Computer-Aided Text Analysis, or: How I Learned to Stop Worrying and Love the Computer

Aaron F. McKenny
Why I love the computer
### Why I love the computer

**Table 2**

*Table from Cohen, 1992*

<table>
<thead>
<tr>
<th>Test</th>
<th>.01</th>
<th>.05</th>
<th>.10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sm</td>
<td>Med</td>
<td>Lg</td>
</tr>
<tr>
<td>1. Mean dif</td>
<td>586</td>
<td>95</td>
<td>38</td>
</tr>
<tr>
<td>2. Sig r</td>
<td>1,163</td>
<td>125</td>
<td>41</td>
</tr>
<tr>
<td>3. r dif</td>
<td>2,339</td>
<td>263</td>
<td>96</td>
</tr>
<tr>
<td>4. P = .5</td>
<td>1,165</td>
<td>127</td>
<td>44</td>
</tr>
<tr>
<td>5. P dif</td>
<td>584</td>
<td>93</td>
<td>36</td>
</tr>
<tr>
<td>6. $x^2$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$1df$</td>
<td>1,168</td>
<td>130</td>
<td>38</td>
</tr>
<tr>
<td>$2df$</td>
<td>1,388</td>
<td>154</td>
<td>56</td>
</tr>
<tr>
<td>$3df$</td>
<td>1,546</td>
<td>172</td>
<td>62</td>
</tr>
<tr>
<td>$4df$</td>
<td>1,675</td>
<td>186</td>
<td>67</td>
</tr>
<tr>
<td>$5df$</td>
<td>1,717</td>
<td>200</td>
<td>71</td>
</tr>
<tr>
<td>$6df$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$7df$</td>
<td>1,717</td>
<td>200</td>
<td>71</td>
</tr>
<tr>
<td>7. ANOVA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$2g$</td>
<td>386</td>
<td>93</td>
<td>38</td>
</tr>
<tr>
<td>$3g$</td>
<td>464</td>
<td>76</td>
<td>30</td>
</tr>
<tr>
<td>$4g$</td>
<td>388</td>
<td>63</td>
<td>25</td>
</tr>
<tr>
<td>$5g$</td>
<td>336</td>
<td>55</td>
<td>22</td>
</tr>
<tr>
<td>$6g$</td>
<td>299</td>
<td>49</td>
<td>20</td>
</tr>
<tr>
<td>$7g$</td>
<td>271</td>
<td>44</td>
<td>18</td>
</tr>
<tr>
<td>8. Mult R</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$2k^*$</td>
<td>698</td>
<td>97</td>
<td>45</td>
</tr>
<tr>
<td>$3k^*$</td>
<td>780</td>
<td>108</td>
<td>50</td>
</tr>
<tr>
<td>$4k^*$</td>
<td>841</td>
<td>118</td>
<td>55</td>
</tr>
<tr>
<td>$5k^*$</td>
<td>901</td>
<td>126</td>
<td>59</td>
</tr>
<tr>
<td>$6k^*$</td>
<td>953</td>
<td>134</td>
<td>63</td>
</tr>
<tr>
<td>$7k^*$</td>
<td>998</td>
<td>141</td>
<td>66</td>
</tr>
<tr>
<td>$8k^*$</td>
<td>1,039</td>
<td>147</td>
<td>69</td>
</tr>
</tbody>
</table>

**Note.** ES = population effect size, Sm = small, Med = medium, Lg = large, dif = difference, ANOVA = analysis of variance. Tests numbered as in Table 1.

*a* Number of groups.  
*b* Number of independent variables.
Manual Coding...

25,000-35,000 pages!

Coder Fatigue

Tenure Clock
Computer-Aided Text Analysis

Organizational Ambidexterity
How it works

• Sample Dictionary: Innovativeness

• Sample narrative to analyze:
  • “The creativity of our research and development team make this organization one of the most innovative in the industry, with patents on over 2,300 inventions.”

Innovativeness: 4
What worried me
Origins of My Recommendations

Construct Validation Using Computer-Aided Text Analysis (CATA)

An Illustration Using Entrepreneurial Orientation

Jeremy C. Short
Texas Tech University
J. Christian Broberg
Wichita State University
Claudia C. Cogilser
Keith H. Brigham
Texas Tech University

Construct validity continues to pose challenges in the organizational sciences. To capture difficult-to-measure constructs of interest, researchers have often relied on content analysis. One content analysis technique, computer-aided text analysis (CATA), is particularly attractive because of the ability to process large samples with high speeds and reliabilities. Unfortunately, inconsistent guidance exists to guide researchers through the use of this tool in a manner compatible with accepted methods used to validate constructs in a rigorous manner. The authors review research using content analysis to examine the extent to which such studies integrate methods for assessing content, external, discriminant, and predictive validity. To provide direction for organizational researchers interested in using CATA to measure theoretically based constructs relevant to the management field, they suggest a number of possible procedures to enhance construct validity. They illustrate these procedures using the construct of entrepreneurial orientation.

Keywords: computer-aided text analysis; construct validation; content analysis; DICTRON; entrepreneurial orientation

Validity refers to evaluating inferences drawn from measures of concepts of interest and considers the extent to which a measure accurately represents that focal concept (Cronbach, 1971). The importance of rigor in construct measurement cannot be overstated. Kertlinger and Lee (2000) argued that the ability to demonstrate appropriate measures of theoretical notions of interest through construct validity “is one of the most significant advances of modern measurement theory and practice” (p. 670).

Despite advancements in the organizational sciences vis-à-vis construct validation, reviews of construct validity in the management field have yielded generally bleak comments about the methods used to establish construct validity. One limitation frequently raised is that researchers using content analysis frequently rely on face validity while ignoring the conceptual content, external, discriminant, and predictive validity of their measures. As a result, the researchers are unable to establish construct validity in a meaningful way.

In this article, the authors argue that researchers using content analysis to measure constructs should follow accepted methods for establishing construct validity. The authors provide a framework for researchers using content analysis to establish construct validity. They illustrate this framework using the construct of entrepreneurial orientation.

Keywords: computer-aided text analysis; construct validation; content analysis; DICTRON; entrepreneurial orientation

Using Computer-Aided Text Analysis to Elevate Constructs: An Illustration Using Psychological Capital

Aaron F. McKenney¹, Jeremy C. Short¹, and G. Tyge Payne¹

Abstract
Applying individual-level constructs to higher levels of analysis can be a fruitful practice in organizational research. Although this practice is beneficial in developing and testing theory, there are measurement and validation concerns that, if improperly addressed, may threaten the validity and utility of the research. This article illustrates how computer-aided text analysis might be utilized to facilitate construct elevation while ensuring proper validation. Specifically, we apply a framework to develop organizational-level operationalizations of individual-level constructs using the psychological capital construct as an example.

Keywords: content analysis, level of analysis, multilevel, psychological capital, construct measurement, computer-aided text analysis

Macro organizational research often “borrows” constructs and associated theories from micro disciplines to investigate how they might apply at higher levels of analysis (Whetten, Felin, & King, 2009). There seems to be good reason for this practice as several micro-level theories and constructs have helped explain phenomena occurring at higher levels (Stass, 1993). For example, the concept of “learning,” once reserved for individuals (e.g., Gagné, 1965), has been expanded by organizational theorists to apply to teams (e.g., Brooks, 1994), organizations (e.g., Huber, 1991), networks (e.g., Knight, 2002), and even institutions (e.g., Siebenliuner & Supple, 2005).

¹Price College of Business, University of Oklahoma, Norman, OK, USA
²Robert College of Business, Texas Tech University, Lubbock, TX, USA

References

Corresponding Author: Jeremy C. Short, Division of Management & Entrepreneurship, Price College of Business, 207 W. Brooks, Rm 256, University of Oklahoma, Norman, OK 73019, USA. Email: jeremy.shor@okstate.edu.
Validity

- Content
- External
- Discriminant
- Predictive
Two-step process

- Deductive word list
  - From theory
  - From existing word lists
  - From the synonym finder

- Inductive word list
  - From sampled texts
Measurement Error

- Random Response
  Something can change while writing the text.
  
  *e.g., get a journal rejection in the middle of writing an email*

- Transient
- Specific Factor
- Algorithm
Measurement Error

- Transient

  Something can change between writing of sequential texts.

  *e.g.*, *Stock market tanks*

- Specific Factor
- Algorithm
Measurement Error

- Random Response
- Transient
- Specific Factor

Contents of the word list involves judgment: omission/selection of words

\textit{e.g., Does “risk” always indicate risk taking?}

- Algorithm
Measurement Error

- Random Response
- Transient
- Specific Factor
- Algorithm

Differences in CATA tool algorithms

e.g., Word list includes phrases but your CATA tool cannot process phrases
### Estimating Reliability

<table>
<thead>
<tr>
<th>Measurement Error Source</th>
<th>Reliability Estimate</th>
<th>Calculation Guidelines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transient Error</td>
<td>Test-Retest Reliability</td>
<td>Collect 2 sequential texts and use CATA to analyze them. Calculate the correlation.</td>
</tr>
<tr>
<td>Specific Factor Error</td>
<td>Parallel Forms Reliability</td>
<td>Manually code 10% of sample. Calculate the correlation.</td>
</tr>
<tr>
<td>Algorithm Error</td>
<td>Interrater Agreement</td>
<td>Run CATA using two tools. Calculate Krippendorff’s alpha.</td>
</tr>
</tbody>
</table>
Is there error to worry about?

<table>
<thead>
<tr>
<th>Error Source</th>
<th>Type of Reliability Estimate</th>
<th>Entrepreneurial Orientation Dimension</th>
<th>Reliability Estimate</th>
<th>Percent of Variance Due to Measurement Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transient error</td>
<td>Test-retest</td>
<td>Autonomy</td>
<td>.32</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Competitive aggressiveness</td>
<td>.43</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Innovativeness</td>
<td>.52</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Proactiveness</td>
<td>.55</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Risk taking</td>
<td>.71</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (Test-retest)</td>
<td>.51</td>
<td><strong>49</strong></td>
</tr>
<tr>
<td>Specific factor error</td>
<td>Parallel forms</td>
<td>Autonomy</td>
<td>.29</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Competitive aggressiveness</td>
<td>.51</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Innovativeness</td>
<td>.35</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Proactiveness</td>
<td>.72</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Risk taking</td>
<td>.30</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (Parallel forms)</td>
<td>.43</td>
<td><strong>57</strong></td>
</tr>
<tr>
<td>Algorithm error</td>
<td>Krippendorff’s alpha</td>
<td>Autonomy</td>
<td>.90</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Competitive aggressiveness</td>
<td>.89</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Innovativeness</td>
<td>.89</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Proactiveness</td>
<td>.88</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Risk taking</td>
<td>.90</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mean (Krippendorff’s alpha)</td>
<td>.89</td>
<td><strong>11</strong></td>
</tr>
</tbody>
</table>
Is that significant?

- Engelen, Neumann, & Schmidt (2016)
  - EO → Tobin’s Q: $r = 0.24$
  - Without measurement error: $r = 0.34$
  - Regression coefficient with error: 0.59
  - Regression coefficient without error: 0.83
  - Difference: 0.24

For a Business Week 1000 firm with $10B in assets, the difference is an increase of $2.4B in market value.
How you too can stop worrying: Transient Error

- Is the construct theoretically stable over time?
- Investigate the possibility of shocks
- Decrease lag between texts
- Text selection
  - Frequency
  - Content consistency (managerial attention)
  - Standardization
  - Author (did it change?)
How you too can stop worrying: Specific Factor Error

Refine the dictionaries

- Deductive/Inductive (Short et al., 2010)
- Manual/CATA (McKenny et al., in press)
- Avoid word roots (e.g., Refin*)
- Include all relevant conjugations
- Replace words with phrases (e.g., “risk” vs “risk taking”)
- Ensure that words are relevant to your context
How you too can stop worrying: Algorithm Error

- Identify CATA tool limitations
- Run CATA using a third tool
- Recreate CATA analysis manually
Thank you

“Based on the findings of this presentation, I hope your conclusion is that measurement error is not a deterrent for reasons which at this moment must be all too obvious.”

~Dr. Strangelove (almost)
Resources

• CAT Scanner CATA software (Free)
  • [http://www.catscanner.net/](http://www.catscanner.net/)

• University of Georgia Content Analysis site
  • [http://www.terry.uga.edu/management/contentanalysis](http://www.terry.uga.edu/management/contentanalysis)
  • Particularly Relevant Presentations
    • Custom Dictionaries in Computer-Aided Text Analysis (McKenny, 2012)
    • Measuring and Validating Constructs Using Computer-Aided Text Analysis (Short, 2012)
Resources

  - Supplemental Materials
