THE MANY HEELS OF ACHILLES: AN ANALYSIS OF SELF-REPORTED LIMITATIONS IN LEADERSHIP RESEARCH

Stéphane Brutus
Concordia University

Kris Duniewicz
Concordia University

Reference: Brutus, S. & Duniewcz, K. (2012). The many heels of Achilles: An analysis of self-reported limitations in leadership research. Leadership Quarterly, 23, 202-212

Acknowledgments: Preparation for this article was supported by a grant from the Social Sciences and Humanities Research Council of Canada. We thank the two anonymous reviewers, David Day, Fran Yammarino, and Gary Johns for their constructive comments on earlier versions of this article.

July 2008

Abstract

The aim of this study was to assess the research published in *The Leadership Quarterly* from its inception in 1990 to 2007. As the foundation for our study, we used self-reported limitations sections of empirical articles as an alternative, novel, and context-sensitive index of state-of-science. Limitations reported in the one-hundred and seventy-four empirical articles published in *The Leadership Quarterly* to date were coded according to traditional threats to validity. Our results indicate that LQ articles mostly report limitations related to external validity issues. Also, a growing concern with internal validity was noted. These findings offer a unique perspective on leadership research, one that paints a considerably different picture than that offered from previous empirical reviews. We discuss the role of self-reported limitations in scientific communication and offer some prescriptions for increasing their value.

The many heels of Achilles: An analysis of self-reported limitations in leadership research

1. Introduction

The scientific process is iterative in nature and advances in any scientific domain are contingent upon the continuous refinement of theories and of the methodology used to test them. For scientific communities, there are clear benefits to gaining some perspective on how research is being conducted in specific domains and to examine how scientific knowledge unfolds. Such epistemological pursuits, also referred to as 'state-of-science', are increasingly common in organizational sciences. Indeed there has been a growing interest in taking stock of the progress attained in our field (e.g., Casper, Eby, Bordeaux, Lockwood, & Lambert, 2007; Duriau, Reger, & Pfaffer, 2007; Johnson, Buehring, Cassell, & Symon, 2006; Kiessling & Harvey, 2005; Williams, Edwards, & Vandenberg, 2003). The strategy adopted in these empirical investigations has been to rely, as an index of state-of-science, on a content analysis of peer-reviewed empirical articles, more specifically of the methods sections. Scandura and Williams (2000), for example, used such an approach to uncover research trends in organizational studies. By and large, this focus on state-of-science assessment has raised researchers' awareness of their areas of study and served, to some extent, to guide research efforts. For example, the investigation of Scandura and Williams (2000) found that research in the human resource management area has mostly relied on field studies. This has led the authors to conclude that this body of knowledge is rich on realism but also lacks in experimental control and, consequently, in internal validity. The present study seeks to further our understanding of leadership research by focusing on a novel operationalization for state-of-science research. Instead of relying on the content analysis of methods sections of articles, we rely on the

limitations sections found in peer-reviewed articles to capture the 'state-of-science' of leadership research.

In comparison to other domains in the applied social sciences, the field of leadership has been particularly reflective. Probably because of the long and rich history of their discipline, leadership researchers have been quite attentive to its evolution (House & Aditya, 1997). As could be expected, leadership research has seen its share of meta-analyses (e.g., Eagly, Johannesen-Schmidt, van Engen, 2003; Bono & Judge, 2004; Burke, Stagl, Klein, Goodwin, Salas, & Halpin, 2006). However, this reflective disposition has also manifested itself through less traditional approaches. Yammarino, Dionne, Chun, and Dansereau (2005), for example, presented a qualitative review of the utilization of levels-of-analysis in leadership research. This review of over three-hundred articles pointed to important issues in the identification and testing of appropriate level-of-analysis in order to develop more complete and integrated theories of leadership. Another recent review by Giambatista, Rowe, and Riaz (2005) assessed the state of leadership succession research by analyzing the content of 61 empirical articles on this topic. These researchers identified some methodological issues that could be addressed, such as providing greater attention to internal validity, in order to advance this body of knowledge. Bryman (2004) reviewed 66 articles on leadership that used a qualitative methodology. His review illustrates the contribution of qualitative research to leadership and, more specifically, how these research efforts have been useful in highlighting the impact of context on leadership and the understanding of leadership in relation to the change process.

Most recently, DeChurch and Lam (2007) reviewed the methodological choices made in leadership research. These authors, using the same framework as Scandura and Williams (2000), analyzed the content of 164 empirical articles on leadership published in the *Journal of Applied*

Psychology, Personnel Psychology, Academy of Management Journal, Organizational Behavior and Human Decision Processes, and Leadership Quarterly (*LQ*) over the past 10 years. Their results show that leadership research is dominated by field studies and that over three-quarters of it relies on full-time workers, mostly managers and military personnel. Although the above-mentioned studies paint a precise and useful picture of the methodological choices made in leadership research they only indirectly address the suitability and appropriateness of these same choices. The current study seeks to provide a more direct and context-sensitive evaluation of leadership research.

As far as we know, no study has ever looked at the content of self-reported limitations. This is surprising because the acknowledgement of limitations is customary in most peer-reviewed journal publications. The *American Psychological Association (APA) Publication Manual* (5th edition), the style manual used in a large number of top-tier journals in the social sciences, states that authors should: "Acknowledge limitations, and address alternative explanations of results" (p.26). Beyond these editorial guidelines, the reporting of limitations is further mediated by the peer-review process. The review process is not only essential to ensure the quality of the papers published (Harrison, 2002; Spector, 1998) but a common task for reviewers is to point out limitations that may have been overlooked, consciously or not, by authors. In a comprehensive exercise aimed at enhancing the peer-review process, Campion (1993) developed a detailed checklist for the review of articles in applied psychology. Of interest are guidelines pertaining to the formulation of limitations such as "Provides a realistic (and adequately self-critical) delineation of limitations and weaknesses" and "Identifies known or suspected boundary conditions or limits of generalizability" (p.717).

McGrath (1981) refers to the research process as "a set of dilemmas that need to be lived with" (p. 179) and self-reported limitations are a reflection of these dilemmas. The predicament of organizational researchers is made all the most acute by the widening boundaries and the multi-disciplinary nature of the discipline, and the methodological inventiveness that characterizes it. Under these circumstances, research objectives are not the only determinant of methodological choices and that other considerations such as norms of practice, political issues and the historical frame of the research topic also influence these choices (Buchanan & Bryman, 2007). It follows that 'state-of-science' efforts and the appreciation of research endeavors are meaningful only when considered in context; a notion that has long been advocated by science historians (e.g., Khun, 1962; Merton, 1973). By focusing solely on objective characteristics of studies (e.g., design, sample size, variable operationalization, etc.), surveys of methodological choices tend to neglect contextual factors. In that regard, they represent a descriptive exercise. Self-reported limitations, on the other hand, let each article (i.e., authors, reviewers, and editors) speak for itself. By taking this more qualitative approach to state-of-science we follow calls for a better appreciation of context in organizational research (Capelli & Scherer, 1991, Johns; 2006). Limitations capture context by positioning each article's weaknesses within the specific scholarly context of the topic, at the time of its publication.

2. A framework of limitations

Previous state-of-science efforts have relied on the work of Cook and Campbell (1976) to analyze the content of empirical articles (e.g., Scandura & Williams, 2000; DeChurch & Lam, 2007). This framework, which delineates threats to validity, can also be used to categorize the content of limitations sections. It proposes various threats to validity that fall under of internal,

external, construct, and statistical conclusion validity. These threats to validity are described in more detail in the next section.

2.1. Internal validity

Internal validity is concerned with causality. In operational terms, causality is determined by an actual covariation between variables, data collection procedures that allow for the determination of cause-and-effect, and the rejection of alternative explanations (Sackett & Larson, 1990). A common threat to internal validity is the use of cross-sectional designs. For example, in their study of demographic similarity, Pelled and Xin (1997) mentioned that: "A second caveat in this study is that the data were cross-sectional, so causality could not be determined" (p.446). Selective survival, which pertains to a systematic attrition of the sample, is also a threat to internal validity. Hirst, Mann, Bain, Pirola-Merlo, and Richver (2004) reported that "While the longitudinal study design allowed us to examine self-reported learning over 12 months, respondent attrition reduced the sample for testing the major hypotheses. […] While we controlled for possible confounds due to respondent attrition, the hypotheses and the model would have been more adequately tested on a larger sample" (p.323). Additional causes of threat to internal validity include interference between testing conditions, cohort effects, counter-balancing issues, changes in measurement tools between pre- and post-test conditions and coding problems.

2.2. Statistical conclusion validity

Statistical conclusion validity refers to the capacity to make inferences based on the statistical evidence presented. Small sample size is the most straightforward instance of this

category. Jung, Chow, and Wu (2003), for example, stated that: "Although we were able to use PLS to accommodate our relatively small sample size […] using data from a larger number of companies will permit more powerful hypothesis tests" (p.540). Variable issues such as a lack of clarity of a key variable or a variable that is simply not measured are also considered here. As an example, in their paper on the integration of top management teams, Carmeli and Schaubroeck (2006) wrote that "[…] it is also plausible that, in addition to perceived decision quality, an unmeasured mediating variable or variables mediated the behavioral integration effect on decline" (p. 450). Other threats to statistical conclusion validity include low internal consistency/reliability and low statistical power.

2.3. Construct validity

Construct validity centers on the fit between the measures employed and the constructs that they purport to represent (Stone-Romero, 1994). Participant bias poses a threat to construct validity. For example, Dasborough (2006) writes in her discussion that "[…] focus group interviews and individual interviews are social situations; hence, interviewees might feel that they must comply with social norms, resulting in positive bias of results" (p. 175). Other threats to construct validity include no control group, rater and confederate biases, the placebo effect, resentment of treatment condition, too much or too little variance, same source variance and inferences made on individuals based on group data or vice versa.

2. 4. External validity

External validity refers to the extent to which the results generalize across time, settings, and individuals (Cook & Campbell, 1976). Common threats to external validity are sampling

issues either related to a specific demographic profile such as that found in Dvir & Shamir (2003) who write that "[…] the study was conducted in a military context during training with relatively young, all-male participants" (p.342). Other threats to external validity are situation-specific results; these are not sample-related and pertain to the context in which the study was conducted. For example, in the discussion of their article comparing transformational and contingent reward leadership, Yammarino, Spangler, and Dubinsky (1998) stated that "[…] examining the generalizability of the results to other superiors and subordinates in other types of jobs and organizations is necessary. […] So, replication of the current results is required in other settings." (p. 51). Other threats to external validity include the file drawer problem, missing data, and low response rate.

3. The evolution of research

The main goal of research lies in the advancement of science and, as a result, 'state-of-science' efforts derive their meaning when taking a longitudinal approach. Over the years, several trends have been noted in management research. For example, in their comparison of the methodologies employed by top-tier management journals in the middle 1980's from that used in the middle 1990's, Scandura and Williams (2000) found a decrease in laboratory experiments and an increase in cross-sectional research efforts. Research methodologies are not insensitive to trends and there are many factors that may underlie the adherence or abandonment of certain methodological alternatives. As possible explanations of an evolution in the choices made by researchers, Scandura and Williams (2000) offer systematic changes in the training of researchers, preferences of editors and review boards, the pressure for more expedient research as a result of tenure pressure, and more conservative human subject review boards. In addition,

the development of survey and statistical software has clearly improved the speed and ease of collecting and interpreting data. Finally, the internet has greatly promoted collaborative efforts, resulting in the progress of scientific inquiry (Amabile, Patterson, Mueller, & Wojcik, 2001). Overall, one would assume that the combination of these trends would also impact the quality of the research produced in the field of leadership.

4. The present study

The main purpose of the study is to appraise the limitations of the empirical research published in *The Leadership Quarterly*. To do so, we pose the following research question: "What are LQ authors reporting as limitations to their research?" In addition, we seek to observe the evolution of this reporting over the past 15 years. We believe that this novel approach to state-of-science will contribute to the extant leadership literature by providing a context-sensitive appreciation of leadership research.

5. Method

5.1. Population and descriptive variables

All of the research articles published in *The Leadership Quarterly* (LQ) from its debut in the spring of 1990 (Volume 1, Issue 1) until the end of 2007 (Volume 18, Issue 5) were used for this study. As a multi-disciplinary journal, LQ has established itself as the major scientific outlet for leadership research (Hunt, 2005; Lowe & Gardner, 2001). It is also ranked as one of the top management journals in various rankings (e.g., Foss & Klein, 2007; Zickar & Highhouse; 2001). Excluded from our sample were theoretical articles, literature reviews, and commentaries. Of the 479 articles published in LQ for the above period, 174 were research articles; these represent the

sample for this study. Table 1 outlines the number of articles published in LQ since 1990. On average, 9.7 empirical articles per year have been published in LQ. The research designs used for the 174 studies were coded (Table 2). One-hundred and twelve studies were based on surveys (64.3%) of which 82 were conducted in the field (47.1% of the total sample). In addition, the sample included 28 experimental studies (16.1%), 23 archival studies (13.2%), 6 studies based on interviews (3.4), and 5 based on observations (4.0%).

Insert Table 1 about here

Insert Table 2 about here

Four different editors have managed LQ over these 16 years (years and amount of empirical articles managed in parentheses): Bernard M. Bass (1990: 7), Robert J. House and Henry L. Tosi, Jr., (1991-1992: 1), Francis J. Yammarino (1992-1998: 42), James G. Hunt (1999-2004: 75), and Michael D. Mumford (2005-2007: 49).

In LQ, limitations are typically presented in two different formats: some authors present them under a clearly identified 'limitations' section while others embed the limitations of their study within the discussion section. Since the content of limitations may possibly be related to the format used in reporting them, it was coded as a variable. Note that when a distinct limitations section was not present, the whole discussion was submitted for coding[1]. In terms of format, less than half of all articles had their limitations clearly distinguished by a heading and

grouped into a separate set of paragraphs (47.7%) whereas the rest had their limitations sections embedded in the discussion section (52.3%).

5.2. Coding Scheme and Procedure

A taxonomy of limitations was developed based on the work of Cook and Campbell (1976) and Scandura and Williams (2000); it can be found in Table 3. Three graduate students were trained to use the taxonomy to code the limitations. Definitions and examples of limitations for each category were introduced and discussed in a training session. In the session, the coders were guided through various examples of limitations or discussion sections. After the training, coders were given a first sample of 30 limitations and discussion sections to code, representing 17% of the articles. Krippendorf's alpha (Hayes & Krippendorff, 2007) was used to evaluate inter-coder reliability for this first set of coding. This measure allows the computation of inter-coder agreement with multiple coders across multiple variables. Krippendorf's alpha is represented by a numerical scale ranging from 0 to 1. The inter-coder reliability across the three coders for this first sample was .78. As a result of some similarities between some of the categories, slight modifications were made to the coding scheme. For example, 'key variable not measured' and 'one or too few measures used to assess construct' were combined into a single category labeled 'variable issues'. An additional training session was conducted to introduce the coders to these changes and to address the categories for which agreement was lower. The coders were instructed to pay particular attention to clearly reported limitations and distinguish them from suggestions for further research. Following this second training, the coders were given a revised coding manual and coded the remaining 126^2 limitation and discussion sections.

Insert Table 3 about here

The resulting reliability across the three coders was .72. As a result of comments by coders, additional modifications were made to the coding scheme. The category 'no control for an important variable' was merged into 'variable issues' and 'time lapse between testing conditions' and 'setting does not assess stable behavior' were merged into 'situation-specific results'. As a result of these modifications, the inter-coder reliability increased to .79.

In order to obtain values for each coded variable, the following decision rule was used: when two or more coders agreed the limitation was recorded.

6. Results

A total of 401 limitations were reported by LQ authors for an average of 2.3 limitations per article (Table 1). The percentages of articles containing each limitation are presented in Table 3. The most commonly reported limitations were sample bias (reported in 46.6 % of articles), variable issues (44.8%), situation-specific results (40.8%), no causality (24.1%), same source variance (20.1%), sample size too small (16.6%), and participant bias (8.6%). Limitations were also collated according to the four general categories of threats to validity. According to this categorization, articles were most likely to identify threats to external validity (63.8%), followed by statistical conclusion validity (51.7%), construct validity (31.0%) and internal validity (27.6%).

6.1. Limitations as a function of time

In order to detect trends in reporting over time, the percentage of articles containing each limitation was computed for three time periods, from 1990 to 1995, from 1996 to 2001, and from 2002-2007 (Table 3). Results indicate a remarkable increase in the reporting of threats to internal validity over the three time periods, 4.5%, 15.1% and 39.4% respectively. In order to test the significance of this trend, we regressed the year of publication onto each of the four categories, controlling for the number of empirical articles published that year. Results suggest a trend toward reporting internal validity as a limitation ($\underline{\beta} = .05$, $\underline{R^2} = .09$, $\underline{p} < .01$) (Table 4); no other significant trends were found.

Insert Table 4 about here

A closer examination of threats to internal validity indicates that the reporting of no causality is mostly responsible for this increase (4.5%, 13.2, and 34.3%). The only other notable trend at the discrete level was found for 'variable issues' that also tended to increase over the years, 22.7%, 43.4%, and 50.5% respectively.

6.2. Limitations as function of design, editor, and reporting format

In order to investigate the effect of research design, editor, and reporting format on the four aggregate threats to validity a multivariate analysis of variance (MANOVA) was conducted. The results of this MANOVA showed that the effects of editor [$F(12, 365) = 1.68$; ns, Wilk's $\lambda = .87$] and reporting format [$F(4, 138) = 1.52$; ns, Wilk's $\lambda = .96$] on these variables were not significant. As expected, design had an effect on the four dependent variables [$F(24, 483) =$

2.11; $\underline{p} < .01$, Wilk's $\lambda = .71$]. More specifically, the design of the study had a significant effect on statistical conclusion validity ($\underline{p} < .01$) and construct validity ($\underline{p} < .05$). Post-hoc tests using least squared difference revealed that there were less limitations pertaining to statistical conclusion validity reported in experimental studies than in surveys and field surveys. In terms of construct validity, these limitations were less present in archival studies than they were in surveys, field surveys, and interviews.

6.3. Reporting format as a function of time and editorship

The evolution of the reporting format of limitations (i.e., embedded or in a separate section) over time and across editor was also tested. In term of fluctuation across time, 18.2% of articles had separate limitation sections in the 1990-1995 period, 43.4% between 1996 and 2001, and 56.6% between 2001 and 2007. A regression analysis confirms the statistical significance of this trend ($\underline{\beta} = .26$, $\underline{R^2} = .06$, $\underline{p} < .001$). An ANOVA also detected an effect for editorship [$F(3, 169) = 3.64$; $\underline{p} < .01$]. The proportion of papers with separate limitations sections varied significantly across editors: Bass (.14), Yammarino (.40), Hunt (.44), and Mumford (.65).

6.4. Clarity of reporting: Inter-coder reliability:

The coding of limitations, and more specifically the inter-coder agreement for each specific limitation, allow us to investigate the clarity with which these limitations are written or identify those that stand out as being very clear for coders and those that do not. Coders agreed the most for limitations pertaining to no causality (.89), same source variance (.89), and small sample size (.84). On the other hand, limitations such as low internal consistency/reliability (.40), participant bias (.57), and unequal sample distribution based on culture (.55) were less

obvious to code and led to more disagreement. Interestingly, the reliability did not increase according to whether or not the limitations were grouped in a specific section of the article. The overall agreement for limitations reported in a distinct section was .81 while it was.79 for limitations embedded throughout the discussion section.

7. Discussion

The present study highlights some interesting facets of the research published in LQ. Our results identify external and statistical conclusion validity as the most important shortcomings of this body of research. In addition, the growing focus on internal validity and causality issues highlights a subtle but steady shift in self-reported limitations. The use of a novel operationalization of state-of-science paints a picture of the 'state-of-leadership' that differs significantly from previous reviews of leadership research (DeChurch & Lam, 2007; Lowe & Gardner, 2001). Self-reported limitations are context-rich and, as a result, our findings are sensitive to the specific concerns of leadership research and to the evolution of the field in general. In the following section we discuss these findings in more detail and offer some prescriptions that we believe may increase the value and usefulness of limitations.

Our results reveal a primary concern for external validity for LQ authors. Despite the abundant use of field surveys, threats to external validity represent the most reported limitations; over 60% of articles surveyed acknowledged such a limitation. While a focus on generalizability is common to all applied social sciences, LQ authors appear particularly sensitive to the contextual specificity of their work and its lack of generalizability. One the great challenges of leadership research has always been to understand the role context and the applicability of findings in different situations (e.g., Bryman, 1996; Hunter et al., 2007; Yukl, 2006). It is

interesting to note that reviews based on the tabulation of methodologies have led us to a different conclusion. Both DeChurch and Lam (2007) and Lowe and Gardner (2001), for example, painted a very positive picture of external validity on the account of the many field studies reported in their reviews. This disparity in findings lends some support to the sensitivity of limitations to the specific setting in which the research takes place.

Another notable finding pertains to the increase in internal validity limitations, the only threat to validity to show an increase over the 18 years of LQ. Interestingly, while our data shows that this increase is mostly due to an increase in causality issues (from 4.5%, 13.2%, and 34.3% across the three time periods), the use of surveys---the main culprit for causality issues---remained relatively stable over time (54.6%, 69.8%, and 63.6%). In other words, leadership researchers have become increasingly concerned with internal validity *despite the fact* that the use of key methodological choice related to this issue has remained constant. Thus, it appears that the increase in threats to internal validity is not due to a change in methodology but rather to a shift in awareness towards this issue. Again, this finding points to the sensitivity of limitations to context. Leadership processes involve complex causal effects, taking the form of reciprocal, and recursive relationships, and spanning levels of analysis (e.g., Hunter et al., 2007). As the leadership area is maturing, disentangling internal validity issues is gradually becoming more of a concern.

The second most-reported limitation by LQ authors pertained to statistical conclusion validity (50.6%). Variable issues, with a frequency count of 44.8%, are mostly responsible for this result. The difficulty of capturing leadership behaviors and the context within which they occur is well-documented and these results are congruent with other reviews of leadership research (e.g., DeChurch & Lam, 2007; Lowe & Gardner, 2001). As stated by Hunter, Bedell-

Avers, & Mumford (2007): "The typical leadership study often fails to take into account the situations, either through the lack of examining potential moderators or by failing to measure and subsequently control for potentially biasing effects" (p.439). This result speaks directly to the operational challenges of conducting quality leadership research.

While the prevalence of specific limitations, their absence can also be telling. For example, the (over) emphasis on dyads in leadership research and, more importantly, the lack of attention given to multi-level issues has been raised by many influential leadership researchers over the years (e.g., House & Aditya, 1997; Yukl, 1999, Yammarino & Dansereau, 2006), yet this limitation was reported in only 3.4% of articles. Although we did not assess the use of multi-level methodologies in our study, Lowe and Garner (2001) did report that a majority of the research published in the first 10 years of LQ remained at the individual level. If one adheres to the notion that "leadership involves at a minimum a leader, a collection of subordinates, and exchanges among them […]" (p.436; Hunter et al., 2007), then we can presume that this limitation is grossly under-represented. Once again, these results point to a selective, context-specific, focus in the assessment of the shortcomings of published research.

7.1 What are self-reported limitations?

The above discussion raises an essential element of our study: the nature of self-reported limitations. What are they exactly? This is a legitimate question because our results indicate that self-reported limitations are not an inclusive inventory of the weaknesses of a study but rather a sample of those weaknesses. In this paper, we argued repeatedly that the appreciation of methodological choices is guided, in part, by contextual factors. In this section we consider the process underlying the generation of limitations.

The process of academic publishing is quite complex and elaborate. It involves multiple steps and many contributors to the finished product. In this long road to print, the self-reporting of limitations poses a real dilemma for authors. One is basically asked to disclose information that will most probably weight negatively in the final editorial decision. The significant pressure stemming from the emphasis on publications in academic reward structures (Gomez-Mejia & Balkin, 1992) and increasingly low acceptance rates for top-journals represent clear motives for not reporting limitations. At the very least, these pressures create a dynamic in which authors take a 'wait and see' approach while editors and reviewers shoulder the task of bringing limitations to light. A rigorous peer-review process is one way to address this issue and limitation sections may well be the most susceptible to the input of reviewers and editors. Gilliland and Cortina (1997), for example, found that the adequacy of research designs is most strongly related to reviewers' evaluations of a manuscript. Reviewers and editors, with their external perspective, are in a very good position to tease out limitations and assess their importance.

However, a unique characteristic of at least some limitations is that they will not always be privy to external evaluators, be it readers, reviewers or editors. Limitations are sometimes related to an anecdotal element of the research endeavor. In these instances, limitations will only be 'visible' to those with intimate knowledge of the mechanics of the study. If not self-reported, or 'confessed', these limitations would never be detected. For example, Fu & Yukl (2000) used scenarios to investigate differences in influence behavior between American and Chinese managers. As a limitation, they reported that: "We found some evidence of interesting cross-cultural differences that we did not report because they involved tactic scales without strong internal consistency. […] The scenario questionnaire will be revised before it is used again to

improve scale reliability, include more tactics, and counterbalance situational variables better within scenarios." (p. 263). Clearly, the reporting of all relevant limitations is an important building block of scientific communication. However, their prejudicious nature makes them a rather unique and vulnerable component of peer-reviewed articles.

So far, our discussion has dealt primarily with the nature of limitations and what they represent. Going a step further, our study begs another question: besides being diagnostic, what impact do limitations have on the research process? Unfortunately, the present study raises more questions about the impact of self-reported limitations than it answers. It is sensible to ponder why, for example, did the increasing concern with internal validity not translate into more experimental designs? We share the view of an anonymous reviewer that limitations sections may be the least read sections of research articles. We discussed the role of limitations with some of our colleagues and many perceive these sections as 'boiler plate' or repositories for general statements about some unimportant flaws of the study. This is unfortunate; our results offer evidence that limitations do tell part of the story, an important one in our opinion.

We were surprised to find very few editorial guidelines concerning the reporting of limitations in the major management journals. As mentioned in the introduction, the *APA Publication Manual* (5th edition) is very terse in its directive to authors regarding limitations. The relative flexibility of these guidelines is in sharp contrast with the trend towards a standardization of research publications in some disciplines. For example, editorial boards of major medical journals such as the *Journal of the American Medical Association*, the *British Medicine Journal*, and the *Lancet* now require authors to adhere to a very strict structure in writing abstracts, discussion sections and, in the case of random clinical trials, whole articles (Doherty & Smith, 1999; Taddio, Pain, Fassos, Boon, Ilersich, & Einarson, 1994). Given their

importance for the advancement of science and relative vulnerability to the peer-review process, we believe that more specific directions are warranted in as to the reporting of limitations. In the following sections, we offer some suggestions as to how the reporting of limitations could be improved.

8. Guidelines for reporting limitations

8.1. Limitations need to be communicated

It is interesting to note that, in comparisons to other scientific areas, LQ authors---and authors in organizational sciences in general---are quite forthcoming when it comes time to acknowledging limitations. A recent study of self-reported limitations in each of the most-cited scientific journals (e.g., *Science, Nature, Proceedings of the National Academy of Science;* Ioannidis, 2007) found that, of the 400 articles analyzed, only five had a separate limitations section (1.3%) and only 67 (16.8%) included limitations. In contrast, only 11.5**%** of the LQ articles surveyed in this paper *did not* acknowledge any limitations and 20.1% reported only one. While our methodology slightly under-estimate the frequency of limitations, we still feel that they are under-reported[3]. If we accept the premise that all empirical research is flawed to some degree, one would expect more limitations to be included in discussion sections. When limitations are not acknowledged by authors, it should be the responsibility of reviewers to highlight them and that of editors to insure that they are reported.

8.2. Limitations need to be communicated…clearly

We were surprised at the variation in the inter-coder reliability across limitations. It is troubling that basic and straightforward limitations such as ‘internal consistency’, ‘participant

bias', and 'unequal sample distribution based on culture' yield such low reliabilities. Reviewers and editors may want to pay more attention to ensure that limitations are clearly communicated. Also, more than half of LQ articles did not use separate headings for limitations. Although our study did not find that placing limitations in a separate section added to their clarity, such a section certainly helps the reader in locating them and may be related to their impact on subsequent research. We did note an increase in the use of these sections in LQ over the years and recommend a continued effort in this direction.

8.3. Focus on the 'why' rather than on the 'why not'

In addition to describing the shortcomings of the study, limitations statements should also delineate the 'why' or the implications of these shortcomings for the research at hand, and possibly, for the area under study in general. In our study, we encountered many descriptions that simply listed limitations, without addressing their relevance. For example, Murphy, Blyth, and Fiedler (1992) state that: "First, the subjects in Study 1 were all undergraduate students, although the subjects in Study 2 from the National Guard were older and had more experience in work-related group interactions." (p. 252). This statement is not very useful for readers as it remain silent as to the rationale behind the limitation and its implication for the results. Another very common type of limitations consists of the following: "*the study had limitation X but it does not really matter because of Y and Z*." Although we understand the need for authors to establish the credibility of their research by explaining that their results stand despite being limited in some way, such an exercise should not come at the expense of an understanding of the influence of the possible limitation on the interpretation of the study (i.e., the 'why').

We did find plenty of examples of detailed, contextual limitations statements in our review. For example, Xin & Tsui (1996) reported that: "The generalizability of these findings may be limited due to the nature of the sample used. […] there may be distinct differences between Asian-Americans managers and Asian managers in Asia. Asian-American managers used in this sample may have been successfully integrated into U.S. society. As a result, their values may now be similar to those in the mainstream U.S. culture." (p. 128). Statements such as this are useful to the reader because it help him or her understand how the limitation may have influenced the results. Again, reviewers and editors can ensure that limitations contain the details necessary to make them informative.

8.4. Authors, not reviewers or editors, are responsible for reporting limitations

While a rigorous peer-review process is quite capable of teasing out a study's major weaknesses, the shared ownership of limitations, in our opinion, is problematic for two basic reasons. First, as stated earlier, external assessors can only catch those limitations that are visible. Other, equally important, 'hidden' limitations will make it to print only if the author chooses to acknowledge them. Unfortunately, a scholarly Fifth Amendment is simply not applicable here because evidence of research shortcomings, self-generated or not, will always remain pertinent in evaluating a research effort. Secondly, as a result of being populated from the outside, in piece-meal fashion, limitations sections end up being disjointed and disconnected from the main thrust of the paper. The reactive stance taken by authors creates limitations that are less likely to be made whole with the paper and an integral part of the story (i.e., the 'why'). Clearly, editors are in the best position to influence authors to disclose, develop, and integrate limitations in their manuscripts.

9. Limitations and future research

Given the topic at hand, we would be remiss not to be particularly mindful of the limitations of our own study. First, this study focused on articles published in a single academic journal. Hence the literature under scrutiny represents only a portion of the leadership research conducted over the past 15 years (file-drawer problem, selective sample). A review of leadership articles from other leading journals may lead to different results and may help tease out an 'LQ' effect, if one exists. If, as argued in this paper, context is so important in determining limitations, journals with different histories, editorial policies, and review procedures may well yield contrasting results. Secondly, many of our interpretations are based on untested assumptions related to the respective roles of the authors, reviewers, and editors in generating limitations. The extent to which limitations are influenced by the review process is still unknown. As noted by an anonymous reviewer, an analysis of the different versions of manuscripts, from the original submission to the published version, would certainly shed some light on this issue. Thirdly, some important categories yielded low coder agreement (low reliability). Although we speculated that some limitations were poorly reported by authors, a closer look at the data also indicates that some of this unreliability may have been caused by elements of the process itself. The fact that the inter-coder reliability decreases after training and the modifications made to the scheme points to this issue. The coding scheme contained a large amount of categories, some of them closely related (e.g., file drawer problem and missing data). The coders often commented on the difficulty in classifying some limitations. In addition, the coding task was spread over a few months and the frame for some categories many have shifted for coders over the course of the task. Finally, the most frequently reported limitations, variable

issues (44.8%) and situation-specific results (40.8%), were the results of a combination of categories in the coding scheme. It is quite probable that our taxonomy was based on uneven categories and that these two covered more of the 'limitation spectrum' than other, more narrowly-defined, ones. A refinement of this taxonomy might be useful for future research in this area.

This study underscores areas of concerns in leadership research but it also brings to light the nature of limitations sections and the purpose that they serve in the advancement of science. Limitations are, and have always been, a staple in the communication of scientific findings in the social sciences and a large majority of the 174 empirical articles published in LQ devoted at least a few lines to describe the study's shortcomings. Limitations represent an important aspect of scientific communication and this study reveals that they can provide an additional and useful perspective on research efforts. We also hope that the study raises the awareness of the role of limitations in science-building. Given the energy and effort necessary to conduct and publish research, we could all benefit from paying a little more attention to the lessons that limitations provide.

References

Amabile, T. A., Patterson, C., Mueller, J., & Wojcik, T. (2001). Academic-practitioner collaboration in management research: A case of cross-profession collaboration. *Academy of Management Journal, 44,* 418-432.

Bono J. E. & Judge, T. A. (2004). Personality and Transformational and Transactional Leadership: A Meta-Analysis. *Journal of Applied Psychology, 89,* 901-910.

Buchanan, D.A., & Bryman, A. (2007). Contextualizing methods choice in organizational research. *Organizational Research Methods, 10,* 483-501.

Burke, C. S., Stagl, C., Klein, C., Goodwin, G. F., Salas E., & Halpin, S. M. (2006). What type of leadership behaviors are functional in teams? A meta-analysis. *Leadership Quarterly, 17,* 288-307.

Bryman, A. (1996). The impact of context: Qualitative research and the study of leadership. *Leadership Quarterly, 4,* 289-304.

Bryman, A. (2004). Qualitative research on leadership: a critical but appreciative review. *Leadership Quarterly, 15,* 729-769.

Campion, M. A. (1993). Article review checklist: A criterion checklist for reviewing research articles in applied psychology. *Personnel Psychology, 46,* 705-718.

Cappelli, P. and Scherer, P. D. (1991). The missing role of context in OB: the need for a meso-level approach. *Research in Organizational Behavior, 13,* 55-110.

Carmeli, A & Schaubroeck, J. (2006). Top management team behavioral integration, decision quality, and organizational decline, *Leadership Quarterly, 17,* 441-453.

Casper, W. J., Eby, L. T., Bordeaux, C., Lockwood, A., & Lambert, D. (2007). A review of research methods in IO/OB work-family research. *Journal of Applied Psychology, 92,* 28-43.

Cook, T. D. & Campbell, D. T. (1976). The design and conduct of quasi-experiments and quasi-experiments in field settings. In M. Dunnette (Ed.) *Handbook of Industrial & Organizational Psychology* (p. 223-336). Chicago: Rand McNally.

Cook, T. D., Campbell, D. T., & Peracchio, L. (1990). Quasi experimentation. In M. Dunnette & L. Hough (Eds.), *Handbook of Industrial & Organizational Psychology: 2nd Edition, Volume 1* (p. 491-576). Palo Alto, CA: Consulting Psychologists Press.

Dasborough, M. T. (2006). Cognitive asymmetry in employee emotional reactions to leadership behaviors. *Leadership Quarterly, 17,* 163-178.

DeChurch, L. A. & Lam, C. F. (2007). *Leadership research methodology*. Paper presented at the 20th Annual Conference of the Society for Industrial and Organizational Psychology, New York, NY.

Doherty, M. & Smith R. (1999). The case for structuring the discussion of scientific papers. *British Medical Journal, 318,* 1224-1225.

Doty, D.H & Glick, W.H. (1998). Common Methods Bias: Does Common Methods Variance Really Bias Results? *Organizational Research Methods*, 1, 374-406.

Dvir, T. & Shamir, B. (2003). Follower developmental characteristics as predicting transformational leadership: a longitudinal field study. *Leadership Quarterly, 14*, 327-344.

Duriau, V. J., Reger, R. K., & Pfaffer, M. D. (2007). A content analysis of the content analysis literature in organization studies. *Organizational Research Methods, 10,* 5-34.

Eagly, A. H., Johannesen-Schmidt, M. C., & van Engen, M.L. (2003). Transformational, transactional, and laissez-faire leadership styles: A meta-analysis comparing women and men. *Psychological Bulletin, 129,* 569-591.

Foss, N. J. & Klein, P. G. (2007). Management journals impact factors: 2005. *Organizations and Markets*. Retrieved July 11, 2007, from http://organizationsandmarkets.com/2007/03/14/management-journal-impact-factors-2005/

Fu, P. P. & Yukl, G. (2000). Perceived effectiveness of influence tactics in the United States and China. *Leadership Quarterly, 11,* 251-266.

Gomez-Mejia, L. & Balkin, D. (1992). "The Determinants of Faculty Pay: An Agency Theory Perspective. *Academy of Management Journal, 35,* 921-955.

Giambatista, R. C., Rowe, W. G., & Riaz, S. (2005). Nothing succeeds like succession: a critical review of leader succession since 1994. *Leadership Quarterly, 16,* 963-991.

Gilliland, S. & Cortina, J. (1997). Reviewer and editor decision making in the journal review process. *Personnel Psychology, 50,* 427-452.

Harrison, D. (2002). Obligations and obfuscations in the review process. *Academy of Management Journal, 46,* 1079-1084.

Hayes, A. F. & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1,* 77-89.

Hirst, G., Mann, L., Bain, P., Pirola-Merlo, A., & Richver, A. (2004). Learning to lead: the development and testing of a model of leadership learning. *Leadership Quarterly, 15,* 311-327.

House, R. J. & Aditya, R. N. (1977). The social scientific study of leadership: Quo vadis? *Journal of Management, 23,* 409-473.

Howell, J.M. (1990). Leadership behaviors, influence tactics, and career experiences of champions of technological innovation. *Leadership Quarterly, 1*, 249-264.

Hunt, J, (2005). Explosion of the leadership field and LQ's changing of the editorial guard. *Leadership Quarterly, 16,* 1-8.

Hunter, S. T., Bedell-Avers, K. E., & Mumford, M. D. (2007). The typical leadership study: assumptions, implications, and potential remedies. *Leadership Quarterly, 18,* 435-446.

Ioannidis, J. P. A. (2007). Limitations are not properly acknowledged in the scientific literature. *Journal of Clinical Epidemiology, 60,* 324-329.

Johns, G. (2006). The essential impact of context on organizational behavior. Academy of Management Review, 31, 386-408.

Johnson, P., Buehring, A., Cassell, C., & Symon, G. (2006). Evaluating qualitative management research: Towards a contingent theory. *International Journal of Management Reviews, 8,* 131-156.

Jung, D. I., Chow, C., & Wu, A. (2003). The role of transformational leadership in enhancing organizational innovation: Hypotheses and some preliminary findings. *Leadership Quarterly, 14*, 525-544.

Kiessling, T. & Harvey, M. (2005). Strategic global human resource management research in the twenty-first century: An endorsement of the mixed-method research methodology. *International Human Resource Management, 16,* 22-45.

Lowe, K. & Gardner, W. (2001). Ten years of The Leadership Quarterly: contributions and challenges for the future. *Leadership Quarterly, 11*, 459-514.

McGrath, J.E. (1981). Dilemmatic: The study of research choices and dilemma. *American Behavioral Scientist, 25,* 179-210.

Merton, R.K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago, IL: University of Chicago Press.

Mook, D.G. (1983). In defense of external invalidity. *American Psychologist, 38,* 379-387.

Murphy, S., Blyth, D., & Fiedler, F. (1992) Cognitive resource theory and the utilization of the leader's and group members' technical competence. *Leadership Quarterly, 3*, 237-255.

Pelled, L. H. & Xin, K. (1997). Birds of a feather: leader-member demographic similarity and organizational attachment in Mexico. *Leadership Quarterly, 8, 433*-450.

Sackett, P. R. & Larson, J. R. (1990). Research strategies and tactics in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.) *Handbook of Industrial and Organizational Psychology : 2nd Edition, Volume 1* (p. 419-490). Palo Alto, CA: Consulting Psychologists Press.

Scandura, T. A. & Williams, E. A. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management Journal, 43*, 1248-1264.

Spector, P. E. (1998). When reviewers become authors: A comment on the journal review process. *Research Methods Forum*, 1-4. Retrieved November 11, 2006 from http://division.aomonline.org/rm/

Spector, P.E. (2006). Method Variance in Organizational Research. Truth or Urban Legend? *Organizational Research Methods*, 9, 221-232.

Stone-Romero, E. F. (1994). Construct validity issues in organizational behavior research. In J. Greenberg (Ed.), *Organizational behavior: The state of science*. Hillsdale, NJ: Erlbaum.

Taddio, A., Pain, T., Fassos, F. F., Boon, H., Ilersich, A. L., & Einarson, T. R. (1994). Quality of nonstructured and structured abstracts of original research articles in the British Medical Journal, the Canadian Medical Association Journal and the Journal of the American Medical Association. *Canadian Medical Association Journal,* 150, 1611-1615.

Terpstra, D. (1981). Relationship between methodological rigor and reported outcomes in organization development evaluation research. *Journal of Applied Psychology, 66,* 541-543.

Williams, L. J., Edwards, J. R., & Vandenberg, R. J. (2003). Recent advances in causal modeling methods for organizational and management research. *Journal of Management, 29,* 903-936.

Xin, K.R. & Tsui, A.S. (1996). Different strokes for different folks? Influence tactics by Asian-American and Caucasian-American managers. *Leadership Quarterly, 7,* 109-132.

Yammarino, F. J., Dionne, S. D., Chun, J. U., & Dansereau, F. (2005). Leadership and levels of analysis: a state-of-science review. *Leadership Quarterly, 16,* 879-919.

Yamarino, F. J. & Dansereau, F. (2006). The Leadership Quarterly special issue on multi-level approaches to leadership. *Leadership Quarterly, 17,* 109-109.

Yammarino, F. J., Spangler, W. D., & Dubinsky, A. J. (1998). Transformational and contingent reward leadership: Individual, dyad and group levels of analysis. *Leadership Quarterly, 9,* 27-54.

Yukl, G. (1999). An evaluation of the weaknesses in transformational and charismatic leadership theories. *Leadership Quarterly, 10,* 285-305.

Yukl, G. (2006). Leadership in organizations, 6th ed. Upper Saddle River, NJ: Prentice Hall.

Zickar, M. J. & Highhouse, S. (2001). Measuring prestige of journals in industrial-.organizational psychology. *The Industrial-Organizational Psychologist, 38.*

Table 1
Number of articles published per year

Year	# of articles	# of empirical articles (% of total)		# limitations reported (average per article)	
1990	16	7	(43.7)	10	(1.4)
1991	15	0	(0)	0	(0.0)
1992	13	1	(7.7)	3	(3.0)
1993	21	4	(19.0)	5	(1.2)
1994	20	5	(25.0)	12	(2.4)
1995	24	5	(20.8)	11	(2.2)
1996	25	8	(32.0)	16	(2.0)
1997	25	8	(32.0)	19	(2.4)
1998	26	12	(46.1)	29	(2.4)
1999	31	6	(19.3)	13	(2.2)
2000	26	9	(34.6)	24	(2.6)
2001	20	10	(50.0)	18	(1.8)
2002	36	18	(50.0)	35	(1.9)
2003	32	12	(37.5)	29	(2.4)
2004	44	20	(45.4)	42	(2.1)
2005	36	12	(33.3)	32	(2.6)
2006	35	19	(54.3)	46	(2.4)
2007	34	18	(52.9)	57	(3.2)
Total	**479**	**174**	**(36.3)**	**401**	**(2.3)**

Average number of limitations per articles is between parentheses.

Table 2
Distribution of research design

Design	Total (174)	1990-1995 (22)	1996-2001 (53)	2002-2007 (99)
Survey	30 (17.2%)	6 (27.3%)	13 (24.5%)	11 (11.1%)
Field Survey	82 (47.1%)	6 (27.3%)	24 (45.3%)	52 (52.5%)
Experimental	28 (16.1%)	4 (18.2%)	10 (18.9%)	14 (14.1%)
Observational	5 (2.9%)	0 (0.0%)	0 (0.0%)	5 (5.1%)
Interviews	5 (2.9%)	3 (13.6%)	0 (0.0%)	2 (2.0%)
Archival	22 (12.6%)	3 (13.6%)	6 (11.3%)	13 (13.1%)

Table 3
Limitations categories and frequency

	Total (174)	**1990-95** (22)	**1996-01** (53)	**2002-07** (99)
Threats to Internal Validity[a]	**27.6%**	**4.5%**	**15.1%**	**39.4%**
1. Interference between testing conditions.	0.6%	0.0%	0.0%	1.0%
2. The cohort effect.	0.6%	0.0%	1.9%	0.0%
3. Selective survival.	1.1%	0.0%	1.9%	1.0%
4. Counter-balancing issues/presentation order.	1.1%	0.0%	0.0%	2.0%
5. Change in measurement tools between pre-test and post-test conditions.	0.0%	0.0%	0.0%	0.0%
6. Coding problems.	0.0%	0.0%	0.0%	0.0%
7. No causality.	24.1%	4.5%	13.2%	34.3%
8. Other threats to internal validity.	0.6%	0.0%	0.0%	1.0%
Threats to Statistical Conclusion Validity[a]	**51.7%**	**36.4%**	**52.8%**	**54.5%**
9. Low internal consistency/reliability.	2.9%	0.0%	5.7%	2.0%
10. Variable issues: Key variable unclear, not measured or a single/too few variables used to measure construct or form conclusion.	44.8%	22.7%	43.4%	50.5%
11. Sample size too small.	16.7%	27.3%	15.1%	15.2%
12. Low power (not related to reliability or sample size).	1.7%	0.0%	5.7%	0.0%
13. Other threats to statistical conclusion validity.	0.6%	0.0%	0.0%	1.0%
Threats to Construct Validity[a]	**31.0%**	**27.3%**	**26.4%**	**34.3%**
14. No control group.	0.0%	0.0%	0.0%	0.0%
15. Participant bias.	8.6%	4.5%	7.5%	10.1%
16. Rater bias.	2.9%	4.5%	7.5%	0.0%
17. Confederate bias.	0.6%	0.0%	0.0%	1.0%
18. Placebo effect.	0.0%	0.0%	0.0%	0.0%
19. Resentment of treatment condition.	0.0%	0.0%	0.0%	0.0%
20. Treatment stimuli did not elicit intended response.	0.0%	0.0%	0.0%	0.0%
21. Too much or too little variance.	4.0%	4.5%	1.9%	5.1%
22. Same source variance.	20.1%	18.2%	15.1%	23.2%
23. Inferences made on individuals based on group data (levels of analysis).	3.4%	4.5%	0.0%	5.1%
24. Other threats to construct validity.	0.0%	0.0%	0.0%	0.0%
Threats to External Validity[a]	**63.8%**	**63.6%**	**66.0%**	**62.6%**
25. Situation-specific results.	40.8%	54.5%	45.3%	35.4%
26. File drawer problem.	1.7%	0.0%	1.9%	2.0%
27. Missing data.	2.3%	0.0%	5.7%	1.0%
28. Low survey response.	2.3%	0.0%	1.9%	3.0%
Unequal sample distribution (*overall*):	(*46.6*)	(*40.9%*)	(*45.2%*)	(*48.5%*)
29. Age.	2.9%	0.0	3.8%	3.0%
30. Gender.	8.0%	4.5%	11.3%	7.1%
31. Culture.	7.5%	9.1%	7.5%	7.1%
32. Student.	11.5%	9.1%	7.5%	14.1%
33. Volunteer.	0.0%	0.0%	0.0%	0.0%
34. Other/selective sample.	16.7%	18.2%	15.1%	17.2%
35. Other threats to external validity.	2.3%	0.0%	5.7%	1.0%

[a] The percentages for the aggregate categories pertain to those articles with one or more related discrete limitation.

Table 4
Summary of linear regression analysis for overall variables in terms of time

Overall Variable	Betas	R^2
Internal Validity	.05	.09**
External Validity	-02	.01
Construct Validity	.02	.00
Statistical Conclusion Validity	.02	.01

N = 174.

** p < .01

Notes

1. At the beginning of the coding process, we checked for the presence of 'stray limitations' or limitations outside limitation sections in a small subset of articles and found none.

2. Note that, upon acceptance of the paper, the database was updated to include recent articles. Due to the unavailability of the original coders, the co-authors coded the limitations found in the last 18 empirical papers (volume 18). The reliability indices pertain only to the original three coders and 156 articles.

3. On average, coders *agreed* on the content of 2.3 limitations per article. Without considering agreement, the three coders identified, on average, 2.4 limitations per article.