INSTITUTE FOR PUBLIC RELATIONS

TEN LESSONS LEARNED FROM STANDARDS THAT FAILED THE TEST

BY KATIE DELAHAYE PAINE
CEO & PUBLISHER
PAINE PUBLISHING, LLC DURHAM, NH
WWW.PAINEPUBLISHING.COM
MARCH 14, 2018

This paper is by the IPR Measurement Commission

THE WHY AND HOW OF SOCIAL MEDIA MEASUREMENT STANDARD

The road from measurement mayhem to standards can be a rocky one, as we learned when we set out to bring some order to the chaos that was social media measurement in 2011. At the time, nearly a billion people were on Facebook and no one really knew what that meant. Twitter was still a toddler, and no one knew who was on it. Nonetheless we were already swimming in new metrics that most of us had never seen before like "engagement rate" and "shares" and "comments." So, it wasn't surprising that half a dozen different industry groups were working on their own set of standards for measuring social media.

Several Institute for Public Relations (IPR) Measurement Commission members thought that the best solution would be to invite all the people working on standards for social media to my farm in New Hampshire and keep them there until we sorted it all out. So we invited representatives from IPR, IAB, WOMMA, DAA, IABC, PRSA[1] plus the clients and their agencies. We called it a "Conclave" and the stated goals were:

1. Eliminate confusion in the marketplace about social media measurement standards
2. Gain consensus around a definition for social media measurement standards
3. Document all efforts in progress to establish "standards" for social media measurement
4. Reduce duplicative and redundant efforts around establishing social media measurement standards

Astonishingly, when the 30 or so participants left that first Conclave gathering in October 2013, we had defined the areas that we would focus on and a deadline by when the standards would be drafted. They addressed six specific issues that the Conclave participants felt needed to be addressed immediately;

1. Content & Sourcing –

At the time, many social listening platforms were only monitoring Twitter or maybe Facebook and very few included YouTube and Instagram. So, the first requirement for a standard was to identify which social platforms were (or were not) included in any measurement report, the definitions used and how any metrics were calculated.

2. Reach & Impressions –

This remains the area of most controversy for social media. Members of the Digital Analytics Association (DAA, known as the time as the Web Analytics Association) were already close to agreement on a standard definition, so the Conclave agreed to adopt their standard. Since then, the Media Research Council, which was chartered by congress to define standards of reach and impression for television, has issued its own standard that incorporated those of the DAA and thus the Conclave accepted as their standards in October 2016. They can be found here.

3. Engagement & Conversation

For engagement, The Conclave developed specific definitions and outlined eight best practices for measuring engagement. While they have not been tested, they have withstood the test of time, since many of the best practices have been adopted by the industry.

[1]Institute for Public Relations (IPR), Interactive Advertising Bureau (IAB), Word of Mouth Marketing Association (WOMMA), Digital Advertising Alliance (DAA), International Association of Business Communicators (IABC), Public Relations Society of America (PRSA)

4. Influence

Concurrent with the Conclave's efforts, the Word of Mouth Marketing Association was also working on its own standard definitions for influence. So, similarly to the DAA, the Conclave adopted WOMMA's definitions and best practices as outlined in their Influencer Handbook.

5. Impact & Value.

While measuring the impact and value of Social Media remains a very hot topic and is no less a necessity now than it was in 2013, the definition and calculations for them are unique to each organization. Therefore, The Conclave limited itself to clear definitions and best practices.

6. Opinion & Advocacy

The basis for the Conclave's standards for Opinion and Advocacy were rooted in the history of similar metrics for traditional media. The IPR Measurement Commission had devoted extensive effort and research into drafting and testing a set of coding standards.

But obviously social media, without the constraints of editors or journalistic ethics posed very different challenges. The first was differentiation between sentiment, opinion and advocacy. In the end the Conclave defined each as:

- Sentiment is a component of opinion and advocacy. Sentiment is the feeling that the author is trying to convey, often measured through context surrounding characterization of an object.
- Opinion is a view or judgment formed about something, not necessarily based on fact or knowledge. Standard indicators of opinion standards have not yet been achieved, but typically opinion is definitively articulated and associated to the speaker.
- Advocacy (n.) vs (v.) is a public statement of support or a recommendation for a cause or policy. Advocacy requires a level of expressed persuasion.

The key distinction between "advocacy" and "opinion," is that advocacy must have a component of recommendation or a call to action (CTA) embedded in it.

The final standards were published in June 2013. You can read them here.

THE TEST

But standards only work if they can be implemented. In real life that means when followed, standards should be able to be used by any organization. If the methodology has been written correctly, it should be able to get the same consistent results for whatever organization is using it

Thus the next step for the Conclave Standards was a real-life analysis using the standards as written, We based our testing framework on the great work done by David Geddes, Julie O'Neil and Marianne Eisenmann testing the tradtional media standards. Luckily, we were able to recruit Julie O'Neil, Ph.D. Associate Professor & Director of Graduate Studies, Strategic Communication, Texas Christian University to supervisor our research and Michelle Hinson, Chief Communications Officer at NxGen Global in Gainesville, Fl to manage the project.

Knowing how different social media treats different types of organizations, we decided to test the standards on a non-profit organization, a federal agency and a consumer company. Goodwill Industries, Southwest Airlines and the U.S. Fish & Wildlife Service agreed to be our guinea pigs and we enlisted Glean.Info to collect the data and assist in the coding test.

Our first step was to create a detailed description of the terms, what types of items qualified for the test, and definitions for the messages and coding criteria. For example, sentiment/opinion was defined as follows:

POSITIVE	**Goodwill:** If the item leaves the reader more likely to donate to, shop at, partner with, volunteer at, or work for Goodwill Industries International, it is considered positive. **U.S. Fish & Wildlife Service:** If the item leaves the reader more likely to support, volunteer at, or work for U.S. Fish & Wildlife Service, it is considered positive. **Southwest:** If the item leaves the reader more likely to purchase an airline ticket, work for, or invest in Southwest Airlines, it is considered positive.
UNDESIRABLE	**Goodwill:** If an item leaves the reader less likely donate to, shop at, partner with, volunteer at or work for Goodwill Industries International, it is considered negative. **U.S. Fish & Wildlife Service:** If an item leaves the reader less likely to support, volunteer at, or work for U.S. Fish & Wildlife Service, it is considered negative. **Southwest:** If an item leaves the reader less likely to purchase an airline ticket, work for, or invest in Southwest Airlines, it is considered negative.
NEUTRAL	If an item does not influence the reader either way and/or contains no sentiment, it is considered neutral.
BALANCED	If an item contains both positive and negative sentiment in equal weights, it is deemed balanced.

We then provided examples for each category. For example, sentiment/opinion might be:

- *I love shopping at Goodwill they have low prices and it helps people.*
- *U.S. Fish and Wildlife Service is a leader in conservation.*
- *Southwest is the best airline ever!*

On the other hand, advocacy would be:

- *If you want to support an organization that creates jobs you should only shop at Goodwill employment placement services and other community-based programs for people having a hard time finding employment.*
- *Write your congressman to properly fund the US Fish & Wildlife Service*
- *You must take Southwest, they're the best.*

In addition to Opinion and Advocacy, we also analyzed for key message content message integrity (i.e. was the message fully communicated or just partially) and visuals as a conveyor of sentiment. Each organization supplied its own messages.

Data was collected from Facebook and Twitter. We did not include posts from the organizations owned pages, but we did included comments on the posts. Initially 113,000 posts were collected which we randomly sampled to get 100 posts for each organization to analyze. To qualify for the study, each item had to meet the following criteria:

- Mention one of the three organizations
- Appear in an item dated between May 1, 2016 – July 31, 2016
- Must be "earned" i.e. neither paid media nor content created by or owned by the brand being studied.

Additionally:

- Straight Retweets of organization-created Tweets did not qualify for the study
- Modified Retweets in which the author comments or expresses an opinion were coded.
- If clarification was needed, coders followed a link provided in the Tweet but only coded for the content in the 140 characters
- Pictures in posts were analyzed in the post to determine sentiment and messaging

Five readers were used to test the results, three with extensive coding experience, two who were relatively new to the process. After the readers completed coding of an initial batch of 40 articles, we compared results between coders, identified problems in the definitions and modified the coding instructions accordingly. That process was repeated three times.
A summary of the intercoder reliability test is below:

Organization	Sentiment	Visual	Message	Message Tone	Integrity	Advocacy
Southwest	.745	.881	.583	.712	.656	.266
US Wildlife	.172	.034	.128	.096	.078	.051
Goodwill	.309	.07	.062	.028	.119	.146

Navy indicates unacceptable reliability. You can read the full report of our findings here.

Obviously, the results indicated that the likelihood of getting reliable consistent results from the standards as written was slim to none, unless you were dealing with a well-known consumer brand. But we learned a lot from the process.

LESSONS LEARNED

Lesson 1: K.I.S.S. is a requirement for standards

The biggest takeaway from our test was that we made it FAR too complicated. Testing the validity of Opinion and Advocacy is complicated enough if you do it for one company. Using three different organizations added too many additional variables. We also added message testing and visuals because we thought it was important, but again, it added unnecessary complexity to the test.

Lesson 2: Sentiment is more consistent and easier to code for consumer brands than for non-profits and government agencies.

Another major lesson was how much easier it was to code consistently for Southwest Airlines than it was for the other two organizations. For whatever the reason, we also found far more opinions and statements of advocacy in Southwest posts than the other two organizations. We postulated that familiarity was a factor, since everyone is familiar with an airline, but not everyone is familiar with Goodwill, plus even fewer have experience with U.S. Fish and Wildlife Service. Our recommendation was to ensure that coders are, in fact, among the target audience or at least trained in the product category for which they are coding.

Lesson 3: Messages aren't likely to appear in social media

Key Message identification received the lowest reliability score across all three organizations. Coders found that messaging and advocacy occur very differently in Twitter than they did in Facebook, so it isn't realistic to code for a single message across all platforms. Interestingly, there were no key messages found outside of comments on owned sites. We suggest that organizations tailor their messages to the platform.

Lesson 4: Message integrity was also a poorly understood concept.

Whether a message was fully or partially communicated was interpreted very differently by the coders, one coder identified partial messages 185 times, and another identified it only 24 times.

Lesson 5: Coding consistently for advocacy is difficult.

When asked to identify instances of expressed advocacy there was a 43-point difference between the numbers found between coders.

Lesson 6: Collaboration is essential

Theoretically, if standards are valid and reliable, there should be no need to confer. However, accuracy and consistency increases when coders are allowed to confer, and it is common practice in many content analysis situations.

Lesson 7: Experienced coders are key.

In our tests the experienced coders were far more consistent than the inexperienced coders. Geddes, Eisenmann and O'Neil found the same to be true in their study.

Lesson 8: If you use multiple coders, you need a single lead coder to ensure consistency.

Five coders operating independently meant that we had to deal with five different sets of questions and someone needed to provide consistent answers.

Lesson 9: Clients need to be an active part of the process

The degree to which coders agreed with each other was closely correlated to the extent to which the organizations were an integral part of the process. The organizations that provided messages and examples and were responsive to questions saw far more reliable results than the one that didn't.

Lesson 10: Examples and strict definitions are needed to make the standards universally useful

In reality the Conclave Standards are really guidelines and best practices and turning guidelines into standards requires very detailed definitions that are customized to each organization and good examples.

NEXT STEPS

The results of the test were presented to the Conclave in October of 2016 and while at first there was general agreement that a rewrite was necessary, there was little agreement on how to make the definitions more specific given the range of opinions and styles in social media and the number of platforms now available. In the end, we decided that the best practices and definitions could stand as is, with the caveat that when implementing them, it is imperative to customize the definitions and instructions for each specific application and organization.