Journal of Management
Vol. 44 No. 7, September 2018 2909–2933
DOI: 10.1177/0149206316657594


# What Doesn't Get Measured Does Exist: Improving the Accuracy of Computer-Aided Text Analysis

Aaron F. McKenny
*University of Central Florida*
Herman Aguinis
*George Washington University*
Jeremy C. Short
Aaron H. Anglin
*University of Oklahoma*

*Computer-aided text analysis (CATA) is a form of content analysis that enables the measurement of constructs by processing text into quantitative data based on the frequency of words. CATA has been proposed as a useful measurement approach with the potential to lead to important theoretical advancements. Ironically, while CATA has been offered to overcome some of the known deficiencies in existing measurement approaches, we have lagged behind in regard to assessing the technique's measurement rigor. Our article addresses this knowledge gap and describes important implications for past as well as future research using CATA. First, we describe three sources of measurement error variance that are particularly relevant to studies using CATA: transient error, specific factor error, and algorithm error. Second, we describe and demonstrate*

*Acknowledgments: This article was accepted under the editorship of Patrick M. Wright. We thank Fred Oswald, two* Journal of Management *anonymous reviewers, and Regan Stevenson for providing highly constructive and useful feedback that allowed us to improve our manuscript substantially. Also, we thank the following three author teams for sharing texts used in their studies: (a) Moss, T. W., Payne, G. T., & Moore, C. B. 2014. Strategic consistency of exploration and exploitation in family businesses.* Family Business Review, *27: 51-71; (b) Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. 2010. Construct validation using computer-aided text analysis (CATA): An illustration using entrepreneurial orientation.* Organizational Research Methods, *13: 320-347; and (c) Zachary, M. A., McKenny, A. F., Short, J. C., & Payne, G. T. 2011. Family business and market orientation: Construct validation and comparative analysis.* Family Business Review, *24: 233-251.*

*Supplemental material for this article is available with the manuscript on the* JOM *website.*

*Corresponding author: Aaron F. McKenny, Department of Management, College of Business Administration, University of Central Florida, P. O. Box 161400, Orlando, FL 32816-1400, USA.*

*E-mail: aaron.mckenny@ucf.edu*

*how to calculate measurement error variance with the entrepreneurial orientation, market orientation, and organizational ambidexterity constructs, offering evidence that past substantive conclusions have been underestimated. Third, we offer best-practice recommendations and demonstrate how to reduce measurement error variance by refining existing CATA measures. In short, we demonstrate that although measurement error variance in CATA has not been measured thus far, it does exist and it affects substantive conclusions. Consequently, our article has implications for theory and practice, as well as how to assess and minimize measurement error in future CATA research with the goal of improving the accuracy of substantive conclusions.*

***Keywords:*** *content analysis; measurement error; reliability; entrepreneurial orientation; market orientation; ambidexterity*

---

Computer-aided text analysis (CATA) is a form of content analysis that enables the measurement of constructs by processing text into quantitative data based on the frequency of words (Short, Broberg, Cogliser, & Brigham, 2010; Short & Palmer, 2008). For example, Pfarrer, Pollock, and Rindova (2010) measured the affective component of firm celebrity using *BusinessWeek* articles about firms and the relative frequency of words such as "love" or "nice." Given the widespread availability of text that can be used to conduct management research across micro and macro levels of analysis, it is not surprising that CATA has recently been used to measure a wide range of constructs, such as organizational psychological capital (McKenny, Short, & Payne, 2013), blame (Gangloff, Connelly, & Shook, 2016), and firmness of resolve (Brett, Olekalns, Friedman, Goates, Anderson, & Lisco, 2007).

The availability of text makes CATA practically and logistically appealing. But even more important from the perspective of theory advancement is that CATA offers psychometric advantages in comparison to more traditional measures, such as self-reports and archival data (Aguinis & Edwards, 2014; Short & Palmer, 2008). Specifically, regarding internal validity, CATA frequently involves data collected in naturally occurring organizational contexts, such as texts included in annual reports (Short et al., 2010). Thus, the data collected allow for greater confidence regarding relations among underlying constructs compared to other types of measures, such as self-reports and archival data (Dalton & Aguinis, 2013). Second, regarding external validity, the use of CATA allows for the collection of a large amount of data across units, enhancing the generalizability of conclusions. In other words, it is easier to collect data across a larger sample of contexts, levels, and circumstances compared to more traditional measures (McKenny, Short, Payne, & Zachary, 2012). Third, regarding construct validity, CATA is less susceptible to threats such as common method variance and endogeneity that are pervasive in more traditional measures because CATA facilitates data collection from different sources. Finally, regarding statistical conclusion validity, CATA facilitates the examination of very large samples of texts, resulting in satisfactory levels of statistical power to test hypotheses, which is not typically the case in management research that relies on more traditional measures (e.g., Aguinis, 1995). In short, CATA offers advantages in terms of internal, external, construct, and statistical conclusion validity, which explains why it is seen as a nascent yet highly promising measurement approach (Aguinis & Vandenberg, 2014; Short & Palmer, 2008).

Although CATA holds considerable promise compared to more traditional measurement approaches, it is ironic that it lags behind with respect to assessing measurement concerns. For

instance, while there is considerable guidance regarding the assessment and mitigation of measurement error variance in survey research (e.g., Schmidt, Le, & Ilies, 2003), there is no such direction regarding CATA. This gap in measurement scrutiny between current CATA research and the standards for other approaches raises the question of how much measurement error exists in studies using CATA and the potential impact on the accuracy of substantive conclusions.

CATA is a promising and useful measurement approach. However, to leverage its strengths, there is a need to understand how to identify and minimize key sources of measurement error. Our manuscript follows in the footsteps of recent *Journal of Management* articles that provide methodological guidance to researchers investigating substantive management phenomena and theories (e.g., Aguinis, Forcum, & Joo, 2013; Aguinis, Gottfredson, & Culpepper, 2013; O'Boyle, Banks, & Gonzalez-Mulé, 2017; Zyphur, Oswald, & Rupp, 2015). Specifically, we demonstrate how to examine and refine CATA measures using three illustrative constructs: entrepreneurial orientation (Lumpkin & Dess, 1996), market orientation (Narver & Slater, 1990), and organizational ambidexterity (March, 1991). In doing so, we demonstrate that measurement issues have affected substantive conclusions of existing CATA research and also present opportunities for improvement. We also demonstrate how to reduce measurement error variance by refining existing CATA measures and offer best-practice recommendations on how to assess and minimize measurement error in future CATA research with the goal of improving the accuracy of substantive conclusions.

## Measurement Error in CATA

Measurement is the quantification of objects or events using a systematic set of rules (Kerlinger & Lee, 2000). In CATA, measurement of organizational phenomena is obtained indirectly through the examination of texts (Duriau, Reger, & Pfarrer, 2007). While indirect measurement is a source of advantage for CATA, it also introduces unique potential sources of measurement error (Short et al., 2010). In self-report measurement instruments, such as surveys, respondents respond to questions that can be carefully worded to solicit answers associated with a specific construct. However, CATA frequently relies on texts that are created without the authors knowing that they will be used to measure a specific construct (Short & Palmer, 2008). As a result, the contents of the texts are frequently broad and variable, introducing the possibility of measurement error.

Measurement error can be classified into two categories: measurement error variance and bias. Measurement error variance reflects the imprecision of a measure to capture only the true variance of the underlying phenomenon and is assessed via reliability estimates (Ree & Carretta, 2006). Bias reflects systematic deviation in observed measurement away from the true score of the underlying phenomenon and is assessed via validity (Kerlinger & Lee, 2000). To address validity issues in CATA, researchers frequently follow existing guidance for validating measures (i.e., McKenny et al., 2013; Short et al., 2010). However, the lack of guidance regarding the assessment and minimization of measurement error variance in CATA limits our knowledge surrounding the precision of these measures. Minimizing the effects of measurement error is a necessary condition for validity because measurement error variance artificially inflates the observed variance of focal constructs (Lord & Novick, 1968; Nunnally & Bernstein, 1994; Shadish, Cook, & Campbell, 2002).

In traditional survey research, there are three key sources of measurement error variance: random response error, transient error, and specific factor error (Schmidt et al., 2003; Schmidt & Hunter, 1999). However, previous research has not estimated the effects of these sources of measurement error variance on measurement using CATA. Next, we describe each of these sources of measurement error variance and their potential applicability to CATA. Also, we describe an additional source of measurement error variance particularly relevant to CATA: algorithm error.

### *Random Response Error*

In survey research, random response error concerns how fluctuations in a study participant's cognitions influence the data provided (Schmidt & Hunter, 1999). For example, if individuals become distracted when answering a question, they may respond differently than when they answer a similar question later in the survey (Sturman, Cheramie, & Cashen, 2005). This fluctuation could influence the words used by the author during the creation of a text, lowering measure reliability if the construct being measured is not related to these fluctuations.

Random response error is commonly assessed using split-half reliability (Schmidt et al., 2003; Schmidt & Hunter, 1999). The direct parallel in content analysis consists of dividing each sampled text in half and analyzing the halves separately. By correlating the results from the first half of the text with the results from the second half of the text, researchers could assess the consistency of language associated with the phenomenon of interest throughout the text. However, individuals tend to write about one topic at a time rather than jumping from topic to topic within a text. For example, in letters to shareholders, CEOs are likely to describe past performance early in the text and future plans later in the text. As a result, splitting the text in half will frequently isolate discussion of certain topics into either the first or the second half of the text, suggesting that this procedure would overestimate random response error. As an alternative procedure, we could consider splitting the text in half vertically. However, this would no longer assess random response error because each of the halves would no longer reflect different points in time during the writing of the text. Furthermore, vertical splits would be sensitive to formatting artifacts such as section headers and page layout artifacts such as columns and would frequently rely on researcher judgment when a word is split by the vertical line. This decision could result in additional need for researcher judgment as many studies using CATA rely on multiple document formats spanning thousands of pages. Because neither operationalization of split-half reliability is likely to accurately capture random response error in CATA, we conclude that this form of reliability estimate is not informative for research using CATA.

### *Transient Error*

Transient error is concerned with persistent temporal factors that could influence a respondent's reported values provided on the entire instrument (Schmidt et al., 2003). For example, respondents' mood may affect their responses on an entire survey (Becker, 2000). In CATA, transient error is also concerned with fluctuations in psychological processes over time and how those changes are manifested in texts. These fluctuations could influence the words used by the author across the creation of multiple texts, increasing measurement error variance.

A test-retest reliability estimate can be used to assess the impact of transient error by examining the correlation between scores collected at two points in time (Schmidt & Hunter, 1996). Many organizational texts, such as letters to shareholders, are generated at regular intervals (often annually), which facilitates the collection of data over time (Short et al., 2010; Zachary, McKenny, Short, & Payne, 2011). Thus, correlating the CATA measurements of texts generated at two points in time indicates the extent to which the language was consistent over time.

### *Specific Factor Error*

When traditional measures are used, specific factor error arises when the content of the measure itself influences the data provided by the respondent (Schmidt et al., 2003; Schmidt & Hunter, 1999). For example, idiosyncrasies in the wording of questions on a survey may influence the resulting score. In CATA, specific factor error is concerned with errors arising from the content of the word lists. Because texts used in content analyses are frequently generated without intervention from researchers, the content of the text can be expected to be free from researcher demand bias (Barr, Stimpert, & Huff, 1992; Duriau et al., 2007). However, when developing CATA measures, researchers identify words that provide evidence of the underlying phenomena on the basis of the judgment of expert raters (Short et al., 2010). Thus, the process of selection and omission of specific words by researchers may introduce specific factor error.

A parallel forms reliability estimate assesses specific factor error by presenting respondents with two equivalent, but different, instruments designed to measure the same construct (Schmidt & Hunter, 1999). In the context of CATA, parallel forms reliability can be estimated by comparing the results of the computer-generated analysis to a manually coded subset of at least 10% of the texts used in the analysis. We suggest using 10% of the sampled texts to obtain stable estimates in content analytic research on the basis of guidance by Wimmer and Dominick (2013). CATA is valuable in its ability to measure large volumes of texts quickly by using predefined word lists (Short & Palmer, 2008). However, CATA's dictionary-based algorithm may not capture context or rhetorical nuance. Accordingly, correlating the results from manual coding with the results from CATA provides a useful quantification of specific factor error.

### *Algorithm Error*

CATA is also subject to a unique source of measurement error that is usually not relevant to more traditional measurement approaches. CATA uses an algorithmic rater (via specific software packages) to evaluate the content of texts (Krippendorff, 2013). Accordingly, if two or more CATA algorithms generate different scores from the same text, "interrater" (i.e., interalgorithm) disagreement quantifies the degree of algorithm error.

To assess the extent of error associated with algorithm error, we suggest using multiple CATA software packages and calculating Krippendorff's alpha interrater agreement estimate, which generalizes several known statistics also referred to as intercoder agreement, interrater reliability, and reliability of coding given sets of units. Krippendorff's alpha is computed as 1 – (observed disagreement/expected disagreement attributable to chance). Our recommendation

**Table 1**
**Sources of Measurement Error Variance in Computer-Aided Text Analysis (CATA) Research**

| Error Source | Reliability Estimate | Calculation Guidelines |
|---|---|---|
| Transient error: Measurement error arising from differences in the language used in texts produced at different points in time. | Test-retest reliability: Assesses the consistency of language from texts produced at two points in time. | Collect two texts for each individual or organization, each produced at different points in time.<br>Use CATA to measure the construct in both samples of texts.<br>Calculate the correlation between the two sets of scores to assess the extent of transient error. |
| Specific factor error: Measurement error arising from choices made in compiling word lists. | Parallel forms reliability: Assesses the extent to which human and CATA coding produces similar scores. | Manually code at least 10% of the texts to be analyzed.<br>Calculate the correlation between the scores generated by the manual and software coding to assess the extent of specific factor error. |
| Algorithm error: Measurement error arising when two CATA software packages produce different scores using the same measures and texts. | Interrater agreement: Assesses the extent to which two CATA software packages produce the same scores. | Conduct CATA with two software packages.<br>Calculate Krippendorff's alpha statistic to assess the extent of algorithm error. |

parallels guidance for other research where multiple third-party judges are used to conduct manual content analyses (e.g., Krippendorff, 2013). An important advantage of Krippendorff's alpha is that it is applicable to any number of algorithms (i.e., "raters"). Also, macros to compute Krippendorff's alpha are available for SPSS, Stata, SAS, and R (Gamer, Lemon, Fellows, & Singh, 2012; Hayes & Krippendorff, 2007; Klein, 2014).

Table 1 includes a definition of the three sources of measurement error variance that we suggest should be examined in CATA research, the reliability estimate that should be used to measure each, and procedures for the computation of each reliability estimate. Next, we use three constructs to demonstrate how to assess measurement error variance in CATA studies: entrepreneurial orientation, market orientation, and organizational ambidexterity. We selected these constructs because they are of interest to a number of management subfields, including business policy and strategy, entrepreneurship, organization theory, and family business.

## Method

For entrepreneurial orientation, we used word lists developed and validated by Short and colleagues (2010). Entrepreneurial orientation is defined as the behaviors and decision-making processes within organizations that facilitate the pursuit of entrepreneurial opportunities and new entry (Lumpkin & Dess, 1996). Entrepreneurial orientation is often conceptualized as a construct composed of five dimensions: autonomy, competitive aggressiveness, innovativeness, proactiveness, and risk taking (Lumpkin & Dess, 1996; Short et al., 2010). We used the dimension definitions offered by Lumpkin and Dess (1996): (a) Autonomy: "the independent action of an individual or a team in bringing forth an idea or a vision and carrying it through to completion" (140); (b) Innovativeness: "a firm's tendency to engage in and

support new ideas, novelty, experimentation, and creative processes that may result in new products, services, or technological processes" (143); (c) Risk taking: "the firm's proclivity to engage in risky projects and managers' preferences for bold versus cautious acts to achieve firm objectives" (146); (d) Proactiveness: "acting in anticipation of future problems, needs, or changes" (146); and (e) Competitive aggressiveness: "a firm's propensity to directly and intensely challenge its competitors to achieve entry or improve position, that is, to outperform industry rivals in the marketplace" (148).

For market orientation, we used word lists developed by Zachary and colleagues (2011). Market orientation is defined as organization-wide creation, coordination, and exploitation of market information in pursuit of competitive advantage (Kohli & Jaworski, 1990; Narver & Slater, 1990). Zachary and colleagues operationalized market orientation using the five-dimensional MKTOR model (i.e., Narver & Slater, 1990). The MKTOR model is composed of three core components—competitor orientation, customer orientation, and interfunctional coordination—and two decision criteria—long-term focus and profitability (Narver & Slater, 1990; Zachary et al., 2011). We used the definitions provided by Zachary and colleagues: (a) Customer orientation: "refers to the degree to which an organization has developed an understanding of its consumer base so as to provide continuous and superior value to present and future customers" (235); (b) Competitor orientation: "requires a business to understand the 'short-term strengths and weaknesses and long-term capabilities and strategies' of competitors (Narver & Slater, 1990, pp. 21-22). Similar to customer orientation, competitor orientation emphasizes the importance of understanding both current as well as potential competitors and their operations (Narver & Slater, 1990)" (236); (c) Interfunctional coordination: "involves the synchronized utilization of resources in such a way as to create superior value for customers and other stakeholders (Narver & Slater, 1990; Webster, 1988). This requires all of a business' structural components to participate in the collection, sharing, and utilization of market information, not just the designated marketing employees (Kohli & Jaworski, 1990; Narver & Slater, 1990)" (236); (d) Long-term focus: "emphasizes the need for businesses to look toward the future as they strive toward the other dimensions of market orientation" (237); and (e) Profitability: "is viewed as the 'overriding objective' of a business (Narver & Slater, 1990, p. 22). In other words, profitability is used as a metric from which a business finds an optimal level of investment in market orientation (Kohli & Jaworski, 1990; Narver & Slater, 1990)" (237).

For organizational ambidexterity, we used word lists developed by Uotila, Maula, Keil, and Zahra (2009). Organizational ambidexterity is defined as the joint pursuit of exploration, or the identification of new opportunities, and exploitation, or the seizing of existing opportunities (Allison, McKenny, & Short, 2014; March, 1991; Raisch & Birkinshaw, 2008). When the exploration and exploitation constructs were first introduced, they were presented in a manner consistent with content analysis, noting that exploration "includes things captured by terms such as search, variation, risk taking, experimentation, play, flexibility, discovery, innovation" and exploitation "includes such things as refinement, choice, production, efficiency, selection, implementation, execution" (March, 1991: 71).

### *Samples*

To enhance the comparability of our study with previous studies using CATA measures of entrepreneurial orientation, market orientation, and organizational ambidexterity, we

obtained our sample of texts from the authors of those studies (i.e., Moss, Payne, & Moore, 2014; Short et al., 2010; Zachary et al., 2011). Because we sought to directly replicate analyses implemented in previous research using CATA measures, we used the same text sampling frame and selection criteria, the same CATA package for estimating algorithm error, and the same measures as in the original studies. The entrepreneurial orientation and market orientation constructs were measured using shareholder letters. The ambidexterity construct was measured using Management Discussion and Analysis (MD&A) sections of 10-K filings. Next, we offer more detailed information on each of these samples of texts.

Shareholder letters are a key component of most large firms' annual reports, providing an outlet for the CEO to speak on behalf of the firm to convey the recent past, current, and anticipated future state of the company (McKenny et al., 2013; Short, Payne, Brigham, Lumpkin, & Broberg, 2009). As a result of their availability and presentation of how managers view the firm, shareholder letters are one of the most widely used texts in management research using content analysis (Duriau et al., 2007). In addition, publicly traded organizations communicate with shareholders regularly through annual reports. Thus, this type of communication provides a valuable sampling frame for content analytic research because it maximizes sample size and increases the availability of texts from multiple time periods.

The sampling frame for our entrepreneurial orientation assessment was shareholder letters from the 450 firms that were listed on the S&P 500 every year from 2001 to 2005. This sampling frame was the same one used by Short and colleagues (2010) in the development of the entrepreneurial orientation CATA measures. Short and colleagues used only shareholder letters from 2005 in their analysis; however, to assess test-retest reliability, we required a 2nd year of texts. Accordingly, we collected a purposive sample of 2006 shareholder letters associated with the same firms for which shareholder letters were available in 2005. Our final sample consisted of 745 shareholder letters from 401 firms representing 169 industries (four-digit Standard Industrial Classification code, or SIC). The length of shareholder letters ranged from 125 to 7,545 words, with an average of 1,717 words ($SD$ = 901).

The sampling frame for our market orientation assessment was shareholder letters from the 224 firms that both were listed on the S&P 500 every year from 2001 to 2005 and could be identified as either a family or nonfamily business. This sampling frame was used by Zachary and colleagues (2011) in the original development of the market orientation CATA measures. Zachary and colleagues used all available shareholder letters from 2001 to 2005 for these 224 firms in their analysis. Our final sample consisted of 1,112 shareholder letters from 224 firms representing 124 industries (four-digit SIC). The length of shareholder letters ranged from 74 to 5,589 words, with an average of 1,611 words ($SD$ = 732).

The sample of texts for our organizational ambidexterity assessment comprised MD&A sections of firms' 10-K filings. This disclosure is used by managers to communicate information about the company, its strategy, financial performance and assumptions, recent activity, and forward-looking statements to current and potential investors (Clarkson, Kao, & Richardson, 1999). Like annual reports, 10-Ks are filed by publicly traded companies on an annual basis, facilitating the collection of longitudinal data. The sampling frame for our organizational ambidexterity assessment was the 205 firms in four high-tech industries (SICs: 2834, 7370, 7372, 7373) for which 10-Ks could be collected every year from 1997 through 2008 and for which family business status could be determined. This sampling frame was used by Moss and colleagues (2014) in their use and refinement of these measures. Our final

sample consisted of 2,460 MD&A statements from 205 firms representing four industries. The length of MD&A statements ranged from 41 to 31,968 words, with an average of 6,617 words (*SD* = 4,423).

### *Assessment of Measurement Error Variance*

We followed each of the calculation guidelines summarized in Table 1 for all three constructs. First, to assess transient error, we calculated correlation coefficients for the dimensions of each construct at two points in time for entrepreneurial orientation and averaging together the correlation for each pair of consecutive years for ambidexterity and market orientation (cf. Schmidt et al., 2003). Second, assessing specific factor error required manual coding. Accordingly, we selected a total of 100 shareholder letters at random across 2 years (50 letters each year) for entrepreneurial orientation and market orientation. This reflects 12.5% of the entrepreneurial orientation sample and 22% of the market orientation sample, exceeding the 10% guideline for estimating parallel forms reliability recommended by Wimmer and Dominick (2013). We selected a total of 64 MD&A statements at random across 2 years (32 statements each year) for ambidexterity. We developed a manual coding scheme using the same definitions of the construct dimensions that were used to develop the CATA measures (i.e., Short et al., 2010; Uotila et al., 2009; Zachary et al., 2011). Furthermore, because the CATA word lists include both words and phrases, we conducted the manual coding at the word or phrase level.

Assessments of algorithm error require that two CATA packages be used with the same set of texts. We used the same software package as in the original studies to maximize comparability: LIWC 2007 (Pennebaker, Booth, & Francis, 2007) for ambidexterity and DICTION 5 (Hart, 2000) for entrepreneurial orientation and market orientation. To add a common second package for these constructs, we relied on the CAT Scanner CATA tool (McKenny, Short, & Newman, 2012). The CATA measures of entrepreneurial orientation and market orientation include phrases in the word lists (Short et al., 2010; Zachary et al., 2011). Furthermore, the original ambidexterity word lists contained stemmed words (e.g., "explor*"; Uotila et al., 2009). CAT Scanner provided a valuable second package because it was designed to accommodate single words, phrases, and word stems in its analysis (McKenny, Short, & Newman, 2012).

## Results

### *Relative Impact of Measurement Error Sources*

Table 2 includes results from implementing the measurement error variance assessment procedures for entrepreneurial orientation. Results for market orientation and organizational ambidexterity are available as Appendices A and B in the online supplemental material. We calculated measurement error variance from each source as (1 – reliability) * 100, which quantifies the percent of observed variance due to measurement error.

Results showed that, overall, the percent of variance due to algorithm error is almost nonexistent in ambidexterity (2%) and is somewhat higher for entrepreneurial orientation (11%) and market orientation (16%). This may be due to the use of phrases in the original entrepreneurial orientation and market orientation measures. CAT Scanner can process phrases with

**Table 2**
**Results of Measurement Error Assessment for Computer-Aided Text Analysis Measures of Entrepreneurial Orientation**

| Error Source | Type of Reliability Estimate | Entrepreneurial Orientation Dimension | Reliability Estimate | Percent of Variance Due to Measurement Error[a] |
|---|---|---|---|---|
| Transient error | Test-retest | Autonomy | .32 | 68 |
| | | Competitive aggressiveness | .43 | 37 |
| | | Innovativeness | .52 | 48 |
| | | Proactiveness | .55 | 45 |
| | | Risk taking | .71 | 29 |
| | | Mean (Test-retest) | **.51** | **49** |
| Specific factor error | Parallel forms | Autonomy | .29 | 71 |
| | | Competitive aggressiveness | .51 | 49 |
| | | Innovativeness | .35 | 65 |
| | | Proactiveness | .72 | 28 |
| | | Risk taking | .30 | 70 |
| | | Mean (Parallel forms) | **.43** | **57** |
| Algorithm error | Krippendorff's alpha | Autonomy | .90 | 10 |
| | | Competitive aggressiveness | .89 | 11 |
| | | Innovativeness | .89 | 11 |
| | | Proactiveness | .88 | 12 |
| | | Risk taking | .90 | 10 |
| | | Mean (Krippendorff's alpha) | **.89** | **11** |

*Note:* The mean reliability estimates across the three sources of error and percent of variance due to measurement error (shown in parenthesis) for each dimension are as follows: Autonomy: .50 (50%); competitive aggressiveness: .61 (39%); innovativeness: .59 (41%); proactiveness: .72 (28%); and risk taking: .64 (36%). Boldface text indicates the mean reliability estimate and percent of variance across all five entrepreneurial orientation dimensions for each type of error.
[a]Percent of variance due to measurement error = (1 – reliability estimate value) * 100.

spaces in them (McKenny, Short, & Newman, 2012). However, DICTION 5 requires that a hyphen or other punctuation be placed in the phrase. If the texts included these phrases using spaces, CAT Scanner would identify them but DICTION 5 would not, decreasing the correlation between the scores obtained with the two packages.

The percent of variance due to specific factor error was high for entrepreneurial orientation (57%) and organizational ambidexterity (81%). The market orientation measure performed somewhat better with respect to specific factor error (34%). This suggests that the original market orientation dictionary has better precision in estimating the market orientation of the analyzed firms. However, the high levels of variance attributable to specific factor error suggest that there may still be opportunities for refining all three measures to improve parallel forms reliability.

The percent of variance due to transient error was high for entrepreneurial orientation (49%) and market orientation (47%) and somewhat lower for organizational ambidexterity (20%). There are several factors that likely drove the considerable differences in the transient error estimated in these documents. The first is the level of formality. Shareholder letters tend to be written in prose with little standardization, whereas MD&A documents tend to be written more formally with greater consistency. For instance, MD&A documents often include

formal statements regarding the firm's accounting policies and procedures. Because these formal policies are relatively stable, the text concerning these policies provides a stable core to the contents of the MD&A statements that is not present in shareholder letters. Furthermore, in our manual coding of the MD&A documents, we identified places where firms appeared to lift text verbatim from the previous year's MD&A document and tweak the language used to reflect the changes in firm activity and outlook over the past year. Second, both texts are influenced by multiple individuals within the firm and can be expected to be reasonably accurate as a result of legislation regulating executive accountability for the contents of corporate disclosures (e.g., Sarbanes-Oxley; Geiger & Taylor, 2003). However, ultimately the viewpoints presented in each narrative are attributable to different parties. Shareholder letters provide the CEO with the opportunity to communicate with shareholders on behalf of the company (Goodman, 1980). Thus, while multiple parties frequently contribute to the document, the nature of the shareholder letter suggests that the CEO's voice will be featured heavily. Furthermore, because shareholder letters tend to be relatively short, the issues salient to the CEO are more likely to be discussed to the potential exclusion of other issues. By contrast, the MD&A document is not attributed to a specific individual and can be quite long.

It is also possible that the transient error we estimated for each construct captures changes in the underlying construct. Ideally, researchers would estimate transient error using texts produced in rapid succession, reducing the likelihood of a change in the underlying phenomenon. This would reduce the likelihood that a change in firm leadership or a change in firm direction would inflate estimates of transient error. However, many organizational texts are produced at fixed, regular intervals outside of researcher control (e.g., shareholder letters, 10-Ks), often leading to a longer lag. In these cases, CATA users should examine whether key antecedents of strategic change, such as a change in leadership, occurred during the assessment period.

While capturing transient error using test-retest reliability estimates is a common procedure, test-retest reliability may also be influenced by the scoring method used. For instance, when using the same documents, test-retest reliability may differ between CATA and human raters. To examine this issue, we estimated test-retest reliability for the results of our manual content analysis as well. Similarly to the CATA results, we found that the test-retest reliability estimates for manual coding were lower for entrepreneurial orientation and market orientation than for ambidexterity. On average, only 14% of the variance in ambidexterity scores was associated with transient error. However, this value was 66% for entrepreneurial orientation and 73% for market orientation. Triangulating with the results from the CATA transient error analysis, these results suggest that the contents of shareholder letters are more variable than MD&A statements.

### *Effects of Uncovered Measurement Error Variance on the Estimation of Substantive Relations*

The aforementioned results uncovered the presence of substantial measurement error variance in all three CATA measures. Next, we calculated the extent to which measurement error variance affects substantive conclusions. To do so, we first located published studies that used CATA to measure entrepreneurial orientation, market orientation, and organizational ambidexterity. We identified relevant studies by searching for articles that cited the

articles that developed the original CATA measures (i.e., Short et al., 2009; Short et al., 2010; Uotila et al., 2009; Zachary et al., 2011). Our initial search identified 18 empirical studies, but we eliminated 9 that did not present correlations, did not use the same measure, used the measure as a moderator, or included other non-CATA data to create composite scores. As a result, our final sample comprised 9 studies.

To understand whether our results regarding measurement error warrant revisiting past substantive conclusions, we used the disattenuation formula, which estimates a measurement error–free correlation based on an observed correlation and reliability estimates for the predictor and criterion as follows:

$$\hat{r}_{x_t y_t} = \frac{\hat{r}_{xy}}{\sqrt{\hat{r}_{xx} \cdot \hat{r}_{yy}}}, \tag{1}$$

where $\hat{r}_{xx}$ and $\hat{r}_{yy}$ are the estimated reliabilities of the measures for $x$ and $y$ respectively, $\hat{r}_{xy}$ is the observed correlation between the measures of $x$ and $y$, and $\hat{r}_{x_t y_t}$ is the estimated true (i.e., measurement error–free) correlation between $x$ and $y$.

*Combining multiple reliability estimates*. Although Equation 1 has been used extensively in management research, particularly within the context of meta-analysis (e.g., Aguinis & Pierce, 1998), it requires the use of a single reliability estimate. Our results indicate that there are two important sources of error in CATA research: transient error and specific factor error.

The coefficient of stability (CS) measures the extent to which a single measurement instrument produces similar results at two points in time. In our study, test-retest reliability estimates are classified as a CS because we examined the stability of CATA measurements over 2 years as follows:

$$\text{CS} = \rho(obs_1, obs_2). \tag{2}$$

The coefficient of equivalence (CE) measures the extent to which two parallel measurement instruments produce similar results at the same point in time. In our study, parallel forms reliability estimates are classified as a CE because we compared CATA and manual coding. Specifically, our measure of CE was calculated as the correlation of manual and CATA coding results for texts within 1 year as follows:

$$\text{CE} = \rho(obs_{CATA}, obs_{Manual}). \tag{3}$$

We combined CS and CE estimates by calculating a composite reliability estimate through the calculation of a coefficient of equivalence and stability (CES; Cronbach, 1947). This is accomplished by correlating the CATA measurement at Time 1 with the manual measurement at Time 2 and the CATA measurement at Time 2 with the manual measurement at Time 1 (cf. Schmidt et al., 2003):

$$\text{CES} = \rho(obs_{CATA,1}, obs_{Manual,2}). \tag{4}$$

$$\text{CES} = \rho(obs_{CATA,2}, obs_{Manual,1}). \tag{5}$$

By correlating the CATA measurement at Time 1 with the manual measurement at Time 2, a discrepancy either in the stability of measurements over time or in the equivalence of measures decreases the overall CES (Schmidt & Hunter, 1999). Consequently, this coefficient is a more complete indicator of measurement reliability and a more appropriate

estimate for correcting for measurement error compared to coefficients that consider only a single source of measurement error (Schmidt & Hunter, 1999).

Equations 4 and 5 would yield the same reliability estimate only if the measures were strictly parallel. However, we know on the basis of our measurement error assessment results that parallel forms reliability estimates are far from perfect. Accordingly, we calculated the two CES estimates and then computed an average of the two. Results of averaging the two CES estimates using Equations 4 and 5 for each construct are included in Table 3 in the CES Reliability Estimate column.

Table 3 includes correlations between the CATA measures and several antecedent and consequent variables as reported in previous research. This table also includes the CES estimate for each measure, a 95% confidence interval for the CES, the disattenuated correlations, and the percent of underestimation in substantive relations caused by measurement error. The percent of underestimation in variance explained is calculated as the percent decrease in observed coefficients of determination (i.e., $r^2$) comparing corrected and uncorrected coefficients.

*Measurement error variance and substantive relations*. Table 3 reveals important and somewhat troubling results regarding published research using the existing CATA measures of entrepreneurial orientation, market orientation, and organizational ambidexterity. First, CES reliability estimates, which combine transient and specific factor error, are quite low for most measures. Second, the highest CES reliability estimates are for the composite entrepreneurial orientation measure (.51) and the innovativeness dimension (.51). Thus, even for the measures with the smallest amount of measurement error (i.e., largest CES value), about 50% of the total variance is random and not substantive in nature. In fact, results in Table 3 show that even for the measures with the highest reliability levels, there is substantial underestimation of the size of substantive relations as indexed by the percent of variance explained. For example, the published studies we examined underestimated the relation between CEO tenure and entrepreneurial orientation by 96%, the relation between entrepreneurial orientation and shareholder value by 100%, and the relation between entrepreneurial orientation and capital raised by 89%.

For the overall ambidexterity measure, the autonomy dimension of entrepreneurial orientation, and the proactiveness dimension of entrepreneurial orientation, we found a zero or near-zero value for the CES. This suggests that the impact of transient and specific factor error combined is so large that the true scores are lost in the error. These results are a likely explanation for why observed correlations involving these variables have been reported to be zero or near zero in past research. In sum, results of our measurement error variance assessment point to the need to revisit past substantive conclusions regarding entrepreneurial orientation, market orientation, and organizational ambidexterity.

One of the advantages of CATA is the ability to process a large number of lengthy texts in a relatively short period of time (Short & Palmer, 2008). But the manual coding of 10% or more of the texts required to calculate the CE may not be practically feasible in many studies. For instance, initial public offering (IPO) prospectuses are commonly examined texts in the entrepreneurship literature (e.g., Payne, Moore, Bell, & Zachary, 2013), and some can be hundreds of pages (e.g., the 2004 Google IPO prospectus was 267 pages), making manually coding a 10% subsample of even a modest sample of texts prohibitively time consuming. In this situation, we recommend

**Table 3**
**Effects of Uncovered Measurement Error Variance on the Estimation of Substantive Relations**

| Article | Variable (Role) | CATA Measure; Observed Correlation *r* | CES Reliability Estimate (95% CI)[a] | CATA Measure; Measurement Error–Corrected Correlation $r_c$ | Percent of Underestimation in Variance Explained in Substantive Relations[b] |
|---|---|---|---|---|---|
| Boling, Pieper, and Covin (in press) | CEO tenure (IV) | Entrepreneurial orientation; –.10 | .51 (.27, .69) | Entrepreneurial orientation; –.14 | 96 |
| Engelen, Neumann, and Schmidt (2016) | Shareholder value (DV) | Entrepreneurial orientation; .24 | .51 (.27, .69) | Entrepreneurial orientation; .34 | 100 |
| Engelen, Neumann, and Schwens (2015) | CEO overconfidence (IV) | Entrepreneurial orientation; .16 | .51 (.27, .69) | Entrepreneurial orientation; .22 | 89 |
| Moss, Neubaum, and Meyskens (2015) | Funding success (DV) | Autonomy; –.00 | –.03 (–.31, .25) | Autonomy; Undefined | Undefined |
| | | Competitive aggressiveness; –.00 | .15 (–.13, .41) | Competitive aggressiveness; –.00 | 0 |
| | | Innovativeness; –.01 | .51 (.27, .69) | Innovativeness; –.01 | 0 |
| | | Proactiveness; –.01 | .00 (–.28, .28) | Proactiveness; Undefined | Undefined |
| | | Risk taking; –.00 | .05 (–.23, .32) | Risk taking; –.00 | 0 |
| Mousa, Wales, and Harper (2015) | Capital raised (DV) | Entrepreneurial orientation; –.26 | .51 (.27, .69) | Entrepreneurial orientation; –.36 | 91 |
| Titus, House, and Covin (2017) | Acquisition use (DV) | Exploration; .16 | .08 (–.28, .42) | Exploration; .56 | 1,125 |
| Ferreira, Raisch, and Klarner (2014) | CEO tenure (IV) | Ambidexterity; –.02 | –.07 (–.40, .29) | Ambidexterity; Undefined | Undefined |
| Luger (2014) | Firm performance (DV) | Ambidexterity; .14 | –.07 (–.40, .29) | Ambidexterity; Undefined | Undefined |
| Zachary, McKenny, Short, and Payne (2011) | Family business status (IV) | Competitor orientation; –.15 | .11 (–.17, .38) | Competitor orientation; –.45 | 800 |
| | | Customer orientation; –.05 | .19 (–.09, .45) | Customer orientation; –.11 | 384 |
| | | Interfunctional coordination; –.07 | .09 (–.19, .36) | Interfunctional coordination; –.23 | 980 |
| | | Long-term focus; –.10 | .18 (–.10, .44) | Long-term focus; –.24 | 476 |
| | | Profitability; –.13 | .17 (–.11, .43) | Profitability; –.32 | 506 |
| | Firm performance (DV) | Market orientation; .10 | .06 (–.22, .33) | Market orientation; .41 | 1,581 |

*Note:* CATA = computer-aided text analysis; IV = independent variable (i.e., antecedent); DV = dependent variable (i.e., consequent); CES = coefficient of equivalence and stability.

[a]CES confidence intervals (CIs) calculated using the Fisher *r* to *z* transformation procedure described by Shen and Lu (2006).

[b]Percent of underestimation in variance explained in substantive relations is calculated as the percent decrease in observed coefficients of determination (i.e., $r^2$) comparing corrected and uncorrected coefficients.

using the Spearman-Brown prophecy formula to predict the reliability of the CE if more texts had been measured. For instance, the Spearman-Brown prophecy formula suggests that if we had manually coded the market orientation for all 224 firms instead of just the 50 randomly selected, the predicted reliability would be an average of .83 rather than .53.

*Refined CATA measures*. Our results regarding the low parallel forms reliability estimates suggest inconsistency between the manual and software analyses. One potential source of inconsistency may be that the raters were either overly inclusive or exclusive in evaluating words for inclusion. For example, Uotila and colleagues (2009) developed the ambidexterity dictionary using only eight word stems. Moss and colleagues (2014) identified several words that seem relevant to exploration (e.g., inventions, pioneer) and exploitation (e.g., marketing, optimization) in many contexts but were not included in the original measures. Another potential source of error arises from differences in context. The original ambidexterity word list was developed in a sample of manufacturing firms (Uotila et al., 2009). However, Moss and colleagues relied on technology industries and identified the words "scientist" and "laboratories" as relevant to exploration in this new context.

To address specific factor error, Moss and colleagues (2014) generated a list of every word used three or more times in any MD&A statement. This word list was evaluated to identify words missing from the measures that would indicate exploration and exploitation in technology ventures and supplemented the measures. By updating the list with other words indicative of exploration and exploitation, we estimated that parallel forms reliability increased from .09 to .87 for exploration and from .30 to .59 for exploitation.

While supplementing the lists with additional words drawn from the narratives is valuable, further refinements can be made through the qualitative comparison of the results of the manual and software analyses. To refine the CATA measures, we used NVivo 11 to identify all occurrences of the dictionary words alongside the manual coding and identified occurrences where the manual and software coding did not match. When a word was frequently used out of context, we removed it from the CATA measure. In some cases, the manual content analysis identified additional words to add to the dictionary. After each change to the measure, we repeated this procedure until no further changes to the measure could be made.

Once the revised measures were finalized, we recalculated all reliability estimates. As expected, the coefficients of equivalence of the refined measures demonstrated considerable improvement compared to the original measures. This indicates that the new word lists now mirror the coding of a human coder more closely. The coefficients of equivalence and stability for the ambidexterity measures also improved significantly. However, because the coefficients of stability remained low for the entrepreneurial and market orientation measures, there were relatively small changes to their coefficients of equivalence and stability. As an additional contribution of our study, we make the revised measures for entrepreneurial orientation, market orientation, and ambidexterity and their reliability estimates available in Appendices C through F in the online supplemental material.[1]

## Discussion

CATA has been proposed as a novel measurement approach with the potential to lead to important theory advancements. However, because measure reliability is necessary for the advancement of research using this technique, this is an issue that needs attention before CATA can be recommended and adopted more broadly.

We first identified and described three sources of error that are particularly relevant for measures developed using CATA. First, transient error arises from phenomena that influence the word choice of the author at the time of writing. These factors may include the emotional

state of the author, the business climate, or the state of the economy at the time a text is written. To assess transient error, we suggest that future research include texts from two points in time and calculate a test-retest reliability estimate. Specific factor error influences CATA measures because of the potentially idiosyncratic choices made by researchers in the process of creating word lists. To assess specific factor error, we recommend that future research manually code at least 10% of the texts being analyzed and then calculate a parallel forms reliability estimate. Algorithm error is relevant for CATA because different software packages use different algorithms for identifying words and determining when a match to a word in the CATA word list is found. To assess algorithm error, we recommend that future CATA studies include the Krippendorff's alpha coefficient from analyzing the same texts using two or more software tools.

Implementing our proposed measurement error variance estimation procedures provided evidence regarding the differential impact of the sources of error we identified. While the three CATA packages evaluated here follow very similar algorithms, differences in each package introduced a small amount of error. In general, this error can be avoided by selecting the appropriate package for the measure being used and considering the use of word stems, numerals, and phrases included in the different CATA packages. Accordingly, finding nontrivial levels of algorithm error would not necessarily suggest that the measure needs to be refined or abandoned because algorithm error may be driven by idiosyncratic features of each package. However, it would suggest a deeper examination of why this error is high. Then, on the basis of this examination, an assessment may be made regarding whether measure refinement or a different CATA package is necessary.

Our results provided evidence that transient and specific factor errors were two key sources of measurement error variance that should be addressed explicitly in future CATA research. Transient error accounted for 16% to 68% of variance in observed scores, whereas specific factor error accounted for about 19% to 91% of variance. However, refining the CATA measures made considerable improvements to specific factor error. After refining the measures, we found that specific factor error accounted only for about 4% to 17% of variance in observed scores.

The reliability estimates for test-retest and parallel forms reported in Table 2 and in online Appendices A and B do not seem to meet the usual .80 benchmark (Nunnally & Bernstein, 1994). However, these results do not suggest that the reliability of CATA measures is lower than that of measures created using more traditional approaches. The reason is that our measurement assessment procedures do not rely on the typical and limited internal consistency reliability estimate (i.e., alpha), which is known to be the upper case for reliability and an overly optimistic estimate (Cho & Kim, 2015). In fact, estimates of reliability for other types of measures based on indexes other than alpha are in line with those we found in our study for CATA measures. For example, the average interrater reliability estimate for ratings of job performance is .52 (Viswesvaran, Ones, & Schmidt, 1996). So, future research could compare the CES reliability estimates we obtained for entrepreneurial orientation, market orientation, and ambidexterity with those for other CATA measures as well as measures using traditional measurement approaches. This type of research would lead to a more thorough and comprehensive understanding of how measurement error affects substantive conclusions compared to a sole focus on a reliability estimate (i.e., Cronbach's alpha) known to underestimate error (Le, Schmidt, & Putka, 2007; Schmidt et al., 2003). Next, we describe additional

implications of our results for theory, the reinterpretation of past research, and the conduct of future research.

### *Best-Practice Recommendations for Improving the Accuracy of CATA Measurement and Research*

Our results provide evidence that reliability concerns should be considered explicitly in the interpretation of past and future research using CATA because measurement error variance is substantial. As shown in Table 2 and online Appendices A and B, transient error and specific factor error accounted for the vast majority of measurement error variance. These results suggest that substantive relations reported in past research have been underestimated. Future research could use our proposed measurement error variance assessment procedures to understand the extent to which previous null findings may be due to unacknowledged measurement error variance. In addition, next we offer best-practice recommendations for improving the accuracy of CATA measurement and research. As a preview, a summary of these recommendations is included in Table 4.

We emphasize that although our proposed procedures allow for an assessment of the presence of measurement error variance in published research, it is always preferable to anticipate and attempt to mitigate measurement error prior to data collection (Aguinis & Vandenberg, 2014; Hunter & Schmidt, 2004). Our test-retest and parallel forms reliability estimates are far from a perfect 1.00. Accordingly, researchers using CATA can consider the following actions to anticipate and minimize the effects of measurement error variance.

*Transient error*. There are two key challenges that may drive transient error in CATA research. The first is where CATA scores demonstrate variability but the language used in the texts is otherwise consistent over time. This may occur when there is a change in the construct being measured. It is not possible to parse out variability due to random error from substantive changes. Accordingly, a valuable assessment to be made is to refer to theory to identify whether the construct is likely to be stable over the assessment period. Statelike constructs, such as optimism, affect, and mood, change more frequently than traitlike constructs, such as strategic orientation, personality, and values. If theory suggests that variability is likely, researchers should attempt to decrease the lag between the sampled texts. This can be accomplished by either collecting the texts more frequently or, if the texts are available only at long intervals (e.g., annual reports), collecting texts that are produced more frequently (e.g., quarterly reports).

Transient error may also be influenced by variation in the construct from exogenous shocks. For instance, when economic conditions change during the time between the creations of texts, the outlook of the author may be affected, increasing the likelihood of changes in otherwise stable constructs. Exogenous shocks, such as the introduction of a lawsuit against the company, the change of management teams/directors, or shareholder activism, may also drive changes in the contents of organizational narratives over time and deflate test-retest reliability estimates. Accordingly, when transient error is significant, CATA users should examine whether any shocks occurred during the assessment period that may have influenced the focal construct.

**Table 4**
**Best-Practice Recommendations for Improving the Accuracy of Computer-Aided Text Analysis (CATA) Measurement and Research**

| Source of Measurement Error Variance | Challenges | Best-Practice Recommendations |
|---|---|---|
| Transient error | CATA scores demonstrate variability over time. | Identify whether the construct is theorized to be stable over the assessment period.<br>Decrease the lag between collected texts.<br>Collect texts that are produced more frequently.<br>Investigate the possibility of shocks within the sampling frame. |
| | The language used in texts varies significantly over time. | Identify whether texts with more standardized contents are available.<br>Identify whether the sampled texts are likely to be influenced by managerial attention rather than the salience of the construct being measured.<br>For individual-level texts, confirm the identity of the author and whether the author changed between texts. |
| Specific factor error | Word lists are either too inclusive or exclusive, resulting in words being used out of context or being missed, respectively. | Iteratively compare the words identified by CATA and manual content analyses and refine the measure to improve alignment.<br>Eliminate word stems and replace them with only the conjugations that fit the construct definition.<br>Eliminate single words that are commonly used out of context and replace them with common short phrases.<br>Identify omitted conjugations of words on the word list that are relevant to the construct.<br>Generate a list of words used in your sampled texts and have judges evaluate whether words should be added to the measure. |
| | The measure was developed in a different context. | Iteratively compare the words identified by CATA and manual content analyses and refine the measure to improve alignment.<br>Generate a deductive list of words thought to indicate the construct in the new context and have judges evaluate whether they should be included in the revised list.<br>Generate a list of words used in your sampled texts and have judges evaluate whether the words reflect the construct in the new context. |
| Algorithm error | Two CATA software packages provide inconsistent scores. | Identify whether the measure uses features idiosyncratic to one package.<br>Select a third package for comparison to both original packages.<br>Recreate a CATA analysis using manual coding and compare results to both packages. |

Variability in the language used in the sampled texts over time is a second key challenge associated with transient error. In our investigation of three different constructs, the two samples using CEO shareholder letters had relatively high transient error: 49% of observed variance for entrepreneurial orientation and 47% of observed variance for market orientation. However, when ambidexterity was measured in MD&A statements, transient error accounted for only 20% of observed variance. There were two likely reasons why the MD&A statements outperformed shareholder letters with respect to transient error. First, the contents of the MD&A statements were more standardized, suggesting that if the underlying construct stayed the same over the assessment period, it was discussed approximately the same amount in both texts. Second, shareholder letters are considerably shorter and feature the voice of the

CEO heavily, suggesting that the contents of these documents may be more subject to the attention of the CEO at the time of the text's creation. For instance, if innovation has been a priority for a firm for many years and this priority has not changed, the CEO may not emphasize this as much as other more timely strategic initiatives since he or she has limited space for communicating with shareholders.

The language used in texts can also be influenced by the authors of these texts. While frequently attributed to the CEO, shareholder letters are often produced by multiple individuals (Barr et al., 1992). The contributions of these individuals embed multiple perspectives of the company into the text, reducing the likelihood that idiosyncrasies of an individual contributor will bias construct measurement. However, having multiple contributors to a text increases the likelihood of measurement error when measuring individual-level constructs. For instance, the use of positive and negative language in a text is commonly used to measure the affective state of the text's author (e.g., Savani & King, 2015). When a text has multiple authors, the presence of positive and negative language cannot be attributed to one individual. Accordingly, when measuring individual-level constructs, researchers should provide reasonable evidence that the only contributor to the text was the individual for whom the measurement is being made and that the author did not change.

*Specific factor error*. Parallel forms reliability estimates ranged from .09 to .81. This wide range of errors arose from two sources. First, the CATA measures included words and phrases that were frequently used out of context and omitted words and phrases that were consistently used to indicate the construct being measured. We addressed this through iteratively removing words from the CATA measures that were frequently used out of context, adding words from the text that were consistently used in context, and recalculating parallel forms reliability estimates. This intervention of a human coder in CATA helped alleviate the threat of specific factor error introduced by the technique's inability to consider the context in which words are used.

Within the iterative word list refinement process, there are several key activities that may help identify words to add or remove. To reduce the frequency of counting words that are used out of context, researchers can eliminate word stems and add the conjugated words that are appropriate for the construct being measured. Illustrating how word stems can cause specific factor error, the original exploitation dictionary used "refine*" to capture how firms make small adjustments to improve existing products and processes. However, the words "refinery" and "refineries" were never used in this context within our sample of texts. Replacing single words that have many meanings with short phrases that have more targeted meanings also decreases the number of words that are counted but are out of context. For example, the original innovativeness measure included the word "new." While "new" frequently signals innovation in shareholder letters, it is also commonly used to communicate phenomena not related to innovation (e.g., new regulations, New York). This suggests that replacing "new" with a number of short phrases such as "new product" and "new technology" might provide a more reliable measurement of innovativeness. To capture omitted words, researchers should identify whether the measures include all conjugations of the words used in the measure and add any omitted conjugations that are indicative of the construct. Researchers should also consider replicating the inductive word list development process advocated by Short and colleagues (2010) to capture other relevant words used in the sample of narratives that reflect the construct being measured.

A second challenge regarding specific factor error in CATA measures relates to the context in which the measure was developed. Different industries, texts, and individuals may use the same words to mean different things or use different words to mean the same thing. For example, the language used in CEO shareholder letters is likely different from the language used in Twitter tweets. Accordingly, measures developed for one context may need to be refined to be reliable in other contexts. An efficient way to refine the measure is to use Short and colleagues' (2010) two-phase CATA measure development process. The first phase calls for words to be identified deductively from theory and existing measures (Short et al., 2010). Researchers should ensure that the words and phrases from existing CATA measures are included. The second phase calls for words to be identified inductively from the sample of texts (Short et al., 2010). At the end of each phase, judges evaluate whether these words are indicative of the construct in the new context. This two-phase process provides an initial refinement of the CATA measure. Further refinement can be accomplished by again iteratively removing words from the CATA measure that are frequently used out of context, adding words from the text that are consistently used in context, and recalculating parallel forms reliability estimates.

*Algorithm error*. Algorithm error is driven by software design choices made by the developers of the CATA software, including the features and limitations of each package. For instance, CAT Scanner can handle phrases with spaces in them but DICTION 5.0 cannot. As a result, CATA measures that include phrases may produce different data across the two packages. Accordingly, researchers should identify whether their CATA measure uses features of one package that are not supported by the other. Other design choices made by the software developers include the handling of punctuation and how words are defined. For example, should hyphenated words be treated as one word or two? These decisions are frequently less apparent than identifying whether features are consistent across packages. However, the impact of these algorithmic discrepancies can still be estimated by selecting a third CATA software package and triangulating the results. Alternatively, the researcher could recreate a CATA analysis by manually coding texts for the words included in the CATA measure. A comparison of these results with the CATA analysis may provide insight into the algorithmic differences among the packages.

## *Implications for Practice*

We use the "customer-centric" approach of converting effect sizes into metrics that can be understood by nonacademic audiences to demonstrate how measurement error influenced the relation between entrepreneurial orientation and shareholder value in a practically relevant way (cf. Aguinis, Werner, Abbott, Angert, Park, & Kohlhausen, 2010). Engelen, Neumann, and Schmidt (2016) reported a correlation of .24, suggesting that entrepreneurial orientation explains 5.76% (i.e., .24 * .24) of variance in shareholder value. However, our procedure indicates that this correlation is actually .34, suggesting that the percentage explained is actually 11.56% (i.e., .34 * .34).

Using a correlation coefficient metric, the difference between uncorrected and corrected measures seems relatively small: "only" .10 correlation points or 5.8% in variance explained. However, we reach a very different conclusion if we use Tobin's Q, the Engelen et al. (2016)

measure of shareholder value. Tobin's Q is calculated as the market value of the firm divided by the replacement value of its assets. A score higher than 1 means that the market expects the firm to outperform its norm, whereas a score below 1 means that a firm is expected to underperform its norm. We can convert the correlation coefficient $r$ to a regression coefficient $b$ for a single predictor regression using the equation $b = r(Sy/Sx)$, where Sy is the standard deviation for shareholder value and Sx is the standard deviation for entrepreneurial orientation. Engelen and colleagues. reported values of Sy and Sx values of .98 and .40, respectively. So, the regression coefficient associated with the observed correlation is .24(.98/.40) = .59, whereas the regression coefficient is actually .34(.98/.40) = .83. In other words, while Engelen and colleagues' study suggested that a 1-point increase in entrepreneurial orientation is associated with a .59 increase in Tobin's Q, it is actually associated with a .83 increase—a difference of .24 points.

Although it is a financial figure that investors understand, a difference of .24 points in Tobin's Q may not be necessarily intuitive or meaningful to a broader audience. So, as an additional way to understand the practical significance of this result, we consider the impact this difference would have on a BusinessWeek 1000 firm with assets of approximately $10 billion (Anderson, Fornell, & Mazvancheryl, 2004). For such a firm, an increase in Tobin's Q of .24 points implies an increase in the firm's market value of approximately $2.4 billion. In short, an assessment of measurement error in CATA measures has important practical implications in terms of our understanding of the relation between entrepreneurial orientation and shareholder value—in the order of billions of dollars.

### *Limitations of CATA*

Measurement using CATA is a potentially valuable alternative to survey research and manual content analysis; however, it is important to consider its limitations, and we do so following best-practice recommendations offered by Brutus, Aguinis, and Wassmer (2013). For instance, CATA is most useful in measuring constructs where single words or short phrases provide evidence of the construct. For instance, the word "creativity" on its own provides an indication of a firm's innovativeness. However, constructs such as exemplification from the impression management literature (e.g., Bolino & Turnley, 1999) may be difficult to operationalize without incorporating the context in which the words are used into the coding. Accordingly, these constructs are better measured using manual content analysis where the context can be readily incorporated into the coding.

Second, CATA is sensitive to impression management (McKenny et al., 2013). For instance, while shareholder letters are used to present information about the company to shareholders, they are also used to shape the impressions of the reader favorably (Barr & Huff, 1997; Staw, McKechnie, & Puffer, 1983). Accordingly, phenomena such as counterproductive workplace behaviors (e.g., Robinson & Bennett, 1995) where impression management is likely to be present may be more accurately measured by questionnaires or interviews than CATA using publicly available texts.

Third, the collection of a sample of texts is central to CATA (Short et al., 2010). One large sample of texts can be used with multiple constructs to publish several studies (e.g., Short et al., 2009; Short et al., 2010; Zachary et al., 2011). While valuable, the time it takes to identify and collect a large sample of texts is considerable. As a result, researchers may be

motivated to keep their texts private. Nevertheless, making data used in a published study available upon request is an important part of scholarly transparency and is a requirement of several management journals (Banks et al., 2016). For content analysis researchers, this suggests that the texts used should be made available. Unfortunately, we contacted authors of CATA studies and found that several were unable or unwilling to share their texts. While their hesitance to share texts is understandable, it is also a problem because replication and verification of previous findings is central to scientific progress. In light of a number of recent article retractions from management journals, providing access to data used in published manuscripts can also help alleviate the concern about the use of questionable research practices (cf. Banks et al., 2016).

To facilitate the sharing of texts while also protecting authors' investments in data collection, we propose two guidelines based on the American Psychological Association's code regarding the sharing of research data (Ethics Code Standard 8.14a): (1) All texts used in a published manuscript using content analysis should be retained and shared upon request for verification purposes unless authors are ethically or legally unable to do so; and (2) The recipient of a shared sample of texts should use the texts only to verify the analyses conducted by the authors unless given permission to conduct additional analyses by the text owner. By following these guidelines, authors' investments in collecting texts for use in future publications are protected while enabling other scholars to verify the conclusions reached in published research using CATA.

Finally, while we advocate for greater attention to measurement precision in CATA research, there may be times where our procedure may not be possible or practical given the research design. For instance, some organizational texts are produced only once during the phenomenon of interest (e.g., crowdfunding campaigns, IPO prospectuses). In these cases, calculating transient error is impractical given the text used. Similarly, some studies that use CATA seek to measure the change in a construct over time (e.g., Allison et al., 2014). In these studies, the variation in the underlying phenomenon and transient error are inseparable. Accordingly, in applying these procedures, we recommend that researchers treat the goals of the research and measurement precision as a trade-off to be balanced.

## Conclusion

CATA offers a novel measurement approach given the known limitations of self-report and archival methods. Accordingly, CATA is becoming a popular measurement approach in management and many other fields. Our article provided evidence that although measurement error variance has not been measured thus far, it does exist. We illustrated this finding with the entrepreneurial orientation, market orientation, and ambidexterity constructs. Our results indicate that existing research using CATA measures may need to be revisited because substantive relations have been underestimated. This underestimation has the potential to derail theory advancements and lead to misguided practices. We offered recommendations on how future research can minimize the effects of transient, specific factor, and algorithm error and demonstrated the significant difference these recommendations can make in terms of the quality of the resulting measures. Overall, we hope that our article will serve as a catalyst for improvements in the use and evaluation of CATA measures in the future and that such improvements will help CATA reach its potential in facilitating theory advancements and useful practical applications.

## Note

1. The CAT Scanner formatted dictionaries are available at http://www.catscanner.net/.

## References


Aguinis, H. 1995. Statistical power problems with moderated multiple regression in management research. *Journal of Management*, 21: 1141-1158.

Aguinis, H., & Edwards, J. R. 2014. Methodological wishes for the next decade and how to make wishes come true. *Journal of Management Studies*, 51: 143-174.

Aguinis, H., Forcum, L. E., & Joo, H. 2013. Using market basket analysis in management research. *Journal of Management*, 39: 1799-1824.

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. 2013. Best-practice recommendations for estimating cross-level interaction effects using multi-level modeling. *Journal of Management*, 39: 1490-1528.

Aguinis, H., & Pierce, C. A. 1998. Testing moderator variable hypotheses meta-analytically. *Journal of Management*, 24: 577-592.

Aguinis, H., & Vandenberg, R. J. 2014. An ounce of prevention is worth a pound of cure: Improving research quality before data collection. *Annual Review of Organizational Psychology and Organizational Behavior*, 1: 569-595.

Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H., & Kohlhausen, D. 2010. Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods*, 13: 515-539.

Allison, T. H., McKenny, A. F., & Short, J. C. 2014. Integrating time into family business research using random coefficient modeling to examine temporal influences on family firm ambidexterity. *Family Business Review*, 27: 20-34.

Anderson, E. W., Fornell, C., & Mazvancheryl, S. K. 2004. Customer satisfaction and shareholder value. *Journal of Marketing*, 68: 172-185.

Banks, G. C., O'Boyle, E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., Abston, K. A., Bennett, A. A., & Adkins, C. L. 2016. Questions about questionable research practices in the field of management: A guest commentary. *Journal of Management*, 42: 5-20.

Barr, P. S., & Huff, A. S. 1997. Seeing isn't believing: Understanding diversity in the timing of strategic response. *Journal of Management Studies*, 34: 337-370.

Barr, P. S., Stimpert, J. L., & Huff, A. S. 1992. Cognitive change, strategic action, and organizational renewal. *Strategic Management Journal*, 13: 15-36.

Becker, G. 2000. How important is transient error in estimating reliability? Going beyond simulation studies. *Psychological Methods*, 5: 370-379.

Boling, J. R., Pieper, T. M., & Covin, J. G. in press. CEO tenure and entrepreneurial orientation within family and nonfamily firms. *Entrepreneurship Theory and Practice*. doi:10.1111/etap.12150

Bolino, M. C., & Turnley, W. H. 1999. Measuring impression management in organizations: A scale development based on the Jones and Pittman taxonomy. *Organizational Research Methods*, 2: 187-206.

Brett, J. M., Olekalns, M., Friedman, R., Goates, N., Anderson, C., & Lisco, C. C. 2007. Sticks and stones: Language, face, and online dispute resolution. *Academy of Management Journal*, 50: 85-99.

Brutus, S., Aguinis, H., & Wassmer, U. 2013. Self-reported limitations and future directions in scholarly reports: Analysis and recommendations. *Journal of Management*, 39: 48-75.

Cho, E., & Kim, S. 2015. Cronbach's coefficient alpha: Well known but poorly understood. *Organizational Research Methods*, 18: 207-230.

Clarkson, P. M., Kao, J. L., & Richardson, G. D. 1999. Evidence that management discussion and analysis (MD&A) is a part of a firm's overall disclosure package. *Contemporary Accounting Research*, 16: 111-134.

Cronbach, L. J. 1947. Test reliability: Its meaning and determination. *Psychometrika*, 12: 1-16.

Dalton, D. R., & Aguinis, H. 2013. Measurement malaise in strategic management studies: The case of corporate governance research. *Organizational Research Methods*, 16: 88-99.

Duriau, V. J., Reger, R. K., & Pfarrer, M. D. 2007. A content analysis of the content analysis literature in organizational studies: Research themes, data sources, and methodological refinements. *Organizational Research Methods*, 10: 5-34.

Engelen, A., Neumann, C., & Schmidt, S. 2016. Should entrepreneurially oriented firms have narcissistic CEOs? *Journal of Management*, 42: 498-721.

Engelen, A., Neumann, C., & Schwens, S. 2015. "Of course I can": The effect of CEO overconfidence on entrepreneurially oriented firms. *Entrepreneurship Theory and Practice*, 39: 1137-1160.

Ferreira, P., Raisch, S., & Klarner, P. 2014. *Staying agile in the saddle: CEO tenure, TMT change, and organizational ambidexterity*. Paper presented at the annual meeting of the Academy of Management, Philadelphia.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. 2012. *Various coefficients of interrater reliability and agreement*. http://www.cran.r-project.org/web/packages/irr/irr.pdf. Accessed July 29, 2015.

Gangloff, K. A., Connelly, B. L., & Shook, C. L. 2016. Of scapegoats and signals: Investor reactions to CEO succession in the aftermath of wrongdoing. *Journal of Management*, 42: 1614-1634.

Geiger, M. A., & Taylor, P. L., III. 2003. CEO and CFO certifications of financial information. *Accounting Horizons*, 17: 357-368.

Goodman, R. 1980. Annual reports serving a dual marketing function—Report as survey. *Public Relations Quarterly*, 36: 21-24.

Hart, R. P. 2000. *DICTION 5.0: The text analysis program*. Thousand Oaks, CA: Sage-Scolari.

Hayes, A. F., & Krippendorff, K. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1: 77-89. (Software available at http://www.afhayes.com/spss-sas-and-mplus-macros-and-code.html)

Hunter, J. E., & Schmidt, F. L. 2004. *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). New York: Academic Press.

Kerlinger, F. N., & Lee, H. B. 2000. *Foundations of behavioral research* (4th ed.). Orlando: Harcourt.

Klein, D. 2014. KALPHA: Stata module to compute Krippendorff's alpha-reliability. *Statistical Software Components*. (Software available at https://ideas.repec.org/c/boc/bocode/s457862.html)

Kohli, A. K., & Jaworski, B. J. 1990. Market orientation: The construct, research propositions, and managerial implications. *Journal of Marketing*, 54: 1-18.

Krippendorff, K. 2013. *Content analysis: An introduction to its methodology* (3rd ed.). Thousand Oaks, CA: Sage.

Le, H., Schmidt, F. L., & Putka, D. J. 2007. The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships. *Organizational Research Methods*, 12: 165-200.

Lord, F. M., & Novick, M. R. 1968. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Luger, J. 2014. *A longitudinal perspective on organizational ambidexterity*. Unpublished doctoral dissertation, University of St. Gallen, Switzerland.

Lumpkin, G. T., & Dess, G. G. 1996. Clarifying the entrepreneurial orientation construct and linking it to performance. *Academy of Management Review*, 21: 135-172.

March, J. 1991. Exploration and exploitation in organizational learning. *Organization Science*, 2: 71-87.

McKenny, A. F., Short, J. C., & Newman, S. M. 2012. CAT Scanner (Version 1.0) [Computer software]. http://www.catscanner.net/

McKenny, A. F., Short, J. C., & Payne, G. T. 2013. Using computer-aided text analysis to elevate constructs: An illustration using psychological capital. *Organizational Research Methods*, 16: 152-184.

McKenny, A. F., Short, J. C., Payne, G. T., & Zachary, M. A. 2012. Assessing espoused performance goals in private family firms using content analysis. *Family Business Review*, 25: 298-317.

Moss, T. W., Neubaum, D. O., & Meyskens, M. 2015. The effect of virtuous and entrepreneurial orientations on microfinance lending and repayment: A signaling theory perspective. *Entrepreneurship Theory and Practice*, 39: 27-52.

Moss, T. W., Payne, G. T., & Moore, C. B. 2014. Strategic consistency of exploration and exploitation in family businesses. *Family Business Review*, 27: 51-71.

Mousa, F.-T., Wales, W. J., & Harper, S. R. 2015. When less is more: EO's influence upon funds raised by young technology firms at IPO. *Journal of Business Research*, 68: 306-313.

Narver, J. C., & Slater, S. F. 1990. The effect of a market orientation on business profitability. *Journal of Marketing*, 54: 20-35.

Nunnally, J. C., & Bernstein, I. H. 1994. *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

O'Boyle, E. H., Banks, G. C., & Gonzalez-Mulé, E. 2017. The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*, 43: 376-399.

Payne, G. T., Moore, C. B., Bell, R. G., & Zachary, M. A. 2013. Signaling organizational virtue: An examination of virtue rhetoric, country-level corruption, and performance of foreign IPOs from emerging and developed economies. *Strategic Entrepreneurship Journal*, 7: 230-251.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. 2007. *LIWC2007: Linguistic inquiry and word count*. Austin: LIWC.net.

Pfarrer, M. D., Pollock, T. G., & Rindova, V. P. 2010. A tale of two assets: The effects of firm reputation and celebrity on earnings surprises and investors' reactions. *Academy of Management Journal*, 53: 1131-1152.

Raisch, S., & Birkinshaw, J. 2008. Organizational ambidexterity: Antecedents, outcomes, and moderators. *Journal of Management*, 34: 375-409.

Ree, M. J., & Carretta, T. R. 2006. The role of measurement error in familiar statistics. *Organizational Research Methods*, 9: 99-112.

Robinson, S. L., & Bennett, R. J. 1995. A typology of deviant workplace behaviors: A multidimensional scaling study. *Academy of Management Journal*, 38: 555-572.

Savani, K., & King, D. 2015. Perceiving outcomes as determined by external forces: The role of event construal in attenuating the outcome bias. *Organizational Behavior and Human Decision Processes*, 130: 136-146.

Schmidt, F. L., & Hunter, J. E. 1996. Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods*, 1: 199-223.

Schmidt, F. L., & Hunter, J. E. 1999. Theory testing and measurement error. *Intelligence*, 27: 183-198.

Schmidt, F. L., Le, H., & Ilies, R. 2003. Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual difference constructs. *Psychological Methods*, 8: 206-224.

Shadish, W. R., Cook, T. D., & Campbell, D. T. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth.

Shen, D., & Lu, Z. 2006. *Computation of correlation coefficient and its confidence interval in SAS*. Proceedings of the Thirty-First Annual SAS Users Group International Conference, San Francisco. http://www2.sas.com/proceedings/sugi31/170-31.pdf. Accessed January 17, 2016.

Short, J. C., Broberg, J. C., Cogliser, C. C., & Brigham, K. H. 2010. Construct validation using computer-aided text analysis (CATA): An illustration using entrepreneurial orientation. *Organizational Research Methods*, 13: 320-347.

Short, J. C., & Palmer, T. B. 2008. The application of DICTION to content analysis research in strategic management. *Organizational Research Methods*, 11: 727-752.

Short, J. C., Payne, G. T., Brigham, K. H., Lumpkin, G. T., & Broberg, J. C. 2009. Family firms and entrepreneurial orientation in publicly traded firms: A comparative analysis of the S&P 500. *Family Business Review*, 22: 9-24.

Staw, B. M., McKechnie, P. I., & Puffer, S. M. 1983. The justification of organizational performance. *Administrative Science Quarterly*, 28: 582-600.

Sturman, M. C., Cheramie, R. A., & Cashen, L. H. 2005. The impact of job complexity and performance measurement on the temporal consistency, stability, and test-retest reliability of employee job performance ratings. *Journal of Applied Psychology*, 90: 269-283.

Titus, V., House, J. M., & Covin, J. G. 2017. The influence of exploration on external corporate venturing activity. *Journal of Management*, 43: 1609-1630.

Uotila, J., Maula, M., Keil, T., & Zahra, S. A. 2009. Exploration, exploitation, and financial performance: Analysis of S&P 500 corporations. *Strategic Management Journal*, 30: 221-231.

Viswesvaran, C., Ones, D. S., & Schmidt, F. L. 1996. Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology*, 81: 557-574.

Wimmer, R., & Dominick, J. 2013. *Mass media research* (10th ed.). Boston: Wadsworth Cengage Learning.

Zachary, M. A., McKenny, A. F., Short, J. C., & Payne, G. T. 2011. Family business and market orientation: Construct validation and comparative analysis. *Family Business Review*, 24: 233-251.

Zyphur, M. J., Oswald, F. L., & Rupp, D. E. 2015. Rendezvous overdue: Bayes analysis meets organizational research. *Journal of Management*, 41: 387-389.