the recent publications are "Motif of Sequence, Motif in Sequence" (2014), "Road Closed/Detour: A Network Analysis of North-South Korea Relations" (2013), "The Dichotomy Unspooled: Outlining the Cultural Geography of Seoul" (co-authored, 2012), "The Other Ride of Paul Revere: The Brokerage Role in the Making of the American Revolution" (2009).