An Examination of the Validity, Reliability and Best Practices Related to the Proposed Standards for Traditional Media

Paper Submitted for the Jackson-Sharpe Award

Marianne Eisenmann, MBA
Chandler Chicco Companies

Julie O'Neil, Ph.D.
Texas Christian University

David Geddes, Ph.D.
Geddes Analytics LLC

Abstract

At the 16th Annual IPRRC in March 2013 the researchers presented the first phase of the study *Assessing the Reliability Metrics Proposed as Standards for Traditional Media Analysis*. Researchers received considerable, thoughtful feedback both during and after the Conference that has informed Phase Two of the research – an enhanced, stronger piece of research building on the previous study. The initial study tested the metrics using *inexperienced* coders, trained using a coding guidebook and instructions developed by the researchers and based on the *Proposed Interim Standards for Metrics in Traditional Media Analysis* guidelines. The Phase One study results yielded low to moderate intercoder reliability based on Krippendorff's *alphas*. The most significant feedback on this work was the recommendation to repeat the study using trained coders to see if/how reliability improves with experience.

The research paper addresses efforts to repeat the methodology from Phase One with two pre-tests followed by independent coding of the identical set of 106 articles about Wal-Mart, using three *experienced* coders. Results indicate that coding for the metrics of standards of traditional media analysis is reliable. Ten of the thirteen media items had moderate to high *alphas*, indicating that the three coders were in agreement the majority of time. The paper documents additional best practices and includes updates and improvements to the coding guidebook gleaned from working with the experienced coders.

The standards document evolved to meet the practitioners' need for replicable, comparable and transparent outcomes in media content analysis. The implementation of Phase Two of this study allows for clearer, more definitive conclusions and provides readers with robust guidance and best practices, and a "ready-made" set of tools in the form of a tested and effective coding guide and coding grid, to implement media content analysis with the necessary transparency in methodology and confidence of replication.

An Examination of the Validity, Reliability, and Best Practices Related to the Proposed Standards for Traditional Media

1. Introduction

Analyzing media coverage has been a common public relations practice since the 1930s (Michaelson & Macleod, 2007). Organizations have varied objectives for engaging in traditional media relations, and likewise many reasons for analyzing the media coverage. Media measurement can evaluate an organization's success, or lack of success, in conveying organizational messages, in countering undesired or incorrect messages, in positioning company or third party spokespeople, and in generating favorable coverage, among other reasons. Public relations professionals analyze media coverage to help demonstrate the value of public relations, provide insights to make better decisions, improve performance, and understand issues.

Communications professionals often face the need to compare the results of multiple public relations campaigns across brands, business units, and geographies. In the absence of an industry-wide methodology for data collection and analysis, in-house communication teams, their public relations agencies, and their media measurement firms use inconsistent definitions and calculations for reporting results. This frustrates management, limits the possibilities for organizational learning, reduces efficiency, and puts budgets at risk. Senior communication leaders want transparent, replicable, credible metrics—similar to those presented by their counterparts in marketing, advertising, or sales—to demonstrate their results.

Practitioners have thus been asking for measurement standards to ensure that all their public relations efforts can be evaluated using consistent definitions and measurements.

In June 2012, the Institute for Public Relations (IPR) released a paper (Eisenmann, Geddes, Paine, Pestana, Walton, & Weiner, 2012) proposing industry standards on how to calculate some of the most fundamental and commonly debated metrics in traditional media analysis: (i) what counts as a media "hit," (ii) impressions, (iii) assessing sentiment, and (iv) gauging quality. In line with the process outlined by the International Organization for Standardization (ISO), the IPR standards were open for comment from industry practitioners who wanted to participate in the development and revision of the proposed standard metrics. In October 2013, four companies that are major buyers of public relations research and measurement services adopted the standards. The corporations are General Electric, McDonald's USA, General Motors and Southwest Airlines.

The ISO process also recommends validation of the standards to demonstrate that the standards actually measure what they say they measure. This research seeks to validate the proposed media standards by measuring the level of agreement among three independent coders when coding media stories based upon the proposed standards and by strengthening the validity of the media analysis codebook, the instrument created to measure the proposed media standards. The purpose of this research is threefold. First, to support public relations practitioners in successfully implementing the proposed standards in their measurement work by providing guidance and best practices on how to set up a detailed coding guidebook and instruction, based upon the proposed traditional media standards. Second, to test the reliability of the proposed

standards based upon a coding analysis of a randomly selected sample of media coverage, providing a path to uncovering best practices to improve the process. Success in defining this pathway should lead to more frequent use of media analysis for measurement in public relations and, more importantly, higher quality, useful results that contribute to communications planning and strategy development. Third, to strengthen the validity of the codebook designed to measure the proposed standards, based upon this research study.

2. Literature Review and Research Purpose

According to public relations historians (Lamme & Miller, 2010; Watson, 2012), media analysis began as early as the late 18th century, when US presidents informally monitored coverage in newspapers to understand public opinion. Watson (2012) says that media analytics proliferated in the mid-20th century, and by the 1990s, measurement and evaluation in general was a popular topic of interest among public relations academics and professionals. Indeed, contemporary books on public relations measurement and evaluation provide excellent guidance and examples of how to use content analysis to analyze media coverage (Paine, 2007; Stacks, 2012; Stacks & Michaelson, 2010; Watson & Noble, 2005). Traditional media analysis includes quantitative measures such as item counts and impressions and qualitative measures such as tone and key message presence, all typically referred to as outputs in public relations measurement and evaluation. Michaelson and Griffin (2005) have proposed alternative approaches focused on content accuracy to track omissions, misstatements, incomplete information, and basic facts. Although more recent developments in media analysis have focused on linking media coverage to business outcomes (Jeffrey, Michaelson, & Stacks, 2006, 2007) and examining the return on investment of media coverage (Likely, Rockland, & Weiner, 2006; Weiner, Arnorsdottir, Lang, & Smith, 2010) analysis of basic media coverage outputs such as impressions, tone and key performance indicators are nonetheless important as a measure of public relations efficiency.

Support for practitioners undertaking media content analysis is important as there is little training available and many times it is the more junior and less experienced team members who are asked to implement the work. Michaelson and Griffin (2005) suggest that media analysis is not frequently used because the evaluation rarely offers any valuable insights or solutions for communication challenges beyond tonality. They also contend that successful results depend on the knowledge and experience of the coders and that rigorous training is needed for consistent and reliable results.

In addition to the many published books and papers on how to do media analysis, entities in the private sector—agencies, corporations, service providers, and consultants— have also developed proprietary systems of media analysis. Widely inconsistent results in media analysis are common, rendering the data useless for comparison among programs and an unreliable source for business decision-making. For example, in 2009 the Central Office of Information (CIO) in the United Kingdom sent an identical brief comprising 138 items of media coverage to five companies for evaluation. They wanted to know how many people consumed the coverage, how much it cost per 1,000 reached and what was the favorability and tone of the coverage was. Despite the fact that these are common measures for public relations, the results were all different and the range within each was very large (CIO).

In an effort to address inconsistencies such as these in public relations measurement, public relations practitioners and academics from around the world have collaborated to create clear and transparent standards and approaches to measuring and evaluating public relations results. The *Oxford English Dictionary* defines a standard as "an idea or thing used as a measure, norm, or model in comparative evaluations" (Michaelson & Stacks, 2011, p. 4). According to Institute for Public Relations president Frank Oviatt (2013), a standard is a published specification in the public domain that provides a common language for comparison purposes. Standards can help foster innovation and increase the credibility of public relations work (Oviatt).

Michaelson and Stacks (2011) contend that standardization of public relations measures requires significantly more than a description of the measure to be included in the analysis. They highlight that the implementation of specific research procedures and protocols that will be applied uniformly and consistently are needed.

However, a public relations standard is not synonymous with a best practice. As explained by Michaelson and Macleod (2007), a best practice is a "technique, method or practice that is more effective than others in reaching an established goal" (p. 3). Standards define *what* needs to be measured whereas a best practice indicates *how* to best meet the objective of the standard (Michaelson & Stacks, 2011).

The Barcelona Declaration of Measurement Principles (2010), developed by the Institute for Public Relations Measurement Commission and the International Association for the Measurement and Evaluation of Communication, and subsequently endorsed by other public relations industry organizations, represented the first step toward creating public relations measurement standards. The Barcelona Principles advocate that (1) advertising value equivalency (AVE) does not equal the value of public relations, (2) that media measurement should include quantity and quality, and (3) that transparency and replicability are paramount to sound measurement, among other principles.

In 2012, a newly formed industry group—the Coalition for Public Relations Research Standards—launched with the purpose of creating a broad platform of standards and best practices of public relations research, measurement, and evaluation. Partners include the Council of Public Relations Firms (CPRF), the Global Alliance, the Public Relations Society of America (PRSA), the International Association for the Measurement and Evaluation of Communications (AMEC), and the Institute for Public Relations.

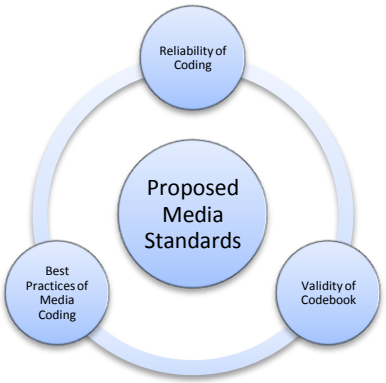
Later in 2012, the Coalition for Public Relations Research Standards created and released the *Proposed Interim Standards for Metrics in Traditional Media Analysis* (Eisenmann et al.), and sought industry input on the proposed standards. In fall 2013, four corporations that are buyers of public relations research and measurement services— General Electric, McDonald's USA, General Motors and Southwest Airlines—adopted the first round of interim standards recommended by the Coalition. In the words of Coalition chair David Geddes, "Basing our process on the recommendations of the International Organization for Standardization (ISO), we said from the beginning that customers like these corporations are the final arbiters of when a standard is ready to move forward" (Four Major Corporations Adopt Public Relations Standards, 2013).

Eisenmann, O'Neil, and Geddes sought to test the reliability of the proposed interim traditional media standards in 2013 by using inexperienced coders, trained using a coding guidebook and instructions developed by the researchers and based on the *Proposed Interim Standards for Metrics in Traditional Media Analysis* guidelines. The research results yielded low to moderate intercoder reliability based on Krippendorf *alphas*. The most significant feedback on this work was the recommendation to repeat the study using trained coders to see if/how reliability improves with experience.

This study addresses this feedback by seeking to accomplish the following objectives (1) to strengthen the validity of the codebook by clarifying and refining the coding descriptions and definitions based upon the proposed standards, (2) to retest the reliability of the proposed standards with three experienced coders using the same corpus of 106 media stories about a single company, and (3) to document best practices for measuring the proposed media standards.

The following model documents this validation process this study seeks to address:

Proposed Media Standards



Reliability: Independent coders who code media stories using the codebook with instructions on how to code for the standards should reach similar conclusions or agreement.

Validity: The measuring instrument, in this case the codebook, must actually measure the concepts described in section 3.1 through 3.4 using proposed media standards.

Best Practices: Coders must be carefully trained on how to use the codebook and a quality control system should be implemented to systematically check on intercoder reliability and provide ongoing feedback.

The specific objectives of this research are as follows:

1. Clarify some of the descriptions in the codebook, which is a set of clear instructions for coding, based upon the proposed standard metrics and the 2013 research study by Eisenmann, O'Neil, and Geddes.
2. Train a team of three analysts, professional full-time media coders, using the 12 training documents. The analysts and researchers will review and discuss their findings as a team, and then revise the codebook as necessary to clarify coding instructions.
3. The analysts team members, working independently, code the test set of 106 documents.

4. Analyze inter-observer agreement using an appropriate statistical test, in this case Krippendorff's *alpha*.
5. Document an appropriate procedure for testing inter-observer agreement in the practice o of media measurement, and make recommendations to the practice.
6. If needed, follow-up with analyst team for feedback on how to further clarify the coding instructions for the proposed standards and then make revisions to the codebook.

3. Standard Definitions and Guidelines for Traditional Media Analysis

3.1 Items for media analysis

An item for media analysis is a "manifest unit of analysis used in content analysis consisting of an entire message itself (e.g., an advertisement, story, press release)" (Stacks and Bowen, 2013). General guidance for what should be included in media analysis is that the item has passed through some form of editorial filter, which is what distinguishes public relations from other forms of marketing. Items for analysis can include many types of communications content, including but not limited to the following:

- An article in print media (e.g. *New York Times*).
- News wire stories from organizations such as Dow Jones, Reuters, Associated Press and Bloomberg. In addition to counting as an item for the news wire, each media outlet running the story counts as a separate item or "hit" because it has different readership. If the wire story is updated multiple times in one day, only count the story once in a 24-hour period using the latest, most updated version.
- Article in the online version of print media (e.g. nytimes.com). An article appearing in both the online and print version of media outlet should both be counted because the readership is different for each channel.
- Article in an online publication (huffingtonpost.com).
- Broadcast segment (TV or radio). In the case of a broadcast segment that repeats during the day, each segment should be counted as an item because audiences change during the day.
- News item on the web site of a broadcast channel or station.
- Analyst report.
- Reprints or syndication of an article. Each appearance in an individual media outlet counts because the readership is different.
- Bylined feature by company executive.

Press release pick-ups from "controlled vehicles" such a posting stories on PR Newswire, Business Newswire and other sites where the content is not "earned" should not be counted as an item for analysis.

Other items for content analysis could include blog posts, comments on blog posts, posts and comments on discussion boards and forums, tweets, Facebook posts and comments, videos and comments posted. These social media channels are not considered part of traditional media and typically have little or no editorial screening, so would not be captured for analysis. However, as this study's goal is to provide a practitioners tool, we have given exception to blog posts, the more notable of which in practice are frequently included in media analysis. This might include the blogs of traditional print media publications such as the *Wall St. Journal* (http://blogs.wsj.com/) or those popular with key target audiences such as Mommy Blogs.

3.2 Impressions

Impressions are the number of people having the opportunity for exposure to a story that has appeared in the media. Impressions are also known as "opportunity to see" (OTS) and usually refer to the total audited circulation of a publication the verified audience-reach of a broadcast vehicle or viewers of an online news story (Stacks & Bowen, 2013). Impressions should not be mistaken for awareness. "Awareness" exists only in people's minds and must be measured using other research tools. Impressions are indicative of the opportunity to see (OTS). Organizations may want to consider OTS as an alternative nomenclature to better clarify what "impressions" mean – potential to see/read and a potential precursor to "awareness."

- For print media, impressions should be based on circulation figures such as those provided by the publication, or through resources such as subscriptions tools such as Cision, or Alliance for Audited Media (formerly Audit Bureau of Circulations) in North America and audit bureau of circulations in the UK, India, Australia, Hong Kong and elsewhere. Multipliers should not be used for calculating impressions.
- For online media – impressions should be calculated by dividing the number of unique visitors per month by the number of days in the month to get the number of daily views. Impressions should be based on the unique URL or sub-domain for the item (e.g., www.yahoo.com vs. finance.Yahoo.com). Unique visitors per month can be sourced through several services, such as Compete.com or Nielsen NetRatings.
- For broadcast – organizations are advised to use the numbers distributed by the broadcast monitoring service provider, i.e. usually Nielsen. For example, a monitoring report for a single clip typically includes the following: Time: 9:30am, Aired On: NBC, Show: Today (6/8), Estimated Audience Number: 5,358,181
- For wire services (AP, Bloomberg, Dow Jones, Reuters, etc.) – no impressions are assigned to stories simply carried by wires services, only to the stories that they generate in other media.

3.3 Tone or sentiment

Tone (or sentiment) measures how a target audience is likely to feel about the individual, company, product or topic after reading/viewing/listening to the item, typically defined as positive, neutral/balanced, or negative.

- Analysis of tone is a subjective aspect of media analysis and there are multiple approaches to assigning tone. The standards recommend that whatever process is defined and applied, the methodology must agreed on from the beginning and must be consistently applied throughout any analysis.
- The approaches for judging tone include "latent analysis," which is to look at the entire article or mention and judge the item as a whole based on the overall tone. A second approach is called "manifest analysis." It looks at an item as a series of sentences or paragraphs, judges each one on its sentiment and then adds up the total number of positives and negatives to obtain an overall score. A third approach is to avoid assessing tone based on the whole story and make the evaluation on the basis of pre-determined positive and negative messages present in the article.
- Likewise there are several approaches for assigning a numeric score for tone. For example, tone could be scored on a three-point scale (positive, neutral or negative), a five-point scale (very positive, somewhat positive, neutral, somewhat negative and very negative) or other similar scales. Another option is to use a 101-point scale ranging from zero (totally negative) to 100 (totally positive). The scoring approach must also be established in advance with defined examples. Typical definitions are:

Positive	An item leaves the reader more likely to support, recommend, and/or work or do business with the organization or brand.
Neutral	An item contains no sentiment-bearing information at all, just reports the facts. If the news is negative, an article can be neutral if it just reports the facts, without any editorial commentary. In an unfavorable environment, neutral may be the best that can be achieved. Coding should be based on whether or not the clip makes people more or less likely to do business with an organization.
Negative	An item leaves the reader less likely to support, and/work or do business with the organization or brand.
Balanced	Requires both positive and negative sentiment-bearing information in roughly equal proportions, and therefore the resulting overall tone is balanced.

- Organizations must define for what or whom they want to determine sentiment. For example, they may seek to understand tone regarding an industry or sector, or sentiment around a specific product or service, an individual or an organization. A single article could mention all of these. As a result, it is necessary to define specifically what element(s) are being targeted for sentiment.

- Organizations must define from whose perspective they are judging the sentiment. It could be the point of view of the general public; a specific stakeholder group such as investors, physicians, teachers or parents.

3.4 Quality Measures

Quality measures should also be included when analyzing each item. Examples include:

- Visuals – percent of items including a photo, chart or logo in the article that will make the article more prominent for the reader.
- Placement – percent of items with preferred placement in the item i.e. front page, first page of a section or website landing page.
- Prominence – percent of items where an organization/program is mentioned in the headline, first paragraph or prominent side-bar, or number of times the organization, brand, or program is mentioned in the item.
- Spokesperson – percent of items including a quote from an organization's spokesperson(s).
- Third party – percent of items including quotes from third parties endorsing a company's organization or program.
- Dominance —(shared vs. sole mention) – (i) percent of items where an organization/program is the sole subject of the item vs. (ii) percent of items where an organization or program shares space with competitors in the same space or a mere passing mention.
- Message Penetration – percent of items that include one or more key message.
- A more advanced approach is to measure message integrity by analyzing message pick-up as full, partial, amplified, or incorrect/negative.

Quality measures can be scored to allow comparisons among those being tracked. If some qualitative factors are more important than others, weighted values could be assigned to reflect this.

4. Research Methodology

In the first phase of this research, the researchers revisited and refined some coding instructions contained in the codebook that was initially developed in 2012 (Eisenmann et al., 2013). Clearer coding instructions were provided for metrics that coders in study one indicated needed clarity. The codebook describes what counts as a media hit, impressions, tone, quality measures, and positive and negative corporate reputation measures, among others. The complete codebook is contained in Figure One.

Insert Figure 1 about here

Notable in the codebook is the approach to assessing sentiment, which was to use latent analysis—assessing the overall tone of the entire article with respect to the company of interest and coding the story on a five-point scale—very positive, somewhat positive, neutral, somewhat negative and very negative. Sentiment is assessed on how the item might influence a reader/viewer's perceptions of the organization and, as a result, his or her decision about doing business with the organization – e.g., buy or recommend its products, apply for a job, etc. Sentiment is not based on the inherent positivity or negativity of the specific news reported and could be impacted by the reporter's approach.

Standard corporate reputation messages, derived from those used by the Reputation Institute and other organizations, were also used in the analysis. These included financial soundness, quality of leadership/management, quality of products and services, innovation, workplace environment and citizenship (the latter includes corporate social responsibility). Reputation messages were coded positive or negative only when prominently present within a story; a mere passing or implied mention was not considered to be substantive enough to be recalled by a reader/viewer and consequently impact reputation. Not all items would be expected to carry a reputation message and these messages are not expected to appear verbatim; therefore, some level of interpretation was required by coders.

Next, the researchers trained three full-time, professional media analysts with an average of 5-6 years of experience how to code media stories using the developed codebook. The initial coding training session lasted approximately two hours.

The media stories used in this study were the same ones that were coded in 2013, when the researchers used systematic random sampling to select 106 media articles about Wal-Mart that appeared during a one-year period July 1, 2011 – June 30**,** 2012. Items for analysis included traditional media items from print, online and broadcast outlets, as well as some blogs, all captured via Factiva and containing at least three mentions of Wal-Mart.

After the training session, the three coders and two of the researchers independently coded twelve stories about Wal-Mart as a part of a pretest to identify and clarify discrepancies. There were some inconsistencies in the pretest, so the researchers spent time with the coders clarifying coding instructions and later revising and clarifying codebook descriptions. A second training was held with the coders to address specific questions and inconsistencies. Again, the researchers spent time reviewing results and answering questions from the coders. Revisions were made to the codebook to elaborate coding instructions for certain media items.

In the next part of the research, the three coders independently coded the 106 randomly selected stories. Coders did not confer with the researchers or fellow coders, even when they had questions.

Once the coding was completed, the researchers calculated intercoder reliability for the variables involved in the coding project. Intercoder reliability refers to the level of agreement among coders when coding a corpus of messages using the same coding instructions and or codebook (Wimmer & Dominick, 2014). Although reliability can be calculated in many ways, Krippendorff *alpha* was calculated because it "can be used regardless of the number of observers, levels of measurement, sample sizes, and presence or absence of missing data" (Hayes & Krippendorff, 2007, p. 77). Krippendorff's *alpha* was calculated using a macro (Hayes &

Krippendorff) with version 19 of the Statistical Package for the Social Sciences (SPSS). Krippendorff *alpha* ranges from 0.00 to 1.00, but should not in any way be considered comparable to percentage agreement. Krippendorff (2004) recommends an *alpha* level of at least .80 as a standard, accepting data in situations where.800 > α ≥ .667 "where tentative conclusions are still acceptable," and rejecting data where .667 ≥ α. (p. 241). Krippendorff emphasizes that neither increasing the sample size or increasing the number of coders will increase intercoder reliability.

Following the statistical analysis, the researchers conferred with the coders to review additional stories and discuss items that presented difficulty with the coders. Based upon those in-depth conversations, the researchers refined coding descriptions for item type, other company/brand mention, and citizenship. The final codebook presented in Figure 1 includes all changes.

5. Results

As indicated by results presented in Table 1, Krippendorff *alphas* ranged from a low of .3849 to a high of .9374. Overall, the *alpha*s indicate moderate to high reliability, thereby suggesting a moderate to high level of agreement on the coding decisions made by the three coders.

From 2013 to 2014, *alphas* improved for 11 of the 13 media variables, indicating the importance of robust training, using a codebook with clear coding instructions, and employing experienced media coders.

Insert Table 1 about here

Not surprisingly, agreement was high for three of the basic, straightforward items in the media analysis project. For example, prominence (whether the first Wal-Mart mention was in the headline, first paragraph, or other paragraphs) had an *alpha* of .9374, shared/sole mention (whether there is a mention of other retail companies in addition to Wal-Mart) had an *alpha* of .8759, and media type (whether the story was a print, online, wire, radio/TV broadcast, or a blog) had an *alpha* of .8116. It was also encouraging that two of the corporate reputation messages—workplace environment and quality of leadership/management—had high *alphas*, .8652 and .8403 respectively. When coding these two items, coders had three choices to select from: no message, positive, or negative. However, because the messages were not verbatim from one story to the next, coding of this variable required interpretation on behalf of the coders.

Sentiment when measured on a five-point had a moderate to high *alpha*, .6775. This indicates that the coders had a mid- to high-level of agreement in terms of coding a story from a latent point of view, which involved interpreting whether the overall attitude about Wal-Mart in the story was very positive, positive, neutral, negative, or very negative. Given that the researcher clarified the coding instructions of sentiment in the codebook and that they spent a fair amount of time training coders how to code for sentiment, the researchers were pleased to see that this year's results improved from .1746 in 2013 to .6775 in 2014. Moreover, when the five-point sentiment scale was collapsed to a three-point scale, the *alpha* increased to .796.

Surprisingly, the experienced coders had low agreement for three media items: (1) item type (*alpha* = .3682), (2) other company/brand mention (*alpha* = .3854) and (3) the citizenship corporate reputational message (*alpha* = .3849). Due to these low scores, the researchers combed through the details of the results to determine areas of coding confusion. Researchers next selected three stories from the corpus of 106 stories to review in detail with the coders to better understand the thought process for the coding decisions of item type, other company/brand mention, and the citizenship corporate reputation message. Once the researchers better understood how and why the coders disagreed on these media items, the researchers went back to the codebook to clarify the coding instructions for these three items.

For item type, coders were originally instructed to code whether a story was best characterized as corporate news, product news, column/opinion to the editor, interview, editorial, feature, or round-up. Out of 106 coding decision, coders disagreed 38 times regarding whether a story was corporate news or a round-up and they disagreed ten times whether a story was best characterized as corporate news or product news.

To improve the codebook, the researchers added some additional description of corporate news, product news, and round-up. For example, they added that when coding for a round-up story or industry overview, "the target company/organization would be mentioned only as an example, not as the sole focus of the item" and for product news they bolded key terms such as "**target company/organization branded** products or services, such as **marketing programs or campaigns** (among others listed in the codebook).

Coders explained to the researchers that they had some difficulty determining whether a company/brand mention was prominent enough to warrant a "sharing of the story." To alleviate this confusion, the researchers changed the coding instructions to: "Is there a mention of other organizations, government bodies, companies or brands (non-retail) as a subject or driver of the story (as opposed to offering comment or analysis)? For example, a company or organization from another industry sector which is being discussed in the same context as Walmart or being compared to Walmart. Or, a government entity imposing or enforcing regulatory action on Walmart, and possibly others in the same category. Other company/organization mention is quite common in the news media. Such a mention should be coded when it is prominent and relevant (e.g. pertinent to understanding the full item and its meaning or impact)."

Conversations with the coders regarding the citizenship corporate reputation message indicated that coders had difficulty differentiating whether a company's CSR activities were implied or explicit in a media story. There was also some confusion regarding whether consumer social media contests were indicative of corporate citizenship. Therefore, to clarify the confusion regarding the citizenship corporate reputation message, the researchers rewrote the coding instructions to the following: "Behavior is/is not socially responsible; does/does not support good causes, contribute/commit to the community beyond selling products; CSR is specific to initiatives or goals that the company has set forward, not solely implied as part of good management. Examples include philanthropic donations, employee volunteerism, community relations involvement, cause-related marketing and cause promotions. A program to engage customers/prospective customers via crowdsourcing or participation in events such as competitions, photo submissions, social media, etc. is not CSR."

The final codebook (see Figure 1) has undergone 14 iterations since this study was initiated in spring 2012. Revisions have been based upon four pre-tests, two coding projects, and two-follow-up discussions with different coders. Examples of how to code for media items are contained in the sample story contained in Figure 2. Commentary is offered in the footnotes of the sample story to identify the elements for coding and to explain the logic of the coding decisions.

Insert Figure 2 about here

6. Discussion and Conclusions

This research has a number of implications essential to the practice of media analysis, remembering that the main objective of standardization is to ensure quality data and comparability of data. First, this research indicates that coding for the metrics as defined by the standards of traditional media analysis—and operationalized in the codebook—is reliable, provided that the coders have a well-developed codebook and sufficient training. Ten of the thirteen media items had moderate to high *alphas*, indicating that the three coders were in agreement the majority of the time. Three of the thirteen items—item type, other company/brand mention, and the citizenship corporate reputational message—had low *alphas*. However, based upon a follow-up meeting with the coders, the researchers were able to clarify the coding instructions for these three items to address some of the coding confusion. The researchers are cautiously optimistic that, based upon these codebook enhancements, coding reliability will improve for these three items in future media analysis projects. Moreover, the stories coded in this research project were fairly substantive, because each story had at least three mentions of Wal-Mart. Some coding projects may involve shorter and perhaps, even more straightforward stories. Therefore, the reliability of coding projects involving less substantive stories might be even higher.

As noted earlier, the 2013 (Eisenmann et al.) study had low to moderate intercoder reliability among the three inexperienced coders. Results of this 2013 study raised the key question of whether some of the standards such as sentiment/tone and corporate reputational messages needed revision or whether the coders lacked sufficient experience and training to code with reliability. The answer to that question is likely a combination of all three. The much improved results of this 2014 study are likely due to greater and improved training, using experienced coders, and clarifying and including more detailed coding decisions for the traditional media metrics in the codebook.

This study indicates the importance of sound training, a well-developed and tested codebook, and the use of Krippendorff's *alpha* as best practices. Human coders need to be carefully trained and ideally have some knowledge of the subject area. In this study, the researchers were a bit surprised that the original plan for one training session with the experienced coders was not sufficient and that an additional training was needed, an indication of the importance of robust training as a best practice. Regardless of the level of coder experience, two rounds of pretests are ideal when initiating a new media analysis project.

It is also important to set realistic expectations with clients, in regards to the time and training needed to secure reliable results. The level of detail needed in a coding project depends upon the client or organizational objectives. Public relations project managers can utilize the codebook to decide coding elements are needed to provide insights so as to not over-complicate the codebook with any unnecessary detail, which could impact reliability. Moreover, measurement agencies and firms should use Krippendorff's *alpha* as part of their training and quality management processes. Clients should expect that agencies and measurement firms provide results of the inter-observer agreement testing.

This study raises important questions about the media analysis training process. In practice, a quality control system should be implemented to systematically check on intercoder reliability and provide ongoing feedback. Using one coder regularly may deliver the most reliable results, but may not be a realistic long-term approach. This study raises important questions about this quality control process. For example, should a project manager of a large media coding project review 10% of coders' stories? Should coders confer after 25 stories? How often should coders be trained? How many stories should be coded as part of the training process? How much disagreement will be tolerated in the training process? Future research might build upon this study to test for the effects of training to better understand best practices.

With respect to relationships between client organizations and measurement teams, clients should not cherry-pick cases where they disagree with the coding of a specific item. This will not improve reliability. Rather, they should provide feedback that can be incorporated into codebook revisions and ongoing training of the coders.

Another key question related to best practices is to determine how many stories should be coded on behalf of a client or organization. The scope of a coding project will be determined by multiple factors, including project objectives, budget, and organizational or client background. For example, when coding stories about the presence of a company like Wal-Mart, it may be best to code more stories to fully capture the range of publications and activities.

Practitioners must also decide whether to code sentiment on a three-point scale (positive, neutral, negative) or a five-point scale (very positive, positive, neutral, negative, very negative). As indicated by the results, *alpha* for the three-point scale was high, .796, whereas the *alpha* for the five-point scale was moderate to high, .675. It may be that when testing for inter-observer agreement, seeking a .796 on the three-point scale is the gold standard, but that practitioners may need to review some project results on the 5-point sentiment scale to discern finer differences.

In summary, this research has helped to validate the proposed standards for traditional media analysis. Public relations practitioners can use and amend the detailed codebook for their specific purposes and borrow from some of the recommended best practices, including training and systematic quality control and feedback. As more practitioners continue to adopt and use the proposed standards, the coding instructions in the codebook can be revised and elaborated. Moving forward, the researchers recommend submitting the same set of stories to one or more automated sentiment scoring companies. Comparing the reliability of stories coded by humans compared to automated sentiment scoring would provide a comparison of the reliability of the two approaches.

References

Barcelona declaration of measurement principles (2010). Presented at the 2nd European Summit on Measurement. Available from http://www.instituteforpr.org/iprwp/wp content/uploads/BarcelonaPrinciplesSlides.pdf

Central Office of Information (UK) (2009). *Standardisation of PR evaluation metrics*. Available from www.cipr.co.uk/sites/.../standardisation%20of%20PR%20metrics.pdf

Eisenmann, M., Geddes, D., Paine, K., Pestana, R., Walton, F., & Weiner, M. (2012). *Proposed interim standards for metrics in traditional media analysis.* Gainesville, FL: Institute for Public Relations. Retrieved from http://www.instituteforpr.org/topics/proposed-interim-standards-for-metrics-in-traditional-media-analysis.

Eisenmann, M., O'Neil, J. & Geddes. D. (2013). Proceedings from the 16th Annual International Public Relations Research Conference: *Testing the Reliability of Metrics Proposed as Standards for Traditional Media Analysis*

Four Major Corporations Adopt Public Relations Standards (2013). Retrieved January 7, 2104 from http://www.instituteforpr.org/releases/four-major-corporations-adopt-public-relations-research-standards/

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

Jeffrey, A., Michaelson, D., & Stacks, D.W. (2006). *Exploring the link between volume of media coverage and business outcomes*. Gainesville, FL: Institute for Public Relations. Retrieved from http://www.instituteforpr.org/topics/media-coverage-business-outcomes/

Jeffrey, A., Michaelson, D., & Stacks, D.W. (2007). *Exploring the link between share of media coverage and business outcomes.* Gainesville, FL: Institute for Public Relations. Retrieved from http://www.instituteforpr.org/topics/media-coverage-share-business-outcomes/

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology (2nd ed).* Thousand Oaks, CA: Sage Publications.

Lamme, M.O., & Miller, K.R., (2010). Removing the spin: Towards a new theory of public relations history. *Journalism & Communication Monographs, 11* (4), 281-362.

Likely, F., Rockland, D., & Weiner, M. (2006). *Perspectives on the ROI of media relations publicity efforts.* Gainesville, FL: Institute for Public Relations. Retrieved from http://www.instituteforpr.org/topics/media-relations-publicity-efforts/

Michaelson, D., & Griffin, T. (2005). *A new model for media content analysis.* Gainesville, FL: Institute for Public Relations. Retrieved from http://www.instituteforpr.org/topics/new-model-for-media-content-analysis/

Michaelson, D., & Macleod, S. (2007). The application of "best practices" in public relations measurement and evaluation systems. *Public Relations Journal, 1*(1), 1-14.

Michaelson, D., & Stacks, D.W. (2011). Standardization in public relations measurement and evaluation. *Public Relations Journal, 5*(2), 1-22.

Oviatt, F. (2013). What are standards? Retrieved from http://www.youtube.com/watch?v=nDMobcQlpo4

Paine, K.D. (2007). *Measuring public relationships: The data-driven communicator's guide to success.* Berlin, NH: KDPaine & Partners, LLC.

Stacks D. W. (2010). *Primer of PR research* (2nd ed.). New York, NY: Guilford.

Stacks, D. & Bowen, S. (2013). *Dictionary of public relations measurement and research: 3rd edition.* Gainesville, FL: Institute for Public Relations. Retrieved from http://www.instituteforpr.org/topics/dictionary-of-public-relations-measurement-and-research/

Stacks, D. W. & Michaelson, D. (2010). *A practitioner's guide to public relations research, measurement, and evaluation*. New York, NY: Business Expert Press.

Watson, T. (2012). The evolution of public relations measurement and evaluation. *Public Relations Review, 38(3),* 390-398.

Watson, T. & Noble, P. (2005). *Evaluating public relations: A best practice guide to public relations planning, research, and evaluation*. London, UK, and Philadelphia, PA: Kogan Page.

Weiner, Arnorsdottir, Lang, & Smith, B.G (2010). *Isolating the effects of media-based public relations on sales.* Gainesville, FL: Institute for Public Relations. Retrieved from http://www.instituteforpr.org/topics/media-based-pr-on-sales/

Wimmer, R. D. & J.R. Dominick (2014). *Mass media research: An introduction* (10th ed.). Boston, MA: Wadsworth.