New avenues for research and practice in communication evaluation and measurement
Guest Editors: Sophia Charlotte Volk and Alexander Buhmann

161 Editorial advisory board
162 Guest editorial
179 Reliable, valid and efficient evaluation of media messages: developing a message testing protocol
Minji Kim and Joseph N. Cappella
198 The status quo of evaluation in public diplomacy: insights from the US State Department
Erich J. Sommerfeldt and Alexander Buhmann
213 Explaining the measurement and evaluation stasis: a thought experiment and a note on functional stupidity
Howard Nothhaft and Hanna Stensson
228 Investigating the role of contextual factors in effectively executing communication evaluation and measurement: a scoping review
Stefania Romenti, Grazia Murtarelli, Angelo Miglietta and Anne Gregory
246 Maturity as a way forward for improving organizations’ communication evaluation and measurement practices: a definition and concept explication
Nathan David Gilkerson, Rebecca Swenson and Fraser Likely
265 Agile communication evaluation and measurement
Betteke Van Ruler


www.emeraldinsight.com/loi/jcom
1. New avenues in communication evaluation and measurement (E&M): towards a research agenda for the 2020s

1.1 Introduction

The question of how public relations and strategic communication contribute to organisations’ success has been a perennial research topic since the 1970s (Stacks, 2017; Gregory and Watson, 2008; Grunig, 2006). Specifically, this question is approached in the sub-field of communication E&M. For more than five decades, scholars have explored how communication effects can be measured alongside different stages (ranging, e.g., from inputs and outputs to outcomes and impacts) and how the value of communication added to the organisational level can be evaluated. According to a recent synthesis of 40 years of literature (Volk, 2016), E&M research has explored numerous questions that can be broadly clustered into the following overarching themes: how to analyse the effectiveness of communication and messages on the output level; how to measure outcomes such as relationships and reputation; how to conceptualise intangible values or capitals and assess the value creation through communication on the impact level; how to further develop E&M methods; and how to assess the state of E&M practices.

As an applied research field, E&M scholars have maintained regular and close relations to the communication industry. Hence, there was and still is a lively and unique exchange between academia and practice (cf. Buhmann et al., 2018). Pioneering scholars, some of them former professionals (earlier on, e.g., Walter Lindemann, Don Stacks, David Michaelson, Donald Wright or Tom Watson, more recently, e.g., Jim Macnamara or Ansgar Zerfass), have stimulated a vital knowledge transfer with international and national membership organisation (e.g. the International Association for Measurement and Evaluation of Communication (AMEC)), E&M consultancies and client organisations, and various task forces or commissions (e.g. the Institute for Public Relations Measurement Commission, or the International Task Force on Standardisation of Communication Planning/Objective Setting and Evaluation/Measurement Models). Particularly, the last decade saw a considerable amount of collaborative projects, searching for more standardised models, metrics, and methods to advance E&M practices (e.g. recently, AMEC Integrated Evaluation Framework, Barcelona Principles 2.0, Social Media Measurement Standards Conclave, DPRG ICV Framework). However, according to a wide range of studies conducted worldwide to explore the state of E&M in the communication practice, the implementation of such E&M methods and standard metrics focussing on communication outcomes and impact remains low (Baskin et al., 2010; Cacciatore et al., 2016; Macnamara and Zerfass, 2017; Michaelson and Stacks, 2011; Zerfass et al., 2017).

Against this background, E&M scholars have lamented a “stasis” in evaluation practice (Gregory and Watson, 2008) or described the practice as caught in a “deadlock” (Macnamara, 2015). The consequent search for causes has centred on a discussion of possible barriers preventing practitioners from conducting (more sophisticated) E&M, including a lack of: time, budget and resources, competencies and knowledge, management demand and support; access to sophisticated methodologies or tools; and common industry models and standards (cf. Buhmann et al., 2018; Swenson et al., 2019).

Despite the considerable progress of E&M scholarship over the past 50 years, in this Special Issue we postulate the assumption that E&M scholarship can profit from connecting to new developments in the quickly changing communication environment as well as to research conducted in neighbouring fields. We furthermore posit that recent E&M research
can profit from new ideas and perspectives, for instance, when it comes to further exploring barriers to E&M implementation in practice – as most of the extant studies rely on purely descriptive and mostly quantitative data from self-disclosure surveys among practitioners: while these studies help, to a degree, paint a picture of the state of E&M and identify potential barriers, they are susceptible to the self-rationalisations of professionals and do not show empirically how strongly alleged barriers actually affect practitioners’ E&M behaviour.

Below, we will first look back and draw attention to current gaps in E&M scholarship. We will then outline the need for new perspectives and for this Special Issue. Following, we will introduce the six contributions to this Special Issue and argue how each of them provides novel ideas for the E&M debate. Finally, looking ahead, we suggest a research agenda for the 2020s in order to inspire future E&M scholarship to shed light on some of the unresolved questions, break new ground and draw insights from related disciplinary perspectives.

2. The current state of E&M research

While practices of public relations measurement can be traced back to the late eighteenth century, scholarly research and theorisation of E&M began to evolve in the late 1960s and made substantial progress through the 1970s and 1980s (for a historical overview, see e.g. Likely and Watson, 2013; Watson, 2012). Over the course of 50 years, questions related to E&M have neither lost momentum nor relevance for scholars. Earlier research, until the 2000s, focussed mainly on analyses of communication effectiveness, methodological refinements, the development of E&M models and frameworks for practice, and empirical studies of evaluation practices. While these topics continue to be prominent in the literature until today, the past two decades were marked by a trend towards importing insights from management and organisation literature, and a stronger orientation towards challenges in measuring outcomes (e.g. relationships or attitude change) as well as impact at the organisational level (e.g. reputation or brand value).

2.1 Current issues and gaps

In view of the long history of communication E&M research, recent works have summarised the debate and highlighted prevalent issues and gaps in the literature (cf. Buhmann and Likely, 2018; Volk, 2016; Likely and Watson, 2013; Macnamara, 2014). While some of these issues are geared more at the state of the practice (such as lack of outcome measurement), others focus more on the state of E&M research. From the latter we can distil the four following overarching gaps in the literature.

Heterogeneous terminologies and non-standardisation in E&M. PR and communication researchers have been closely involved in the development of various frameworks for the practice that represent the (ideal) evaluation process, such as Cutlip et al.’s (1985) often-cited Planning, Implementation, Impact Model, Lindenmann’s (1997) PR Effectiveness Yardstick or Watson’s (1997) Short Term Model and Continuing Model of Evaluation to name just a few of the early models. This has been done by employing a plethora of concepts and literature for distinguishing main phases, stages and units of evaluation. Yet, in spite of recent efforts aimed at synthesis, definitional inconsistencies remain and key constructs and terms – even “evaluation” and “measurement” – are used synonymously or confusingly (Macnamara and Likely, 2017; Schriner et al., 2017).

The prevalent heterogeneity of available frameworks and measures has led to a growing critique of a lack of standards (Macnamara, 2014; Michaelson and Stacks, 2011; Ragas and Laskin, 2014), also voiced by leading practitioners (Marklein and Paine, 2012). Countering the criticism, scholars have recently called for proposing more integrated and holistic models as well as developing unified “inventories” of measures and metrics (cf. Macnamara
and Likely, 2017; Schriner et al., 2017). Nevertheless, little is known to date about how standardisation processes evolve and what actually makes standards stick (or not).

Overemphasis on the instrumental focus. Most research in E&M follows a functionalist, positivist and partially normative approach (Volk, 2016; Macnamara, 2014), which can be attested, e.g., by quantitative practitioner surveys assessing “the state of the field” or the development of applicable best practice models for E&M. The prevalence of rationalistic and instrumental assumptions has recently been problematised for being too narrow, control focussed and organisation centric, combined with a call for more open, continuous, dynamic and expanded approaches (Macnamara and Gregory, 2018). Yet, critical inquiries and qualitative and interpretive approaches remain strongly underrepresented in the E&M literature (Macnamara, 2014).

Moving beyond instrumental towards critical questions could provide novel insights for the evaluation body of knowledge. However, few works have so far explored the possible “pathologies” in specific evaluation practices and of evaluation as a whole. For instance, previous research has shown that communicators may feign expertise in evaluation, whitewash data and produce invalid or unreliable reports (Place, 2015). From this perspective, evaluation becomes rather a form of self-justification that fixates on performance, rationality and objectivity to stabilise some particular ideology. It is conceivable that evaluators shade, overgeneralise, or tamper results to support particular goals or decisions in compliance within an established power structure. Such “pseudoevaluations” are a recognised issue in evaluation in many other domains (Stufflebeam and Coryn, 2014), but are rarely addressed in PR and communication E&M literature.

Disregard of intervening variables and prevented/hidden/indirect effects. Alongside the dominant orientation towards research with application value for practice comes the risk of overly simplistic, mechanistic and linear assumptions of communication effects. Many extant E&M models and frameworks ignore intervening variables, do not rigorously separate antecedents from outcomes of communication processes and postulate unwarranted direct cause–effect relationships. The perspective of the recipient and the social and organisational context, in which communication messages and effects are embedded, are largely overlooked (Buhmann and Likely, 2018). For instance, the mainstream E&M literature has focussed on intended messaging effects, but has remained widely silent on elusive hidden, indirect or prevented effects of communication. This is quite surprising, since preventing communication – i.e. a critical media coverage – is a legitimate and often-pursued goal of communication professionals. Following Wehmeier and Nothhaft (2013, pp. 123-124), one may even claim that PR E&M scholars may run a two-sided risk in developing E&M models and frameworks that oversimplify regarding third and intervening variables: first, such models may fail to reliably measure communication effects in practice, and hence will not aid practitioners, possibly decreasing their confidence in academic knowledge. Second, E&M scholarship risks to be perceived as an industry-driven, atheoretical field, downgrading its standing in the wider domain of PR scholarship.

Siloed debate and disconnect from other disciplines. Further criticism has evolved around the siloed approaches to E&M theorising and research. Even though the past two decades were marked by a growing import of managerial thinking (e.g. the resource-based theory of the firm or stakeholder theory, see Volk, 2016), the contemporary debate remains quite fragmented and unconnected to recent advancements in its parent discipline – communication science – as well as its embedding in the organisation context – studied in management, organisation and marketing/advertising research. One recent attempt to reach out for insights from one of the neighbouring fields has been put forward by Macnamara and Likely (2017), who suggested a disciplinary “home visit” to programme
evaluation that could inform the search for standards and overcome the long-standing stasis in practice. Strikingly, the majority of E&M literature has neglected to import highly relevant knowledge generated in the neighbouring disciplines and communication subfields.

2.2 How far have we come?
Going beyond a list of current issues and gaps in E&M research, and attempting to further assess how far E&M research has come we draw on Nothhaft et al. (2018), who refer to Shneider’s (2009) four-stage model for describing the evolutionary life cycle progress of scientific fields. Shneider (2009) suggests four stages to model the typical life cycle: in brief, Stage 1 describes the invention of a discipline and the development of a new scientific language and respective theoretical frameworks. In Stage 2, inventive scholars build on the insights of the first-stagers and develop them further to achieve breakthroughs in terms of methods and techniques. Stage 3 constitutes the apex of a discipline where methods and theories begin to inform other disciplines, standards are developed and tolerance for mistakes is low. In Stage 4, disciplines are reaching maturity and aim at safeguarding the body of knowledge in order to claim a state of the art, with a focus on refining pedagogy (Shneider, 2009, pp. 218-221; Nothhaft et al., 2018, pp. 354-355).

Of course, transferring this four-stage model has to be made with caution, not only because E&M research is rather a sub-field of PR and strategic communication scholarship than a discipline in its own right, but also because the disciplinary progress and status of the field as a whole is indeed debatable (see e.g. Nothhaft et al. 2018). According to Nothhaft et al. (2018), however, the E&M debate can be characterised as a model example of Stage 2, in which “energetic contributors […] not only develop the new language further, but develop methods and techniques that yield concrete, increasingly precise results” (p. 355). They argue that the E&M field comprises a “rather small and tightly knit” (p. 356) community, in which “colleagues work in cooperation with the industry to develop systems that are not only academically sound but work in practice” (p. 355), but they also criticise the “unsatisfactory transfer between academia and practice and practice more inspired by academia rather than academics taking the lead” (p. 356).

Indeed, the previous review of the achievements and shortcoming in the E&M field is in line with Nothhaft et al.’s assessment. We see a fairly well-defined but small community of scholars and professionals (Likely, 2018), who work on standardised models and metrics and a unified terminology – which would be one characteristic of a Stage 3 discipline. Nevertheless, E&M scholarship has remained a niche research field within the broader PR/strategic communication field – especially when compared to the popular subfields of crisis communication or OPR and dialogic theory – and consequently, a marginal side topic in the broader communication discipline. A commonly agreed theoretical basis is still missing (Likely and Watson, 2013; Volk, 2016), although there are constant conceptual developments. Little knowledge has so far been imported from neighbouring disciplines and E&M research is still far from exporting knowledge or methods and informing other disciplines. Hence, E&M research has clearly not yet reached Stage 3; however, considering the range of popular textbooks and handbooks available, there are certainly attempts to safeguard the produced knowledge, which is characteristic for Stage 4. Looking ahead, in order to progress towards Stage 3, new conceptual approaches, refined methodologies and the inclusion of knowledge produced in related disciplines are necessary for the E&M community.

3. Rationale of the Special Issue
This Special Issue emerged from a panel at the International Communication Association’s (ICA) 2018 Annual Conference in Prague, which was organised by the first author, titled “New voices in PR evaluation: innovative approaches and new research avenues for a field in stasis”. The rationale of this panel was that “new voices” are needed to move E&M out of
the diagnosed “deadlock” or “stasis” and to stimulate future progress in scholarship. The
panellists – some of them contributors to this Special Issue – presented provocative ideas
and new perspectives mostly inspired by work outside of the core PR research domain.

The underlying assumption was that, in spite of considerable scholarly efforts, the broad
majority of E&M research has explored a fairly focussed set of research questions with and
largely sustains from challenging existing explanations or developing alternative
assumptions. In following Alvesson and Sandberg’s (2014) discussion of “boxed-in
research”, we posit that much E&M research has been conducted “in the box”. While the
positive development of the E&M community towards sharing a common research interest
and coherent body of knowledge is generally appreciated, signalling the further maturation of
this sub-discipline, we share concerns voiced by Alvesson and Sandberg regarding the
limitations of boxed-in research (see also Werder et al., 2018, pp. 335-336). Despite the many
advantages of boxed-in research, Alvesson and Sandberg (2014, p. 974) raise their concerns
against a possible shortage of imaginative and interesting research, overspecialisation, silo
mentality, fragmentation, box identity, suspiciousness, intra-box communication, unnecessary
polarisation among researchers and unquestioning attitudes. Research conducted within a
specific box is “characterized by a strong and narrow focus on some issues within a
well-defined, specialized intellectual terrain” (Alvesson and Sandberg, 2014, p. 971). Some of
these negative consequences of boxed-in research may apply to E&M scholarship just as
well – for instance, the occurrence of a silo mentality and fragmentation, or the shortage of
interesting research that questions dominant concepts or attitudes. Indeed, this is the key
argument put forward by Nothhaft and Stensson (2019) in this Special Issue, who criticise the
“circumspection and narrowness” of the debate about barriers to E&M implementation, and
claim that the academic debate shows symptoms of “functional stupidity”.

In view of this assessment, we see a need to support more box-changing research that
has the potential to broaden the debate. Box breaking research occurs when scholars
“broaden their intellectual territory and research competence by moving beyond their
specific research boxes to also consider the resources and ideas of other research boxes and
intellectual terrains” (Alvesson and Sandberg, 2014, p. 968). This exactly is the ambition for
this Special Issue: to provide a platform for unconventional and alternative arguments,
interdisciplinary and integrative perspectives. We therefore particularly invited
submissions that would introduce new arguments and highlight promising starting
points for a more “diverse” future of E&M scholarship.

4. Articles in the Special Issue

Against the backdrop of the above discussion on the state of communication E&M, we turn
to the six articles contained in this Special Issue. In our view each article makes important
contributions to the current advancement of the debate, opening up new and interesting
angles for understanding the practices and contexts of E&M.

Table I provides a summary of each contribution’s primary purpose, theme, key
literature, approach and implications. In contrast to the overall trend towards a fixation on
(quantitative) empirical research in communication science as well as in strategic
communication/public relations, most of the invited articles in this volume set a
counterpoint and present conceptual arguments, partially supported by qualitative insights.
In terms of disciplinary orientation and key literature, most contributors drew considerable
inspiration from organisation research and management science. With regard to the themes,
four of the six articles shed light on the application of E&M in communication practice in
general or in specific organisations, examining different barriers that might hinder the
implementation or advancement of E&M, i.e. organisational, motivational, cultural,
structural, technical or alignment-related barriers. These articles have significant overlaps
at the outset, but propose quite alternative explanations for the existence of barriers, and
<table>
<thead>
<tr>
<th>Article</th>
<th>Theme</th>
<th>Primary purpose</th>
<th>Key literature/theory</th>
<th>Method/Approach</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim and Cappella: “Reliable, valid and efficient evaluation of media messages: developing a message testing protocol”</td>
<td>Message effectiveness, methodology</td>
<td>To propose a reliable, valid and efficient standard process for the evaluation of media messages, based on recent advances in message effects research and campaign design research</td>
<td>Message effects, message testing, persuasion</td>
<td>Meta-analysis</td>
<td>Use perceived message effectiveness (PME) and perceived argument strength (PAS) as strong proxies allowing for inferences about relative effectiveness when many messages need to be evaluated quickly and efficiently</td>
</tr>
<tr>
<td>Sommerfeldt and Buhmann: “The Status quo of evaluation in public diplomacy: insights from the US State Department”</td>
<td>E&amp;M practices</td>
<td>To understand the barriers and context around the reluctance in the practice to evaluate public diplomacy programmes</td>
<td>Public diplomacy concepts</td>
<td>Qualitative interviews</td>
<td>Assess cases of reluctant application of M&amp;E by addressing potentially disjointed systems and structures, lack of development and focus, devaluation of public diplomacy efforts, lack of resources, capital-field tensions, reliance on outputs and anecdotes, as well as misalignment of public diplomacy goals</td>
</tr>
<tr>
<td>Nothhaft and Stensson: “Explaining the measurement and evaluation stasis: a thought experiment and a note on functional stupidity”</td>
<td>E&amp;M practices</td>
<td>To explain the E&amp;M stasis in practice through consideration of practitioners’ self-interest as business people and the industry’s anomaly of expectation-inflation and overpromising</td>
<td>Principal-agent theory, theory of lemon markets, functional stupidity</td>
<td>Thought experiment, qualitative interviews</td>
<td>Test the plausibility of motivational barriers at the individual level as an explanatory factor for the stasis of E&amp;M practices; avoid narrow and circumspective research assumptions and functionally stupid explanations in theorising about barriers in practice</td>
</tr>
<tr>
<td>Murtarelli, Romenti, Miglietta, and Gregory: “Investigating the role of contextual factors in effectively executing communication evaluation and measurement: a scoping review”</td>
<td>E&amp;M practices</td>
<td>To conceptualise the role of internal contextual factors affecting the E&amp;M management process as an often-overlooked barrier of E&amp;M practices</td>
<td>Programme evaluation theory, theory-based evaluation, theory of change</td>
<td>Narrative scoping review</td>
<td>Consider the so far often neglected contextual factors influencing the practice of E&amp;M: (a) evaluative capacity and history (organisational context); (b) evaluative culture and leadership (cultural context); (c) stakeholder-evaluator relationship (relational context); (d) evaluation communicative network (communicative context)</td>
</tr>
<tr>
<td>Gilkerson, Swenson and Likely: “Maturity as a way forward for improving E&amp;M practices”</td>
<td>E&amp;M practices</td>
<td>To explicate the concept of maturity and propose a maturity model that</td>
<td>Maturity models and concepts</td>
<td>Conceptual review</td>
<td>Use the dimensions of a “holistic approach” (i.e. levels of analysis), “investment” (time, budget, tools, skills), “alignment” (process, integration) and “culture”</td>
</tr>
</tbody>
</table>

Table 1. Overview of articles in this Special Issue

Guest editorial
<table>
<thead>
<tr>
<th>Article</th>
<th>Theme</th>
<th>Primary purpose</th>
<th>Key literature/ theory</th>
<th>Method/ Approach</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>organisations' communication evaluation and measurement practices:</td>
<td>Methodology, E&amp;M</td>
<td>highlights the key dimensions of E&amp;M maturity in organisations and can serve as</td>
<td>Agility concepts,</td>
<td>Development debate</td>
<td>(motivations, relationships, standards) to assess E&amp;M maturity in organisations as a step towards, increasing both accountability and credibility for the work done by the communications function</td>
</tr>
<tr>
<td>a definition and concept explication</td>
<td>practices</td>
<td>a basis for maturity assessments</td>
<td>evaluation theory,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Van Ruler: “Agile communication evaluation and measurement”</td>
<td></td>
<td>To question the fundamental assumptions of summative, goal-dependent and</td>
<td>action research,</td>
<td></td>
<td>Reconsider E&amp;M theory and practice against the radical assumptions of agility, moving towards goal-free and on-going evaluations oriented towards users/needs and qualitative measurement methods, such as action research and sense-making methodology</td>
</tr>
<tr>
<td></td>
<td></td>
<td>organisation-centric evaluation in view of agility concepts, and to present</td>
<td>sense-making</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>alternative measurement methods addressing the needs of agile organisations</td>
<td>methodology,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Weberian idea of</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>“Verstehen”</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table I.
outline different implications for how our theorising on barriers should develop further. The other two articles focus specifically on methodological approaches to conducting E&M. While one article investigates the basic mechanisms of message effects, going back to the roots of communication science, the other article challenges the fundamental assumptions of E&M scholarship in view of the changing organisational and communicative environment.

4.1 Annotation of the articles in this Special Issue

The first study revives an older, largely abandoned debate in the E&M literature. Evaluation of message effectiveness has been one of the initial starting points of E&M research, at least since the 1970s. In fact, message effectiveness was among the most researched topics in the early 1980s and 1990s (Volk, 2016). However, the topic received little attention over the past decades and the field has not kept abreast to the discussion on message testing as it evolved especially in message effect, campaign design and persuasion research. Kim and Cappella (2019) introduce and discuss an efficient, reliable and valid message testing protocol covering how to conceptualise and evaluate the content and format of messages; procedures for acquiring and testing messages; and the application of robust measures of perceived message effectiveness and perceived argument strength. The particular strength of the approach lies in the ability to facilitate a selection of candidate messages for subsequent deeper testing, for various types of communication campaigns and for research in theory-testing contexts – avoiding the limitations of using a single instance of a message to represent a category (i.e. case-category confound). Ultimately, the contribution also shows that strong inspiration that can be drawn from adjacent fields, such as health communication, that have developed very sophisticated approaches with high potential for adaptation within the domain of communication E&M.

Sommerfeldt and Buhmann (2019) posit that, despite increased calls for enhanced monitoring and evaluation in public diplomacy, the state of practice remains grim. To better understand this situation, they undertook a qualitative study uncovering the perceptions around evaluation from the voices of those who must practice it. The authors conducted 25 interviews with public diplomacy officers with the US Department of State, both in Washington, DC as well as at posts around the world. The study shows a previously not discussed tension between diplomacy practitioners in Washington, DC and those in the field stationed outside the USA. Specifically, this pertains to a lack of joined clarity about the goals of specific public diplomacy programmes and public diplomacy as a function in general – which form the basis of targeted evaluations. The practice-level insights from this study contribute to a better understanding of the factors that may drive or hamper evaluation in day-to-day public diplomacy work. Ultimately, the article also shows that, while communication E&M face some persistent challenges which reach across professional fields (such as lack of resources, reliance on outputs or misalignment of goals), there are also unique challenges (such as capital-field tensions) that pertain to field-specific contexts and that need to be understood more closely to a particular domain of communication practice.

Nothhaft and Stensson (2019) propose a new angle to explaining the evaluation “deadlock” or “stasis” and the discrepancy between practitioners’ words and actions when it comes to E&M. While, at present, most explanations tend to focus on lack of knowledge or inadequate systems or frameworks, the authors – by means of a thought experiment and qualitative interviews with practitioners – highlight practitioners’ self-interest as well as their passive or active (but covert) forms of resistance against E&M. Using the functional stupidity concept, the authors argue that the current academic debate is conducted in narrow and circumspect ways and that we, as academics, should “take the blinders off” to allow for alternative explanations. That is, that a critical mass of actors benefits from talking about E&M, yet stands to lose from doing it. The authors posit that if the long-time neglect of E&M has led to expectation-inflation and overpromising, even well-performing actors might shy away from...
rigorous E&M, fearing being measured against promotional, not realistic standards. At the same time, on the level of industry discourse, these practitioners would still advocate for E&M in principle, so as to avoid the suspicion of underperformance.

Romenti et al. (2019) posit that while most work aims to raise the level of sophistication in practice – discussing best practice models and methods – in real life, most organisations experience issues in implementing and managing E&M that remain largely unexplored. Bringing together insights from both the programme evaluation and performance measurement literature, the paper works out key organisational, relational, cultural and communicative factors, which constitute the context of management of the evaluation process and ultimately come to bear on any E&M implementation. The identified contextual factors are clustered into four groups: evaluative capacity and history (organisational context); evaluative culture and leadership (cultural context); stakeholder–evaluator relationship (relational context); and evaluation communicative network (communicative context). Thus, the authors provide new arguments for additional and alternative dimensions in the E&M barriers/drivers literature (e.g. Buhmann and Brønn, 2018; Macnamara, 2015).

Gilkerson et al. (2019) argue that a stronger focus on the maturity concept in E&M has the potential to advance the field by increasing both accountability and credibility for the work done by the communications function. While general “maturity stages” (adolescent, advanced and mature) have been used several times to assess the advancement of E&M in organisations, this literature has not defined maturity and its dimensions. Drawing from previous work on maturity models within other fields, recent communication scholarship and industry practice, the authors propose a new theoretical conceptualisation of maturity, including the construct’s core dimensions and sub-dimensions. E&M maturity is conceptualised into four essential elements: holistic approach, investment, alignment and culture. The contribution of E&M efforts is represented as the direct support of corporate strategy, and ultimately increased value, from the communications function. Operational elements of maturity include levels of analysis, time, budget, tools, skills, process, integration, motivations, relationships and standards.

Van Ruler (2019) posits that, while most E&M models and approaches rely on the assumption that specific and quantifiable goals precede the development of any practical E&M, the reality of organisational processes can be a lot more fluid and agile, often with fast-moving targets, which calls for an introduction of the concept of agility. Based on a literature review, the author highlights central tenants of agile E&M, such as that what works is more important than what was agreed upon in advance, that agile evaluation is always formative (as opposed to summative), and that qualitative methods (e.g. action research or sense-making approaches) are key to supporting formative and fluid practices. These points challenge widely established (organisation-centric) concepts of E&M and brings the needs of users and the context of a rapidly changing environment into the equation.

4.2 Comment on the box-changing contributions of the articles

In Figure 1, we visualise how the articles in this Special Issue present research that can be characterised as box changing. According to Alvesson and Sandberg (2014), a box-changing work is rooted in a specific box, but “reaches outwards for new ideas, theories or methods that can be used to change the box in some significant way” (p. 980). The aim is hence not to further refine the existing body of knowledge, but instead to “identify crucial problems that can lead to substantial rethinking about central elements in the existing literature”. Alvesson and Sandberg categorise this as “a radical reform from within” (p. 980), but acknowledge that box changers still remain boxed-in to a certain extent, observable, e.g. in terms of a similar style, references, vocabulary or community, where scholars attend the same conferences or publish in the same journals. One step forward in this metaphor is box jumping research, which occurs when scholars leave their primary box to work with two or
several boxes and engage with significantly different topics, methods or theories simultaneously (cf. Alvesson and Sandberg, 2014, pp. 980-982). Even a step further, box transcendence, occurs when scholars aim for the “synthesis, confrontation, concept blending or bricolage of various metaphors or empirical materials”.

In light of the above, all articles in this Special Issue meet the criteria of box-changing research. They take the E&M literature as a primary reference point and then reach outwards for new contexts, ideas and theories or for new methods and draw inspiration from neighbouring fields. Using these new perspectives, they challenge the E&M box or rework box-specific assumptions. However, they also show characteristics of boxed-in research, since they share the same vocabulary, references and scholarly community. One exception is the work by Kim and Cappella (2019) who come from media effect and health communication research, but bridge disciplinary silos by engaging with the E&M literature.

The four articles focussing on E&M barriers in communication practice reach out for new contexts, ideas and theories and provoke a substantial rethinking of how we theorise the diagnosed stasis. Nothhaft and Stensson (2019) question the dominating assumption that E&M practices are inhibited by a lack of budget, time, knowledge or methods, and instead propose an alternative and provocative assumption that rigorous E&M might simply not be part of practitioners’ self-interests. Using a thought experiment they draw on the concepts such as “lemon markets”, principal–agent theory, and functional stupidity to emphasise that this explanation for the E&M stasis has been overlooked by scholars so far.

The works by Romenti et al. (2019) and Gilkerson et al. (2019) identify further crucial problems in the current debate about barriers – the omission of the organisational context in which communication E&M is embedded – and reach out towards programme evaluation, performance measurement literature, and maturity conceptualisations and models for inspiration. Instead of merely adding to the literature on barriers by
conducting another quantitative study diagnosing a stasis in practice, they introduce novel theory-based explications. Their findings imply that future investigations into E&M barriers need to consider further contextual factors and go beyond self-disclosed surveys of communication professionals.

Sommerfeldt and Buhmann (2019) add to the research on barriers from a much-needed qualitative perspective, producing insights form the PR-related field of public diplomacy. Their findings highlight resemblances with studies conducted in the professional fields of strategic communication and public relations more generally, and they point out specifically to so far overlooked tensions regarding E&M between central and peripheral units of organisation.

Van Ruler (2019), in contrast to the common emphasis on the relevance of goal-setting for evaluations (cf. e.g. Hon, 1998 but also Sommerfeldt and Buhmann in this Special Issue), challenges the very fundamental assumptions of “mainstream” E&M research, and criticises dominating summative and organisation-centric approaches to E&M. Van Ruler instead reaches out to agility research and calls for a shift towards action research or sense-making methodologies.

5. A research agenda for the 2020s

The articles collected in this Special Issue open up promising new and alternative avenues for future research and demonstrate the potential for the continuous reflection of already established assumptions in the “E&M research box”. Looking ahead, we believe first and foremost that research in the 2020s can profit from increased efforts particularly to understanding and explaining E&M and its manifestations in practice, rather than aiming at improving practices. Of course, producing theories with a “cash value” (Toth, 2002) is a legitimate goal for an applied research field such as E&M and there is indubitably a need in practice for scientifically sound but efficient methods, measures, models, frameworks and tools (Wehmeier and Nothhaft, 2013). Grunig (2006, p. 152) for instance famously claimed “public relations scholars need to develop both positive and normative theories – to understand how public relations is practiced and to improve its practice [...].” But improving practices and fostering professionalism through the development of positivist, prescriptive, functionalist or normative scholarship should not be the only goal. Fundamental research into the underlying logics of E&M practices and critical and interpretative examinations are just as relevant – and currently still starkly underemphasised.

Following this line of thought, and responding to earlier calls (e.g. Wehmeier and Nothhaft, 2013; Aggerholm and Asmuß, 2016), we advocate for research employing a practice perspective on E&M theorising and studies (e.g. see Whittington, 2007; Czarniawska, 2008; Jarzabkowski et al., 2007; Sandberg and Tsoukas, 2011). We believe that a turn towards a practice perspective would allow scholars to better reflect the logics and (ir)rationalities of practice and, e.g., shed new light on the often-discussed barriers of E&M practices. By observing and studying how E&M practitioners act and interact in the organisation on a day-to-day basis and with which motivations and consequences, scholars would gain a deeper understanding of the social mechanisms underlying E&M practices (Sandberg and Tsoukas, 2011; Aggerholm and Asmuß, 2016).

Looking ahead, below we outline a research agenda with eight possible directions for future research for the 2020s. This agenda builds on both the existing and established discussion on the state of the field in general as well as on the suggestions proposed in consequence of the 2018 ICA panel discussions, and the work presented in this Special Issue:

1. Towards clarifying key terminology and understanding standardisation processes:

   - How can vague, inconsistent definitions and diverging approaches to operationalisation of key constructs be harmonised? What should be the basic definitional elements of key E&M terms in the field?
• How do standards as formulated rules for common and voluntary uses actually evolve from their early development to later application and compliance (or not)? What is the role of central actors in standard setting and what are dynamics of the process of standardisation in E&M?

• What is the value of scalable (or customised) standards for E&M practice? What is the value of accessible standardized databases in which E&M data collected across organisations could be integrated, synthesised, and compared?

(2) Towards fundamental research aimed at understanding E&M practices:

• How are E&M frameworks/models and methods actually used in the practice (with what motivations, to what end)? How do practitioners get away with avoiding evaluation or using invalid or vanity metrics? (Macnamara, 2018).

• What is the power/pressure context in which E&M practices evolve (or not)? Which variables explain the variability in the adoption of (mature) E&M systems?

• What are the drivers and barriers to conduct E&M, at the industry level (i.e. anomaly of overpromising), organisational level (i.e. evaluation culture, management support, evaluation capacity, history, structures, etc.), department level (i.e. leadership, alignment, stakeholder relations, life cycle, etc.) and individual level (i.e. motivations, expertise, etc.)? (Romenti and Murtarelli, 2018)

• How do E&M fads and fashions evolve and spread (and disappear)? How do challenges of implementing evaluation correspond to other related practical challenges, i.e. the formulation of strategically aligned goals and their operationalisation into measurable objectives?

(3) Towards understanding intervening variables in E&M:

• Which theoretical frameworks and concepts from communication science (i.e. priming, cultivation, agenda setting, information processing, persuasion, etc.) can be integrated in E&M frameworks and models to better conceptualise the nature of communication processes?

• Which studies can inform the modelling of relations between antecedents and outcomes of communication in order to make more cautious interpretations of causality?

• How can we better explain the impact of mediating factors, including, e.g. psychological factors (e.g. influence of personality traits on perceptions and emotions, cognitive dissonance) or social factors (e.g. influence of social networks and group integration on attitudes)?

(4) Towards understanding prevented or hidden communication effects:

• Which frameworks and models (i.e. crisis/risk communication, conflict resolution, etc.) provide a basis for conceptualising and measuring prevented communication effects? How do hidden or indirect communication effects manifest empirically?

• How can internal evaluation services (i.e. media intelligence, strategic insight analysis) provided by communication departments to inform organisational decision making be integrated in conceptualisations of E&M frameworks? How can the value of counselling and advising organisational leaders, based on communication E&M insights, be conceptualised?
Going beyond positivist epistemology:

- Which paradoxes or conflicting patterns of E&M have been discussed in the literature and can be observed in practice? In which contexts do communication E&M logics conflict with the overall logic of the organisation in relation to evaluation (i.e. reluctance to quantification vs measurement-driven cultures, etc.)? (Romenti and Murtarelli, 2018)

- How are E&M results used for ritual, legitimacy purposes of signalling rationality (i.e. pseudoevaluations, myths of rationality)? How do E&M tools and results shape and change perceptions of organisational realities (i.e. performativity)? Which uses of E&M results are harmful for the organisation (i.e. overly trust in data, blind spots, functionally stupid evaluations, etc.)? (e.g. see Falkheimer et al., 2016; Wehmeier, 2006)

Incorporating critical thinking:

- Which perspectives (i.e. critical, sociocultural, postmodern, etc.) can help to overcome the extant functionalist organisation-centric assumptions? What are the shortcomings and deficiencies of contemporary theorising about the fundamentals of measurement of communication effects (e.g. Thummes, 2009)? How can we better conceptualise the (unintended) societal implications of communication effects?

- Which ethical quandaries exist in E&M practices (i.e. misrepresentation or whitewashing of data, deceiving data display, use of vanity metrics, e.g. Place, 2015)? In which cultural situations or industry contexts do unethical E&M practices evolve? What are the hidden motivations of practitioners and which consequences arise from unethical E&M behaviour?

Exploring and criticising existing measures and metrics:

- Which metrics assist in the measurement and valuation of immaterial capitals (i.e. social capital, human capital and intellectual capital) (e.g. see De Beer, 2014)?

- Which measurement methods and metrics are applicable and useful in agile organisational contexts with dynamic communication goals and fluid organisational targets?

- How can the cognitive and behavioural measures typically used in communication E&M be made compatible with overarching organisational business, performance and valuation metrics? How can advancements in other professional fields (i.e. online marketing/advertising, data science and artificial intelligence, accounting and controlling standards, etc.) inform the further development of metrics for communication E&M (i.e. social media metrics, big data analytics, etc.)?

Bridging disciplinary siloes and aiming for synthesis:

- Which recent discussions in the neighbouring disciplines can be integrated in our research field to resolve the continuing fragmentation into siloed debates and make our research more compatible, particularly from…?
  - media effect research, media psychology, public opinion/audience/reception research, health communication and public diplomacy;
  - programme evaluation and performance measurement/management;
– critical organisation and management research;
– strategy and alignment research;
– (online/social) marketing measurement, market research, brand evaluation and advertising effectiveness research;
– accounting and controlling research; and
– data science and data analytics.

- And vice versa, how can we enrich and inform the evaluation debates carried out in neighbouring fields with findings from our field (i.e. by re-conceptualising simplistic understandings of communication processes)?

- Which economic theories help to better reflect the economic dimension of the organisational setting in which communication effects and E&M takes place?

- Which implications will blurring boundaries and increasing convergence among communication fields, i.e. marketing, branding, advertising, journalism and strategic communication, have for E&M?

Hopefully, this collection of eight directions for future research as well as the arguments provided in the six articles in this Special Issue might inspire scholars to walk some new avenues in the 2020s. We hope that further work will follow the signposts for box-changing research and consider the resources and ideas of other disciplines’ research boxes and intellectual terrains, thus potentially contributing to further opening up the E&M research box.

Sophia Charlotte Volk
Institute of Communication and Media Studies, University of Leipzig, Leipzig, Germany, and

Alexander Buhmann
BI Norwegian Business School, Oslo, Norway

Acknowledgements
This Special Issue is based on the collective efforts of many authors, panellists and attendees of the ICA panel in Prague (2018), as well as reviewers, who provided valuable feedback to earlier conference papers and presentations and later drafts of the articles. The authors would like to thank everyone who contributed in this Special Issue: all submitting authors and reviewers, who brought great dedication and expertise to each manuscript. The authors are also very grateful to the Journal Editor Jesper Falkheimer, who approached the authors with the idea for this Special Issue following the 2018 ICA panel, and the Editorial Assistant Hui Zhao, who expertly and patiently guided the authors through the whole process. The authors furthermore thank the members of the IPR Task Force on Standardisation of Communication Planning/Objective Setting and Evaluation/Measurement Models, which had provided the forum for the initial 2018 ICA panel idea and has been, since its founding in 2015, a platform where many of the issues addressed in this Special Issue have been first discussed, in particular: Fraser Likely (President and Managing Partner, Likely Communication Strategies), Forrest Anderson (CEO, Forrest W. Anderson Consulting), Dr Mark-Steffen Buchele (CEO, Buchele CC), Dr David Geddes (Managing Director Geddes Analytics LLC), Dr Nathan Gilkerson (Marquette University), Dr Jim Macnamara (University of Technology Sydney), Dr Rebecca Swenson (University of Minnesota) and Michael Ziviani (CEO, Precise Value).
References


**Further reading**


Reliable, valid and efficient evaluation of media messages

Developing a message testing protocol

Minji Kim
Center for Tobacco Control Research and Education,
University of California San Francisco, San Francisco, California, USA, and
Joseph N. Cappella
Annenberg School for Communication,
University of Pennsylvania, Philadelphia, Pennsylvania, USA

Abstract

Purpose – In the field of public relations and communication management, message evaluation has been one of the starting points for evaluation and measurement research at least since the 1970s. Reliable and valid message evaluation has a central role in message effects research and campaign design in other disciplines as well as communication science. The purpose of this paper is to offer a message testing protocol to efficiently acquire valid and reliable message evaluation data.

Design/methodology/approach – A message testing protocol is described in terms of how to conceptualize and evaluate the content and format of messages, in terms of procedures for acquiring and testing messages and in terms of using efficient, reliable and valid measures of perceived message effectiveness (PME) and perceived argument strength (PAS). The evidence supporting the reliability and validity of PME and PAS measures is reviewed.

Findings – The message testing protocol developed and reported is an efficient, reliable and valid approach for testing large numbers of messages.

Research limitations/implications – Researchers’ ability to select candidate messages for subsequent deeper testing, for various types of communication campaigns, and for research in theory testing contexts is facilitated. Avoiding the limitations of using a single instance of a message to represent a category (also known as the case-category confound) is reduced.

Practical implications – Communication campaign designers are armed with tools to assess messages and campaign concepts quickly and efficiently, reducing pre-testing time and resources while identifying “best-in-show” examples and prototypes.

Originality/value – Message structures are conceptualized in terms of content and format features using theoretically driven constructs. Measures of PAS and PME are reviewed for their reliability, construct and predictive validity, finding that the measures are acceptable surrogates for actual effectiveness for a wide variety of messages and applications. Coupled with procedures that reduce confounding by randomly nesting messages within respondents and respondents to messages, the measures used and protocol deployed offer an efficient and utilitarian approach to message testing and modeling.

Keywords Communication strategy, Health communication, Marketing communications, Persuasion

Paper type Research paper

From the 1970s to the 1980s, scholarly discussions in the field of public relations evaluation and measurement research started to evolve in the direction of evaluating message effectiveness, often in the context of internal communication or PR campaign design. Grunig’s groundbreaking “Excellence Study” introduced a typology for PR evaluation and measurement, of which message evaluation was the first of five ascending levels (Grunig et al., 2002, p. 91). According to
the typology, evaluation should be conducted at the, first, message or publication level; second, program or campaign level; third, function or department level; fourth, organizational or enterprise level; and fifth, societal level. Although research into the first two levels – message and campaign effectiveness – proliferated until the 1990s, since then it has lost scholarly attention in favor of investigations directed toward evaluating communication effects on the level of the communication function and linking these to organizational effects such as performance (Likely and Watson, 2013, p. 156).

Current discussion of message effectiveness, evaluation and measurement in the field of PR and strategic communication has become scarce. This is quite surprising, given the starting points of the debate, but also because message effects research is an important topic in neighboring fields such as media effects research, media psychology and mass communication. However, the knowledge produced in these fields has been seldom referenced in public relations and strategic communication research. This silo effect of fields so closely aligned is unfortunate. It seems reasonable and valuable to re-connect an historically important debate in public relations and strategic communication back to its roots in communication science with the hope that researchers in both silos will be informed.

Assessing the effectiveness of messages is crucial in many areas of communication research from campaign design to theory testing. Serious, careful evaluation is very expensive in terms of time, financial and human resources but, at the same time, it is necessary to avoid expenditures of campaign funds on weak messages or theory tests with messages that fail to achieve change or to control for confounders linked to message exposure. This research has sought to advance efficient, reliable and valid measures of message effectiveness as a complement to “gold standard” message evaluation that focuses on long-term behavioral consequences for a small set of focal messages in comparison to control groups. Instead, our protocol and measures allow the ranking of many messages on their likely persuasiveness (Bigsby et al., 2013) and argument strength (Zhao and Cappella, 2016), inviting the selection of candidate messages for theory testing, campaign use, or more intensive testing. Efficient procedures for evaluating the effectiveness of messages for the purposes of persuasion would be a useful tool for those designing public health campaigns, political and social issue campaigns, and certainly for researchers seeking to evaluate various types of message effects as they undertake theory testing.

Public health campaigns (Noar, 2009; Hornik, 2002; Hornik and Yanovitzky, 2003; Slater, 2006; Wakefield et al., 2010) and public relations and strategic communication campaigns (Buhmann et al., 2018; Macnamara, 2018; Volk, 2016) often select message-based interventions that need to be effective before in-depth and behavioral testing. Our approach complements campaign planning, development and evaluation in these domains by focusing on messages and their persuasive value to those in the targeted populations. The approach described in detail, in what follows, does not address the stasis that Macnamara (2018) laments in strategic communication nor the need for measures of reputation and relationships that Volk (2016) finds lacking in PR campaign evaluations. The measurement of message impact that is proposed is responsive to James Grunig’s admonition about a more careful and serious measurement that Buhmann et al. (2018) remind us is now 35 years old. In this paper, a protocol is offered that can be followed to study or evaluate messages that are candidates for inclusion in communication campaigns or research studies. Message evaluation is just a piece, though a crucial piece, of campaign design and implementation. The method of description and evaluation of candidate messages is then provided.

**Message testing for theory, design and campaigns**

Interest in message assessment and message effects research in the field of communication has undergone a renewal. About 40 years ago, Jackson’s (1992) book and Bradac’s (1989)
compendium provided impetus for message effects research and principles of study design and analyses of messages. Interest in message effects research and methods has seen a significant revival in published scholarship (O'Keefe, 2015, 2018; Reeves et al., 2016; Slater et al., 2015; Nabi, 2018) as well as in conference activities (e.g. a 2019 ICA pre conference on message testing; a similar panel at the 2018 University of Kentucky Health Communication conference).

The research on message testing has focused primarily on drug use, smoking of combustible tobaccos and now modified risk tobacco products. However, these narrow contexts are built around a broad-based method. This method is summarized in sections below, treating a conceptualization of messages and message data, allowing synthesis through analysis; a protocol for message testing including specifically efficient, reliable, and valid measures of message persuasiveness and argument quality; and prototypical procedures followed in various types of empirical testings. The discussion concludes with limitations and it addresses counterarguments that have been raised.

Conceptual and operational approaches to messages
The kinds of persuasive messages that have been treated in our empirical studies include video-based Public Service Announcements (PSAs), still image advertisements about public health, and text-only persuasive appeals. The approach to complex stimulus materials is through analysis and re-synthesis. Analysis breaks messages into their executional features (e.g. pace of scene changes, presence of human characters) and content features (e.g. self-harm vs other-harm, episodic and thematic frames).

Executional features
These are elements of a message that identify how content (what is said) is delivered to the audience. For example, a climate threat can be delivered in expository form or through testimonials or via a narrative account (Murphy et al., 2015; Shen et al., 2015; Zillman and Brosius, 2000). The core claims would remain approximately the same but the vehicles of delivery – how the content is executed or delivered – change.

There are, of course, an infinity of executional features attributable to messages. Our work has been guided by two broad sets of such features: message sensation value (MSV) (Kang et al., 2006; Strasser et al., 2009) and information introduced both audio and visual (I²) (Lang et al., 1993, 2006).

MSV describes a set of message features that can function to attract attention. The features are derived from theory about the sensation-seeking personality (Zuckerman, 1979; Zuckerman and Kuhlman, 2000) and are presumed to elicit sensory, affective and arousal responses (Éverett and Palmgreen, 1995; Palmgreen et al., 1991; Harrington et al., 2003). These structural features include formal video features (e.g. cuts, edits, special visual effects, etc.), formal audio features (e.g. sound effects, music, voiceover, etc.), and content format features (e.g. act out vs talking head, surprise ending) (Morgan et al., 2003). Empirical studies have generally supported the impact of certain MSV features (e.g. fast-paced editing) on arousal, using both physiological measures such as heart rate and skin conductance (Detenber et al., 2000; Lang, 1990, 1993, 1999) and self-reported measures (Éverett, and Palmgreen, 1995; Stephenson and Palmgreen, 2001). Some individual features of MSV, such as edits, cuts (Basil, 2014; Geiger and Reeves, 1993; Niederdeppe, 2005), visual graphics (Thorson and Lang, 1992), pace (Bolls et al., 2003; Yoon et al., 1999) and emotional messages (Hitchon and Thorson, 1995; Lang et al., 1996) have shown effects on attention, arousal, memory and cognitive capacity (Lang, 1990; Kang et al., 2006). Video messages can be readily and reliably coded for their sensation value (Kang et al., 2006; Strasser et al., 2009), providing important information about a message’s ability to attract attention and place processing demands on the audience.
Information Introduced ($I^2$) is a measure of televised message complexity (Lang et al., 2006). The measure is designed to tap the cognitive resources required and employed in the processing of messages. Coding rules are derived from structural changes known to produce cognitive and physiological responses. Camera changes (CCs) produce orienting responses (Lang et al., 1993; Geiger and Reeves, 1993) but the information introduced following a CC can alter demands on the resources viewers allocate. After each CC, $I^2$ codes the following: object change, new object, relatedness to the prior image, distance closer to focal object, perspective change and emotion change. Lang et al. (2006) detail the cognitive and physiological bases for each of these codes. The result is a measure that tracks the sometimes rapid changes in cognitive demand by a video message.

Information introduced offers a more continuous measure of cognitive demand than MSV even though the two are similar. The theoretical and empirical relationships between these measures require scrutiny, but together they offer insight into the structural relations within complex video messages.

Content features
These are elements of a message that refer to what is said. Like with executional features, the number of codable content features is infinite. They can include themes (e.g. consequences for self vs others), types of appeals (evidentiary vs authoritative), framing of content (Cappella and Jamieson, 1997; Rothman et al., 2006) and perhaps, most importantly, argument strength among many other possibilities. Argument strength appears to be the most commonly manipulated message feature in persuasion research (Johnson et al., 2005). When audiences process information centrally (Petty and Cacioppo, 1986), strong arguments produce more attitude and belief change than weak ones (Johnson et al., 2004; Johnson and Eagly, 1989). The most widely used measure of argument strength is thought listing, which assesses an audience’s cognitive responses to persuasive messages (Brock, 1967; Petty and Cacioppo, 1986; Petty and Wegener, 1998). Messages that elicit predominantly favorable thoughts are likely to change attitudes in the advocated direction, whereas messages eliciting predominantly unfavorable thoughts typically do not. An alternative to thought listing is introduced below as an efficient, reliable and valid measure of argument strength (Zhao et al., 2011).

Overall perceived message effectiveness (PME)
In addition to coding executional and content features and to rating the explicit and implicit arguments in messages, the message’s effectiveness needs to be assessed overall. McGuire’s (1989) influential description of the persuasion process implies that there are several necessary conditions that must be met if a persuasive undertaking is to be effective. These include exposure, attention, comprehension, storage, retrieval and acceptance. Our measure of PME focuses on acceptance or the perception of the convincingness of the message. This focus is not to deny the importance of other elements of persuasion in a message’s effectiveness overall, but privileges the final step in McGuire’s (1989) sequence – namely, its ability to generate cognitive acceptance by the targeted audience. The measure we have been using to assess PME will be described in more detail below in terms of its validity and reliability.

Message testing data structures
In Figure 1, the data generated by our conceptual and operational approach to message testing are described. A given message (textual, video, pictorial, etc.) is coded for its executional and content features, including ratings of argument quality. Coding is done by trained assistants, adhering to normal standards for reliability (Krippendorff, 1980). Rating of arguments in the message — whether implicit or explicit[1] — is done by members of the target population. Their aggregate ratings of argument strength are treated as a message-level
characteristic, similar to any other coded feature. PME scores for the message as a whole are obtained from members of the target population as well, but they are distinct from the sample asked to provide argument quality ratings.

So the data provided from message analysis include executional and content features. These features can be stored alongside PME evaluations for every evaluator exposed to the message. With the testing of many instances of the targeted category of persuasive messages, the data set available will potentially be very large, depending on how many messages each person evaluates and how many executional and content features are coded.

Data of this sort are very flexible, allowing message-level analyses of PME, with or without consideration of the coded features, and person-level analyses, with or without consideration of coded features. Ultimately, the reconstitution of the message’s overall effectiveness can be obtained from aggregate or individual data predicting PME from executional and content features and their interactions. (Some examples are presented later.)

The conceptualization of messages offered can be made efficient in a variety of ways. We next describe a protocol for message testing that is both efficient as well as reliable and valid, ultimately allowing a data array for message testing that is extensive and useful.

A protocol for message testing

A variety of techniques have been employed in the literature to carry out message testing with varying degrees of depth, breadth, focus, cost and verisimilitude to the ultimate context of application. Some of these include cognitive testing, talk-aloud procedures, in-depth interviews, among other approaches (Schwarz and Sudman, 1996). Two common ways to evaluate messages prior to actual implementation are focus group interviews and quantitative surveys[2]. With focus group interviews, small groups (usually 7–9 people at a time) from the target audience discuss the message(s) with a moderator, and the content is subject to qualitative thematic analyses for the messages’ strengths and weaknesses. Sometimes, the focus group members see a single example message and discuss ways to make it more effective for the target audience (e.g. Cates et al., 2004); in other forms, participants see multiple stimuli, and results are used to compare the audience's reactions across the messages (e.g. Gallopel-Morvan et al., 2011). Although focus group interviews provide insight into detailed audience responses, they should be conducted and examined with care (Atkin and Freimuth, 2001). Focus groups that are carefully moderated maintain awareness of sample size limitations, and the presence of well-known group biases – such as censoring, conformity (Carey, 1994) and group polarization (Sim, 1998) – can provide useful information.

Figure 1. A schematic summary of message content and delivery features plus external evaluation of overall message via PME by members of the targeted population.
The gold standard for message testing would be to conduct a test in the field with the appropriate target population. For example, PSAs to be used in a communication campaign can be compared on the outcome variables of interest against an appropriate control group. For messages aimed at behavior change (assuming the proper design and controls), this is perhaps the most accurate way to measure message effectiveness. Unfortunately, this “gold standard” strategy requires running at least a small-scale campaign in the field — several mini-campaigns if there are multiple messages to be tested — and then evaluating its success in changing behavior. This approach can take long periods of time and it can be very costly, rendering it impractical and expensive. The gold standard needs an alternative approach: simpler and more efficient with sufficient validity and reliability to allow message selection from an initial pool, perhaps allowing later testing against a higher, behavioral standard.

The obvious problem with the “gold standard” protocol is its inefficiency. Researchers and campaign designers need to know about the effectiveness of messages before testing the final campaign and before deploying campaign resources, rendering long-term observation of behavioral change virtually impossible while maintaining timeliness. Even preliminary versions of messages (e.g. “concepts”) need to be evaluated for their effectiveness, so the best message concept is chosen (and the least effective avoided) for production and implementation.

The protocol that we have been using lies somewhere between the small sample, qualitative, short term, verbal reports of focus groups and the resource-intensive, quantitative, long-term behavioral measures of the gold standard of actual effectiveness (AE). Our protocol involves large samples of respondents and messages, it is quantitative, it is efficient in data acquisition and in measurement, and it is both reliable and valid if not overinterpreted.

Our protocol has the following components (Sketched in Figure 2):

1. Valid and reliable measures of the core outcomes that are also efficient, allowing rapid responses to multiple instances of testable messages. The core measures developed include perceived argument strength (PAS) and PME, both of which tap into message acceptance.

---

**Figure 2.**
Sequence of random selection and random assignments of messages and raters for PME. Generates message by person data matrix with many structural zeros.
A large pool of messages to be evaluated that meets the needs of the proposed campaign or research project. Variation in message features and outcomes is a necessity.

Evaluators of the messages from the target population.

Each evaluator being exposed to multiple messages selected randomly from the pool of messages and presented in random order.

Enough evaluations by respondents per message to ensure stable estimates of outcomes.

Before elaborating each of the elements of the protocol, its advantages should be noted. It develops simple, efficient and valid measure of PAS and PME in order to be able to assess a large number of messages quickly, yielding a reasonably accurate, approximate rank order in terms of their potential to create acceptance – persuasive and convincing – in the sense of McGuire (1989). It allows the evaluation of many messages, which is important at the early phases of message triage. Simple random selection and random ordering in presentation – double randomization – avoid biases when message scores are aggregated. Coupled with the message coding procedures described earlier, the message testing data made available to those responsible for selection will be extensive and detailed, allowing careful selection for effectiveness and for comparability on competing features.

Obviously, the stated protocol will be an effective and efficient one if its implicit assumptions and claims are met. Some features can be obtained by design – target populations, randomization, numerous cases, variable messages. But do we have reliable and valid and efficient measures?

An efficient, valid and reliable measure of perceived argument strength (PAS)

Zhao et al. (2011) conducted a series of studies to develop an efficient measure of argument quality to replace the “thought listing” standard developed by Petty and Cacioppo (1986). The procedure first extracted arguments used in a large number of real-world anti-drug and anti-smoking campaign PSAs. These were evaluated on a wide variety of potential scale items in samples of adolescents and adult smokers, respectively. The items were analyzed for their dimensionality, internal consistency, and construct validity. Single factor solutions provided adequate fit for both samples with strong internal consistency.

In both studies of Zhao et al., average PAS scores for individual participants (who evaluated a random subset of all the arguments) correlated significantly with subsequent behavioral intention ($r = 0.35$ and $-0.44$, respectively), providing evidence of predictive validity. Zhao and Nan (2010) showed results that were consistent: congruence between PAS ratings and typical message responses, such as liking, message-evoked emotions, attitude, and intention. In a test of discriminant validity, the need for cognition and PAS ratings ($r = -0.03$) were unrelated. The need for cognition should not influence assessment of argument strength because it is only supposed to affect level of elaboration.

The key validation test conducted by Zhao et al. (2011) evaluated PAS against thought listing, which is the standard measure of strong vs weak arguments in support of a senior comprehensive exam policy (Petty and Cacioppo, 1986). The student respondents completed thought listing for a random half of the arguments and rated the other half using the PAS scale. Thus, for each argument, two separate indicators of argument strength were obtained, one based on thought favorability, the other, the PAS scale, from two different subgroups randomly assigned. These two indicators correlated strongly across arguments ($r = 0.77$), providing evidence of convergence between the traditional thought-listing procedure and the PAS scale.

Versions of the scale have been used (or adapted) in numerous testing contexts. The testing procedures usually involve multiple messages per condition that differ in PAS. Importantly, the messages selected to be high and low in PAS are evaluated by one set of
respondents from the target population, and then the testing of the effects of strong and weak arguments is carried on a separate but comparable population. Those evaluating the arguments for PAS are never the same people, responding cognitively or behaviorally to the strong and weak messages. The two groups are completely distinct.

Messages varying in PAS have had predictable main effects on heart rate and skin conductance (Strasser et al., 2009; Kang et al., 2009), cotinine levels (Wang et al., 2013), neural activation (Weber et al., 2014), and genetic differences and brain reactivity (Falcone et al., 2011). Similarly, messages varying in PAS have had predictable moderating effects on visual attention (Sanders-Jackson et al., 2011), objective memory (Lee and Cappella, 2013), and physiological responding (Kang et al., 2009). These biobehavioral outcomes provide solid predictive validity data for the PAS scale on outcomes that are not simple self-reports. The full measure is available in Zhao et al. (2011) and a shortened version is available in Zhao and Cappella (2016) or by request from the authors.

**An efficient, reliable and valid measure of PME**

The gold standard for message testing assesses some form of AE related to behavior, attitudes, beliefs, or some outcome designated as pertinent. However, behavioral testing is quite difficult in general and sometimes virtually impossible in the short run (e.g. changing dietary habits) or with large numbers of competing messages. One class of alternatives is to develop measures of PME rather than the AE of messages. PME focuses on ratings of whole, intact, often "real world" messages, whereas PAS focuses only on the argumentative content of the message, textual or visual.

Early work by Fishbein et al. (2002) initiated a four item scale to evaluate 30 anti-drug PSAs in part to argue the importance of some form of message testing in advance of a large scale anti-drug campaign being contemplated. This early scale showed some predictive validity, differentiating hard drug PSAs from those focused on softer drugs such as marijuana. Dillard et al. (2007) advanced scale development, considerably carrying out five experimental studies to evaluate the causal direction between PME and AE. They concluded that PME "may be viewed as a causal antecedent of AE" (p. 467). A meta-analysis of the PME–AE relationship from 40 studies reports a highly significant effect size of $r = 0.41$ (95% CI (0.38, 0.43)) (adjusted for attenuation) (Dillard et al., 2007).

Our own measure of PME is built on the work of Dillard and his colleagues, crafting a scale with high internal reliability and good construct and predictive validity. Three tests of the predictive validity of PME are worth describing. Bigsby et al. (2013) exposed smokers to four anti-smoking ads, randomly selected from a large pool. Since the anti-smoking ads varied in their degree of PME, the amount of aggregate "persuasive impact" of the set of ads for any one person varied from person to person. To get a measure of the aggregate impact of the four ads experienced by each person, two summations were carried out. The first averaged the PME score for each ad across those who viewed and rated the ad. The second sum aggregated the total aggregate PME across the ads seen by that person wherein each PME score for each ad was the average from the first sum. The distribution of PME as "persuasive impact" across the set of raters is depicted in Figure 3. The distribution is fairly normal, indicating a good distribution on aggregated PME.

The degree of aggregate PME predicted the individual respondents’ intention to quit and to reduce consumption of tobacco. This validation test is important in several respects. The outcome behavior is non-trivial – intention to quit and reduce tobacco consumption. Such an outcome is not likely to result from exposure to a single tobacco control ad. In addition, the outcome measure – an individual’s intention – is not a possible causal predictor of the aggregate PME to which the respondent is exposed. Basically an individual’s intention to quit could not predict the group’s aggregate evaluation of the PSAs’ persuasiveness. So causal
direction is preserved. This is direct evidence that rated PME causally affects a non-trivial outcome, namely intentions to reduce and even quit tobacco use.

A similar procedure was conducted with exposure to a random selection of tobacco warning labels and intentions to quit (Morgan et al., 2019). These data are not presented here, as the study is not yet published, but the procedures and results are very similar. In a third study, anti-smoking PSAs tailored to be high in PME for the individual smoker predicted intentions and subsequent quitting behaviors (Kim et al., accepted). The study by Kim used reported quit attempts at 2–3 week follow-up as an outcome, taking predictive validity one step further.

Multiple stimuli
Jackson’s (1992) message effects research argues that the use of multiple messages to represent an underlying construct is a necessity to avoid what has been called the case-category confound. The basic idea, reiterated by O’Keefe (2015), is that when researchers use a single message to test an underlying construct of which the message is an instance, other factors that covary with the construct might be responsible for the observed effects.

Although single message experiments are employed by the majority of studies in media psychology (Reeves et al., 2016), this approach restricts the ability to generalize to a larger category of messages as well as to limit inferences about construct validity. Using several instances – each representing the underlying construct but in different messages that vary in ways other than the core construct (e.g. Pechmann and Reibling, 2006) – increases the likelihood that the observed effect is due to the construct and not some other irrelevant factor. Other recent reviews advance the discussion on stimulus sampling and research design using multiple messages while also advocating the core idea that testing using multiple messages within and between studies is a good practice (Slater et al., 2015).

When analyzing data from multiple message experiments, the statistical comparisons should be made at the construct level – ideally with planned contrasts – not between all
possible instance pairs. In this way, the practice of testing and using multiple messages helps avoid false positives and the replication problem (Open Science Collaboration, 2015) that has plagued social, behavioral and medical sciences by using multiple instances of the same construct rather than a single instance as if it were the prototype of the construct being tested.

Accepting the multiple message approach in message effects studies to test theory-derived hypotheses requires evaluating many messages for their effectiveness rather than arbitrarily using a single example of the broader category. The researcher must begin with a pool of messages, the larger the better, from which to choose. Testing them all – or a significant sample of them – requires an efficient testing method.

**Evaluators from the target population**

One obvious requirement in message testing is to use evaluators who are members of the targeted population. A reasonable assumption is that audiences will vary in how they respond to messages, especially if the message is not intended for them or is viewed as irrelevant. Recently developed resources from commercial vendors facilitate access to these samples using nationally representative or opt-in online panels at costs that are not prohibitive. Although the development of crowdsourcing resources such as Amazon’s Mechanical Turk also allows access to targeted segments of populations, one must exert care in the filtering process and be vigilant to inattention by participants (Berinsky et al., 2012; Chandler et al., 2014).

**Multiple messages per evaluator**

One way to increase the efficiency of message testing would be to show each evaluator more than one message, with the messages in the evaluation set presented in random order, different for every respondent. For example, suppose that we needed to evaluate 100 PSAs. Obtaining 50 evaluations per PSA, with one exposure per evaluator, would require 5,000 participants. But if each evaluator sees 5 PSAs randomly selected from 100 (and presented in a unique random order for each rater), then 1,000 participants would provide 50 evaluations per PSA for a total of 5,000 evaluations. Survey duration is the obvious criterion in deciding the number of messages per participant, as it affects not only the study budget but also participants’ fatigue that can result in reduced reliability. We have showed four messages per evaluator in the case of 30-second video messages (e.g. Kim et al., 2016); when using textual arguments, which are only one or two sentences long, we have used varying numbers between 6 and 12 (e.g. Zhao et al., 2011).

When many messages are being evaluated in a study and each respondent is receiving two or more messages with random selection, then both messages and respondents are random variables, in that messages are randomly assigned to people and people are randomly assigned to messages (see Figure 2). The order of presentation is also random. Although the double randomization is a boon to inference, it does require some fancier statistical analyses (Raudenbush and Bryk, 2002). With messages randomly assigned to people, any natural variation that accompanies the random assignment (e.g. degree of message effectiveness) allows inference to person outcomes (for example belief changes). When people are randomly assigned to messages, then aggregate PME scores per message are unbiased – except to the extent that the sample as a whole has biases – and so message-based PME scores are representative of the sample within sampling error. The design with double randomization has real benefits for different types of inferences about PME (the same comments apply to PAS with appropriate double randomization).

What the multiple message exposure does not address even with double randomization is whether messages evaluated in any position other than the first position are affected upward or downward in PME systemically. We have addressed this question in
unpublished work, finding a little change in PME for messages, regardless of the sequential position in which they appear (Kim et al., accepted).

**Stable estimates**

To achieve valid results, it is important to have a sufficient number of evaluators per message. As the number of (representative) evaluations per message increases, so will the validity of the message’s aggregate rating in comparison to the population mean (assuming a representative sample of evaluators and random selection). One important question for an efficient message evaluation protocol is the point of diminishing returns. How small can the sample of evaluators be and still have message ratings that are in close agreement with population scores? Reconsider the example presented above – testing 100 PSAs, showing five messages per evaluator. If only 25 evaluations are necessary to obtain stable estimates of a message’s perceived effectiveness, it could be obtained from only 500 participants ($500 \times 5 = 2,500$ evaluations = 25 per message); if the cut-off is closer to 50 per message, the total sample size doubles.

Researchers have used a wide range of sample sizes, no doubt, in part, because little information exists about the relative impact of sample size for any message domain – from 18 (Durkin et al., 2009) to ~120 evaluators per message (Nonnemaker et al., 2010; Parvanta et al., 2013). If sample sizes are too small, then estimates of PME (or any other evaluation measures) will be unstable and susceptible to chance factors in replication; if sample sizes are too large, resources have been wasted and ethical issues arise about the misuse of respondent time and energy. We have addressed this question in research that is under review, finding that modest samples are sufficient to yield stable estimates (Kim et al., accepted).

**Prototypical research using the message testing protocol**

The protocol described above is particularly useful in the following cases: when many messages have to be evaluated quickly and efficiently in order to complete a kind of message triage prior to more intensive testing; when selection of a homogeneous or heterogeneous set of messages that fit some criteria is needed; or when selection of a set of messages all fitting within a construct category is needed, thereby ensuring that an experimental test meets standards of message variability and generalizability. The applications to be pursued are up to the creativity of researchers. A few general applications are described.

**Predictors of PME or PAS as an outcome**

In a recent study by Sutton et al. (2018), a large set of pictorial tobacco graphic warning labels and pictorial tobacco control ads were evaluated for the features that predicted their effectiveness. Two message features with strong and consistent effects were emotion-evoking images (e.g. diseased body parts) and messages with testimonial formats. Other research had reported similar effects, but Sutton and colleagues were able to show these results across some 360 image sets, far greater and more generalizable than other studies have been able to show. Studies like Sutton et al. use the protocol and the valid measures of PME and PAS to efficiently test a very large number of messages for the impact of design features on outcomes directly involved in belief, attitude, and intention changes. Instead of concluding that one or two example messages with testimonials have an effect, the evidence base is able to be expanded substantially, thereby increasing generalizability, robustness, and avoiding the case-category confound.

**Using PAS or PME to select campaign messages**

Studies by Nonnemaker et al. (2010) and by Gibson et al. (2015) used variants of PAS and PME to evaluate FDA’s proposed graphic warning labels, yielding fairly consistent results
about which of the nine was least effective and the relative effectiveness across potential target groups. The proposed protocol and its measurement scales are especially useful in the early stages of public health campaigns wherein triage of possible message interventions is being considered. Most public health campaigns conduct some message testing to eliminate ineffective and promote more effective messages before final testing and implementation (Zhao et al., 2016; Parvanta et al., 2013; Kelder et al., 2002). This early testing helps to set the stage efficiently for more intensive testing of those messages surviving triage.

In campaigns in public relations and strategic communication, the message testing protocol proposed here would only be relevant to the media influence applications of the much broader landscape that has included problem identification, planning, implementation, evaluation, feedback and re-planning with subsequent modification of implementation (Buhmann et al., 2018; Macnamara, 2018; Volk, 2016).

**Using PAS and PME to select multiple messages for theory testing**

In order to carry out generalizable tests of those aspects of messages that are persuasive, it is necessary to use multiple instances of messages representing a theoretically tested construct. Using one message per condition runs afoul of the case-category confound (Jackson, 1992), thereby limiting generalizability to the one or two messages tested and limiting the robustness of conclusions able to be drawn. For example, in a study by Wang et al. (2013), anti-tobacco PSAs with strong arguments yielded observable differences in levels of cotinine (a chemical indicator of tobacco use in the body) compared to ones with weaker arguments. The differences in the PSA’s argument were established by an independent sample of smokers using PAS scales. In order to make an inference about the importance of argument strength, rather than a much more limited inference about arguments in a single, specific message, multiple instances of messages with strong and multiple instances of weaker arguments were used in the test.

Similar procedures yielded differences in objective memory about PSA content (Lee and Cappella, 2013) and boomerang effects to PSAs with weak arguments and smoking cues (Kang et al., 2009). The moderating role of argument strength in these two studies is rendered more robust by using multiple instances carefully tested in advance on different but comparable target populations.

Each of the above classes of studies depends on having to study and evaluate multiple messages in order to achieve its ends. Our protocol serves such goals reasonably well, because of the efficiencies in measurement and procedure that have been introduced and thoroughly evaluated.

**Limitations and counterarguments**

Our protocol for message testing is not to be confused with broader campaign development and evaluation procedures, whether from communication and public health (Noar, 2009; Hornik, 2002; Hornik and Yanovitzky, 2003; Slater, 2006; Wakefield et al., 2010) or from public relations and strategic communication (Buhmann et al., 2018; Macnamara, 2018; Volk, 2016) wherein robust literature and debates reside. Message testing is often a part of campaign development and implementation – sometimes ignored or neglected – but functions more as a complement to the full dimensionality of campaign problem identification, planning, implementation, evaluation, feedback and re-planning with subsequent modification of interventions.

As with any measurement tool, trade-offs come into play, for example, between comprehensiveness and focus and between precision and efficiency. The measure of PME has emphasized the acceptance element of the persuasion sequence, ignoring other aspects of effectiveness such as comprehensibility, sustained attention or retrievability. It was a conscious choice and it assumes that acceptance is the most difficult criterion to achieve in the persuasion process. The others are not unimportant, but acceptance offers the most challenges.
The most significant criticism leveled at PME and PAS is that they represent some aspect of the audience’s psychological judgments of argument strength and message convincingness but not AE. The argument – made most vigorously by O’Keefe (2018) – is essentially that AE is always to be preferred to perceived effectiveness, in that it provides evidence of actual rather than judged persuasive impact.

The problem is not the evidentiary superiority of observed effects versus judged effects. Evidence from AE will always be superior to judged effectiveness. But the gold standard of message testing for effectiveness is one that is slow, costly, and simply difficult to carry out, except in limited circumstances. For example, if there are a very few messages eligible to be assessed or if behavior change is not needed, then direct effects testing is likely be preferred. But with a large number of eligible messages or the need to have many examples in theory testing applications, short cuts are much needed so as to proceed to the next step of more intensive testing.

Other components and features of PME might be added to increase its reliability and predictive validity (also see Davis et al., 2016; Dillard et al., 2007; Noar et al., 2018; Yzer et al., 2015 for elaborated discussions of and cautions about perceived effectiveness measures), but the idea of an efficient, reliable and valid – if imperfect – measure of message effectiveness is still behind other proposed measures. Perfect predictive validity is bought at a heavy cost. Efficient measurement with reasonable predictive validity optimizes between competing demands.

Conclusion
The protocol for message testing and evaluation that is presented here and that has been deployed in many studies uses measures of PME and PAS. The proposal is not naive enough to claim that these measures are equivalent to gold standard tests, but it is believed that there are sufficient data on predictive and construct validity, showing that these efficient – if imperfect – measures are reasonable substitutes to separate the wheat from the chaff in message testing. PME and PAS measures cannot be treated as a replacement for more direct measures and certainly not a replacement for behavioral change, but they produce sufficient reliability and validity to allow sound inferences about relative effectiveness when many messages need to be evaluated quickly and efficiently.

The story about effective and efficient message testing is not over but it has just begun. In order to offer and evaluate useful and generalizable measures, a wide variety of scholarship, problems and researchers need to be involved in both critical commentary and constructive advancement. Given the historical importance of message measurement in public relations and strategic communication, as well as the current resurgence of interest in these problems, knocking down the walls restricting communication and collaboration between fields such as public relations research and communication effects research are both welcome and necessary.

Notes
1. An explicit argument makes a specific claim relating two concepts, e.g. “Gargling with Listerine reduce the chances of getting a cold” vs an implicit argument “Wet feet? Look out for a cold. Gargle with Listerine quick” (Dyer and Kuehl, 1978).
2. There are many other ways to evaluate messages using neural and physiological measurements, implicit association tests (IAT), eye tracking and so on. But these are either more labor intensive or ask different questions than message persuasiveness, which is focused by this manuscript, and therefore these are not discussed here.
References


Further reading


About the authors
Minji Kim (PhD, University of Pennsylvania) is Postdoctoral Fellow at the Center for Tobacco Control Research and Education, UCSF. Her research focuses on persuasion, message effects and tailored/targeted health communication.

Joseph N. Cappella (PhD, Michigan State University) is Gerald R. Miller Professor of Communication at the Annenberg School for Communication at the University of Pennsylvania. His research has focused on the effects of messages on the public in the political and health domains. He is Fellow of the International Communication Association (ICA), the National Communication Association and past president of ICA. Joseph N. Cappella is the corresponding author and can be contacted at: joseph.cappella@asc.upenn.edu
The status quo of evaluation in public diplomacy: insights from the US State Department

Erich J. Sommerfeldt
Department of Communication,
University of Maryland, College Park, Maryland, USA, and
Alexander Buhmann
Department of Communication and Culture,
BI Norwegian Business School, Oslo, Norway

Abstract

Purpose – In recent years, expectations for demonstrating the impact of public diplomacy programs have dramatically increased. Despite increased calls for enhanced monitoring and evaluation, what texts exist on the subject suggest the state of practice is grim. However, while the current debate is based mostly on practice reports, conceptual work from academics or anecdotal evidence, we are missing empirical insights on current views of monitoring and evaluation from practitioners. Such a practice-level perspective is central for better understanding factors that may actually drive or hamper performance evaluation in day-to-day public diplomacy work. The purpose of this paper is to update knowledge on the state of evaluation practice within public diplomacy from the perspectives of practitioners themselves.

Design/methodology/approach – This study assesses the state of evaluation in public diplomacy through qualitative interviews with public diplomacy officers working for the US Department of State – a method heretofore unused in studies of the topic. In total, 25 in-depth interviews were conducted with officers in Washington, DC and at posts around the world.

Findings – The interviews suggest that practitioners see evaluation as underfunded despite increased demands for accountability. Further, the results show a previously not discussed tension between diplomacy practitioners in Washington, DC and those in the field. Practitioners are also unclear about the goals of public diplomacy, which has implications for the enactment of targeted evaluations.

Originality/value – The research uncovers the perceptions of evaluation from the voices of those who must practice it, and elaborates on the common obstacles in the enactment of public diplomacy, the influence of multiple actors and stakeholders on evaluation practice, as well as the perceived goals of public diplomacy programming. No empirical research has considered the state of evaluation practice. Moreover, the study uses qualitative interview data from public diplomacy officers themselves, an under-used method in public diplomacy research. The findings provide insights that contribute to future public diplomacy strategy and performance management.

Keywords Evaluation, Measurement, Organizational performance, Public diplomacy, Communication goals, Foreign policy

Public diplomacy is a communication domain that involves the conveyance of information to foreign publics and the building of long-term relationships to create an “enabling environment” for government policies (Nye, 2008, p. 101). According to the most recently available data, the US Government spent $2.03bn dollars on public diplomacy initiatives in 2016 (Powers, 2017). While this amount is but a small fraction of that spent on defense and social services in the USA, the billions allocated to public diplomacy nonetheless requires the demonstration of high and measurable returns to Congress and other important stakeholders (Brown, 2017). As such, monitoring and evaluation – as the US State Department has labeled...
the practice of evaluation and measurement (Department of State, 2017) – has been positioned
as vital policy connected to several public diplomacy functions, including program planning,
providing evidence for the impact of public diplomacy on informing and influencing foreign
publics, and, ultimately, showing support for strategic goal attainment in foreign affairs.

Public diplomacy as a field is increasingly shaped by demands for quantified
performance management and evidence-based decision making. This is a global trend that
affects many other public sector domains (Christensen et al., 2018; Van de Walle and
Roberts, 2008). Especially in the past decade, around the world, calls for increased efforts in
public diplomacy evaluation have grown strong (Cull, 2014; Pamment, 2012). However,
recent literature on monitoring and evaluation in public diplomacy still paints a picture of
low satisfaction with the state of the field (e.g. ACPD, 2014, 2018; Banks, 2011; Gonzalez,
2015). Public diplomacy practitioners must demonstrate the impact of their work, but have
little knowledge or resources with which to do so.

Related to this is a broader challenge in the public diplomacy field: there is often no
agreement on what the goals and objectives of public diplomacy initiatives are – neither in
the field, nor in specific institutions (Sevin, 2015). This is further complicated by the fact that
the field or domain of public diplomacy is made up of a multitude of actors with partially
diverging interests (White, 2015). Simply put, without a clear understanding of public
diplomacy goals evaluation becomes difficult to effectively enact. Moreover, overlapping
and competing approaches to evaluation in US diplomacy across the various bureaus and
operating locations of the US State Department further complicate the already muddied
responsibility of public diplomacy practitioners to prove the value of their work.

Given these challenges, the purpose of this paper is twofold. First, we update and add to
discussions of the state of contemporary evaluation practice within the field of US public
diplomacy. Our approach is different than and complementary to other recent efforts in that
we utilize a qualitative interview methodology to understand practice through the voices of
public diplomacy practitioners themselves. While the current debate is based mostly on
reports produced by the Advisory Commission on Public Diplomacy – an independent office
of the State Department charged with appraising US diplomacy efforts – conceptual work
by engaged academics (e.g. Sevin, 2017; Pamment, 2014), or anecdotal pieces by
practitioners (Banks, 2011; Gonzalez, 2015), we are missing arguments that use insights
from empirical research to assess current views and perceptions of practitioners. However, a
deeper understanding of this practice-level perspective is central for understanding factors
that may drive or hamper performance evaluation in day-to-day public diplomacy. Second,
we draw upon the perceptions of public diplomacy practitioners to better understand their
views of public diplomacy goals and the challenges of executing evaluation within a multi-
stakeholder environment.

Literature review
Evaluation is a central element of all strategic communication fields, such as public
diplomacy, public relations, and health communication, among others (Buhmann and Likely,
2018). Generally speaking, evaluation refers to the systematic assessment of the value of an
object and can serve two equal purposes: accountability (were objectives met? – often the main
purpose of reporting) and improvement (how were objectives met?) (Stufflebeam and Coryn,
2014). In public diplomacy practice, as recent works suggest, the focus is primarily on
reporting and accountability to important stakeholders such as Congress (Gonzalez, 2015).

Practically, evaluations are commonly modeled by setting up various evaluative stages
for which objectives are formulated and success measures are defined. The development of
evaluation models in strategic communication domains dates back decades and all current
approaches resemble, more or less, the structure of common “logic models,” which, in their
most basic form, distinguish between inputs (program resources), activities, outputs

Status quo of evaluation in public diplomacy

199
(the products that result from the activities), and finally outcomes (the mid- and long-term change or impact that results from the program). Implicit in widely accepted approaches to evaluation is the assumption that goals are derived from the overall mission, goals, and objectives of the organization (Hon, 1998). Thus, through evaluation, practitioners can determine and demonstrate tangible outcomes in the extent to which their efforts contribute to overall organizational goal achievement.

For public diplomacy, these outcomes ideally tie into larger foreign policy goals which, in the USA, are set by officials in the State Department (Sevin, 2015). As a strategic element of the foreign policy toolkit, public diplomacy has to help achieve foreign policy goals and advance national interests (Sevin, 2017). However, the current practice of evaluation is not only strongly focused on reporting, but furthermore, reports mainly on the output (not outcome) level (cf. Pamment, 2014). That said, part of the problem with evaluating most strategic communication programs is that they encompass many different activities, goals are complex and objectives will vary (Hon, 1998). Comor and Bean (2012), in reviewing current public diplomacy practice geared at complex intangible outcomes (such as “dialogue” or “engagement”) on the one hand, and extant public diplomacy performance reviews done by the US Government Accountability Office criticize the often reductive evaluations based on polling or focus groups. They stressed that despite sizable investments (communication initiatives worth $10bn between 2001 and 2009), evidence for the impact of public diplomacy programs is limited. Moreover, as Sevin (2017) has noted, based on Pahlavi's (2007) and Pamment's (2014) findings, the “reporting culture is also observed in the design and execution stages” of public diplomacy projects – he gives a recent example:

An illustrative project was undertaken by the Bureau of International Information Programs at the U.S. State Department to increase its social media outreach. The Bureau followed a reporting understanding, the focus was strictly limited to increasing the number of users. Relying on Facebook advertisements, over $600,000 was spent between 2011 and 2013 to increase the number of users subscribed to the State Department feeds [...] This campaign was an output success as it managed to raise the number of Facebook fans [...] however, such a success does not necessarily mean that it was effective in helping the United States advance its national interests. (Sevin, 2017, p. 883)

Such a strong focus on outputs and accountability is common to many strategic communication domains. The lack of rigorous measurement in related fields like public relations, and a general lack of monitoring and evaluation implementation, has continued despite considerable growth within the field (Macnamara and Zerfass, 2017). A central part of this problem is the continued focus on outputs, as opposed to outcomes (Macnamara and Zerfass, 2017; Wright and Hinson, 2009), the perception that evaluation is too time consuming or expensive (Gregory and Watson, 2008), or the perception that the often complex and intangible outcomes are actually unmeasurable. Such views hinder the development of public relations as a management function (Hon, 1998) – the same argument could be made for public diplomacy. Thus, despite considerable research on advancing evaluation in strategic communication domains, “evaluation has become and remains something of a ‘holy grail’” (L’Etang, 2008, p. 26) – and given what extant writings tell us, this observation holds true for public diplomacy as well.

**Actors and obstacles in the evaluation of US diplomacy**

While dedicated offices for the research and evaluation of public diplomacy exist, most notably the Research and Evaluation Unit (REU) within the Office of the Undersecretary for Public Diplomacy, official Department of State (2017) policy requires that “all bureaus and independent offices must develop a monitoring plan for their programs or projects” and that all “bureaus and independent offices should conduct evaluations to examine the performance and...
outcomes of their programs, projects, and processes” (pp. 4-5). Bureaus such as Education and Cultural Affairs and International Information Programs also have dedicated research and evaluation staff, but other bureaus and independent offices do not have such offices or staffing resources. Official policy also prescribes general standards for the development of monitoring and evaluation plans, but specific guidance as to the creation of indicators, or offering common metrics, does not appear to exist. Moreover, public diplomacy officers receive evaluation education from the State Department’s Foreign Service Institute (FSI) – the training grounds for the US foreign affairs community. Little is known about such trainings, save for the fact that a publicly searchable course directory for the FSI includes a course called “Foreign Assistance Program Monitoring and Evaluation.” The course is two hours in length. What little else know about the state of evaluation in contemporary public diplomacy mostly comes from white papers published by the US State Department itself (e.g. ACPD, 2014, 2018; Banks, 2011), or from commentaries written by experts within the public diplomacy practitioner community (e.g. Brown, 2017; Gonzalez, 2015). Several challenges for evaluation practice in public diplomacy can be distilled from these recent projects. First, political fluctuations, changes in legislatures and parallel changes in preferences and approaches to evaluation undermine attempts to establish a consistent approach. Second, there continues to be a lack of resources dedicated to evaluation (while sophisticated evaluations are time- and cost-intensive) and, at the same time, for many public diplomacy practitioners in the field there is an increasing pressure to do more with less. Third, the long-term nature of public diplomacy programs, with complex intangible outcomes, and related challenges with attribution and causality as well as possible third-factor influences make monitoring and evaluation difficult. Fourth, there is no established culture of evaluation within the State Department, and there is lack of knowledge and skills in evaluation models and methods, which heightens the general reluctance of public diplomacy professionals toward adopting sophisticated monitoring and evaluation. Finally, there are no standard monitoring and evaluation models/terminology within public diplomacy practice, resulting in a heterogeneity of approaches and activities within the broader public diplomacy domain and accordingly specialized and segregated cultures of, and approaches to, monitoring and evaluation.

Assessing practitioner perceptions
While the aforementioned obstacles provide a solid ground on which to base claims about current public diplomacy practices, nonetheless we know little about the perspectives of practitioners themselves when it comes to enacting evaluation in practice. However, gaining such perceptions is vital to better understand drivers and barriers in monitoring and evaluation practice (Buhmann and Brønn, 2018). This leads us to our first research question:

*RQ1*. How do US public diplomacy officers perceive the current state of evaluation in practice?

Obviously, the challenges facing sophisticated monitoring and evaluation practice in public diplomacy are many. However, a particularly thorny challenge for the evaluation of public diplomacy is the potential plurality and “fuzziness” of diplomacy goals. This is a problem for many public service domains, which face challenges of goal conflict and managing multidimensional performance data (Christensen et al., 2018). Furthermore, the goals and explicit functions of public diplomacy itself are varied. Fitzpatrick (2010) reviewed the scholarly literature on public diplomacy and suggested six functional categories of public diplomacy research and practice, namely, public diplomacy as advocacy, public diplomacy as communication/information, as relational, as promotional, as warfare/foreign policy propaganda and public diplomacy as a political strategy. While such a typology is heuristically useful for empirical research on diplomacy programming and communication messages, it lacks insights from practitioners themselves, who may
have differing opinions about the goals of public diplomacy – particularly when put in context of performance evaluation.

The low consistency of goals and lack of clarity in public diplomacy has been addressed from the conceptual side. Recent research has suggested several models to clarify public diplomacy objectives that can serve as a basis for improved performance evaluation. These works have suggested interrelated targets, such as soft power, reputation or relationships (cf. Pamment, 2014), as well as interrelated levels of impact, such as public opinion, relationship dynamics and public debates (cf. Sevin, 2015). However, we do not know how these recent conceptual works relate to extant ideas around public diplomacy goals in the practice. This is relevant if we want to understand if and how such conceptual advancements can find their way into the practice. This leads us to our second research question:

*RQ2. How do public diplomacy practitioners perceive the goals of public diplomacy?*

**Method**

Despite research on evaluation in related fields, such as public relations, little is known about the perceptions of public diplomacy practitioners toward evaluation and the general state of the field. To date, no empirical research on evaluation in US public diplomacy – or, for that matter, any subject related to public diplomacy – has utilized the voices of practitioners themselves. Thus, we take a qualitative perspective to determine how evaluation is perceived in public diplomacy by the people who practice it.

**Sampling and interview participants**

In total, 25 individuals working for the US Department of State agreed to be interviewed for this study. Interviewees were identified via the snowball sampling technique – a widely used qualitative method to identify samples with specialized experiences. As public diplomacy is performed across various bureaus in the Department of State, to help achieve maximum variation in the sample (Miles et al., 2014), interviewees were asked to identify contacts at various bureaus and at posts overseas whom they knew had experience with monitoring and evaluation. The final sample of interviewees included 7 males and 18 females. The gender imbalance in our sample is not unusual or of significant concern, as the field of public diplomacy has become increasingly feminized (Towns and Niklasson, 2016). The interviewees ranged in age from 25 to 67 years. Interviewees worked in public diplomacy positions for various bureaus in the State Department, including Near Eastern Affairs (NEA), African Affairs (AF), European and Eurasian Affairs (EUR), Education and Cultural Affairs (ECA), International Information Programs (IIP), Public Affairs (PA) and in the Office of the Undersecretary for Public Diplomacy. All interviewees had at least two years of experience in public diplomacy work, ranging to 22 years, and all claimed to have experience with public diplomacy monitoring and evaluation. Job descriptions of the participants ranged from public affairs officers at embassies abroad, directors of diplomacy programs, evaluation officers and directors of research. The sample included ten people currently overseas at posts abroad in Europe, Asia, Africa and South America. Most of those interviewed in Washington had also served in public diplomacy positions outside the USA.

**Interview procedure**

In total, 15 interviews were conducted in-person with public diplomacy practitioners in Washington, DC, and 10 with practitioners posted overseas via Skype. Interviewees consented to participate on the condition they remain anonymous and that quotes from the interviews would remain unattributed. All but two of the interviews were audio-recorded – the two interviewees not recorded consented to participate but declined to be recorded due to their
position within the State Department. In these cases, extensive notes were taken. The duration of the interviews ranged from 32 to 74 min. We reached data saturation when the interviews began to show repetitive patterns (Glaser and Strauss, 2017).

The interviews were guided by a semi-structured topics guide of 12 general questions with multiple follow-up probes (Lindlof and Taylor, 2011). The guide was developed to examine participants’ general views of, and specific observations about, monitoring and evaluation practice in public diplomacy. First, participants were asked to describe “their experience with monitoring and evaluation in public diplomacy” to ascertain their level of involvement with evaluation in the past, and probed as to the specific evaluation practices in which they have engaged. Participants were asked to describe the State Department’s “general approach to evaluation in public diplomacy practice,” as well their satisfaction with specific evaluation practices, like the conceptualization of goals and objectives, and the use of various kinds of measurement indicators. The interviewees were asked to articulate what they “believe to be the goals of public diplomacy.” They were also asked about the differences in expectations for evaluation across the various bureaus of the State Department or offices charged with evaluating public diplomacy programming. Finally, they were asked about any additional opportunities or challenges they perceived for evaluation in contemporary practice.

Analysis

The audio recordings of the interviews were transcribed and the resulting transcripts as well as field notes were analyzed with the assistance of qualitative data software NVivo. The research questions asked after public diplomacy practitioners’ perceptions of the current state of evaluation as well as their views of the goals of evaluation. To address these questions, a dual-approach thematic analysis was conducted where the codes of interest emerged from the data through open coding as well as based on established concepts within existing evaluation frameworks (Blair, 2015; Boyatzis, 1998). Analysis began with primary coding to establish general themes emerging from the transcriptions, and was followed by axial coding in which the continued organizing of sub-themes emerged from the initial data reduction process (Miles et al., 2014). The sub-themes that emerged from the coding process are presented in the following section.

Results

Perceived state of evaluation in public diplomacy

The first research question asked about the perceptions of public diplomacy practitioners regarding the current state of evaluation practice in US public diplomacy. The responses to questions about the state of evaluation within US public diplomacy can be best represented through the following sub-themes.

Disjointed systems and structures. Interviewees generally bemoaned the lack of consistency in evaluation approaches across the various bureaus and offices of the State Department. Because bureaus enact various kinds of public diplomacy programming, evaluation is approached quite differently depending on the location within the State Department. A program director located in Washington, DC observed:

It’s a hundred percent varied. It’s so different. ECA [Education & Cultural Affairs] is so much more bureaucratic […] The difference is it’s got these huge programs and huge budgets and a lot of them, like I said, the funding doesn’t come from a central budget source. It comes from the Hill [Congress]. So with that comes different reporting responsibilities and stuff whereas IIPs [International Information Programs] programs are more central and it’s a lot more products, videos, and things like that.

Given that different bureaus use different metrics, and that practitioners will sometimes jump from bureau to bureau, the inconsistency with which evaluation is practiced across
entities of the State Department becomes a noticeable point of confusion. Indeed, when asked if there are consistent approaches to evaluation, one Washington-based interviewee responded: “No, there’s all kinds, because for public diplomacy, you have REU [Research and Evaluation Unit], you have ECA [Education & Cultural Affairs], which is education cultural, and then you have IIP [International Information Programs], who all do their own […] I’ve heard there’s a little turf battle on who does what.” This problem becomes particularly amplified at the field level, where public diplomacy practitioners at posts may be implementing programs sponsored by different bureaus with different evaluations.

Furthermore, interviewees frequently used the term “piecemeal” to describe the efforts of their various bureaus to enact evaluation within programs. One public affairs officer at an overseas post noted:

The offices, or bureaucracy, set up to do it is very piecemeal, disjointed, and lacks much coherency. There are individual offices, or maybe individual officers who, at a post, have been better at it, but, for the most part, it is mostly talk and not a lot of good concrete action on making programs measurable, and then actually measuring them.

Relatedly, while there may be expertise in evaluation to be found within the State Department – particularly within offices such as the REU, located within the Office of the Undersecretary for Public Diplomacy – when advice comes to programs on the ground, its utility can be questioned. As one public diplomacy officer overseas stated: [...] in my experience, most evaluation teams, the recommendations that they provide, tend to be very topical and don’t really address the challenges of operating in these environments, in these restrictive environments, in these short-term environments.

Lack of development and focus. In public diplomacy, monitoring and evaluation is regarded as not a particularly well-developed area of practice, nor one that receives a great deal of emphasis. Whatever emphasis it does receive from top brass in the State Department is largely seen as “lip-service,” which reinforces a general view that evaluation is a “tick the box exercise” for public diplomacy. Some reported a perception that while superior officers might emphasize the importance of evaluation and demand regular reports of impact, when such reports are delivered they “sit on the boss’s desk and collect dust.” Moreover, the lack of emphasis on monitoring and evaluation is nowhere more notable than in the perceived lack of training public diplomacy practitioners received on the subject. Many claimed to have received minimal training at the FSI – at the very most a day and a half. This training, they noted, also did not prepare them for the realities of executing evaluations at under-resourced posts.

The interviewees also often compared public diplomacy to development communication, claiming that entities which administer and measure development communication programs, like the US Agency for International Development, have a far easier time engaging in evaluation due to the arguably more concrete outcomes of development activities (e.g. number of wells dug, vaccinations given, radio stations built, etc.). As a consequence, practitioners believed development communication to be more advanced in monitoring and evaluation practice than in public diplomacy.

Devaluation of public diplomacy efforts. The public diplomacy practitioners interviewed expressed a general frustration with a perceived devaluation of public diplomacy by others within the State Department, they noted, is likely a contributing factor behind the lack of focus and attention to evaluation in public diplomacy. Furthermore, the interviewees noted that many within the State Department, including public diplomacy officers, do not understand the “real goals” of US public diplomacy efforts (a topic we address more thoroughly below). As a result, after more than 20 years since the US Information Agency – the entity formerly responsible for diplomacy programs – was merged with the State Department, there is still an
unclear view of public diplomacy’s place within the institution. One Washington, DC-based participant satirically described the views of others toward public diplomacy in the following way:

Oh, they’re just gonna go and do another party. Oh they’re doing another reception. Oh, they’re just sending another person on exchange. That’s nice. That’s great. But it’s hard for people who focus on other things to sort of see the policy imperative. So one of the areas where I feel like public diplomacy practitioners need to spend a little time, or that the thinkers back here in Washington need to think about is how can we make sure that everyone else is seeing the value of what we do.

Though practitioners continually struggle to enact rigorous evaluation, they recognize that monitoring and evaluation is a means through which they can escape their derogated position. One participant suggested that monitoring and evaluation can actually help to shift the poor perception of diplomacy within the State Department, proving that diplomacy is more than just “jazz hands” and show the importance of their work.

Moreover, while public diplomacy work may be misunderstood and devalued within the State Department at large, it is also sometimes subjugated in the field at the embassy level. While public diplomacy is seen as “touchy-feely” and devalued in Washington, DC and at posts abroad, “[public diplomacy] wouldn’t be as hard a sell if we could actually gather like this massive amount of data and prove it.”

Lack of resources. A common theme in texts describing evaluation practice in public diplomacy is the lack of dedicated financial and staff resources. Our research echoes those previously identified problems (Banks, 2011; Gonzalez, 2015). Practitioners lamented the funds available for evaluation, and some questioned why they would spend money on evaluation research when they could use such funds for more program development. Time was also frequently noted as a hindrance to engaging in evaluations, as practitioners are regularly overwhelmed with responsibilities, and evaluation therefore takes a “lesser priority.” Relatedly, the mere number of staff at some US embassies was noted as a restriction to enacting evaluation, not to mention conceiving and executing programming. As one senior-level DC-based participant claimed: “Take little Moldova. We have three American officers there, Russia has 1,600 people on the ground.” With such vast differences in staffing resources being channeled to certain posts, practitioners noted there is little way for evaluation to make much headway in practice. At many posts, there is “one officer doing everything.” Given these pressures, evaluation was often noted as one of the first things to be “put off” or “shelved.”

Capital-field tensions. An unexpected theme to emerge from the discussion was the perceived differences in expectations for the enactment of evaluation in public diplomacy programs between offices in Washington, DC and those at post overseas. Most of those interviewed had served at posts abroad, and noted differences with how evaluation is discussed based on position – whether one is stationed in Washington and overseas public diplomacy programming, or whether one is on the ground implementing them. For one, the expectations of Washington actors coupled with the demands of the work at post contribute to problems of time allocated to evaluation at embassies abroad: “We do things that Washington wants us to do, things that the embassy wants us to do, and then things that we have planned for ourselves, this means that there’s not a lot of time for monitoring and stuff.” Practitioners at posts are faced with being held accountable to both their direct supervisors at posts (often the head public affairs officer or the ambassador) as well as offices in Washington.

Practitioners in the field are aware that there are research offices such as the REU that exist to assist in program design and evaluation. And those interviewees who worked for such offices are willing and eager to assist those in the field. But, the perceived distance and availability of such assistance was questioned by officers at posts abroad. Echoing this
sentiment, other interviewees expressed a desire for there to be a local staff member at post to help evaluate programs. Currently, most US embassies and other types of posts do not have a resident evaluation specialist. As one public affairs officer abroad noted, “[…] in a perfect world, I would love for every embassy to have a social scientist or a regional social scientist who was, you know, could really help us be shaping [evaluations] on a regular basis.”

Interestingly, one interviewee based in Washington suggested that it is not the “job” of public diplomacy practitioners at post to evaluate programs, believing they are not equipped with the skills or resources to do so. Refuting this point, most public diplomacy officers abroad believed they have received increased pressure from “Washington” to demonstrate impact and satisfy important stakeholders.

Finally, the interviewees stressed an awareness of the political winds that shape public diplomacy goals as well as funding priorities. While some speculated that quality evaluations were necessary to “prove the value” of certain public diplomacy programs and thereby preserve funding, others believed evaluations play little to no role in funding decisions. As one participant claimed:

Political will defines budgeting. And frankly, if you’re looking at the DC versus post dichotomy, it’s all about how much money you’ve got, or how much staffing you get, or how much support you get, or who you get out to visit. And that has to do with political will and the events of the day. It doesn’t have to do with someone looking at your post-project evaluations and saying you did a great job.

Reliance on outputs, anecdotes. While some practitioners indicated a desire to use sophisticated evaluation metrics of attitude and behavior change, the primary way in which participants described modern public diplomacy monitoring and evaluation efforts is in the reporting of a few output indicators coupled with positive anecdotes chronicling participants experiences in a program. Practitioners noted the relative ease with which output indicators are gathered through tracking participation rates or monitoring social media metrics. Such indicators were the ones most frequently reported by practitioners because, by their own admission, these indicators are the easiest to acquire and report. Most participants showed a familiarity with outcome measures of attitudinal and behavior change, but believed they lacked the time, resources and skill to plan and execute such metrics.

Anecdotes are also frequently used as evidence of impact, as practitioners believed such stories would satisfy superiors’ demands for program accountability and sufficiently persuade lawmakers. Furthermore, like output indicators, anecdotes require relatively little time and skill to gather. As such, these are the forms of evidence most frequently used in public diplomacy evaluations. Indeed, in many cases, utilizing anecdotal quotes, likes and comments was noted by a public affairs officer overseas as “about the best we’re doing on some of our evaluation for a specific program.” Indeed, the only consistent “metrics” or reporting schemes that could be discerned across all of the evaluation programs discussed by the participants appeared to be the use of anecdotal quotes and basic output indicators such as likes and shares.

Even in the rare cases when outcome indicators like attitude and behavior change are tracked, they occur only in the short term, and are not linked to larger public diplomacy goals. As one interviewee at a post overseas explained:

Most often our outcomes measure the PD programs, so in the short term people’s eyes were opened to something new, it changed their attitudes, they have a greater appreciation of anything from human rights to transparency to fighting corruption and things like that. But then we don’t typically have actually what would be more direct outcome that we’d like to measure, which is, they’ve actually supported some sort of legislative reform that makes it harder to commit corruption, or that makes it easier to investigate corruption. Something like that.

So, while practitioners might rarely identify short-term outcomes of public diplomacy programming, there is still a lack of connection to larger policy goals. As one participant
People struggle with what are the real outcomes here? Where’s the beef? Where’s the real substance of why this was important?” Until practitioners can describe the “real outcomes,” reliance on outputs and short-term changes will likely continue to be the norm.

**Perceptions of public diplomacy goals**

The second research question concerned how public diplomacy practitioners perceive the goals of modern diplomacy practice. Setting clear goals and objectives is a well-accepted part of communication planning and evaluation. The primary way in which practitioners characterized the goals of public diplomacy centered around “building influence.” Influence, however, was thought of in several different ways. First, practitioners desired to influence other societies to fundamentally change certain aspects of it, such as diminishing the likelihood of “embracing a violent extremist ideology,” or simply to make a country’s government more transparent and accountable – in other words, influencing a society to promote democratization and good governance.

Second, influence was cast as “building networks of people that support or have a deep understanding of the U.S. perspective.” Such network building often occurs through public diplomacy programming that exposes foreign publics to US arts and culture that purportedly “opens people up to seeing that we’re friendly, and we like to get to know them, and we want to share our culture and the great things about the United States with them.” Thus, while some practitioners suggested that a significant goal of public diplomacy programs is simply to “build a relationship with people who might not otherwise been in contact with an American,” most often relationship building was seen as a way to advance the strategic interests of the USA:

> […] the goal is to achieve U.S. foreign policy objectives abroad by creating connections between the people of the U.S. and the people of other countries as well as creating greater understanding of the United States among people of other countries so that our foreign policy is more likely to be successful in those countries.

The definition offered above is consistent with what many academic texts would agree is the explicit purpose of public diplomacy – building relationships with foreign publics so as to ease the facilitation of American foreign policy objectives in that nation (Cull, 2014; Nye, 2008). But how “building relationships” can be translated into manageable goals and objectives for specific public diplomacy programs remains a significant point of confusion for public diplomacy practitioners: “I would say that public diplomacy has moved away from the USIA days where generalized exposure and people-to-people connections that are sort of more open and open-ended and less defined as sort of a clear set of specific goals […] those days are over.” As illustrated in the preceding quote, it is not enough to build support and “relationship networks” in contemporary public diplomacy, but practitioners are unclear as how to progress further. Moreover, relationship building and cultural exchanges were thought only to yield dividends in the long term – making the measurement of such activities effects in the short term quite difficult. Indeed, measuring immediate outcomes was called “an unrealistic goal for public diplomacy.”

While influence and relationships are terms that surfaced in the discussion of public diplomacy goals, many respondents puzzled over how to articulate the goals of modern public diplomacy. As one DC-based program officer working for a regional bureau, when asked about the goals of public diplomacy, stated, “I don’t know. That’s a good question. I think that’s one of the problems we haven’t defined […] but each individual, depending on where they sit in the organization, seems to have a different idea of how to define that.” To that effect, and perhaps unsurprisingly, views on public diplomacy goals and evaluation seemed to vary by the location of the individual within the State Department structure. But, generally speaking, participants agreed that public diplomacy goals lack clarity. As a
consequence, how to articulate objectives and propose appropriate measurement indicators also suffers. Still, there was recognition that tying public diplomacy activities and resulting measurements to larger goals is required in modern practice. A diplomacy officer overseas commented:

Public diplomacy is now expected to support policy and to be clearly tied to specific policies and/or mission goals. And we can no longer say we’re going to take a bunch of American artists and Tunisian artists and put them together for two weeks and we know, “we have faith” that good things will come out of this.

But, while practitioners recognized that evaluation must be tied to policy goals, no clear solution as to how accomplish this was offered. One final quote illustrates this struggle:

What people want to know is how public diplomacy is advancing foreign policy, right? That’s what it’s all about. That’s our banner. Yet the way in which we approach evaluation, it really is focused a lot more on the outputs ultimately and very intermediate kind of outcomes that don’t get it back […] we are going to have a hard time telling that story of really wide public diplomacy is relevant advancing foreign policy without more substantive evaluations. I just don’t think we’re set up to do that right now.

Discussion
As noted earlier, extant empirical research on public diplomacy in general, and evaluation in public diplomacy in particular, has not considered the voices of the people responsible for enacting diplomacy programs. While policy papers and statements from high-ranking government officials provide evidence of the official stance of the US Government in terms of its public diplomacy goals, the ways in which those who execute public diplomacy describe its enactment sheds a different light on the practice of public diplomacy. We believe this approach has pointed out several implications and future directions for public diplomacy research.

Structure hinders evaluation consistency
The structure of the US State Department and the various bureaus and offices that execute public diplomacy programs has no small amount of influence in how the purpose of public diplomacy is perceived. Indeed, as the participants reported, how the goals of public diplomacy are articulated and measured depends, to a great extent, on the bureau for which they worked, or on whose behalf they executed programs. The disparate nature of these programs and the expectations of those overseeing them pose no small amount of difficulty in enacting common metrics and indicators.

The interviews revealed that while there is an ostensible “central” monitoring and evaluation body for public diplomacy in the State Department (REU), this office has limited resources, oversight, and influence on programs executed outside the diplomacy branch of the department. Moreover, the REU lacks the resources to collect and analyze evaluation reports from across the various bureaus of the State Department, let alone the programs executed under the Department’s dedicated diplomacy arm. This hinders the capacity of the REU to work toward uniformity in evaluation procedures for evaluation programming. Moreover, as much of the evaluation of diplomacy work occurs on the ground, by overworked practitioners without the assistance of a local monitoring and evaluation office, the quality of evaluations and their consistency is unsurprisingly varied.

Moreover, enthusiasm for evaluation and the expected level of sophistication also varied. The Education and Cultural Affairs bureau in particular was criticized for overly simplistic evaluations of long-standing diplomacy initiatives like the Fulbright Scholarship Program. That said, others suggested that output indicators and anecdotes were enough to satisfy
superiors – who might, on the surface, stress accountability through regular evaluation reports, but might also have a similar lack of training in evaluation. Accountability is perhaps currently “enough” for superiors and stakeholders, suggesting that actual “learning” through evaluation and examination of program outcomes is yet to occur, nor is it likely to in the foreseeable future.

**Conflicting views of Washington’s influence**

A theme heretofore unexplored in the public diplomacy literature is the potential tensions that may arise between the expectations of Washington, DC stakeholders and practitioners in the field. Washington-based officers often oversee public diplomacy programming, or assist in the research and evaluation of programs at the post level. Thus, they often regarded Congress or the Undersecretary of Public Diplomacy as the only stakeholders to which they felt beholden, suggesting their perceived direction of accountability is largely linear. In contrast, practitioners in the field must juggle the expectations of, and report to, offices back in Washington, Congress, as well as the ambassador for whom they directly work.

While consideration of management officials and stakeholders at the level of Congress is a key focus for public diplomacy practitioners, to advance evaluation practice, public diplomacy officers should be able to identify the specific goals of programs at both the program and societal level. While practitioners may know their goals for a specific public diplomacy initiative, they remain unaware of how such programs tie into larger public diplomacy goals or of programs’ long-term social impacts. The result of this being that anecdotes and short-term output indicators are the normal standard by which practitioners communicate accountability. Some practitioners believed that producing insightful and quality evaluations has no impact on future funding (instead tying funding to political will), and others believed the opposite.

**Goal articulation and evaluation capacity**

Setting clear goals and objectives is a well-accepted part of communication planning and evaluation (Smith, 2013). It is difficult to know what you have accomplished if you have no idea of where you started and what you set out to change. Without concrete, measurable objectives, the ability of public diplomacy practitioners to assess the true effects of their work is limited. While the academic literature may have reached a relatively shared view of the general purpose and function of public diplomacy, the explicit goals of public diplomacy are less clear among practitioners when it comes to the various specific “day-to-day” programs.

Indeed, practitioners are less than clear – and even confused – about the actual goals of public diplomacy, particularly in the context of assessing the contributions of their work the State Department based on measurable objectives. Positioning this finding generally within Fitzpatrick’s (2010) categorization of public diplomacy goals, the interviewees articulated their views of public diplomacy goals as largely “relational” and “informational,” and to a lesser extent, as “advocacy.” Thus, while outside observers might describe the goals of public diplomacy as multifaceted, practitioners possibly view the function to be much more focused – depending on the specific role they assume and units with which they are affiliated.

**Limitations and future directions**

The road ahead for public diplomacy is an uncertain one. Budget cuts to the State Department, and proposed mergers of the PA and International Information Programs bureaus (Gramer, 2018) do not portend development on monitoring and evaluation anytime soon. Despite the many challenges for enhancing evaluation practice in public diplomacy, it should be noted that some interviewees believe the practice to be improving: “We’re getting better about it every year. I think when I first got here, it wasn’t so good 10 years ago. I think
it’s better. We are coming up with results that mean something.” Practitioners value evaluation and want more of it. Thus, there remains potential for monitoring and evaluation to flourish as a constructive part of public diplomacy practice.

Our study is not without limitations. Though generalizability is not the goal of qualitative research, we cannot guarantee our results represent the views of all public diplomacy practitioners, particularly those outside the USA. Snowball sampling, while a useful technique to gain access to individuals with specific knowledge sets, may also have reduce the range of views represented. Future work should also carefully consider the educational and professional backgrounds of public diplomacy practitioners, as education, training and experience with evaluation often arose as an indicator of practitioners’ level of comfort with monitoring and evaluation.

Our research has focused on the State Department specifically as the single main institution of US public diplomacy. To a degree, however, we argue that our insights can contribute to recent comparative work, looking of further cases such as the UK and Sweden (Pamment, 2012). Specifically, these works have highlighted that the literature is still lacking organizational-level insights into monitoring and evaluation practices – something that our results on practice-level perceptions in the State Department helps address. However, in a wider view of the practice, which includes also non-state (corporate) actors that may work in consort with government entities (cf. White, 2015), the questions addressed here are complemented with yet another layer as evaluations then may involve the (potentially incommensurable) goals of multiple entities. In light of the recent conceptual work on public diplomacy goal frameworks (e.g. Sevin, 2015, 2017; Pamment, 2014; Fitzpatrick, 2010), our findings suggest that one current emphasis in the practice could be to apply and “break down” these more macro-level concepts for the plethora of different projects and programs that are run by the Department of State. Furthermore, it would be interesting to examine how academic descriptions of public diplomacy targets (i.e. Fitzpatrick, 2010) are interpreted and applied in the agencies and on field level. As our findings also point to a perceived disconnect between the center and periphery of the practice (“DC vs post”), it is also important to understand if such more “holistic” stories about diplomacy goals have the potential to improve the alignment between the levels of civil and foreign service.

References


**Further reading**


**Corresponding author**

Erich J. Sommerfeldt can be contacted at: esommerf@umd.edu

---

For instructions on how to order reprints of this article, please visit our website: [www.emeraldpublishing.com/licensing/reprints.htm](http://www.emeraldpublishing.com/licensing/reprints.htm)

Or contact us for further details: permissions@emeraldinsight.com
Explaining the measurement and evaluation stasis
A thought experiment and a note on functional stupidity

Howard Nothhaft
University of Lund, Campus Helsingborg,
Helsingborg, Sweden, and
Hanna Stensson
More PR, Göteborg, Sweden

Abstract
Purpose – The purpose of this paper is to explain the “evaluation deadlock” or “stasis” diagnosed by many authors. The explanation relies on a thought experiment.
Design/methodology/approach – The paper is conceptual and builds on a thought experiment inspired by qualitative research such as interviews with communication consultants in Sweden. It makes use of principal–agent theory and Akerlof’s theory of lemon markets.
Findings – A plausible explanation for the evaluation stasis requires consideration of practitioners’ self-interest as businesspeople. The deadlock is explained by an anomaly in practitioner populations and passive or active but covert resistance. If the long-time neglect of measurement and evaluation has led to expectation inflation and overpromising, even well-performing actors might shy away from rigorous measurement and evaluation practices in their own mandates, since they fear being measured against promotional, not realistic standards. At the same time, on the level of industry discourse, these practitioners would still advocate for measurement and evaluation in principle, so as to avoid the suspicion of underperformance.
Research limitations/implications – The paper suggests an explanation for further empirical investigation. It does not attempt to demonstrate anything else than that the suggestion is plausible and that it warrants further investigation.
Practical implications – The scientific community engaged in the measurement and evaluation debate appears puzzled by the discrepancy between practitioners’ words and actions. The authors hope that the paper contributes to a more realistic and thus more constructive dialogue between practitioners and academics in the measurement and evaluation debate.
Originality/value – Inspired by Alvesson and Spicer’s concept of functional stupidity, the paper argues that attempts to explain the evaluation stasis have been marked by circumspection and narrowness. At present, explanations for the evaluation stasis tend to focus on lack of knowledge or inadequate systems or frameworks. The paper offers a more comprehensive explanation.
Keywords Evaluation, Measurement, Public relations, Communication practitioner

Stuck, stasis and deadlock
Academics engaged in the debate seem in agreement that measurement and evaluation in public relations are “stuck” (Gregory and White, 2006), in a state of “deadlock” (Macnamara, 2014) or “stasis” (Macnamara and Zerfass, 2017). Scholars and professional associations have suggested “myriad models” (Place, 2015, p. 120) as well as more or less rigorous frameworks (such as the evaluation framework proposed by AMEC, the Association for Measurement and Evaluation in Communication, or the Barcelona Principles). Nevertheless, despite widespread agreement about the topic’s strategic importance (Gregory and Watson, 2008), actual adoption by practitioners lags, at times appearing lacklustre, even reluctant. Thorson et al. (2015) summarise: “In general, surveys of practitioners tend to show a great deal of excitement about the future of PR measurement, but they also show that this excitement is not translating into action over time” (p. 10).
Research conducted by Jesper Falkheimer, Mats Heide, Charlotte Simonsson, Sara von Platen, Howard Nothhaft and Rickard Andersson in the so-called “Communicative Organizations-Project” suggests that the industry in Sweden is no exception (Falkheimer et al., 2017). When asking roughly 500 communication practitioners in 11 organisations about areas in the greatest need of development and improvement as well as resources invested in the same areas, a curious discrepancy emerged. While 10.4 per cent of respondents \((n = 493)\) expressed the opinion that “evaluation of communication” is in greatest need of development and improvement, only 0.6 per cent of respondents \((n = 489)\) stated that they devoted most resources in form of time and money to it. The evaluation of communication was considered the third highest priority concerning development and improvement, but came last of 18 items, even behind miscellaneous “others”, in terms of resources invested (in the self-assessment of practitioners). In other words, the results reproduce what other studies found before: practitioners rate measurement and evaluation as highly important, but effectively do very little about it.

The scholarly community has suggested many explanations for this discrepancy, and the resulting deadlock and stasis. Volk (2018) initiated a renewed debate by drawing attention to the following issues in the measurement and evaluation debate (see Volk, 2018 for the following):

- confusing terminology and inconsistent measures;
- downplay of confounding variables and intervening effects;
- overreliance on positivist epistemology;
- under investigation of cost reduction through communication;
- barriers and obstacles at the organisational level, department level and individual level; and
- disciplinary silos.

There is little doubt that these factors play a role, and that they, taken together, explain the situation to a degree. However, this contribution argues that one complex is missing. It suggests furthermore that the omission is interesting, as it reveals a mild form of functional stupidity on the side of the scholars. Practitioners, conversely, play a far more demanding game than theorists acknowledge.

The factor missing in the debate is the individual practitioner’s interest. Laskin (2016) suggests that “the issue of measurement and evaluation still remains a mystery for many public relations professionals whether it is because of lack of knowledge or lack of interest” (p. 2). What Laskin seems to have in mind is disinterest, however. Our contribution suggests, in contrast, that the reason for the stasis is self-interest. A critical mass of practitioners – not all, but enough – do not want rigorous measurement and evaluation, because they have very little to gain from it. Note the qualification “critical mass” here. Recent research by Swenson et al. (2019) demonstrates that there are practitioners with successful measurement and evaluation programmes (although the conditions seem quite demanding), so the question is about dynamics in populations. So our proposition is that a sizeable practitioner population undermines, by action or inaction, any initiative that tries to establish binding standards or a more rigorous practice.

Individual practitioner interest as an explanation of the evaluation stasis

The resistance of practitioners, actively-covert or passive, would explain the deadlock, but before one engages in empirical investigation, a basic plausibility test is in order. Why should some public relations practitioners publicly bemoan the inadequacy of existing measurement and evaluation, as they currently do, if they not really believed the industry
would benefit from a more rigorous practice, as they apparently also seem to do? Why has there been, in the words of Thorson et al. (2015, p. 9), “more talk than action”? It should not be forgotten that there is a lively and fruitful exchange between practitioners and academics in measurement and evaluation. This should not be the case if there were a mere “conspiracy” of practitioners to stall progress in evaluation and measurement simply.

A thought experiment, theoretical concepts and distinctions

Our plausibility test will take the form of a Gedankenexperiment or thought experiment. In the words of Yeates (2004), a thought experiment “is a device with which one performs an intentional, structured process of intellectual deliberation in order to speculate, within a specifiable problem domain, about potential consequents (or antecedents) for a designated antecedent (or consequent)” (p. 150). Brown and Fehige (2014), in the Stanford Encyclopedia of Philosophy, define succinctly: “Thought experiments are devices of the imagination used to investigate the nature of things” (p. 1).

Although the term invokes associations of unempirical philosophising, thought experiments have played a pivotal role in the development of modern science (Feyerabend, 1993; Cohen, 2005), and as deiknymi they constitute the most ancient form of mathematical proof. Contrary to popular belief, Galileo’s (1564–1642) demonstration that heavier objects do not fall faster than lighter ones did not simply involve a climb up the Pisa tower, but in Discorsi e dimostrazioni matematiche (1635) takes the form of hypothetical speculation about combining a heavier and a lighter object (Cohen, 2005). Perhaps the most famous thought experiment is the case of Schrödinger’s cat (1935), a thought experiment proposed by physicist and Nobel laureate Erwin Schrödinger (1887–1961) to illustrate the counter-intuitive consequences of applying the Copenhagen interpretation of quantum physics to normal objects. Brown and Fehige (2014) go so far to state that “in our own time, the creation of quantum mechanics and relativity are almost unthinkable without the crucial role played by thought experiments” (p. 3). Our aim is somewhat more modest, of course.

Principal–agent theory and the theory of lemon markets

In conducting a thought experiment, we follow Swedberg’s (2012) advice to theorise by asking what process would produce the observed result (2012, p. 29). Our theorising requires some distinctions and a few theoretical concepts capturing asymmetry, however, like principal and agent or Akerlof’s idea of “lemon markets”. As is perhaps well known, principal–agent theory captures (for an overview see Eisenhardt, 1998) the idea that those tasked with actually executing a task, the agents, normally have an information advantage vis-à-vis the actor who commissions the task, the principal. Principal-agent theory has been used to theoretically model the change from an entrepreneurial to a managerial economy, in which a caste of professional managers, as agents, conducts business on behalf of owners, the principals. The theory posits that the managers will use their information advantage to further their own interests as opposed to the interests of the owners, e.g. optimise manager salaries and managerial power as opposed to return on investment. The challenge for principals, thus, is to find a solution for this problem, the so-called agency cost, e.g. by designing contracts, so that manager remuneration reflects return on investment.

The concept of lemon markets was developed by Akerlof (1970) in a well-known paper, “The Market for Lemons: Quality Uncertainty and the Market Mechanism”, which led to the Nobel Memorial Prize in Economics, jointly received with Michael Spence and Joseph Stiglitz in 2001. Akerlof illustrated the formal theory by looking at the market for used cars in the USA. Because buyer protection was rudimentary at the time of writing (a couple of years after publication, some states in the USA actually introduced “known defects”-laws), buyers of used cars ran a considerable risk of ending up with a “lemon”, i.e. a car with defects that would become obvious only after the buy. Akerlof’s key insight was that the quality
uncertainty would lead to a dynamic that drove out the good cars or “peaches”. Because of the information asymmetry, the best bet of any buyer would always be to assume at most average quality and tailor their offer accordingly. This, in turn, makes it impossible for sellers of peaches to get a fair price. As a consequence, peaches completely disappear from a market with no buyer protection and high-quality uncertainty; good used cars are simply not traded openly but passed on amongst relatives and friends.

**Principals and agents**

With public relations, principal–agent theory and lemon markets come into play because public relations are a professional service commissioned by principals and executed by agents. The measurement and evaluation debate as a whole reflects the fact that the information asymmetry is considerable, furthermore, as true effects of communication, and consequently quality of the service, are notoriously hard to establish. In their roles as agents, public relations-personnel have a strong individual incentive to turn this information asymmetry to their advantage. In their role as principals, conversely, the interest is to prevent its exploitation. For our argumentation, it becomes necessary, thus, to distinguish the roles of practitioners on various levels, acting either as principals or agents, sellers or buyers, i.e.:

1. consultants in agencies who perform public relations, i.e. operative work, as a service and as agents of communication directors;
2. in-house teams who do the same, i.e. perform operative public relations as a service, albeit with a different incentive structure, as agents of the communication directors;
3. specialized measurement and evaluation consultants who provide the service of evaluating and measuring other actors’ work, be they in-house teams or agencies, as agents of the communication director or management board;
4. communication directors who direct and are principals to in-house teams and consultants, and act as buyers of public relations services as agents of the management board; and
5. and, finally, management boards who are not only internal clients of in-house teams and agencies but also the principals of the communication directors.

The five groups are not only different and have different interests, but act as either principal or agent in their relations with each other. Consultants and operative in-house practitioners are in a way at the bottom of the food chain; they are always agents and need to make sure that they deliver the best possible results and impress. The management board at the top of the food chain is in a very similar position, as they are always principals. Top managers want to make sure that they get the best results and that they are not only impressed but rightfully impressed.

**Individual and collective interest in measurement and evaluation**

The distinction between roles and interests as principals or agents is not all. What is equally important, second, is a distinction between the interests of individual practitioners on one side (micro-level), the industry’s interest as a whole on the other (macro level). This is where lemon markets come into play. While practitioners have an individual incentive to exploit the quality uncertainty that is due to inadequate measurement and evaluation, they also have a collective incentive not to let the market for their service deteriorate into a lemon market. Lemon markets make it impossible to get a good price for decent work. The individual has an interest in the industry’s reputation, thus, as long as it does not transform into a competitive disadvantage for her own business. This does not only apply to
consultants, but it also applies to communication directors whose prestige in the organisation depends at least partly on the prestige of their “profession”, the communication industry. It is here that another distinction becomes important, namely the distinction between general discourse or “talking the talk” on one side, individual action or “walking the walk” on the other. What one says in public debate and what one does in one’s own mandates does not necessarily have to match. A discussion in the profession, especially one that invariably leads to the conclusion “that we are not there yet”, can be used to stall progress as much as it can further it.

Overperformers and underperformers

Finally, as a last distinction, public relations practitioners are not a homogenous population. The European Communication Monitor (Zerfass et al., 2018, pp. 108-115) demonstrates that there are vast differences between public relations practitioners, even in the self-selected sample of the ECM, with salaries in the band between lower than EUR 30,000 on the one hand, above and beyond EUR 300,001 on the other. However, most important for us is to distinguish between the self-perception as overperformers and underperformers. Overperformers are actors who suspect they deliver more than they are paid and credited for. Underperformers, in contrast, suspect that they deliver less or, perhaps a more common view, that rigorous measurement in its current state does not capture the real value of their work due to the “performance predicament” pointed out by Garnett (1997), i.e. the fact that benefits of communication are far harder to measure than costs. One can, of course, ask how actors would know in which category they fall, given that no evaluation is in place. The actors’ convictions are construed as entirely subjective, however. In order to avoid the connotation of “tricking” clients, it is perhaps easiest to view the stance as a competitive strategy: overperformers think they would thrive on a market with quality certainty ensured by rigorous measurement and evaluation; underperformers do not. Given unclear goals and ambivalence of expectations, it would be very difficult to establish who is really, objectively and in every single case, an overperformer or underperformer. For our purposes, it is not necessary, as the only criterion is whether actors believe that industry-wide, comparable, standardised, i.e. “hard” measurement and evaluation and a practice of rigorously comparing results promised and delivered would strengthen their competitive position overall.

The real complexity with a competitive strategy based on one’s perception as underperformer or overperformers does not primarily lie in its subjectivity, but in the vague connection between communication performance and business success. Overperformers are not necessarily successful service providers, nor are underperformers necessarily unsuccessful. A factually underperforming agency head might be very good at reassuring clients; therefore, the agency is successful in conveying a sense of achievement that is valued by the client but would not show up in a rigorous evaluation. Alternatively, maybe the agency head is simply well-connected in a political way. Presumably, it is this particular constellation that overperforming but unsuccessful practitioners, perhaps new to the game, hope to address by joining arms with scholars who demand more rigorous and objective standards. Unfortunately, successful overperforming companies, i.e. established high performers, do not necessarily pursue the same strategy. Although they could expect small gains from rigorous evaluation that highlighted their overperformance, they might prefer the existing status quo as it pays handsomely already, and at smaller risk. This remains particularly so, as long as there is no pressure from the side of the buyer of communication services.

A thought experiment

With the distinctions drawn so far in mind, we are equipped to answer the question why public relations practitioners have an interest in publicly complaining about the
inadequacy of existing measurement and evaluation while simultaneously stalling meaningful progress towards rigour in their own mandates. Our argument takes into account that the constellations in which practitioners work, measure and evaluate might be very different. We begin, however, with the simplest, least rigorous constellation. Only then do we offer thoughts on how to develop the model further to explain more complex and rigorous settings.

Maybe the most common and also least rigorous constellation is that of an in-house team, which is headed by a director and hires the services of an agency. The agency as the seller of services then self-evaluates and presents its achievements to the in-house team as well as the director in her role as buyer and principal of the service. The director then presents the results to the management board, where she acts not as principal, but as agent in a principal–agent relationship. The most rigorous constellation, in contrast, would be a setup where the consultants or in-house teams are under the control of the director of communication, while evaluation and measurement remain in the hands of a specialised consultant hired independently of the director, remains under control of the management board.

If all public relations services were to be self-evaluated by the agencies as agents, how would the industry end up in the situation it apparently is in? The thought experiment clarifies our explanation. Figure 1 illustrates what one could expect in a public relations industry without distortions. If one postulates a population of ten practitioners with comparable business models or careers, it is reasonable to assume that eight practitioners would be of average quality, one of distinctly high quality, one of distinctly low quality. The *de facto* performance of the population would probably distribute along the same lines,
as market forces push out low performers, and exceptional performance becomes increasingly rare the more the practice professionalises (i.e. agencies know their worth). It is important to note, here, that the public relations industry is modelled as de facto performing in accordance with real expectations of clients. Communication is not modelled as a lemon market in Akerlof’s sense. There is no reason to assume that the industry is crowded with overperformers or peaches; yet there is no reason to expect a stable population of consistent underperformance, or lemons either. Average performance is determined by the prices paid for the service, as clients find they cannot get more for the same money.

The problem lies with the practitioners’ beliefs about what rigorous, objective evaluation would reveal about their performance. Again, in a normal distribution, one would assume that only a small percentage of the population suspects that they are underperforming (note that Figure 1 displays an average performer, dark grey circle in the lowest box, who mistakenly believes to be underperforming). Another, for reasons of optimism slightly larger percentage, would assume, rightly or wrongly, that they are overperforming. When it comes to their strategies in handling their individual mandates, those who believe to be overperformers have a strong incentive to push for rigorous and objective evaluation (hard evaluation), as that would reveal the true quality of their work and give them an advantage in competition with average performers. Underperformers, in contrast, are incentivised to argue for no evaluation at all (or evaluation that is so vague that it is practically equivalent to non-evaluation). Average performers could go the same way as underperformers or argue for a reasonable and negotiable evaluation (soft evaluation), i.e. a scheme that is maybe not objective but suffices to reassure clients. In Figure 1, the black arrows show that four average performers followed the later path, while the other three adopted the strategy of underperformers. That nearly half of the average performers shy away even from soft evaluation is quite a pessimistic assumption, and makes our case as strong as possible.

When it comes to discourse or talk about the role of measurement and evaluation, the normal thing to expect would be a congruence of individual action and demands for the industry, i.e. those who do not implement evaluation in their own mandates also speak out publicly against it. However, an important feature here is that outspoken advocacy against evaluation is risky in the current climate of new public management and audit society (Zerfass, 2005), as it suggests that the advocate is an underperformer. Why would someone argue against demonstrating one’s value? The reason must be that this particular agent is underperforming, i.e. tricking clients. The societal macro level also plays a role, in other words. This is the reason, presumably, why outspoken public advocates against evaluation are rare; thus, for reasons of simplicity, advocacy against evaluation is not modelled as a viable option. For underperformers, the best position is to commit conditionally, i.e. to wholeheartedly agree to measurement and evaluation in principle while emphasising that the present models are unrealistic, not yet there. Figure 1 arrives at the conclusion, thus, that there would be unanimous agreement in the discourse that measurement and evaluation are important, with only disagreements as to whether one should go for hard and objective schemes or more reasonable, negotiable, soft ones. This constellation is by and large what we see in reality: a perennial industry discourse of being misunderstood, underestimated and in the process “of becoming” (cf. Stensson, 2017).

On the level of individual mandates, the model predicts a preponderance of actors who argue for either hard or soft evaluation, with non-evaluated mandates in the minority. In the model, despite rather pessimistic assumptions about the measurement-affinities of practitioners, the distribution is 60/40. As our argumentation is a thought experiment, the exact number is not important. What is important is that self-assessed average performers – in defending against underperformers and deterioration of prices – should largely gravitate towards an effective evaluation practice that robustly reveals that they are
delivering what they promise. Promotional rhetoric aside, delivering what one promises is hard enough in a functioning market. In any case, together with the overperformers, the dynamic should be robust enough to drive measurement and evaluation forward.

The anomaly of the public relations industry

One can argue, of course, whether the majority of mandates are reasonably or only vaguely evaluated, but since there is a dissatisfaction with the current situation, it is perhaps more fruitful to ask what depresses the dynamic. Figure 2 makes the case that the depression is due to the proportion of practitioners who fear that rigorous and objective evaluation would reveal underperformance. For reasons we come to later, the argument is that the proportion is abnormally high in the public relations industry; this is marked as anomalous in Figure 2. Note that Figure 2 does not claim that the de facto quality of the industry’s work is low. Public relations is not a lemon market where the peaches have disappeared. That would be extraordinary, as the distribution of quality in a given market segment should group near to the average.

Figure 2 only depicts that four out of eight de facto average performers (five all in all, but one, represented by the black circle, is a genuine underperformer), instead of only one in Figure 1, are worried that rigorous and objective standards if applied, would reveal their work as not delivering what is promised. Consequently, their mandate- or micro-level strategy is to push for non-evaluation. If we assume that only one performer takes the safer route of pursuing the same strategy, as represented by the downward black arrow, the balance suddenly becomes 60/40 against any form of evaluation. Again, this does not mean that public relations practitioners are con men, or constitutionally more worried

Figure 2.
Public relations industry with anomaly
about underperformance than other professionals. The question simply does not arise in industries with hard performance standards and quality certainty. Sales reps do not worry about being revealed as missing their sales targets; they worry quite straightforwardly about missing them.

Overpromising as the origin of the anomaly

As with lemon markets, the ultimate origin of the anomaly lies in the fact that success and performance in public relations are not immediately obvious, i.e. quality uncertainty. In contrast to Akerlof’s used car market, however, our argument is that lack of credible measurement and evaluation over years and decades have led to inflation of expectations and a practice of continuous overpromising by the public relations industry. Take press clippings. Although it is clear that the circulation of a newspaper does not equate to exposure (very few people read a whole newspaper), the industry tended to treat the figures as equivalents. With those making the greatest promises winning the mandates, and with very little in terms of objective evaluation to keep outrageous claims in check, a spiral of ever greater promises is hardly surprising. For the practitioner caught in the spiral, it becomes very hard to break out (except perhaps by arriving at agreements with reasonable individual clients and managers), since making “realistic promises” is indistinguishable from promising lacklustre performance. In a typical Nash-equilibrium (Nash, 1950; Nash 1951), no individual practitioner has the interest to be the only or first player to call out the game of overpromising.

A spiral does not necessarily equate to a corrupted and dishonest industry. It is likely that professional buyers and sellers of services, i.e. industry insiders, are aware of the discrepancy between powerpoint presentations and reality. This is particularly true for the gap between the output-oriented performance agencies actually demonstrate, from press clippings to click rates, and the extraordinary effects in terms of behavioural change that are supposed to flow from it on outcome, outflow and outtake level. For many practitioners the worry is not that objective evaluation would show that they are doing a bad job to the buyers of the service. Judging by the implicit, “real” standards of the industry, which heads of agencies and directors of communications largely share, their clients and managers would probably agree that they are doing okay. The worry is that the sudden introduction of objective evaluation would reveal the industry-wide discrepancy between promised and actual effects to principals on the level of management boards, in the worst of cases leading to further erosion of prices. Xavier et al. (2006) already pointed to the dangers of an overpromising dynamic driven by output-oriented evaluation: “when practitioners either decide to or are required to demonstrate their organisational value, they fail to recognise the limitation of such methods in demonstrating higher-level goals and therefore make success claims that are unsubstantiated by the evaluation data” (p. 7).

Breaking out of the spiral by rigorous measurement and evaluation?

The communication industry, to repeat, is not modelled as a lemon market. Although there are information asymmetry and quality uncertainty, agencies do not knowingly try to sell faulty goods like used cars close to breakdown. Moreover, public relations is a service that can be more or less successful in achieving what the agent promises, sometimes irrespective of the investment the agency makes. This observation, the volatility of public relations, might provide a clue as to why the more rigorous constellations – where specialized consultancies perform measurement and evaluation research independently of the agency – did not solve the problem so far.

The question consists of two related issues. The first is why independent evaluation, as opposed to self-evaluation, has not become industry standard. This question is particularly interesting since other management functions are increasingly measured and evaluated by
management consultancies (McKinsey, PWC, KPMG or BCG), and against robust key performance indicators. The second is why apparently nothing short of the most rigorous constellation – measurement and evaluation services hired independently and under control of the management board – sufficed, so far, to break the industry out of the overpromising spiral?

A heretic answer to the first question might be that public relations are simply not important, i.e. expensive enough, to require independent auditing as standard. Maybe management boards are not interested in rigorous measurement and evaluation simply because the gain is not worth the effort? Could it be, for example, that most top managers and high-level organisational decision makers think of public relations in terms of “satisficing”, where good enough is just that … good enough? Could it be, moreover, that top managers took to heart what communication practitioners told them, namely that the communication is extremely complex and highly volatile, but drew a very different conclusion? The conclusion namely that optimising and fine-tuning one’s public relations is not worthwhile as the best one can achieve under conditions of volatility, is to keep things more or less under control – and to do so, rudimentary measurement and evaluation is more than enough.

What comes on top of a possible satisficing attitude, and explains why only independent measurement and evaluation stands any chance, is the fact that buyers and sellers of communication services are not only in a principal–agent relationship but share a common interest as communication agents facing the management principal. Communication agencies and heads of communication share a common interest in avoiding total measurability, since it not only comes with the danger of revealing an industry-wide performance discrepancy, but could also lead to erosion of prices, and the trivialisation of communication services. By diminishing the information asymmetry, i.e. the advantage the executing agent has over the principal, measurement and evaluation might reveal the agency cost, i.e. the cost that the agent’s self-interest inflicts on the principal.

Some tentative empirical evidence
The status of our contribution is conceptual; the suggested population dynamics are abstractions of real dynamics, which are far more complex, of course. However, we do believe that the thought experiment demonstrates the basic plausibility of our suggestion to a degree that warrants further empirical investigation. There is fragmentary qualitative evidence, moreover, that not only inspired the thought experiment but further strengthens our suggestion.

One source of inspiration was unpublished background interviews conducted in the construction industry in the context of another project. The interviews suggested that the frequent delays and cost explosions in public building projects are partly caused by a system that forces contractors into overpromising in their initial tender, so as to keep projected costs low for public consumption and political approval. Although all insider parties are more or less aware that the project is impossible to deliver at the projected cost, the tacit agreement is that the contractor takes the public censure in exchange for securing the contract. Whether true or not, the case of the construction industry suggested that the explanation of anomalies may not only lie in the accumulation of individual causes but may be systemic in nature.

A second source of inspiration was the second author’s master thesis supervised by the first author. Stensson (2017) explored the evaluation deadlock in the Swedish public relations industry by conducting 15 anonymous interviews with practitioners. What she found in interviews, and what she consequently models, amounts to systemic buyer-seller-dynamics. Apart from interviewees who simply thought that evaluation was not “fun” and that money
spent on “cleaning up” rather than “action” was wasted (p. 46), the interviews support the assumption of systemic overpromising in three ways.

Some practitioners explicitly expressed concern about living up to expectations set by former agencies of their clients, often very optimistically, and the necessity to play the same game: “we consequently set a bar that we, of course, had to reach, to give them just as much ‘bang for the buck’” (Stensson, 2017, p. 36). Along the same line of thought, practitioners continue to use AVE (13 of 15 in Stensson’s study did), although they are fully aware that it does not reflect the real value of public relations work. At a certain point, quantity becomes quality, as one practitioner explains:

Somewhere, there is a value, but there is no way to set a great number on it. […] I must admit that focus tends to be more on quantity and that you almost drool when you see those 25 million. It is also a great sales pitch for us to say “Hey, you get 26 million but it only cost you 140 thousand”. Of course, this gives an acknowledgement from the client that it is worth the money and for us that we have delivered something. (Stensson, 2017, p. 32)

Other consultants emphasised the pressure on communication directors, the buyers of the service, to justify spending on public relations, which in turn required consultants to give their clients the evaluation figures they needed for an internal presentation. One consultant stated that his primary mission was to make his client “a hero” (p. 38); whether genuine long-time effects were achieved was secondary. Another interviewee stated in a similar vein: “Everything goes back to making the client satisfied no matter the price. If that means that we must multiply it [advertising cost] with seven or whatever […] to increase a client’s reach, we will do it” (Stensson, 2017, p. 37). Another admitted that he, while well aware of the pressures to look good internally, would not otherwise put his name on figures he provided for some clients, as they were alternative facts, not true (2017, p. 35).

Finally, very much in line with our argument, some consultants expressed a fear not only of the comparison with others but with reality:

If you were to do this before and after-measurement about how people think about the brand or about awareness, and end up noticing “Well, this didn’t really make any difference”; then of course it is easier just to give a rough estimate to make it look better”.

Interestingly, the same consultant also feared the comparison with previous service providers:

“[…] and if you get a new client that earlier have reach[ed] 90 billion, then, our 90 million is not that much to boast about. (Stensson, 2017, p. 36).

Conclusion

The status of our contribution, to repeat, is theoretical. The contribution suggests an explanation for deadlock and stasis and tries to make the case that it warrants empirical investigation. The empirical investigation might show that the explanation is wrong or incomplete. In any case, the explanation has to be expanded to satisfactorily cover the more sophisticated and rigorous configuration involving evaluation services.

Stasis is likely to stay

Formulated as a game of individual interest on mandate-level and collective interest on industry level, the measurement and evaluation deadlock becomes a function of practitioner populations. As long as there is no critical mass of unsuccessful overperformers, who would benefit from objective standards yet do not benefit from the status quo already (as successful overperformers might), adoption of rigorous practice or even binding standards is likely to lag, with no end to “wearisome” debates (Stensson, 2017, p. 1). Even with pressure from the client side, the dynamic is not straightforward as the buyer does not only want to know
whether she got “bang for the buck” on the principal side but also wants to be “a hero” on the agent side, i.e. when reporting to her principal[1].

A meaningful move is even more unlikely as long as there remains a critical mass of successful overpromisers, i.e. actors who do well in the current system. As has been argued, even high-performing service-providers are forced into a long-established and deeply rooted game of overpromising – well known to insiders, but easily misrepresented by “objective” scholarly investigation.

The result is a fear that objective and rigorous evaluation based on promotional, as opposed to realistic standards, would show even high-quality work by implicit industry standards, to be underperforming by explicit, promotional standards, i.e. fall far short from what has become customary and accepted talk. This concern, we suggest, is the driver of a silent resistance enacted in individual mandates, between buyers and sellers. The resistance is silent, because the industry talk, in contrast to individual action, is characterised by widespread agreement that measurement and evaluation are highly desirable, a pathway to professionalism. The underlying cause for the widespread disconnect between discourse and action, it has been argued, is the individual actor’s worry that outspoken resistance against evaluation marks the potential underperformer.

The industry discourse of “becoming”, in contrast, suits almost everyone – except commercially unsuccessful overperformers and pioneers of rigour (often communication directors who come from a different industry), who consequently link arms with scholars to push for rigorous measurement and evaluation practice (hard evaluation).

**Functional stupidity**

What do the dynamics have to do with functional stupidity? Functional stupidity is a concept developed by critical management scholars Mats Alvesson and André Spicer, initially in a scholarly article (Alvesson and Spicer, 2012), later in a monograph (Alvesson and Spicer, 2016). It was a key motif in an early version of our argumentation. Alvesson and Spicer define functional stupidity as the “inability and/or unwillingness to use cognitive and reflective capacities in anything other than narrow and circumspect ways” (Alvesson and Spicer, 2012, p. 1201). The symptoms of functional stupidity, in the authors’ words, “involve[s] a lack of reflexivity, a disinclination to require or provide justification, and avoidance of substantive reasoning” (Alvesson and Spicer, 2012, p. 1201). In the popular monograph, Alvesson and Spicer elaborate:

There are three telltale aspects of functional stupidity: not thinking about your assumptions (what we call reflexivity), not asking why you are doing something (justification), and not considering the consequences or wider meaning of your actions (substantive reasoning). (Alvesson and Spicer, 2016, KL 1132-1135)

In a way, the idea rephrases the popular saying that folks believe all kinds of things if their livelihoods depend on them. Similar to Herbert Simon’s concept of “bounded rationality” (Simon, 1957), functional stupidity captures the idea that organisations, contrary to the rhetoric of the knowledge society, require intelligent people to don blinders that mask out the very questions intelligent people would typically ask. Yet Alvesson and Spicer make very clear that functional stupidity is by no means only negative. As “knowledge-intensive organisations” are far from what their shiny rhetoric claims they are, functional stupidity serves many organisations well: “Quick feet and hands combined with a friendly smile matter more for the economy than the intellect of a relatively small group of elite knowledge workers” (Alvesson and Spicer, 2016, KL 164-165). Well-known and outspoken critics of the higher education industry, Alvesson and Spicer do not exclude their own. Academia, they stipulate, “could be seen as a hothouse of functional stupidity” (2012, p. 1014). Admittedly, neither the article nor the book makes entirely clear whether functional stupidity is meant to be a precise and operationalizable concept or a provocative figure of speech.
We do not suggest here that practitioners are functionally stupid. On the contrary, our modelling suggests that successful practitioners are smartly coping with the macro climate of an audit society affecting an industry that has gotten away effectively unaudited for very long. Functional stupidity comes into play because we, the scholars engaged in the measurement and evaluation debate, currently suffer from a mild form of it – or at least we behave as if we do. The current attempts to explain the deadlock, in any case, seem “narrow and circumspect”. Take the discussion of results in Xavier et al. (2006), for example, where it was found that out of 60 practitioners not one stated that he or she avoided evaluation in order to avoid accountability. The authors are circumspect, when they suggest that self-reporting might have played a role here and that practitioners were “not prepared to acknowledge that accountability is an issue when in fact it is” (2006, p. 6). Alternatively, take the example of expertise: Zerfass et al. established that practitioners often “lack the necessary expertise to conduct reliable evaluation and measurement” (2017, abstract). Interestingly, the industry discourse tends to leave it at that. However, is lack of expertise a satisfactory explanation on its own? Even if one leaves shortcomings in education and training undisputed, and accepts that at least 20 years of dedicated education in public relations have not led to any improvement, is it plausible to suggest that a highly dynamic industry like communication cannot quickly catch up with the newest trends in measurement and evaluation? What is reproduced here, it seems, is the paradigm of the knowledge society, in which unique knowledge bases are hard-to-imitate resources. Alvesson and Spicer challenge exactly that presumption: “Often it is the flexible use of disciplined and hardworking young people who can be plugged into any project that gives the edge to large management consultancy firms. Not their unique knowledge base” (Alvesson and Spicer, 2016, KL486-487). Measurement and evaluation are not trivial, but they are not rocket science either. Against that background, it becomes doubly important to ask why disciplined, hardworking young professionals in the public relations industry adapt to many other trends, yet mysteriously cannot come to grips with evaluation. Knowledge and expertise can be one, but it cannot be the only explanation. If mastery of measurement and evaluation were the key to the survival and success of agencies, we suspect the industry would soon excel. A more plausible assumption, thus, is that it is not the key – at least not yet, despite the rhetoric – and that the industry is currently served best by its discourse of “becoming”.

In the case of the examples, it is to the authors’ credit that they do not mix empirical data with theoretical speculation. On a general and collective level, however, a broader view and less circumspection would benefit the measurement and evaluation debate. For academics who work in close cooperation with practice, it is, of course, functional to avoid avenues of enquiry that might upset good relations with the industry. Also, it is eminently functional for a scientific community to stick to the belief that interviewees always tell the truth, and that the solution always lies in more research, more education. However, there is danger in pre-emptively masking out alternative lines of investigation.

Note
1. The employment of independent evaluation services, e.g. Prime or Cision, change that dynamic, but do not fundamentally upset it as long as the buyer of communication and evaluation services remains the same entity. Moreover, measurement and evaluation service providers, similar to agencies, also have a “hidden interest” in keeping their services non-trivial by protecting their inner logic as a black box. A more sophisticated and comprehensive future model should take this into account. The authors thank Sophia C. Volk for drawing attention to this dynamic.
References


Further reading


Corresponding author
Howard Nothhaft can be contacted at: Howard.Nothhaft@isk.lu.se
Investigating the role of contextual factors in effectively executing communication evaluation and measurement

A scoping review

Stefania Romenti and Grazia Murtarelli
Department of Business, Law, Economics and Consumer Behaviour, International University of Languages and Media, Milan, Italy

Angelo Miglietta
Department of Humanities, International University of Languages and Media, Milan, Italy, and
Anne Gregory
The Business School, The University of Huddersfield, Huddersfield, UK

Abstract

Purpose – Evaluation and measurement (E&M) remains a critical and debated topic among communication scholars and practice. Substantial research and professional efforts have been devoted to discussing what should be measured and which methods should be applied. Most of the E&M models seem to carry a positivist imprint. But, in real-life, organizations could not have clear aims, enough resources, or adequate informative systems to support E&M. Moreover, several contextual factors could affect the implementation of E&M management processes. The communication literature rarely highlights these factors. To fulfill this gap, the purpose of this paper is to theorize the contextual factors relevant to the management of the evaluation process.

Design/methodology/approach – A scoping literature review was carried out exploring the role of contextual factors and impact of contextual factors on E&M management processes. More specifically, the review examines the contribution provided by program evaluation and performance measurement (PM) fields of research.

Findings – The paper provides a scoping review of program evaluation and PM approaches. Additionally, it explains how both streams of thought argued the importance of contextual factors, such as organizational, relational, cultural and communicative factors, for the success of any evaluation processes. The study underlined that the main evaluation models used in the field of communication have overlooked these studies and put on evidence the role of contextual factors in effectively executing communication E&M.

Originality/value – The paper enriches the dominant rationale concerning the E&M management processes by incorporating literature on: program evaluation; and PM. The analysis could provide useful insights also from a professional perspective, by helping practitioners for a contextual assessment of strategic communication programs and activities.

Keywords Evaluation, Organizational effectiveness, Measurement, Organizational performance

Paper type Research paper

Introduction

Since the 1960s, evaluation and measurement (E&M) remains a critical and much debated topic among communication scholars. Substantial research and professional effort have been devoted to discussing what should be measured and which methods should be applied.

Focusing on these questions is crucial, but in the meantime, it assumes that the effectiveness of E&M models’ design and their successful execution mostly depend on the choice of the right metrics and research methods. In fact, most of the E&M models and frameworks seem to carry a positivist imprint, by conceptualizing communication programs
as rational actions based on the program logic methodology objectives-means(inputs)-outputs-outcomes-impacts. This positivist rationale also implies that individuals operate with a goal-oriented, planning-focused and information-driven culture within their organizations. Contra to this supposition, in real-life organizations often do not have clear aims, adequate resources, or adequate information systems to support E&M processes.

E&M processes can follow non-linear and uncertain paths in response to contextual factors (Pawson and Tilley, 1997; Swenson et al., 2019). E&M processes can be very challenging, requiring a complex blend of elements, which may at times demand the resolution of conflicting needs. Morgan (2007), for instance, describes E&M as political processes, characterized by political alliances, bargaining, conflicts and compromises. Baskin and Aronoff (1988), for instance, analyzed open and closed system evaluation, by recognizing and examining the role of factors outside the control of communication which could affect public relations’ program results.

The influence of context in the E&M processes has been included in some of the models and frameworks more recently developed by communication scholars – for example, the integrated evaluation framework by AMEC (AMEC, 2017) and the Integrated Evaluation Model by Macnamara (2018a, b). Nevertheless, the communication literature has largely overlooked these influential contextual factors. The debate so far has been focused on identifying barriers preventing the advancement in E&M implementation (Swenson et al., 2019). Much scholarship has explored the impact of the expertise and role of chief communication officers within successful measurement programs, but it underestimates the role of organizational context (Swenson et al., 2019).

In this conceptual article we argue that a plethora of contextual factors, such as for instance, concerned parties’ expectations or evaluative management styles, could affect the implementation of any E&M process. More specifically, this paper aims at theorizing the contextual factors relevant to the management of the evaluation process. This paper explores those factors which need to be taken into account to effectively support initial and ongoing E&M implementation within the communication area.

The paper enriches the dominant extant academic debate concerning the E&M process in two ways. First, it provides a scoping literature review of the theoretical underpinnings of performance measurement (PM) and program evaluation research, which helps close an identified research gap in communication literature. Despite the differences between the two approaches, they are complementary ways of acquiring and analyzing information that is intended to reduce uncertainty around programs (McDavid and Hawthorn, 2006). These streams of thought underline a new research problem, namely, how to effectively implement and execute an E&M system. Second, it proposes a categorization schema for the factors shaping the E&M models’ implementation.

In order to achieve its aim, this paper is structured as it follows. The first part provides the results of a scoping literature review focused on the extant communication E&M models and identifies how both program evaluation and PM streams of thought argue the importance of contextual factors, such as organizational, relational, cultural and communicative factors, in the success of any evaluation process.

In the second part, insights from program evaluation and PM fields of research are combined and discussed and insights are developed for a contextual assessment of strategic communication E&M programs and activities. Finally, limitations and future applications are discussed in the conclusive section.

**Literature review**

Communication studies have shown that in the design phase of communication programs the most fruitful approach might combine E&M methodologies, by highlighting their complementary role (McDavid and Hawthorn, 2006; McDavid et al., 2012). On the one hand,
evaluation consists of systematic collection, analysis and interpretation of data and information concerning activities under scrutiny, in order to pass judgment on their functioning as well as to guide further decisions and actions for all the involved actors (Patton, 1998; Rossi et al., 2003). On the other hand, “measurement is the process of translating constructs into procedures for collecting data” (McDavid and Hawthorn, 2006, p. 159). Combining E&M within a single research design represents something more than a mere methodological exercise. It calls for a union between two divergent scientific domains: program evaluation and PM. Since the 1960s, the communication literature largely has borrowed concepts, techniques and methodologies from these two scientific domains. Among the inherited concepts, in both approaches, a certain weight has been given to the context within which E&M processes are embedded. For this reason, the paper provides a scoping review (Rumrill et al., 2010) of the two different scientific domains of program evaluation and PM.

As pointed out by Macnamara and Likely (2017) program evaluation is based on the assessment of “how programs can be designed and implemented to achieve their objectives” (p. 9), while PM is based on measuring program outputs and outcomes, without trying to recognize the extent to which the program has affected the outcomes (Nielsen and Randall, 2013). Despite differences, both approaches provide useful insights concerning the role of contextual factors in the effectiveness of E&M management processes.

A scoping literature review is a useful method to synthesize research findings embedded in diverse disciplinary areas and to identify potential gaps in the research literature (Arksey and O’Malley, 2005; Rumrill et al., 2010). This paper follows the procedures discussed by Arksey and O’Malley (2005) and illustrated by Rumrill et al. (2010) that can be summarized as it follows: identify the initial research questions; identify the relevant studies; refine the inclusion and/or exclusion criteria for selecting studies; organize data to represent the descriptive elements of the scoping review; and provide a narrative description for reporting results.

The paper is framed around the following research questions:

**RQ1.** What is the contribution provided by program evaluation and PM fields of research concerning the role and impact of contextual factors on E&M management processes?

**RQ2.** Which insights can be drawn with regard to contextual factors influencing the execution of communication E&M management processes?

In order to answer these research questions, a search of the literature was performed across the EBSCO – Academic Search Complete database. The paper takes into consideration relevant studies and research conducted in the program evaluation and PM fields. Review includes studies which investigate and evidence the role of internal contextual factors in E&M management processes. In anticipation of the findings, the identified literatures focused on internal contextual factors that could drive how evaluation is undertaken: evaluation questions and measures; evaluation approach and design; role of the evaluator and role of stakeholders; evaluative culture; final reports; information technology support and use of results.

The following sections provide a narrative description of the scoping review and they address three main cluster of results: first, an analysis of the state of the art of the communication E&M research; second, the contributions published in the program evaluation domain; third, analysis of studies and research in the PM research area.

**Communication E&M research**

The growing relevance of E&M as a strategic activity for communication professionals has given rise to a considerable amount of research during the last 40 years (Watson and Likely, 2013; Volk, 2016; Zerfass et al., 2017; Macnamara, 2018a, b).
A wide range of studies has underlined a number of common concepts that underlie E&M processes. For instance, multiple purposes can be fulfilled through E&M, primarily, formative and summative (Scriven, 1996; Macnamara, 1999). For formative purposes, E&M help to optimize the financial resources allocated to communication activities or help to improve and prioritize communication strategies. For summative purposes, E&M processes can assist communication managers in obtaining internal legitimacy by making communication value more explicit and the results clearly beneficial in the eyes of the entire organization.

The importance of goal setting is an additional element common to E&M models to date. For instance, among the earliest models, the Short and The Continuing Model of Evaluation by Watson (2001) attributed a close relationship between the definition of communication objectives and the choice of a specific evaluation model. Setting down clear and specific objectives is also the first of the ten Barcelona principles, a declaration set down in 2010, revised in 2015 (AMEC, 2015) and agreed by more than 200 professionals from across the world.

Communication scholars have also adopted a multi-level approach which identifies the stages of E&M. Hence, the concept of the “logic model” has gained predominance (Julian, 1997). Logic models assume that inputs feed activities and in turn generate outputs, outcomes and impacts. Many communication frameworks follow the logic model approach to define the stages of E&M (i.e. outputs, outcomes and impacts) such as, for instance, the Yardstick Model (Lindenmann, 1997), the Preparation, Implementation and Impact model (Cutlip et al., 2000) and the Macro Model of PR Evaluation (Macnamara, 1999), the GCS Evaluation Framework (GCS, 2016), the AMEC integrated evaluation framework (AMEC, 2017) and the Integrated Evaluation Model (Macnamara, 2015, 2018a, b). The main assumption underlying such frameworks is that the choice of metrics matters. Communication evaluators should replicate the logic model's stages and identify proper metrics at each level. Other frameworks, such as the communication balanced scorecard (Fleisher and Mahaffy, 1997; Ritter, 2003; Vos and Schoemaker, 2004; Zerfass, 2010), share the same basic assumption and offer a set of communication performance metrics which acknowledge a closer link between value created by communication activities and firm's goals.

Communication E&M models put a great emphasis on the methodological rigour and the use of research methods and techniques to collect valid and reliable results. To this end, traditionally quantitative research methods are given priority over qualitative ones, because collecting quantitative data seems to provide a more objective approach to reality (Daymon and Holloway, 2010). This way of examining reality emphasizes unbiased assessment. Reality is seen as an entity that can be observed irrespective of the individuals who experience it, so factors are set aside from their natural context for measurement purposes (Morgan, 2007). In distinction to this, qualitative methods are characterized by an interpretative approach to the analysis of social reality: they take into account and explore the intentions, motivations and experiences of all involved actors (Daymon and Holloway, 2010; Grunig et al., 2002). For this reason, qualitative methods highlight the relational dimension of the communication process. It is helpful to gain basic knowledge of unusual phenomena and provide some interpretations of them (Ritchie, 2003).

Using different approaches enables practitioners to describe, foresee and control the totality of the communication process. An effective evaluation model does not limit the use of qualitative methods to the experimentation stage, but merges them with quantitative ones to exploit the benefit of both (Daymon and Holloway, 2010).

One of the first evaluation models that underlines the role of contextual factors for implementing an E&M process is the Measurement Tree model, elaborated in 2003 by the Institute of Public Relations. The measurement model is graphically represented as a tree. The roots are the communication objectives. The trunk stands for communication strategies and tactics. The bark covering the trunk represents services, products and other company
aspects, which are made public. The sap going through the trunk embodies all those elements of communication that can only be accessed by staff and employees. Branches refer to the target audience, leaves are related to the public for which media relations plans are designed, and flowers represent outcomes—all the changes in opinions, attitudes and behaviors that communication has produced in target audiences (Jin and Cameron, 2007). Finally, the tree is planted into a natural environment, which refers to the market the organization operates in and the organizational climate where the communication plan has been implemented. The Measurement Tree model is one of the early evaluation models which take account of the organizational context and the external environment and considers them as factors influencing results. Nevertheless, the measurement tree model does not help communicators deepen their understanding of which specific environmental factors can impact the implementation of E&M management processes; it provides just a generic outline of the environment in which they are embedded.

Other recent and well-known models, such as the GCS Evaluation Framework (GCS, 2016), the AMEC integrated evaluation framework (AMEC, 2017) and the Integrated Evaluation Model (Macnamara, 2018a, b), value both internal and external organizational contexts where the E&M management processes take place and recognize this can affect their design, but they do not investigate in detail the nature of these factors.

Overall, current studies focused on E&M management processes and their implementation in the communication field lack a detailed analysis of those contextual factors affecting execution of E&M processes that could be organizational, relational, cultural and communicative. By reviewing the research in the fields of program evaluation and PM these contextual factors can be identified and categorized.

Program evaluation research

The conceptual roots of most communication E&M models lie within the expanded domain of program evaluation, which places special emphasis on the application of social research methods to address evaluation questions and formulate sound judgments (Fitzpatrick et al., 2004). The origin of the scholarly inquiry on evaluation research can be traced back to the end of the 1950s. Evaluation research, or program evaluation, can be defined as “a social science activity directed at collecting, analyzing, interpreting and communicating information about the workings and effectiveness of social programs” (Rossi et al., 2003, p. 2). Social programs are designed to ameliorate a social problem (i.e. delinquency prevention programs) or improve social conditions (i.e. public housing programs), in specific areas such as public health, welfare and education.

Until the 1980s, evaluation research theory focused attention on the distance between the actual inputs and the expected impacts of a program, the so-called “black box.” Since the main purpose of program evaluation consists of understanding how much of the observed impacts of a program could be attributed to the invested inputs, the experimental technique has always been considered as the research design par excellence (Cooper et al., 2006). In experimental studies, both the internal validity (causality links among inputs and outcomes) and the external validity (generalization of results) of a program are the main concerns. By definition, using an experimental design means to implicitly assume that every program occurs in a vacuum. In other words, nothing, with the exception of inputs, can affect the expected outcomes and impacts of a program.

Since the 1980s, so-called theory-based approaches challenged these assumptions by positing that black boxes should have been opened and by arguing that evaluation practices should have been aimed at examining the interrelationships among inputs, outputs and outcomes (Weiss, 1997). The relevant literature on theory-based evaluation (TBE) between 1980 and 1990 yielded a new wave of “opened black boxes” approaches. The conceptual rationale underlying TBE is based on the notion of “bounded rationality,” which can be
defined as “the limits upon the ability of human beings to adapt optimally, or even satisfactorily, to complex environments” (Simon, 1991, p. 132). Therefore, programs could function completely differently from being intended, since outcomes and impacts will depend on how the mechanisms that are supposed to be at work within a program, will be enacted in a given context. Most of the evaluation models assume that individuals operate with a goal-oriented, planning-focused and information-driven culture within their organizations. But, in real-life situations people experience quite different situations.

Thus, evaluators should be aware of the contextual factors affecting the outcomes and impacts of a program, for instance, concerned parties’ expectations are crucial to inhibiting impacts from happening or enabling a program to produce intended results. According to TBE, evaluators must take into account the legitimate concerns of those individuals who can affect (or be affected by) evaluation’s results (Pawson and Tilley, 1997; Scriven, 1996). This means evaluators need to be aware of possible evaluation questions that could arise from stakeholder interests during the communication program and which may impact on the measures that have to be taken into account in evaluation.

Pawson and Tilley (1997) argue that the principal goal of evaluation research is to understand what works best, for whom, in what circumstances and why. Hence, theories, not methods, should be put at the center of any evaluation design and implementation. No method is seen as the single gold standard, but TBE approaches combine all methods that could be suitable without privileging any one of them. TBE reflected the importance of addressing the needs of those who have an interest in the evaluation process (Scriven, 1996).

Other scholars are primarily concerned with the methodological aspects of evaluation and more specifically the evaluation approach and design (Weiss and Bucuvalas, 1980; Fetterman et al., 1996) and argue that priority should be accorded to information holders, who ensure the timely collection of reliable data.

Morgan (2007) states that evaluation aims could be affected by eventual compromises developed between competing coalitions. Sometimes aims could appear ambiguous and weak in an attempt to satisfy stakeholders with different needs. Here, Morgan’s (2007) vision of how assessment processes take place within organizations appears realistic and concrete. The author claims a “pragmatic approach,” as a new paradigm leading social research methods, which emphasizes the relevance of shared meaning and joint actions in evaluation management processes. As Morgan (2017) argues “the essential emphasis is on actual behavior (lines of action), the beliefs that stand behind those behaviors (warranted assertions), and the consequences that are likely to follow from different behaviors (workability)” (p. 67).

Rossi et al. (2003) use the concept of “social ecology of evaluation” to address the social arena where evaluations are implemented, since organizational members’ attitudes toward evaluation could vary. Referring to this practice, additional contextual factors are again related to the role of stakeholders involved in the program evaluation process and to the role of evaluators themselves (Fetterman et al., 1996; Askew et al., 2012; Patton, 1998).

Fetterman et al. (1996) address the concept of empowerment evaluation, “an approach to gathering, analyzing, and using data about a program and its outcomes that actively involves key stakeholders in all aspects of the evaluation process” (p. 814). The rationale underlying this perspective lies in the crucial role of collective and collaborative interpretation, discussion and reflection on evaluation findings. The more closely stakeholders are engaged, the more likely they are to use results to improve their work.

Over the last few years, constructivist approaches to evaluation which enhance engagement have acquired more relevance, such as, for example, participatory, collaborative and empowerment evaluations (Askew et al., 2012).

Since evaluation processes refer to multiple stakeholders, the aspiration of satisfying all stakeholders’ expectations appears unrealistic and so the evaluator’s role can be seen as
making decisions about the order and priority of stakeholders. Taking a utilization-focused
evaluation approach, Patton (1998) points out that evaluation should provide strategic
support to decision-making processes and therefore the final users of results are given more
priority among stakeholders.

Turning to the role of evaluators, this can be affected by the evaluative culture (Julian,
1997; Jackson, 2013; Mayne, 2008). According to TBE, logic models can be most useful when
integrated with the theory of change. Logic models are concerned with how programs
ideally work, by representing them as sequences of steps from inputs to the final impacts
(Julian, 1997) whereas theory of change establishes a clear understanding of how a program
actually works. Theory of change refers to “the construction of a model that specifies
(usually visually) the underlying logic, assumptions, influences, causal linkages and
outcomes of a development program or project” (Jackson, 2013, p. 97). Through the theory of
change, increasingly it has become clear in TBE studies that the contextual factor of
evaluative culture, affected for instance by different management styles, can impact the
functioning of a program, and in turn, its effects. In fact, programs are the results of the
aggregation of a larger set of assumptions, risks, and contextual factors.

Moreover, any organization features a more or less explicit evaluative culture, which
“denotes an organizational culture that deliberately seeks out information on its performance
in order to use that information to learn how to better manage and deliver its programmes and
services, and thereby improve its performance” (Mayne, 2008, p. 1). According to Mayne
(2008) a number of factors foster an evaluative culture: committed organizational leaders,
regular communications on results management, a balance between accountability and
learning, capacity for results management, supportive structures and policies.

Evidently, an organizational culture focused on quantitative measures and/or
economic outputs will encourage professionals to apply strategies assigning an economic
value to communication results. Therefore, auditing the evaluative culture is a necessary
condition for program evaluation because evaluation itself is a contextual process and is
affected by the organizational situation in which it takes place. Auditing the evaluative
culture is also important in order to understand the environment in which evaluation will
take place and to make changes to solve problems as a result of that environment that
might affect the evaluation.

Finally, contextual factors relate to the support of information technology, to the way
final reports are created and to using results (McDavid and Hawthorn, 2006; Macnamara,

According to McDavid and Hawthorn (2006), in order to implement an E&M
management process, it is necessary to take into account and establish the multi-channel
ways of communicating to facilitate the sharing of information, measurement problem
identification, and solving. Information technology has to be tailored to the program that
needs to be evaluated and final reports and the use of results have to be customized and
negotiated by stakeholders. Program evaluation scholars evidence the crucial role of
iterative planning that uses results and findings collected at each stage of the program
evaluation process in order to adjust strategies and tactics (Macnamara and Likely, 2017).
Knowlton and Phillips (2013) for instance depict program evaluation as a cycle
characterized by an iterative approach.

Performance measurement research
At the turn of the twenty-first century in the USA, PM became a critical and much debated
issue, since it was connected primarily with financial accountability to measure costs and
outputs of governmental programs. PM can be distinguished from performance
management. PM “provides the basis for an organisation to assess how well it is
progressing towards its predetermined objectives, helps to identify areas of strengths and
weaknesses, and decides on future initiatives, with the goal of improving organisational performance” (Amaratunga and Baldry, 2002, p. 218). Performance management can be defined as “the use of performance measurement information to effect positive change in organizational culture, systems, and processes by helping to set agreed-upon performance goals, allocating and prioritizing resources, informing managers” (Sachs, 2011, p. 58).

As previously discussed, PM and program evaluation are complementary ways of acquiring and analyzing information about a program and they rely on a common core of methodologies (McDavid et al., 2012).

Concerning the contextual factor of evaluation questions and measures, while program evaluation allows users to address questions of why observed programs results occurred, PM is more focused on cost-effectiveness and describes patterns of actual programs and organizational results (McDavid et al., 2012). In other words, PM describes what is going on, but it does not explain why it is happening. Measures usually derive from strategies and are clearly identified and simple to understand and use (Lynch and Cross, 1991; Maskell, 1989; Neely et al., 1996). Similarly, also the evaluation approach and design in the PM literature are considered as a routine to replicate (Neely et al., 2000).

Within the performance measure field, managers play the role of evaluators, that means they determine the main aspects on which program and employee are assessed and they identify the consequences related to the measured performances (Kerssens-van Drongelen and Fisscher, 2003).

Moreover, since the introduction of PM is an important change to the organization’s way of doing business, it is important that there are organizational leaders who are considered champions of this in order to provide continuous support to the process and to exert the power of evaluator. However, measurement sometimes increases uncertainty among managers since results are unpredictable and therefore some of them resist the introduction of PM.

Concerning the role of stakeholders, results are shared among them. According to Band (1990), performance measures should have support from management and stimulate engagement among stakeholders in order to create a feedback loop that enriches the interpretation of results.

Over the last 20 years, the PM literature has moved from considering the structural and technical features of the measurement process to examining contextual factors, such as the evaluative culture and management style, as well as other behavioral aspects (Neely, 2005). Here, context relates to organizational, relational, social and cultural factors that could affect how the measurement process is implemented and how measurement results are used to improve organizational performance. When a measurement system is context-sensitive it has reached a high level of maturity: “Maturity, in the context of PM, is defined as the ability to respond to the environment in an appropriate manner through performance measurement & management practices” (Bititci et al., 2015, p. 25). A positive relationship exists between the maturity of the performance system and the quality of obtained performance. If the measurement system is mature, it is context-optimized, results are used effectively to support managerial routines based on communication, which in turn facilitates informed-decision making and helps to establish an organizational learning culture (Neely, 2005). Scholars also emphasize the role of cultural differences among those involved in evaluation processes which is crucial in shaping attitudes and orientations toward E&M (Mattila, 1999; Winstead, 1999). Kravchuk and Schack (1996) suggest that the most appropriate metaphor to build a performance culture is the learning organization by Senge.

In order to learn from results, organizations need to rely on information technology procedures that not only support data collection, but also facilitate the top-down, bottom-up and horizontal sharing of information and help with problem identification and problem solving (Wettstein and Kueng, 2002).
Results obtained through information technology procedures are translated into final reports that are developed through routinized processes (Douwe et al., 1996). The use of such results ideally evolves over time as management guides the PM process (Hatry, 2006).

**Discussion**

To draw together what has been explored in the previous section, Table I compares the different roles that contextual factors play within program evaluation and PM perspectives. While considering their centrality, each of the two perspectives conceptualizes in a different way the contextual factors in the evaluation process.

**Role of contextual factors in program evaluation and performance measurement (RQ1)**

In sum they begin to answer RQ1. Concerning the first two factors in Table I (evaluation questions and measures, and evaluation approaches and designs), program evaluation includes *ad hoc*, tailor-made and customized processes taking into account organizational evaluation needs.

PM refers to routinized and regularized processes, which are performed on an ongoing basis. Program evaluation holds in high regard the criteria used to select measures since they have to be fully aligned to the purposes of the measurement process. With PM the selection of measures takes place on a one-off basis and their collection is seen as a routine practice. According to program evaluation *ad hoc* measures are identified for each evaluation process. Program evaluation attaches great importance to the fact that the research methods used have to be built on sound scientific grounds in order to ensure the replicability of the evaluation process. However, scientific approaches entail methodological skills, longer implementation time and more resources.

<table>
<thead>
<tr>
<th>Contextual factors</th>
<th>Program evaluation literature</th>
<th>Performance measurement literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Evaluation questions and measures</td>
<td>Questions emerge from stakeholder interests at one point in time (issue specific) (Pawson and Tilley, 1997; Scriven, 1996)</td>
<td>Routinised, concerning general management issues (McDavid et al., 2012; Lynch and Cross, 1991; Maskell, 1989; Neely et al., 1996)</td>
</tr>
<tr>
<td>2. Evaluation approach and design</td>
<td>Selected at the beginning of the process and driven by <em>ad hoc</em> needs on communication projects (Weiss and Bucuvavas, 1980; Fetterman et al., 1996; Morgan, 2007; Rossi et al., 2003)</td>
<td>Assumed, measurement is a routine; designed for providing regular information for organization (Neely et al., 2000)</td>
</tr>
<tr>
<td>3. Role of evaluator</td>
<td>Evaluator as an external auditor (Patton, 1998)</td>
<td>Managers are evaluators (Kerssens-Van Drongelen and Fisscher, 2003)</td>
</tr>
<tr>
<td>4. Role of stakeholders</td>
<td>Collaborative, empowering, participatory (Fetterman et al., 1996; Askew et al., 2012)</td>
<td>Results are shared among evaluation stakeholders (Band, 1990)</td>
</tr>
<tr>
<td>5. Evaluative culture</td>
<td>Evaluator should analyze evaluation culture (Julian, 1997; Jackson, 2013; Mayne, 2008)</td>
<td>Aligned between measurement and organizational systems (Neely, 2005; Bititci et al., 2015; Mattila, 1999; Winsted, 1999; Kravchuk and Schack, 1996)</td>
</tr>
<tr>
<td>6. Final reports</td>
<td>Reports are customized (McDavid and Hawthorn, 2006; Macnamara and Likely, 2017; Knowlton and Phillips, 2013)</td>
<td>Reports are developed through routinized processes (Douwe et al., 1996)</td>
</tr>
<tr>
<td>7. Information technology support</td>
<td>Seldom used, tailored to the project evaluated (McDavid and Hawthorn, 2006; Macnamara and Likely, 2017; Knowlton and Phillips, 2013)</td>
<td>Relied on IT procedures for ongoing collection of data (Wettstein and Kueng, 2002)</td>
</tr>
<tr>
<td>8. Uses of results</td>
<td>Uses are usually negotiated up front (McDavid and Hawthorn, 2006; Macnamara and Likely, 2017; Knowlton and Phillips, 2013)</td>
<td>Uses evolve over time due to management’s needs (Hatry, 2006)</td>
</tr>
</tbody>
</table>

Table I. Comparison between the role of contextual factors in program evaluation and performance measurement studies.
Concerning the second set of factors (role of evaluator, role of stakeholders and evaluative culture), the two perspectives have some similarities and differences. In program evaluation, the role of the evaluator is to guarantee the right research design, according to his/her expertise. However, in PM, given the evaluation process is essentially a management activity, the manager plays the role of an evaluator in charge of selecting, collecting, and interpreting metrics as well as of strategically managing the whole measurement process.

Identifying stakeholders’ expectations of the E&M process is crucial. Individuals other than communication managers and technicians might be interested in the evaluation process, so they have to be properly informed and their needs met. Program evaluation identifies stakeholders at the beginning of the process in order to understand their needs and expectations, but in PM, stakeholders are informed about results through periodic reports.

Organizational members’ attitudes toward evaluation could vary. All organizations feature a more or less explicit evaluative culture, which includes the set of measures that the management regards as important as well as the goals to be measured through the evaluation (e.g. a tool for punishing or rewarding staff for improved performance). Auditing the evaluative culture is also important to understand the current state of affairs and to make changes to solve any issues that might affect the evaluation. Similarly, PM should be aligned with the managerial systems used by organization otherwise it could create separate information chains and prompt criticism by organizational subjects. Addressing the last group of factors (IT support, final report and use of results), the communication of results is an essential part of measurement and evaluation processes. How the results are communicated could be different according to the used approach. Within program evaluation, reports are created ad hoc, while in PM, reporting activity is continuous, systematic, and follows standardized procedures. Moreover, in PM, information systems provide data from different sources, which are automatically integrated, elaborated, and systematized in standard reports. The use of information systems is less systematic within program evaluation.

Influencing contextual factors for E&M management processes in the communication field (RQ2)
The analysis of the role of contextual factors in program evaluation and PM studies reveals a new research issue which deserves attention: the need to investigate how to effectively implement and execute E&M management processes and what those contextual factors are which constitute potential sources of strength and weakness, by going beyond looking at the individual competencies and demographics of communicators as well as the time and budget available for E&M processes.

Specifically, the contextual factors previously discussed have been clustered in four groups. Each cluster is related to the perspective driving how evaluation is undertaken within the internal context. E&M processes can be affected by multiple factors, such as organizational characteristics, cultural traits, relational attributes and communicative features characterizing the internal context. This paper argues that these factors included in each cluster need to be addressed in order to guarantee effective implementation of E&M management processes. Figure 1 synthesizes the underlying clusters: evaluative capacity and history (organizational context); evaluative culture and leadership (cultural context); stakeholder-evaluator relationship (relational context); and evaluation communicative network (communicative context).

Evaluation capacity and history (organizational context)
Evaluation capacity is the knowledge, behavioral and affective skills that enable employees to think more positively about evaluation (Duffy, 2007; Preskill and Boyle, 2008;
Stockdill et al., 2002). Building evaluation capacity is an essential step for the effective execution of E&M management processes, because it enables people to embrace systematic enquiry. Systematic enquiry lies in routinely asking questions and seeking answers (Preskill and Boyle, 2008). As previously noted, participative and collaborative forms of E&M are becoming more popular (Patton, 1998). Participation encourages learning, but evaluative capacity needs further strategies if it is to be developed. Dedicated training, technical assistance, written materials, communities of practice, meetings are concrete examples of strategies to teach people how to conduct evaluation (Preskill and Boyle, 2008).

Understanding at what stage in the organizational lifecycle the evaluation process is carried out is pivotal for communication managers. For instance, the needs of a company that wants to evaluate internal communication in the normal course of business will be different from those of a company that experiences a significant process of change to its identity and core values. Accordingly, the evaluation process will depend on the situation in which the organization finds itself and the strategies put in place.

The first action that should be done to devise an effective evaluation management process is to analyze current and past communication activities and processes and the procedures put in place to do that. Neglecting what has already been done, setting a benchmark, is important so that new data can be interpreted. Existing evaluations should be scrutinized by looking at the measures in place without being distracted by the way results are collected. Prior to auditing, it is also advisable to look at case histories and studies from other organizations that involve topics the organization under examination might be interested in. In other words, the evaluator should assess if other organizations have dealt with the same questions at the time of evaluation and what answers have been given in order to build on or move away from their approach.

Evaluative culture and leadership (cultural context)
An evaluative culture is a set of “shared and persisting positive beliefs about the value of evaluation to the organization and enduring commitment to evaluation practice and the use of findings among its members” (Preskill and Boyle, 2008, p. 451). Since any organizational culture consists of unwritten and unspoken guidelines on how to get along in an...
organization and conveys a sense of identity to employees (Mayne, 2008), an evaluative culture is a part of an organizational culture that values evidence-based learning.

Evaluative culture can be characterized as a culture of results (Siddiquee, 2010), a culture of performance (Peters, 2004) and a culture of inquiry (Dana et al., 2002). Marr (2006) makes a distinction between “Command & Control” vs “Enabled Learning Environment” cultural approaches to evaluation. In “Command & Control” cultures, the focus is exclusively on the attained outcomes, without any interest in setting up effective learning circles. Such a cultural approach might generate misbehaviors, frustration and fear of results because people are exclusively assessed on results without any learning on how to improve their performance. Conversely, creating an “Enabled Learning Environment” means working on the results to gain insights and seek out new ways of undertaking daily activity.

Leadership has been considered an important topic in the field of management and has a huge literature which cannot be fully considered here, however, evidence indicates that there is a close relationship between leadership style and the effective implementation of E&M management processes (Obiwuru et al., 2011). Managers can adopt a different leadership style in implementing E&M management processes depending on whether: others help them to achieve their objectives more efficiently and effectively; results are linked to valued rewards; and, they have the resources needed at their disposal to accomplish the desired aims. The literature distinguishes between transformational and transactional leaders: “Transformational leaders are likely to be more effective than transactional leaders at inspiring followers to identify with a mission while rallying them to work together to achieve higher performance” (Kahai et al., 2003, p. 500).

Communication managers hence need to reflect upon evaluation culture and leadership commitment to E&M processes. They play a pivotal role in shaping evaluative culture by helping communication staff to engage with E&M management processes and coordinating collective efforts to ensure the right quality of data is collected. If the leaders’ commitment to E&M is high, it indicates that the organization places importance on outcomes and on their use in decision-making.

Stakeholder-evaluator relationship (relational context)
Every evaluation represents a nexus in a set of political and social relationships. While developing an evaluation plan, each evaluator should carefully analyze the context of evaluation to determine who the evaluation stakeholders are, what they want and why they want what they want (Rossi et al., 2003). Stakeholders’ expectations about performance are also at the center of PM research. “Good performance management requires that a manager knows when expectations have been accomplished and is able to demonstrate this. Clear and concrete expectations about the outputs and outcomes to be accomplished enable a manager to do this, and to know how successful a program has been and where improvements are required” (Mayne, 2004, p. 35).

For nurturing effective relationships with those with interests in the programs high quantity contact is vital in communication E&M, this concerns specifically top managers and other internal partners. More specifically, Torres et al. (1997) suggest maintaining close contact with stakeholders along the different phases of the evaluation process through ongoing communication, periodic meetings and informal conversations.

At the same time, the evaluator should prioritize stakeholders since the type of relationships varies depending on the stakeholder considered. Patton (1998) assumes that each evaluation process should point out who will use results. Dissemination of results should be clearly addressed to primary stakeholders (e.g. top managers, collaborators, internal partners) and secondary stakeholders (e.g. journalists, customers, service providers or agencies), by using a language suitable to their concerns and depending on what aim the evaluation has. In terms of the mode of relationships, more and more organizations call for
evaluators to work collaboratively with stakeholders, by implementing participatory or engaging forms of evaluation (Patton, 1998) and empowerment evaluation models (Fetterman et al., 1996).

Evaluation communicative network (communicative context)
Evaluation rests on collecting real-time and up-to-date information derived from a variety of sources. For this reason, the evaluator should primarily identify the sources and set up an information network to gather data, providing network members with tools and guidelines to make data usable and reliable. Sustainable E&M practices require systems, processes, policies and plans that help embed evaluation work into the way the organization accomplishes its strategic mission and goals (Preskill and Boyle, 2006, p. 444). Currents trends in communication field related to big data, digital transformation and artificial intelligence (Forbes Communication Council, 2018) have posited new challenges to communication departments and practitioners on how data can be analyzed and what technological and organizational infrastructure are needed to strategically convert these insights into actionable decisions.

As noted previously, the identification of metrics in use is one of the main streams of communication research. It is crucial to understand the metrics used by other functions, if, for instance, they are more financial or non-financial related, quantitative or qualitative. It might also be relevant to understand the CEO’s satisfaction with E&M processes employed by other functions. In some cases, management prefers dealing with numbers so the outcomes of the evaluation process should be expressed in numerical terms if possible. In other cases, words are preferred when reporting results (Antonioni, 1996; Waldman et al., 1998). Interestingly, Fersl and Bruskiewicz (2000) found that, when provided with the same results in a quantitative and then a qualitative format, managers pay more attention to the latter.

Conclusions, limitations and future trends
E&M is a critical topic of inquiry among communication scholars and practitioners. Considerable research and professional efforts have been addressed so far to discussing what should be measured and which methods should be implemented. Most of the E&M models seem to carry a positivist imprint, but, in real organizational life, organizations could not have clearly defined aims, sufficient resources or adequate informative systems to support E&M management processes.

Our scoping review has shown that E&M of strategic communication practices mainly borrow principles from two disciplinary domains: program evaluation and PM. By analyzing the legacy of studies on program evaluation and PM and combining the insights with communication research, we identified multiple contextual factors which impact the execution of an E&M management process: evaluation questions and measures; evaluation approach and design; role of the evaluator and role of stakeholders; evaluative culture; final reports; information technology support and use of results.

However, strategic communication literature has so far largely neglected to consider contextual factors when exploring barriers as obstacles of E&M implementation in practices. To fulfill this gap, this paper has theorized the contextual factors relevant to the management of the evaluation process. In fact, only a few studies in communication have focused on contextual factors affecting E&M management process success or failure. Against this background, this paper begins to fill an apparent gap by categorizing those factors affecting any E&M management process into four clusters. Thereby, this paper extends current thinking on communication E&M in three ways. First, it explicitly includes the importance of a fit between the main features of the organizational evaluative environment and the E&M management process. Second, it explicitly addresses the role of relationships among evaluators and evaluation stakeholders. Finally, it explores the role of technology and the use of E&M results in order to manage reliable and meaningful E&M management processes.
The findings of this scoping review demonstrate that there are great opportunities for additional research deepening knowledge about the implications of contextual factors within evaluation management processes in the strategic communication area.

For what concerns implications for practice, it is argued that methodological sophistication and scientific rigour are frequently the first elements considered when setting up a plan for measuring communication outcomes. While there is no doubt that these are critical, practitioners should bear in mind that effective E&M strategies are built upon factors including organizational, cultural, relational and communicative ones. Several studies in the field of program evaluation and PM have already evidenced that measurement is not a linear process. But, communication professionals have not yet fully embraced the importance of developing a full awareness of which factors, besides the methodological ones, can influence the effectiveness of an evaluation management process. Hence, it is argued that considering contextual factors, such as the organizational (evaluative capacity and history), cultural (evaluative culture and leadership), relational (stakeholder-evaluator relationship) and communicative ones (evaluation communicative network), will help communication managers better deal with stakeholders’ expectations of evaluation while at the same time gaining trust in the eyes of those organizational managers demanding that communication activities and programs are evaluated. The absence of empirical data represents the first limitation of the article that future research could address: the contextual factor-based clusters identified need to be tested empirically ideally via qualitative and quantitative case studies. Second, our scoping review has not covered the entire literature about evaluation program, PM and performance management streams of research: future research could enrich the list of contextual factors that could affect the E&M processes. Since our review has focused on examining internal context, future studies could address the analysis of external context by identifying contextual factors that could affect E&M processes outside the organization.

References


Investigating the role of contextual factors


Further reading


Institute for Public Relations (2003), Guidelines for Measuring Effectiveness of PR Programs and Activities, Commission on PR Measurement and Evaluation, Institute for Public Relations, Gainesville, FL.


Corresponding author
Grazia Murtarelli can be contacted at: grazia.murtarelli@iulm.it

For instructions on how to order reprints of this article, please visit our website: www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com
Maturity as a way forward for improving organizations’ communication evaluation and measurement practices
A definition and concept explication

Nathan David Gilkerson
Strategic Communication, Marquette University, Milwaukee, Wisconsin, USA

Rebecca Swenson
College of Food, Agricultural, and Natural Resource Sciences, University of Minnesota, St Paul, Minnesota, USA, and

Fraser Likely
Fraser Likely PR/C Performance, University of Ottawa, Ottawa, Canada

Abstract

Purpose – The purpose of this paper is to propose an explication of the concept of “maturity,” as it applies to communication evaluation and measurement (E&M) practice, along with contextualization of recent maturity model adoption within academic and professional communities.

Design/methodology/approach – Drawing from previous work on maturity models within other fields, recent communication scholarship and industry practice, this paper fills a gap in the literature by offering a theoretical conceptualization of communication E&M maturity, including the construct’s core dimensions and sub-dimensions.

Findings – Communication E&M maturity is conceptualized into four essential elements: holistic approach, investment, alignment and culture. The contribution of E&M efforts is represented as the direct support of corporate strategy, and ultimately increased value, from the communications function. Operational elements of maturity include levels of analysis, time, budget, tools, skills, process, integration, motivations, relationships and standards.

Originality/value – In exploring the factors necessary for “mature” E&M programs, and specifically emphasizing the need for a holistic approach, along with sufficient investment and alignment, and conducive cultural factors, the research builds upon existing work examining how communication can serve to inform corporate strategy and create value for an organization. Greater understanding and application of the maturity concept has the potential to advance the field by increasing both accountability and credibility for the work done by the communications function.

Keywords Evaluation, Corporate communication, Measurement, Public relations, Communication management

Paper type Conceptual paper

1. Introduction

Much has been written in the past decade, since it was first highlighted by researchers Anne Gregory and Tom Watson (2008), about the perceived “stasis” afflicting the realm of public relations and communication evaluation and measurement (E&M). Scholars have recently noted the field can look back on more than four decades of study on how practitioners and educators have evolved their efforts to determine the value and measure the effectiveness of public relations, and to promote more sophisticated evaluation within the profession (Volk, 2016; Likely and Watson, 2013). Meanwhile, others have compellingly argued that true evolution (i.e. improvement and innovation) in the area has stalled and perhaps even turned into deadlock, requiring new models and approaches to advance the profession (Macnamara, 2015).

Risking topic fatigue in some circles, the profession continues to engage in debate about the relative strengths and benefits of various E&M models and frameworks, alongside
earnest discussion regarding industry best practices and topics such as measurement standardization and strategies for encouraging practical implementation. For the vast majority of communication leaders within a specific company or organization, however, the key questions most commonly posed related to E&M practice focus on where their company stands among its peers and competitors — and what specific steps can be taken to improve the organization’s E&M processes and outcomes. These questions point directly to the maturity of an organization’s E&M program, and aspects such as the relative investment and sophistication a company applies to its communication evaluation efforts.

A sign of its perceived utility, the concept of communication E&M maturity has become a focal point of independent but parallel initiatives occurring within both industry and academia. In late 2018, the International Association for Measurement and Evaluation of Communication (AMEC), an industry trade body comprised of vendors, consultants and service providers, announced a major new initiative promoting its “M3” (pronounced “M-Cubed”) E&M maturity mapping tool, which is designed to allow users to complete self-assessments of their company or organization’s measurement sophistication and practices. Concurrently, a group of academics and practitioners (Swenson et al., 2018) published research similarly proposing the application of a multi-stage maturity model process to evaluate and advance E&M implementation.

While the notion of E&M maturity rests at the center of these initiatives, the core concept has yet to be fully explicated or defined. Although Swenson et al. highlight the opportunity the application of maturity may offer the field of E&M, the literature does not explicitly define the concept. Similarly, AMEC’s promotions surrounding the M3 tool imply maturity is linked to E&M program “sophistication,” but otherwise the concept is not defined. Thus, the following questions guide this research:

RQ1. What are the dimensions and sub-dimensions of E&M maturity?

RQ2. How might a specific E&M maturity model help to improve E&M practice?

This paper explores the utility of maturity-based programs by examining the concept’s origins and application in other fields. A review of the communication literature focused on barriers to successful E&M is then followed by an analysis of Swenson et al.’s (2018) maturity model, which is based upon the overcoming of these barriers. Informed by this scholarship, the research offers a formal definition and concept explication of “maturity” relative to communication E&M. A discussion of past industry efforts in advancing E&M practice, including drives for best practices and standardization, provides important context. Finally, the paper connects the maturity concept to current industry practice, including AMEC’s M3 tool, and discusses the unique dynamics of communication E&M along with opportunities for future maturity-focused research.

2. Literature review

Evaluation & measurement: terminology and reviews of existing scholarship

Within the communications industry, the terms E&M are often used interchangeably, with “measurement” sometimes serving as an umbrella term to broadly describe E&M efforts and activities (Macnamara, 2018b). In his 2018 book on the subject, however, Macnamara notes that measurement is simply the “taking of measures” (i.e. collecting and analyzing data), while evaluation is the use of measurement-based information to learn, and make a judgement or assessment about something – often its relative impact, success or effectiveness. Thus, in communication, measurement without intentional and considered evaluation carries limited value for an organization; evaluation of measurement data is what actually creates insights, informs strategic decision making and provides value to a business and communications function. Put another way, while measurement is a necessary component to the process,
especially considering increasing expectations from management and calls for accountability, “evaluation is essential” (Macnamara, 2018b). Several articles offer thorough reviews, assessments and historical accounts of the progression of communication E&M work (Volk, 2016; Macnamara and Likely, 2017; Likely and Watson, 2013; Watson, 2012; Lindenmann, 2005). Additional research has examined the evolution, and to some extent proliferation, of various models of communication E&M over the past several decades (Watson, 2012; Macnamara, 2014; Macnamara and Likely, 2017; Buhmann and Likely, 2018; Macnamara, 2018a).

Maturity and maturity model applications

As is evident from the robust literature and evolution of theoretical models, communication practitioners and leading academics alike have long advocated the communication field more fully embrace a culture of consistent, reliable and trustworthy E&M. Leaders have noted the adoption of E&M best practices not only promotes professional accountability, but strengthens the credibility of the industry as a whole while driving improved value from the communication function for clients and employers (Gregory and Watson, 2008; Macnamara, 2015). As noted earlier, however, the concept of maturity has only recently been applied to assessing and helping advance E&M practice. Before addressing the research questions specific to communication E&M maturity, an understanding of the origins of maturity models and their applications within other industries is useful.

Notions of evolution and a process of “completion” and advancement over time are evident in various academic publications that define the term “maturity model.” A piece in the journal Performance Improvement uses this definition: “A maturity model is a structured collection of elements that describe the characteristics of effective processes at different stages of development. It also suggests points of demarcation between stages and methods of transitioning from one stage to another” (Pullen, 2007). Similarly, an article focused on knowledge management emphasizes the versatility of the maturity concept, explaining how “Maturity models describe the development of an entity over time. This entity can be anything of interest: a human being, an organizational function, etc.” (Klimko, 2001).

An assessment technique that first gained traction in the field of software engineering, and has since been adopted by a wide range of industries, the application of maturity-model-style tools can help provide an organization with a simple yet effective method of evaluating the quality of its processes (Wendler, 2012). The modern concept of a maturity model can be traced back to 1979, with the publication of a book on quality management by Philip Crosby (1979), who proposed a quality management process maturity grid and outlined how maturity stages could be conceptualized as progressive building blocks, and thus also serve as a simple analysis and assessment tool. That same year, an article on the maturation of data processing was also published, and proposed six distinct and progressive stages of growth necessary to reach process maturity (Nolan, 1979). The basic “stage” technique has since been adapted to a wide variety of fields – and the approach has been used in numerous academic studies conceptualizing, developing and proposing custom models in varying applications. A 2012 meta-analysis of maturity model research that examined the content of more than 230 journal articles found maturity-model-focused publications in more than 20 different domains, with the large majority appearing in venues focused on software development and information technology, as well as government, engineering and education sectors, and business specialties such as supply chain management (Wendler, 2012).

One of the most widely used maturity models is commonly known and frequently referred to as the capability maturity model, or CMM. First developed with funding support from the US Department of Defense, the model’s original description notes “the CMM framework represents a path of improvements to increased software-process capability” (Paulk et al., 1993, p. 24). More recent overviews of the CMM (2018) describe it as an assessment tool that utilizes a “five-level evolutionary path of increasingly organized and systematically more
Maturity as a way forward

mature processes.” The representation of the CMM as the path, or the way forward, conveys the maturity model’s perceived connection to the notion of movement and progress. Today, a version of the framework serves as the foundation of services sold by the CMM Institute (2018), a diversified company offering products designed to “elevate business performance and provide insights for baselining and optimizing key organizational capabilities” – and used by over “10,000 organizations in 106 countries around the world.”

An important consideration of maturity is whether the concept promotes an idea of progression toward a final, perfect or “complete” ideal (i.e. full maturity), or if it conveys a continuous and never finished process. In his 2012 analysis of maturity model research, Wendler identified two points of view found in developed models: the life cycle perspective and the potential performance perspective. Life-cycle-style-models emphasize a defined evolutionary course and specify and denote a “final” stage, or state, of maturity. Potential performance models, while similarly structured with a progressive development path, are distinct in that the stages focus on potential improvements that may be achieved by progression: “Every stage holds an inherent effectiveness and self-evident value. The user has to decide by himself which level of maturity (i.e. completeness, perfection) is best for the situation” (Wendler, 2012). In considering the features and structures of various models, Wendler determined most contemporary models follow the potential performance approach vs the life-cycle design.

Key basic components of all maturity models include the definition of a set of stages, or levels, in a simplified manner (Klimko, 2001), which are presented to the user in a sequential or hierarchical progression, and designed to “be closely connected to organizational structures and activities” (Gottschalk and Solli-Seether, 2009). While some models are quite simple and represent just a single criterion and are thus described as one-dimensional, most models used currently are multi-dimensional and consider several elements and perspectives, such as corporate processes, organizational units and problem domains (Lyytinen, 1991). While interest in the development and usage of maturity models has grown substantially in the past two decades, Wendler (2012) in his meta-analysis determined that “many maturity models suffer a lack of a proper validation of their structure and applicability,” with qualitative approaches such as case studies (and interviews) used predominantly as the typical methods of validation, when such efforts are employed at all. A longitudinal and multi-method approach using empirical data at various stages of the model, and the application of insights from an ongoing combination of qualitative and quantitative assessments, is recommended for effective refinement and validation of a custom maturity model tool (Wendler, 2012). Reflecting on the purpose of maturity models, Wendler (2012) is careful to point out that while the tools do offer useful benefits, “they are no ‘silver bullets’” and cannot solve all problems.

A maturity model developed around overcoming E&M barriers

A significant portion of the existing literature on communication E&M has focused on perceived “barriers” to implementing successful programs. The earlier-referenced Swenson et al.’s (2018) maturity model (discussed in detail below), was developed through an in-depth examination of the literature on barriers, and with a general understanding that advancing maturity, as it relates to communication E&M, is directly related to an organization’s ability to work past, or overcome, these various barriers. The discussion below categorizes each area, or type, of barrier, and briefly reviews relevant literature, before exploring the development of the Swenson et al. model. Barriers to E&M identified in the literature include: lack of time and budget; lack of an organizational performance measurement culture; lack of competencies; lack of an appropriate set of measurement tools; lack of accepted best practices or standards around units of analysis, stages of evaluation, frameworks and terminology; and lack of involvement in the management’s strategic planning efforts.

Lack of time and budget. Lack of time is often cited in research and falls into one of two categories: no time to measure, or monitoring, tracking and measuring takes too much
staff time. For budget, explanations include: budgets are too tight to allocate resources for E&M or more sophisticated techniques are too costly. Lack of time and budget were regularly reported as primary barriers in earlier studies (Watson, 1994; Xavier et al., 2006; Wright et al., 2009) and put forward by practitioners (Lindenmann, 1990; Gaunt and Wright, 2004; AMEC, 2016a, b, c). Conversely, Macnamara (2015, 2017) and Macnamara and Zerfass (2017) suggest that these supposed barriers have less relevance today and should be dismissed, arguing that there are numerous low-cost options and the overall spend on strategic communication has grown.

Lack of organizational performance measurement culture. This barrier is presented as a lack of performance measurement culture at the level of the organization or business unit. It can also be understood as a lack of management demand for a strategic communication performance measurement program, or a lack of management support or interest in its results (Baskin et al., 2010). This has been expressed as the fear that management will question the value of communication performance results (Wright et al., 2009), low confidence in the value of available communication metrics (Gaunt and Wright, 2004) and a deep frustration with the management’s misunderstanding of strategic communication (Watson and Simmons, 2004). Other scholars have pointed to organizations that are not innovative or proactive, and thus not likely to have a self-evaluative culture (O’Neil, 2013; Thorson et al., 2015; AMEC, 2016a, b, c).

Lack of knowledge and competencies. Multiple studies have found that departmental expertise and analytical confidence was indeed missing (Gaunt and Wright, 2004; Baskin et al., 2010; Cacciatore et al., 2016; Zerfass et al., 2017). Watson and Simmons (2004) suggest that a lack of confidence among practitioners is “illustrated by an inability to make a case for evaluation budgets with their clients or managers” (p. 3). Research has also linked intention to conduct E&M work to practitioner attitudes and perceived behavior control related to E&M expertise and skills (Buhmann and Brønn, 2018).

Lack of tools. This barrier – variously described as tools, techniques or methodologies – is often explained as: the department’s technicians employ only basic process, output and outtake measurement tools; the organization itself does not employ more sophisticated performance measurement techniques; or other functions have ownership of more sophisticated measurement tools but the communication department lacks access (O’Neil, 2013; Tench et al., 2017). Ownership includes those functions involved with customer relationship management, brand image, employee engagement, stakeholder management, corporate reputation and corporate social responsibility. Many studies have also focused on the sophistication level of research methodologies employed by practitioners (Gregory, 2001; Baskin et al., 2010; O’Neil, 2013; Place, 2015). Some studies suggest that a practitioner’s emphasis on basic tools at the expense of advanced tools derives from either a lack of expertise (Xavier et al., 2006; Baskin et al., 2010; Michaelson and Stacks, 2011) or a misunderstanding of program objective setting and its complexities (Gregory, 2001; O’Neil, 2013).

Lack of standards. A lack of industry standards is a more recent identifiable barrier. Practitioners and scholars often articulate it as a lack of consensus on terminology and their definitions, a lack of consensus on a “best” evaluation model, and a lack of consensus on the various units of analysis for E&M. Various studies report this barrier (Wright et al., 2009; AMEC, 2011; Ragan, 2013; Thorson et al., 2015; AMEC, 2016a, b, c) and a number of researchers have commented on the state of industry standardization (Michaelson and Stacks, 2011; Macnamara, 2014, 2017; Macnamara and Likely, 2017). Researchers such as Watson (1997), Macnamara (2014) and Buhmann and Likely (2018) commented on the lack of shared definitions, even for terms like measurement and evaluation. Thorson et al. (2015) found that 25 percent of respondents in their survey of senior-level communication practitioners in the USA were trying to standardize in some way.
Lack of strategic management. With this barrier, research focuses on how communication planning goals are tied to client program objectives. It also addresses if the communication department is aligned with the strategizing, planning and decision-making functions of the organization and its business units, and if there is a strong link with organizational financial goals. Various studies have examined the placement of E&M within strategic management, particularly with regard to planning systems (Baerns, 2008; Baskin et al., 2010; Tench et al., 2017; Zerfass et al., 2017). Others have looked at the return on investment or concepts such as market mixed modeling as part of the process of linking communication results to financial goals (Gaunt and Wright, 2004; Baskin et al., 2010; Likely and Watson, 2013; Zerfass et al., 2017). Zerfass et al. (2017) report “the value of data for managing strategic communication seems to be overlooked by many communication departments today” (p. 12).

Drawing from the literature on barriers, and informed by a series of interviews conducted among large-company communication executives viewed as E&M industry leaders, Swenson et al. (2018) examined how organizations regarded among their peers as implementing highly effective measurement programs have successfully overcome the known barriers. As part of their analysis, Swenson et al. (2018) developed and proposed an assessment tool, described as a “scalable maturity model that aids in the development, formalization, and optimization of strategic communication measurement and evaluation” (Table I).

The Swenson et al. maturity model is multi-dimensional in its structured representation of four evolving stages or sequential categories (early, mid-, advanced- and optimal growth), seven common “barriers” to overcome, including time, budget, culture, knowledge/skills, research tools, industry standards and strategic management, and four distinct increasing “levels” of perspective (channel/media/product/message, campaign/program, organizational and societal). Along with outlining short descriptions of each distinct categorical stage, a unique contribution of the model is its final levels of analysis dimension, which promotes a holistic understanding of evaluation and guides the user’s attention through the spectrum of lower-level “tactical” perspectives (within early maturity), to the higher-level organizational concerns and societal, big-picture “license to operate” considerations. This perspective supports and informs the holistic approach core dimension of E&M maturity.

Although not explicitly named within the model, another key finding from Swenson et al.’s (2018) research was the crucial role of alignment within various aspects of successful E&M programs. Though often described by participants using varying anecdotes and terminologies – including the application of E&M for driving key adjustments (to tactics, and strategies), integrating communication programs (with broader business priorities) and effectively positioning the communication function within the larger organizational structure – alignment, in its various forms, was regularly viewed as an essential element of E&M that helped to bring value to the organization. This emphasis on alignment directly connects to recent scholarship examining and explicating the concept of alignment and its varied role as a “central aspect” within strategic communication (Volk and Zerfass, 2018). In detailing the various types and applications of alignment within (and outside) an organization, and its processes, Volk and Zerfass conceptualize and present a comprehensive definition of alignment, emphasizing effective alignment is essential to successful strategic communication. This understanding of the necessity of alignment similarly positions the concept as a core dimension of E&M maturity.

3. Toward a definition and conceptualization of E&M maturity

Scholars and practitioners alike have long emphasized the importance of clarity and shared understanding as it relates to terminology and key constructs. Despite its central application
to recent initiatives, communication E&M maturity has not been formally defined or explicated. Following a similar process used in past communication research for explication of core concepts (Kioussis, 2002; Volk and Zerfass, 2018), including the exploration of general historical background, identification of existing definitions, scrutiny of relevant literature, and review and determination of relationships between key elements and concepts (Chaffee, 1991), the construct of E&M “maturity” was closely examined.

The communication literature demonstrates a holistic approach for E&M is paramount, in that it is required for understanding and recognizing the complexity of the public relations function. Research also shows that without sufficient investment, however, none of the process occurs to begin with — and, similarly, without alignment (i.e. the application and usage of E&M efforts), the investment of both personnel and operational costs is of negligible value. Finally, scholarship has concluded that an organizational culture that is supportive and receptive to the benefits of E&M — and, ideally, demanding of its rigorous presence — promotes the necessary investment of resources and opportunities for alignment from its generated insights. Working in concert, the combined elements of a holistic approach, investment, alignment and culture allow communication E&M efforts to help inform corporate strategy. The ultimate support of strategy is essential because, as

---

**Table I.** Maturity model for communication department evaluation and measurement program sophistication

<table>
<thead>
<tr>
<th>Maturation Model</th>
<th>Time</th>
<th>Budget</th>
<th>Culture</th>
<th>Knowledge</th>
<th>Skills</th>
<th>Research Tools</th>
<th>Industry Standards</th>
<th>Strategic Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial Growth</td>
<td>Time is not an issue. We have sufficient budget to meet the costs of communication and business impact evaluation.</td>
<td>We report actionable insights on the communication costs and business impact of our projects.</td>
<td>We report on the communication costs and business impact of our projects.</td>
<td>We report on the communication costs and business impact of our projects.</td>
<td>We report on the communication costs and business impact of our projects.</td>
<td>We report on the communication costs and business impact of our projects.</td>
<td>We report on the communication costs and business impact of our projects.</td>
<td>We report on the communication costs and business impact of our projects.</td>
</tr>
<tr>
<td>Advanced Growth</td>
<td>Time is not an issue. We have sufficient budget to meet the costs of evaluation and measurement program sophistication.</td>
<td>We report on the communication costs and business impact of our projects.</td>
<td>We report on the communication costs and business impact of our projects.</td>
<td>We report on the communication costs and business impact of our projects.</td>
<td>We report on the communication costs and business impact of our projects.</td>
<td>We report on the communication costs and business impact of our projects.</td>
<td>We report on the communication costs and business impact of our projects.</td>
<td>We report on the communication costs and business impact of our projects.</td>
</tr>
<tr>
<td>Sluggish Growth</td>
<td>Time is not an issue. We have sufficient budget to meet the costs of output and outcome measurement.</td>
<td>We report on the communication costs and business impact of our projects.</td>
<td>We report on the communication costs and business impact of our projects.</td>
<td>We report on the communication costs and business impact of our projects.</td>
<td>We report on the communication costs and business impact of our projects.</td>
<td>We report on the communication costs and business impact of our projects.</td>
<td>We report on the communication costs and business impact of our projects.</td>
<td>We report on the communication costs and business impact of our projects.</td>
</tr>
</tbody>
</table>

---
visualized by Zerfass and Viertmann (2017), corporate strategy rests at the center of how communication can create value for an organization. Thus, drawing from and building upon the existing literature, including a synthesis of the scholarship on barriers, a definition for the concept of maturity is proposed:

**Maturity** within communication evaluation and measurement (E&M), dependent upon a holistic approach and understanding of the activity, is the relative degree of investment (in E&M efforts), alignment of activities and decision-making (within and outside the unit and organization), and organizational culture conducive to utilizing insights. The overall level of maturity serves to create value through the support of corporate strategy.

This conceptualization of E&M maturity can be thought of as a “four-legged stool,” in that each core component (a holistic approach, investment, alignment and culture) captures an inherently unique, but also essential, aspect of the phenomenon. As represented in Figure 1, E&M maturity depends on the presence and degree of each of the four necessary components. The definition allows for qualitative assessment of the its components, in particular aspects such as alignment or culture, which are essential for mature E&M but may be dynamic and difficult to quantify.

The above designation of four distinctive elements of E&M maturity speaks to the first part of RQ1, which asks about both the dimensions and sub-dimension of E&M maturity: the four legs of the stool are the main dimensions of the construct.

Addressing the second part of RQ1, which explores the sub-dimensions of E&M maturity, the proposed conceptualization draws more deeply upon the reviewed literature on E&M barriers. Extending the basic conceptualization of maturity outlined above to operational levels, the four core dimensions of E&M can be broken down into distinct and measurable elements. Designation of these sub-dimensions of E&M maturity allows for a simple visualization and understanding of the construct. As illustrated in Figure 2, the presence of a holistic approach can be operationalized through the adoption of levels of analysis (see Buhmann and Likely, 2018; Swenson *et al.*, 2018, for descriptions of various levels: messages, products and media/channels; campaigns/programs; organizational; society), and accepts that evaluation, and not “measurement,” is the primary goal and focus of activities, and allows communicators to see how value can be determined at various places or levels of production and output. The importance is the need to have a sophisticated understanding of the goods and services a function produces and their effect on various
aspects of organizational strategizing, planning and execution processes. Each level of analysis has its own measurement regime.

Arguably, the most straightforward dimension, investment, can be measured through basic aspects such as allocation of time and budget for E&M activities, use of specific tools (e.g., software platforms, etc.), and hiring and designating of skilled people capable of conducting analysis and insights generation work. As demonstrated in the review of existing E&M literature, lack of time and budget have historically been among the most frequently cited factors by practitioners as perceived barriers to implementing effective E&M programs (Watson, 1994; Xavier et al., 2006; Wright et al., 2009). Similarly, research has found professionals commonly view access to measurement tools, combined with trained staff with specialized evaluation expertise, as inherent E&M challenges (Gregory, 2001; Baskin et al., 2010). The grouping of these four sub-dimensions (time, budget, tools and skills) of E&M Maturity under the broader investment dimension speaks to the ability, or opportunity, organizations have for strengthening and prioritizing these elements of their E&M program, with an understanding that increasing levels (or building up capabilities) in these respective areas can pay dividends relative to insight generation and strategic success.

Alignment, as articulated by Volk and Zerfass (2018), is a highly complex and wide-ranging concept, yet it can also be broken down into two simple aspects: outcomes and processes. For the purposes of operationalization of maturity, one component of alignment has been labeled to represent the ongoing process of aligning organizational efforts based upon E&M insights, while the other, integration, represents the level of alignment that has occurred. Integration could be thought of as an assessment of the tangible “nuts and bolts” elements to implementing alignment-focused actions. Both are thus connected to the strategic management decisions made within the organization. In their examination of the construct, Volk and Zerfass (2018) defined primary alignment as rooted in the connection between communication strategy and overall organizational strategy, and secondary alignment as the rationality of various communication activities, compared against overall communication strategy and organizational goals. The Volk and Zerfass explication and framework of the concept also distinguishes between ideas such as internal and external communication alignment, alignment within and across organizational functions, and alignment in management and organizational processes. This conceptualization of alignment works in unison with the proposed sub-dimensions of E&M maturity, since E&M is inherent at least in some way to achieving each type of alignment. In breaking alignment into the two sub-dimensions within E&M maturity, the notion of process can be thought of as the degree of ongoing efforts being put into (each unique aspect of)
alignment – while integration is the level of successful alignment, throughout the organization, which has been achieved.

Finally, as the fourth component, elements of culture include the motivations (or lack thereof) within the organization, and its people, for conducting rigorous and effective E&M; the relationships between and among personnel, whether it be executive leadership dynamics or attitudes of lower-level support staff; and, finally, the adoption and regular usage of designated standards for collecting, measuring and reporting communication activities and metrics. As noted earlier, the broad category of organizational culture has been highlighted by numerous scholars examining E&M barriers, with research focusing on whether specific environments value and demand performance measurement within the communications function. Research has examined aspects such as accountability expectations from organizational leadership and management’s awareness and understanding of the dynamics of strategic communication (Wright et al., 2009; O’Neil, 2013). Interview-based research has shown that relationships among colleagues can influence E&M practice, in particular rapport between communication function leaders and organizational management related to aspects of trust and credibility (Swenson et al., 2018). The conceptual dimension of culture also encompasses the organization’s attitude toward and application of communication measurement standards, a function of the professionalism (and rigor and consistency) with which E&M efforts are conducted. (Discussed in detail below, standardization of E&M practices has seen limited success.)

4. Discussion: maturity models and advancing E&M practice in industry

While the refinement of models and explication of conceptual definitions is an important exercise, a likely question from communication leaders working in industry aligns with RQ2, which specifically asks how maturity model-based tools might help to guide E&M practice forward. Before exploring that question, however, it is worthwhile to examine the evolution of past E&M initiatives from professional groups and associations, including some of the challenges such efforts have faced while working to advance E&M adoption in the field. The current initiative to leverage maturity as a driver of E&M is best understood with historical context via a review of recent industry-sponsored E&M efforts.

Referenced earlier, a key international entity within this area of the industry, particularly among traditional media, social media and digital measurement consultants, vendors and service providers, has been AMEC, a trade body which was originally formed and known as the Association of Media Evaluation Companies (Watson, 2012). AMEC (2010) has provided industry leadership with a series of measurement-focused initiatives in the past decade, including the original 2010 issuing of “The Barcelona Declaration of Measurement Principles,” a list of seven basic measurement principles, or broad “Best Practices,” to assist communicators in assessing and adopting communication evaluation efforts focused on things like outcomes vs outputs. An outgrowth of the principles, in 2011, AMEC also launched a measurement effectiveness assessment instrument titled the “Valid Metrics Framework” – which was followed up, a year later, with an updated version incorporating social media considerations (AMEC, 2011, 2012). Several years later, AMEC also promoted a re-launch of the Barcelona Principles, with the newer “2.0” version addressing the heightened importance of social media, and promoting a stronger emphasis on evaluating outcomes (vs only measuring outputs and outtakes) and connecting evaluation results to overall organizational performance (AMEC, 2015a, b).

AMEC (2016c) launched its “Integrated Evaluation Framework” (IEF), an online tool which walks the user through a sequentialized process of assessing an organization’s communications efforts within categories, including Objectives, Inputs, Activities, Outputs, Outtakes, Outcomes, and Organizational Impact, and applying “drop down” style menu options to help determine and develop evaluation strategies (AMEC, 2016c). As part of the
introduction of the IEF, AMEC (2016b) highlighted how the tool’s design was rooted in social science research and developed by “look[ing] beyond PR evaluation models” and applying knowledge from other fields and disciplines such as performance management. An AMEC (2016a) web page tells readers how the framework “shows how to operationalise” the Barcelona Principles and demonstrates how to turn Principles into action, and to finally prove the value of our work.” Most recently, in late 2018, AMEC (2018) launched its “M3 Measurement Maturity Mapper” (described below), an online tool which positioned the existing elements of the organization’s recommended metrics and framework within the broader context of maturity. From 2010 to 2018, AMEC’s focus moved from measuring effectiveness of traditional media, social media and digital communications to evaluating campaign planning inputs, campaign outcomes and the effect of those outcomes on business unit and organizational strategies and, ultimately, organizational goals. This evolution of perspective is reflective of the conceptualization of E&M maturity proposed here, in that AMEC’s more recent initiatives, particularly the M3, have adopted a more holistic approach.

The challenge of establishing “standards” of E&M
Another key element, as well as perhaps one of the most debated and challenging aspects, of the push toward improving communication E&M practice has been the mixed success of academic and industry attempts to designate and agree upon specific “standards” of public relations and communications research metrics and terminology. As Macnamara (2014) noted in 2014, “despite a move towards standards, there is lack of consistency and agreement within industry even on basic metrics” (p. 13) – and part of the reason the field is “going round in circles” is connected to weak engagement between practitioners and academics attempting to develop standards, a dynamic which unfortunately has “resulted in insular debates and simplistic solutions” (p. 23). Early industry and scholarly efforts were made attempting to define and agree upon numerous measurement related terms, which multiplied with the advent of social media (Lindenmann, 1997; Stacks, 2006; Michaelson and Stacks, 2011; Paarlberg, 2013; Paine, 2018). With newer labels like “views,” “tone” (or “sentiment”), and “engagement,” being added on top of traditional output focused metrics, such as audience, reach and impressions, digital technology – and the widespread usage of social channels for public relations campaigns – compounded the need (and spurred industry desire) for common definitions of terminology and standard measurement approaches for communication metrics.

In 2011, a co-sponsored initiative, supported by industry organizations including the IPR, the Council of PR Firms, the Global Alliance for Public Relations and Communication Management, PRSA and AMEC, was formed to address the issue, with the establishment of the Coalition for Public Relations Research Standards (Macnamara, 2014). Following the guidelines and six-step adoption process recommended by the International Standards Organization (ISO) (IPR, 2018a), the Coalition collaborated with business and communication industry leaders, and in 2012 released a 12-page document outlining definitions and proposed methods for determining five basic standards related to traditional media analysis – with the initial guidelines “designed to be used in the interim pending industry feedback” (IPR, 2012). A year later, IPR announced four large US-based companies, including McDonald’s USA, General Electric, General Motors and Southwest Airlines (all “major buyers of public relations research and measurement services”) had agreed to adopt the first round of standards within their organizations (IPR, 2013a, b; Bradley, 2013). In his critical analysis of progress within the standards movement, Macnamara (2014) acknowledged other efforts around this time, including AMEC’s (2012) introduction of its Valid Metrics for Social Media standards, and work done by a wide-ranging group of industry representatives known as the “Social Media Measurement Standards Conclave” (or #SMMStandards Conclave).

Despite a 2012 AMEC European Summit presentation titled “The March to Standards” (Marklein and Paine, 2012), evidence shows achieving standardization within the industry in
the years since these efforts has been challenging and seen mixed results (Buhmann et al., 2018; Paine, 2018). Buhmann et al. state that “Standards development is advancing on two levels consistent with ISO recommendations: the level of technical standards for the measurement of communication, in other words, ‘how to standards,’ as in how to measure impressions and how to calculate ROI; and the level of process standards for evaluation as in the broader frameworks (of which measurement itself is just one important element)” (p. 115). Initial work under the Coalition for Public Relations Research Standards banner on traditional media analysis standards (e.g. see Eisenmann et al., 2015) and on social media standards (Paine, 2018) has stalled. The Coalition itself never did examine E&M frameworks.

As leading academics have noted, there are complex factors involved, especially among competing agencies and measurement service providers – many of which utilize “black box” approaches, based on proprietary data collection and analysis techniques, in selling clients their measurement services – which may be serving in some ways to create disincentives or structural impediments to the industry’s willingness, or ability, to adopt proposed standards (Macnamara, 2014; Buhmann et al., 2018). As Paine (2018) suggests, the work to date on social media measurement standards are more “guidelines and best practices” than actual standards. The same can be said for traditional media measurement standards.

Complexities, however, and the inherent challenges in identifying and implementing standards, are not solely reserved to industry, as was the determination of a recent “task force” group, made up of academics and industry practitioners, charged with examining existing public relations planning, objective-setting and measurement models, in order to offer a recommendation on model standardization (Likely, 2017; IPR, 2018c). Members of the Task Force on Standardization of Communication Planning/Objective Setting and Evaluation/Measurement Models (Buhmann et al., 2018), with a mandate to explore standards, produced a number of peer reviewed research papers as well as blog posts directed to professionals but this output did not focus on standards per se, given the Task Force’s determination after extensive study that measurement standards were probably impossible and that evaluation standards were more likely best left at evaluation best practice models. While limited progress has been made, identification of a single standardized (or universal) “model” (or system for conducting measurement and evaluations for use within the field), based upon existing tools, has been elusive. The industry saga of establishing E&M-focused standards is linked to today’s focus on maturity because maturity model-based tools do not necessarily require agreement related to specific processes or terminologies, but instead focus on the nature of an organization’s own internal practices.

Maturity as a driver of E&M – an industry association push for a maturity “mapper”

The proposed definition and conceptualization of maturity is useful when considering RQ2, and the broader question of how maturity models might help to guide E&M practice. From industry, AMEC’s (2018) launch of an interactive online tool, the “M3 Measurement Maturity Mapper” represents the most noteworthy application of maturity in the profession to date, and the ambition from industry leaders the concept will serve to help advance communication E&M adoption beyond the success of prior initiatives. The tool guides the user, presumably a corporate communication executive or measurement consultant working with a client, through a detailed questionnaire that asks questions about the user’s organization, such as size, industry, global region and current measurement activities related to earned, owned, shared (including social) and paid media. The M3 prompts users to indicate basic information, such as how frequently they report various metrics, such as volume of coverage (including “advertising value equivalencies”), coverage quality, social media posts and total impressions, before probing about more sophisticated activities such as the use of KPIs, evaluation benchmarks, measurement for planning, and the regular review and adjustment of organizational goals, objectives and strategy based upon evaluation-generated insights.
Connected to the above discussions, and this study’s conceptualization of E&M maturity, the Mapper also gauges aspects of organizational measurement culture, explores internal relationship dynamics, and asks users about the application of innovative technologies and “more advanced tools and techniques,” such as big data analysis, randomized control trials, and the tracking of cultural and societal trends (AMEC, 2018). Upon completion, the user is provided with a series of benchmarked percentages indicating their company’s total overall performance and relative scores in reporting, planning and impact categories. A final report highlights specific strengths of the user’s organization and recommends actions to improve the relative maturity level within each area.

In unveiling the new Maturity Mapper at the organization’s 2018 Global Summit event in Barcelona, AMEC representatives were careful to repeatedly stress to audience members that it was NOT “a model,” but instead was a “tool,” a diagnostic device, or just a “Mapper.” The implicit message, which was humorously depicted in an infographic summarizing the conference session – including numerous declarations (“It’s a TOOL!!” “Nope, not a model.” “Really: NOT a Model.”) – was that AMEC was aware of and sensitive to industry fatigue and past baggage from previously introduced models of measurement. Still, the M3’s design and use of progressive stages (basic, standard, advanced, fully integrated), and varied assessment elements of reporting, planning and impact, positions it squarely in the realm of traditional multi-dimensional maturity models.

The tool’s incorporation of culture-based questions, assessment of overall E&M investment, and items measuring the organization’s integration of E&M activities (i.e. alignment) also correspond directly with this study’s defined dimensions and sub-dimensions of E&M maturity. Furthermore, the M3’s climbing themed imagery (showing the pathway of measurement maturity as a journey to the top of the mountain) extends the map metaphor and conveys that the tool’s purpose is to guide the user forward. By emphasizing the benefits and benchmarking abilities of the M3, AMEC’s Mapper tool, with its easy-to-follow and understand interactive survey design, circumvents much of the perceived tedium and aversion associated with models while providing users a simple method for self-assessment and guidance for increasing measurement and evaluation sophistication. In addition, the M3’s reliance on self-assessment indicators, and use of relatively broad survey item terminology, similarly sidesteps much of the debate, confusion and controversy related to industry standards.

In reference to RQ2, it is too early to assess whether maturity-focused initiatives such as AMEC’s M3 will succeed in advancing E&M practice within the communications industry, but the tool does appear to be designed well to speak directly to the key questions and concerns of practitioners: where does my organization stand compared to peers and competitors, and what can we do to improve? In framing the conversation around E&M through the concept of maturity, tools like the M3 (and the Swenson et al. maturity model) explicitly show practitioners that the evaluation of communication is an ongoing progressive process, and one that can be strategically invested in and continually improved upon. The potential performance perspective of maturity model design allows users to assess levels of E&M that are achievable and optimal, and intentionally pursue steps necessary to improve. Maturity tools like AMEC’s M3, and its feature allowing users to benchmark their own organization’s E&M practices among industry peers, may harness the demands of a competitive marketplace, helping motivate communication executives to invest in measurement efforts – or perhaps provide data and persuasive “ammunition” to make budgetary requests to management for E&M prioritization.

Maturity models, and the explication of the concept’s key dimensions proposed here, should help users re-envision E&M barriers instead as potential strategic opportunities. Beyond “best practices,” which often focus just on certain aspect of E&M, or formal models which can preoccupy users with things like stages and categories, the concept of maturity acknowledges there may not be one best way to conduct E&M efforts – and that practical solutions are typically unique and specific to an organization’s goals and capabilities. Conceptualizing
maturity as reliant upon a holistic approach to M&E, combined with essential investment, alignment and culture, should help practitioners better understand the “big picture” dynamics within their evaluation efforts, while an appreciation of the concept’s sub-dimensions may support day-to-day decision making. Perhaps the most promising aspect of maturity model approaches for advancing E&M practice is that the maturity assessment process can help to simplify and demystify the (at times, esoteric and technical) topic and provide professionals with a practical method for initiating discussion around ways to improve.

A candid reflection and assessment of the complex nature of E&M culture

While maturity-focused initiatives give reason for optimism, it is also worthwhile to thoughtfully revisit the topic of culture, both at the organization-level and industry-wide, and the important role that dynamics such as relationships and motivations may play in the “real world” implementation of E&M efforts. Recent research by Romenti and Murtarelli (2018) addressed potential “conflicting logics” between communication departments and management leaders, which could influence the implementation and maturity level of E&M systems within an organization. In their contribution to this Special Issue, Murtarelli et al. (2019) argued for more research that closely examines contextual factors of effective E&M management processes, including, e.g. evaluative capacity and history, evaluation culture and leadership, stakeholder–evaluator relationships, and evaluation communicative network.

The wider perspective of the Murtarelli et al. research aligns closely with the proposed definition and conceptualization of E&M maturity, in that it speaks directly to notions of adopting a holistic approach to measurement, with multiple levels of analysis, and the role of organizational culture and professional business relationships. If a company does not value, or is even culturally resistant to, improving E&M efforts, spearheading such an initiative may be minimally beneficial or even present professional risks. In a similar vein, Swenson et al. (2018) also concluded that maturity of evaluation practice was linked to key cultural and behavioral aspects, including the overall measurement culture within an organization, strength of relationships with senior management, authority given to the communication department, and the ability to build and leverage reports with value, including actionable insights to influence organizational decision making. In other words, motivations and ability to advance E&M maturity are likely very dependent upon relationships and organizational culture. It might also depend upon the level of skill of individual communicators, as Buhmann and Brønn (2018) pointed out. In a study that used the theory of planned behavior to test factors that prevent or support measurement and evaluation efforts among communicators, they found that attitude toward outcome measurement and evaluation and perceived behavioral control, specifically skill or capability with measurement and evaluation, were the strongest drivers of practitioners’ intention to measure and evaluate communication initiatives (Buhmann and Brønn, 2018).

Other scholars have also speculated on ways in which behavioral components might stunt maturity. In a provocative challenge to many industry assumptions, Nothhaft and Stenås (2019) in this Special Issue suggest that communication practitioners’ self-interest as business people might hinder progress in E&M practices. According to their argumentation, the communication industry has continually been overpromising on communication outcomes and under-delivering on actual communication performance. Following this line of thought, maturity of communication E&M might be deadlocked because communication practitioners need to play this game to survive, and in reality, more rigorous practices might impede practitioners’ competitive advantage and demonstrable success (Nothhaft and Stenås, 2019). The perspective that advancing E&M maturity may not necessarily be in the best interest of many communication professionals – or, even at a macro level, the broader communications industry as a whole – is a thought-provoking consideration. The arguments posed by Nothhaft and Stenås are worth further study and reflection, and can be applied against several of the dimensions of E&M maturity including aspects of motivations and holistic approach.
5. Conclusion and directions for future research

In offering a definition of E&M maturity, and a practical concept explication of the dimensions and sub-dimensions of the construct, informed by and framed around existing research into E&M barriers, this research aims to help clarify understanding of the topic for both academics and professionals alike – and contribute to the ongoing conversation around how E&M practices can be improved, hopefully helping to strengthen the credibility of the broader communication field. Looking down from a higher perspective, while there has certainly been a “long and winding road” (Macnamara, 2014, p. 8), it does seem as though progress has recently been made, and more momentum may now be taking hold within both academic and industry circles for advancing communication E&M. As Buhmann et al. (2018) note, “the last decade saw more consensus within the greater E&M community of practice than all the previous three decades beforehand.”

Future research might offer an in-depth study of maturity and how concepts provided here are managed in practice. Longitudinal studies that assess the impact of applying a maturity lens to advancing the communication research, measurement and evaluation practices of diverse organizations is one promising avenue for scholarship. An empirical and comparative look at conditions that support maturity would also strengthen the theoretical foundations of evaluation and help translate insights into practice. Finally, opportunity exists for more conceptual exploration into the foundational elements of E&M, and the preliminary and ongoing research component of the process, which often receives scant attention and discussion, suffers from implicit assumptions regarding its presence, or lacks theoretical connections within professional and academic conceptualizations.

Tracing the progression of events from an early emphasis on best practices, including the Barcelona Principles declarations, to the important but challenged movement toward consensus definitions of key terms and industry standards, to today’s focus and interest in advancing E&M maturity, instead of just “reinventing the wheel” (Macnamara and Likely, 2017), perhaps the field is beginning to roll forward. As many have noted, the future viability and success of the industry is at stake. It is still important, however, to remember that, as experts from other fields have warned, the maturity model is not a silver bullet (Wendler, 2012) – and past initiatives have suffered from unrealistic expectations due to a preoccupation with the idea that a single “mythical magic formula” or silver bullet solution exists (Macnamara, 2014). Only time and careful evaluation and measurement will tell whether the notion of maturity is to become useful to the cause. It is the modest hope of the researchers that the contributions made in this study, if nothing else, will help to further the conversation on maturity, and may serve to promote both its understanding and its usage.

References


Further reading


Corresponding author
Nathan David Gilkerson can be contacted at: nathan.gilkerson@marquette.edu

For instructions on how to order reprints of this article, please visit our website: www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: permissions@emeraldinsight.com
Abstract

Purpose – The purpose of this paper is to analyze what the concept of agility means for communication evaluation and measurement and to challenge assumptions of goal-oriented and organization-centric approaches to evaluation and measurement.

Design/methodology/approach – This paper is a development debate based on a literature review, regarding agility, evaluation theory, communication evaluation approaches and what agility means for communication measurement.

Findings – Agility teaches that what works is more important than what was agreed upon in advance, so it is with more emphasis on needs rather than objectives. Regarding evaluation, the findings show that in today’s communication evaluation theory, evaluation is equated with summative evaluation of smart designed and fixed objectives. In agility, evaluation is always formative, to foster development and improvement within an ongoing activity. Consequently smart objectives are no longer valid as fixed benchmarks and ex ante and ex post evaluations do not exist; instead evaluation is an on-going and forward looking activity during action. Regarding measurement, the basic focus in agility on user needs implies that qualitative methods are more obvious than quantitative. The classic Weberian idea of “Verstehen” is helpful to understand how to focus on needs rather than objectives. This paper finally explores the merits of action research and sense-making methodology as applicable measurements in which “Verstehen” is the basis.

Research limitations/implications – Agility is a very radical concept. The practical and theoretical implications of agile evaluation and measurement mean a total change for practice as well as for communication measurement and evaluation theory building.

Originality/value – The value of this paper is that it is the first to include agility into communication evaluation and measurement and that it, consequently, moves beyond organization-centric concepts of evaluation and measurement by bringing the often overlooked user needs into the game.

Keywords Evaluation, Measurement, Communication strategy, Communication management, Communication goals, Communication planning

Paper type Conceptual paper

Introduction

“New technologies, evolving customer preferences and changing employee expectations are fundamentally challenging established ways of working in more and more sectors. It’s time to move beyond a rigid hierarchy, siloed business units, crippling bureaucracy and an increasingly unwieldy matrix. Agile organizations combine the efficiencies of scale with the speed, flexibility and resilience to compete and win in today’s world,” is the claim of one of the biggest consultancies in the world, McKinsey, on its opening page. Becoming an agile organization allows an organization to better respond and adapt to customer/client/patient/member needs and increase speed of execution (Van Solingen, 2018).

The term agile comes from Greek (the mythological figure Achilles) and the Latin word agilis and referred to the ability of a person to change the body efficiently and rapidly (Muller, 1921). Sheppard and Young (2006) propose to define agility in sports as “a rapid whole-body movement with change of velocity or direction in response to a stimulus.” In today’s world of management agility is rather a mindset or philosophy or even a buzz-word used for describing continuous adaptation and speeding up. The ultimate claim of organizational agility is that by better adapting and rapid (re)action, an agile organization can deliver higher returns faster to shareholders.
It is difficult to find a concise definition of organizational agility. Arteta and Giachetti (2004) indicate that the central aspect of agility refers to an ability of the enterprise to respond to a change. Singh et al. (2013), describe organizational agility as the ability of a firm to sense and respond to the environment by intentionally changing magnitude of variety and/or the rate at which it generates this variety relative to its competitors. Zerfass et al. (2018, p. 7) describe it as: “Agility is the overall capability of an organization to respond to and take advantage of the changes initiated by the drivers in the internal and external environments. It includes the ability to identify relevant changes and to respond proactively, efficiently and effectively, employing the right personnel based on competence, not hierarchical status. Additionally, it includes the ability to implement flexible structures and processes suited to the immediate tasks at hand and to employ the appropriate resources in the shortest possible time.” All in all, definitions of organizational agility describe it as the ability to adapt to circumstances and do it rapidly and smart.

Agility is practiced via all kinds of planning methods, like Scrum which is used most (Van Ruler, 2014, 2015), and the upcoming approach of design thinking (Kolko, 2015) and all kinds of practices like collaboration with stakeholders and flexible teams. Most experts on agility claim that it is a completely different way of planning. Consequently, it will also affect the way of evaluation of communication effects in the organizational context. The research question of this article is therefore:

RQ1. What does agility mean for communication measurement and evaluation?

In this first section of this development debate an overview is given of the principles of agility in organizations and the use of agility in communication management, second a review is given of the literature on evaluation theory and how evaluation is conceptualized in communication management. Third these findings are compared with what is needed for agile communication evaluation; finally methods for agile communication measurement will be discussed; in this section action research and sense-making methodology will be presented as appropriate methodologies for agile communication evaluation.

On the concept and principles of agility in organizations

The first goal of this conceptual article is to discuss the principles of agility in organizations. PWC (2012) and with them many consultancies as well as management scholars claim that businesses as well as other organizations need to re-think the way they operate; they favor the so-called agile organization. The agile organization strives to make change a routine part of organizational life, in order to reduce or eliminate the organizational trauma that paralyzes many businesses attempting to adapt to new markets and environments. Because change is perpetual, PWC claims, the agile enterprise is able to nimbly adjust to and take advantage of emerging opportunities.

In management the need to adjust is far from new. It has – although in different wordings – been discussed ever since the beginning of the twentieth century, found its way in management as contingency theory and culminated in more recent concepts, like the learning organization (Senge, 1990). However, agile stands not only for the ability to adjust but also for being able to do it rapidly (Van Ruler, 2014). This was not the claim of earlier adaptation concepts like contingency and the learning organization.

This combination of adjustability and speed can be traced back to Takeuchi and Nonaka (1986) who described how successful Japanese companies altered their ways of producing by constant multi-level and multi-functional learning and transfer of learning in short cycles. They used the metaphor of scrum, borrowed from rugby sport, to describe the methodology of these companies, of what we now know as the agile organization. Some years later, Nagel (1992) wrote a report of a panel discussion among executives from enterprises in the USA with a rather similar outcome, called “the 21st Century Manufacturing Enterprise Strategy.”
They are the first who used the concept of agility. They focused on the necessity for business to speed up to agility: “Those nations that focus now on speeding the transition to agile manufacturing will become the strongest competitors in the global marketplace” (p. 2).

It is in the software industry that agility has become a well-developed concept, with the Agile Manifesto (see www.agilemanifesto.org) as its bible. The promises of becoming agile are enormous. Agile and scrum guru Sutherland (2015) has called his most recent book “The art of doing twice in half the time.” Waldock (2015, p. 6) claims that “companies reported better team morale, better time efficiency due to the time management team and an improved responsiveness to ‘mission creep’.” Organizations like the Project Management Institute (2012) refer to 30 per cent more profit for agile companies than companies with low agility can get. However, according to Van Solingen (2018, p. 23) agility is not a “silver bullet. It is not the solution to every problem.” Agility works best for organizations in complex environments, he claims. Many authors describe our world as a complex, or even a volatile, uncertain, complex and ambiguous world (Bennett and Lemoine, 2014). That would mean that agility can be profitable for many organizations.

If becoming agile is indeed becoming the new norm for organizations, it will, consequently, also become the new norm for its functions, like communication management.

**Agility in communication management**

Agile communication management (public relations, corporate communication, etc.) is not a widely studied field yet. So far, there is one article on agile public relations planning, published in a special issue on public relations in the digital age (Van Ruler, 2015). In her article, Van Ruler shows that agile communication management urges for another planning model and she introduced scrum as a promising way of working in communication management. Recently the Academic Society for Management and Communication of the University of Leipzig started a research project on corporate communication in agile organizations (Zerfass et al., 2018, see also Dühring and Zerfass, 2018). But that is all, so far.

In essence, agility in organizations and teams means working in short cycles of planning, in two directions: inside-out and outside-in, as a two-way street. In communication management, this demand of a two-way street is an old discussion. Grunig (2009) describes how Cutlip – in the 1950s – showed to his students that what he called public relations[1], should help organizations to interact with the environment, and should, therefore, always be two-way communication. Grunig (1975, 1976) constructed the theoretical foundations for this concept of public relations by introducing his so called multi-systems theory of organizational communication, e.g. by theorizing that organizational communication is part of organizational behavior and, as such, relates to the organizational theory-in-use and the style of decision making. According to Grunig, only a certain style of decision making (interactive) and a certain concept of organization and management (open) delivers the best organizational communication. Later Grunig and his research colleagues altered this basic idea into the excellence theory of public relations in which accommodation to stakeholder demands is the preferred strategy (Grunig, 1992).

In the 1990s, Cancel et al. (1997) introduced contingency theory into public relations. The basic idea of this theory is that there is no best strategy; what to do is contingent upon a variety of internal and external variables. These variables affect whether an organization accommodates a public or advocates solely the perspective of the organization (Reber, 2013, p. 187). Contingency theory brought the stance the organization takes in an organization-public relationship to the fore and solved the problem of normativity in the excellence theory, as Reber explains. In agility, stance can never be seen as a variable because user needs are the only ones to explain the strategy Agility must, consequently, be seen as rather radical in its normativity. This implies that the idea behind agility will also change the meta-theoretical reflections about organizations and communication and hence about evaluation and measurement.
Additionally, many authors in public relations and communication management repeatedly argued that communication management is double sided. On the one hand, it is concerned with initiating communication processes with the aim of conveying the company’s point of view and influencing stakeholders. On the other hand, communication management monitors relevant stakeholders and communication processes within the organization and in the organizational environment and brings the analyses onto the decision making table (for an overview, see Dühring, 2012). Nevertheless, as Heath (2000, p. 2) concludes in the first edition of his seminal SAGE Handbook of Public Relations, an evening at a banquet of communication management professionals shows that they are more interested in the first concern than in the latter. The European Communication Monitor and numerous other, smaller, research projects indeed show that professionals in communication management are often more concerned with influencing stakeholders than with bringing their perspectives in the organizational decision-making process. Finally, this two-way street of communication management has for long been an ideal without much practicality and the field cannot lean on solid agile communication management theory and methodology so far. An empirical research project in Germany shows that many departments have only just begun to reflect on what it means (Dühring and Zerfass, 2018). Agility will, therefore, be quite a change for communication departments.

In this article the focus is on what the concept of agility means for evaluation and measurement in communication management. Key element of the methodology of agile working is the concept of evaluation in the form of iterative testing to be able to adjust to user needs as quick, efficiently and effectively as possible (Jongerius et al., 2013; Schwaber, 2014; Van Solingen, 2018; Sutherland, 2015). This brings us first of all to evaluation theory.

A review of evaluation theory

The second goal of this conceptual article is to review evaluation theory. To evaluate means literally “extracting the value out of something,” e-valu(e)-ate. It is normally defined as “to judge or make a judgment about the value or significance of something,” Macnamara (2018, p. 18). So, it is not only taking value out of something but also interpreting the value, or said differently, giving value to something.

In social sciences a distinction is made between summative and formative evaluation. It was Scriven (1974) who suggested this distinction for the educational process, but he was convinced that it was useful for all disciplines (Scriven, 1991). In his definition, formative evaluation is intended to foster development and improvement within an ongoing activity, e.g. formative evaluation helps to look forward while acting. Summative evaluation, in contrast, is used to assess whether the results of the object being evaluated (program, intervention, person, etc.) meet the stated goals.

Scriven not only suggested to differentiate between summative and formative but also between goal-based and goal-free evaluation (Scriven, 1974, 1991). If we combine these different forms of evaluation, we may conclude that goal-based evaluation is making sure that you are working toward reaching your goals or not (formative) or having achieved your goals (summative). Goal-free evaluation is meant to obtain all kinds of insights for decision making and uncover unanticipated effects (formative) or to review and renew the planning process itself (summative).

McCoy (2006) investigated the scientific discussion on evaluation in her dissertation and states that Scriven preferred to describe evaluation as “the study of the merit, worth, or significance of various entities” (p. 11) because he wanted to move away from the comparison of results to goals. He questioned the efficacy of criteria-based evaluation and relegated goals/objectives to minimal importance. In his view, evaluations exist to make value judgments on whether a program was or is of use to its stakeholders. In other words he calls for more emphasis on needs rather than objectives, as McCoy rightly claims in her dissertation. So, Scriven was in favor of goal-free evaluation. Goal-free evaluation prevents
tunnel vision and is a good means to manage the unexpected, as Weick and Sutcliffe (2007, p. 8). In order to structure the ideas of Scriven on evaluation, we can make a matrix with four different aspects of evaluation (see Figure 1).

Formative evaluation was also a basic concept in Kurt Lewin’s (see also Burnes, 2004) approach to change, although he did not name it formative evaluation. In his frequently cited article “Action Research and Minority Problems” he argues that professionals should continuously reflect on their actions while trying to change something. In his model of professional work he claims that every action must lead to observation of the reactions and to reflection on these reactions, thus benefiting the planning of the following action. Planning, Lewin says, normally starts with a general idea. Fact-finding about the situation is then required to alter the idea into a plan and to start acting. Highly developed manufacturers, he says, then do more fact-finding, first to find out whether the action was successful, but also to learn for next actions and, even, to alter the overall plan if necessary.” Rational social management, therefore, proceeds in a spiral of planning, executing, reconnaissance or fact-finding about the result of the action” (Lewin, 1946, p. 38). Consequently, goals should not be smart but flexible.

This model of professional work of Lewin formed the basis of the plan, do, check, act (PDCA) circle, known as the Deming cycle for continuous improvement and is equivalent to the idea of iterative testing and refinement of actions in agility (Deming, 1982). Agility, therefore, urges for a dominance of formative goal-free evaluation.

A critique of the adoption of evaluation theory in communication management planning

Searching through the literature on communication management planning methods and the role of evaluation in it (e.g. Cornelissen, 2017; Ferguson, 1999; Potter, 1997; Smith, 2013; Szyska and Dürig, 2008; Vos et al., 2004; Weintrob and Pinkleton, 2001), studying the extensive research project on evaluation and measurement of Macnamara (2018), analyzing the German project on Communication Controlling (Pfannenberg and Zerfass, 2010; see also Huhn et al., 2011), the AMEC Barcelona Principles Statement from 2010 and the 2016 AMEC Integrated Evaluation Framework in which the smart objective is still the basis and evaluation is still on the effects (see amec.org), my conclusion is that communication management evaluation is almost solely oriented at evaluating outputs, outtakes and outcomes in relation to earlier set goals. So, in communication management planning methodology, evaluation is mostly equated with what Scriven has called goal-based summative evaluation.

Sometimes, formative evaluation is mentioned in communication management planning methodology, but in a very particular way. Smith (2013) for example differentiates between formative and summative evaluation, in so far that formative evaluation needs to be done in order to develop a plan as such. That is why formative evaluation is phase 1 in his model, sometimes also referred to as the stage of diagnosis. Consequently, formative evaluation is not meant to cope with change during acting but to form the plan itself. In this phase it is...
goal-free, not by choice but because there is no goal set yet. Evaluation comes back as the fourth and final phase in the planning model of Smith, and is used to check whether the objectives are met. This evaluation is to be equated with summative, goal-based evaluation, although only named “evaluation.”

Zerfass introduced all kinds of management tools into the communication management practice, as for example the corporate communications scorecard as a framework for managing and evaluating communication strategies (Zerfass, 2008). He was also one of the steering scholars in the German task force on value creation through communication and the development of the framework of communication controlling (Pfannenberg and Zerfass, 2010; see also Zerfass, 2010). Also this model focuses on goal-based summative evaluation.

In the first edition of the International Encyclopedia of Strategy Communication, Buhmann and Likely (2018, p. 625) explain what evaluation in the field of communication management is meant for. They claim that evaluation serves two purposes: accountability (were objectives met?) and improvement (how were objectives met?). They, too, emphasize formative evaluation, but like Smith, this is only applicable in the stage of situational analysis and strategic planning, not in the stage of implementation (p. 637). After goals are set, they do not talk about summative evaluation, but about evaluation as such. Consequently, also Buhmann and Likely are most of all concerned with goal-based summative evaluation as the only form of evaluation.

Macnamara (2014) came up with another model, the so-called measurement, analysis, insights, evaluation model, in which he explains that evaluation is not orientated toward looking back, but toward identifying value in order to fine-tune the campaign in an early stage. Yet, in his book on evaluation, Macnamara (2018) describes the evaluation phase in this model in a more details but different way, namely, as focused on the effectiveness of the communication program (p. 104). “This forward-looking approach – designed to contribute to future business or organization strategy, as well as inform future communication strategy and effectiveness – helps to bridge the gap between public communication, such as PR and corporate communication, and organizational outcomes,” Macnamara says (p. 105). It seems that Macnamara sees evaluation not as input for next actions within the current campaign but as input for next campaigns.

In his extensive research project on evaluation in public communication, Macnamara (2018) describes formative evaluation as “the use of ex ante research before planning and designing a campaign or project to identify factors such as pre-existing levels of awareness, pre-existing perceptions and attitudes, audience interests and concerns, and communication channel preferences” (p. 19). Then, formative evaluation is the first phase of the planning process and most of all as a measurement at time 0, in order to be able to find change at the end of the campaign or project. It is not meant for learning during the execution phase.

Macnamara refers to process evaluation as the method to monitor progression and to track to identify whether or not milestones are being met. At first sight it equals with what Scriven called formative evaluation and looks at what Lewin meant with reconnaissance to benefiting the planning of the following action and if necessary the overall plan. But for Macnamara (2018, p. 39), evaluation always requires smart formulated objectives, because otherwise the objectives are not measurable. That makes his approach to evaluation ultimately goal-based and oriented at reaching or having reached pre-set goals, the objectives are a given. So, although it is formative, process evaluation is not the same as what Lewin and Scriven were looking for in their description of fact-finding and reconnaissance or formative evaluation in order to alter actions and eventually the overall plan.

There are some exceptions in the field of communication management. For example Watson (1996, 2001) developed a more dynamic planning model, called the continuing model, “to cope with ever changing circumstances” (p. 267). This model offers elements, he claims, that have not been included in more linear step-by-step models. “The elements of the
continuing model are an initial stage of research, the setting of objectives and choice of program effects, followed by the strategy selection and tactical choices. As the program continues, there are multiple levels of formal and informal analysis from which judgments can be made on progress in terms of success or ‘staying alive’. The judgments are fed back to each of the program elements. These iterative loops assist the practitioner in validating the initial research and adding new data, adjusting the objectives and strategy, monitoring the progress to create the desired attitudinal or behavioral effects, and helping with the adjustment or variation of tactics” (p. 267). Watson advocated to replace the standard single post-intervention evaluation in most of the existing planning models (p. 268). Nevertheless, Watson’s continuum model has not become a standard in the communication management branch; most models still focus on goal-free formative evaluation as the initial stage of the planning process and summative goal-based evaluation as the form of evaluation of actions and pre-set objectives.

Christensen et al. (2008) also scrutinized the notion of a communication planning model as a fixed input-throughput-output model that expect members of the organization to operate in the service of the communication plan, while it should be the other way around. This equals the basic focus of agility toward user needs. At the same time, Nothhaft and Wehmeier (2007) focused on the concept of complexity and presented socio-cybernetics as a non-linear meta-theory for communication management. Although the authors do not focus on evaluation, iteration is part of their approach. Earlier, Murphy (1996) and Gilpin and Murphy (2006) argued that most public relations planning is very mechanistic and detailed, and oversimplifies the complexity. Murphy (1996) was focused on chaos theory and proposed to view public relations as a non-linear process. Gilpin and Murphy (2006) are more specific about the concept of evaluation as a learning tool and suggest to use the concept of the enacting organization (from Weick), which assumes the world to be indeterminate, yet actively engaged with it rather than passively reacting to occurrences (p. 376). Enacting organizations are heavily investing in learning by doing, laying an active part in the entire interpretation cycle, from awareness to sense making to learning and back again, as Gilpin and Murphy explain. It helps organizations to construct mechanisms to interpret ambiguous events and to provide meaning and direction for participants. These processes are flexible rather than rigid, situated rather than sweeping, gently indicative rather than strictly prescriptive. Gilpin and Murphy propagate improvised teamwork. Improvisation deals with the unforeseen, but should not be mistaken for unskilled decision-making. “Rather than attempting to identify every potential threat and response, (crisis) managers and researcher are free to actively explore new areas of interrelating. This approach therefore embraces the emergent properties of complex systems, instead of striving to contain, control, and predict them” (p. 389). This equals with what is now called agility. But such a conceptualization of the role of evaluation in planning models is still marginal.

Consequence of the summative goal-based view on evaluation in communication management planning methodology is that accommodating change during implementation is not part of the planning model; change is rather seen as a negative attack on continuity and the ability to reach one’s goals. Nor is any attention for the unexpected, apart from the first phase in which there is no goal set yet. It is true that in the explanation of models, authors sometimes stipulate the iterative character of their models, but that is not enough. It should be part of the model itself.

As explained earlier (Van Ruler, 2015) existing planning models in communication management provide an illusion of stability and control, and can be equated with models from the classical planning school in strategy development theory (Van Ruler, 2018). In all these planning models, evaluation is equated with goal-based summative evaluation, formative evaluation is only used ahead of the start of any implementation, being the first phase in the planning process. Evaluation in agile ways of working is the opposite.
Agile forms of communication evaluation and measurement

The third goal of this conceptual paper is to find out what agility means for communication evaluation and measurement.

Many researchers in public relations do argue that we should alter our perspective: “Most important,” Grunig (2009, p. 4) for example says, “is that we abandon the illusion of control.” Murphy (1996), Gilpin and Murphy (2006), Nothhaft and Wehmeier (2007) said about the same. This is exactly what agility claims. Agility teaches us that what works is more important than what was agreed upon in advance. When using agile methods such as scrum, accommodating change is more important than following a strict plan, Jongerius et al. (2013) claim. If change should be in the lead and acting is dependent on the outcome of the evaluation, there is a problem of manageability of the work. Scrum is the most used planning method in agile working (Van Solingen, 2018). Scrum teams work with the “epic” as the overall goal of the project and a so-called “backlog,” which is a list of requirements and issues to be handled. This list is not a fixed one, as the team can add to it (or discard) during the process. Scrum experts believe that it makes no sense to change the work every hour. That is why they work in so-called sprints of one to four weeks. During a sprint, the work should not be changed. If change is needed, the scrum team should break off their sprint and start all over again in a new sprint. That is how agile teams keep the work manageable.

Jongerius et al. (2013) also state that scrum is not only suitable for software development but also “extremely suitable for improving the entire process of concept, design and development” (p. 18). “Scrum is about being relevant for end-users, about freedom in exchange for commitment, about eliminating waste, about self-propelled teams, about time boxing, and about results” (pp. 19-20) and as such is of interest for any project (see also Schwaber, 2014). The basis of scrum is iterative planning. This phenomenon is well known within crisis communication (see for example Coombs, 2011) and press relations (see, e.g. Watson and Noble, 2005) and seems to work well in these areas, but is not normal practice in regular daily work in communication management, so far.

Regarding the research question of this paper, the most interesting aspect of scrum is the prominence of formative evaluation during implementation. This is partly because scrum is very results-oriented and monitors its results on an iterative basis (Schwaber and Sutherland, 2013) but also because ongoing reflecting and monitoring can easily be included in decision making what to do how in the next sprint, including altering objectives. Scrum starts with so-called “generalized ends” (see for an explanation of this concept, e.g. Watson and Noble, 2005, p. 160), sometimes also called “qualitative objectives” (Van Eck et al., 2013). Specific objectives are being settled during the process, at the start of each iterative time-boxed sprint and its focus depends on the permanent testing of prototypes and the monitoring of change. Smart objectives are no longer in place as fixed benchmarks in the planning process and ex ante and ex post evaluations do not exist anymore. Evaluation is an ongoing, iterative element in the planning process of continuously testing what works and what does not (what meets the needs of the user and not so much the goals). Summative evaluation is then the sum total of all formative evaluations and is, consequently, also goal-free. In the end, evaluation in scrum is the independent variable, while the objectives as well as the actions are the dependent ones. So, basically, evaluation is the essence of agile ways of working, and evaluation is always formative and always goal-free, to foster development and improvement within an ongoing activity, as had been explained by Scriven already, and with more emphasis on needs rather than objectives, as McCoy (2006) promoted (p. 11). So, agile evaluation is the opposite of mainstream communication management evaluation.

Agile forms of measurement

The next question is what kind of measurements would fit agile ways of working in communication evaluation and measurement? “Measurement is the collection and analysis
There are a lot of metrics in agile working that are focused on the efficiency of the team, like velocity and time blocked per work item (Tranter, 2018). In this paper the focus is on measurement of the user needs instead of the agile team itself.

A basic principle in agile working is that a product (idea, intervention or whatever) always needs to be tested for its applicability by the (potential) user. “Whereas quantitative research produces aggregated and averaged data about broad groups and topics, qualitative research seeks to provide deep insights into human thinking, perceptions, attitudes, and interpretations in particular contexts and situations” (Macnamara, 2018, p. 200), and that is exactly what agility is focusing on. Consequently, qualitative observation and interviewing are basic methods for agile measurement.

Since agile planning can be seen as more oriented at needs than at objectives, we also need to focus on the theoretical embedding of the measures. Communication scholar Brenda Dervin prefers, therefore, to focus on methodology instead of measurement per se. “Methodology embraces a broad spectrum of activity that has implications for looking at research comprehensively as communication practice”.

What kind of methodology would then match agile forms of measurement and focus on user needs as agility claims? The methodology of Weber in which “Verstehen” is one of the key concepts, is helpful to understand this basic idea of focusing on needs rather than objectives. Verstehen, often translated as interpretative understanding (Eliaeson, 2002, p. 41) refers to understanding the meaning of action from the actor’s point of view. Verstehen has nothing to do with intuition, it is an act of rational interpretation (p. 42), but inductive instead of deductive and opposite to descriptive understanding following basic psychological laws that allows scholars to predict. It is, therefore, opposite to positivism. Interpretivists are scholars who are interested in the ways communities, cultures, or individuals create meaning from their own actions, rituals, interactions and experiences (Littlejohn and Foss, 2009, p. 557). Agility is focused on creating meaning in order to adapt to change. That is why the theoretical concept of interpretative understanding matches the concept of agility. Agility breaks with the classical functionalistic, positivistic idea of communication management and promotes a more socio-cultural approach in which Verstehen is a basic principle and interpretivism is the meta-theoretical basis (see also Edwards, 2018).

Putnam and Pacanowski (1983) were one of the first who showed how qualitative methods and narratives are rich forms of interpretative inquiry. That is why I want to introduce two methodological approaches that fit into the concept of interpretative understanding, because they are qualitative and narrative and oriented at understanding: action research and sense-making methodology and as such interesting for agile communication evaluation and measurement.

**Action research**

Lewin was the first to coin the term action research. Since then, it is an approach that is developed by researchers in social change who emphasize that research should really matter to the people who will be affected by it. Key in action research is the inclusion of the stakeholders as co-researchers (Greenwood and Levin, 1998, p. 3). “Action research explicitly rejects the separation between thought and action that underlies the pure-applied distinction that has characterized social research for a number of generations,” Greenwood and Levin claim. They have used action research for quite some time in industrial and community development in western countries.

In communication management, this methodology has been used by Pieczka and Wood (2013) in their ICA price-winning case study on action research and community change as a public relations practice. They show a typical feature of action research: its cyclical nature,
represented as a cycle or a spiral with an iteration of four distinct phases: planning, action, observation or reflection, and conceptualization. The inquiry cycles move between reflection and action. In management jargon we could name this an agile way of working.

Lewin was in favor of experimentation in natural settings. Greenwood and Levin (1998) state that action research is now evolved into co-generative learning in ongoing dialogues (p. 18). They promote to see action research most of all as a continuous and participative learning process. That is very much in harmony with agility, although in agility the co-creation is not so much emphasized in its methodology, yet. Every change process has an open starting point and often no absolute final goal, they say. So, preparing pre-set smart objectives makes no sense in this methodology. They claim that not only Lewin gave a foundation to action research, but also Dewey who stated that all knowledge testing and proofs are experimental activities (p. 73) and that scientific knowing is a product of continuous cycles of action and reflection (p. 74). This fits in very well with agility in which actions are seen as cycles of experiments to test whether this is what will fulfill needs. Moreover, Greenwood and Levin (p. 75) state that action research “generally takes on much more complex problems than conventional social sciences do.” This, too, matches with the statement of Van Solingen (2018) that agility is most of all applicable to complex situations. When using agile methods, accommodating change is more important than following a strict plan, as the Agile Manifesto dictates. It is not about being relevant to clients but about being relevant to users of the interventions (products, messages) made, just as action research methodology dictates.

Although action research is explicitly developed for social change, the basic principles of action research could provide a very interesting methodology for agile communication measurement. Key elements of this approach and the philosophy of agility are the same. What is not fully covered yet, is the specific measurement methodology. For this, sense-making methodology can be helpful.

Sense-making methodology
Sense-making methodology, as developed in the 1970s by communication scholar Dervin, could be helpful to understand and give a theoretical and methodological foundation of the principles of agile measurement practice in communication management. Dervin developed this methodology as a critique of the transmission model in communication science (Foreman-Wernet, 2003). “Communication research based on the transmission model tends to focus on the messages send by senders and the resulting effects or impacts of these messages on receivers,” as her student and successor Foreman-Wernet explains (p. 5). Dervin suggested an alternative communication model in which messages are not things but constructions that are tied to the specific times, places, and perspectives of their creators. Such messages are understood to be of value to receivers only to the extent that they can be understood within the context of receivers’ lives. Dervin developed the sense-making methodology to study information needs, seeking and use and is concerned with how people make sense of their own worlds (Dervin, 2003a). “Perhaps most fundamental of sense-making’s meta-theoretical assumptions is the idea of the human, a body-mind-heart-spirit living in a time-space, moving from a past, in a present, to a future, anchored in material conditions,” Dervin teaches (p. 138). This is exactly what agile measurement is looking for: it must help people to formulate their own stories in their own context.

Dervin prefers to see communication as dialogue, which requires open-endedness or reciprocity in an organization’s approach to the receivers of their messages. In order to fully acknowledge reciprocity, organizations need to learn to listen and to address differences and contests in human being’s understandings and experiences, as Foreman-Wernet (2003) shows. This includes, consequently, forms of measurement that will give room to dialogue, e.g. the open interview. Users, as Dervin used to call receivers of messages of organizations,
have to be interpreted in their own context and have to be given room to tell their stories. This matches in a wonderful way with the basic idea of the use of user stories as the one and only basis for agile action (Van Solingen, 2018). In agile software development user stories are normally made up by the product-owner and not by the users themselves. User stories are conversations about functionalities, developed from the user’s perspective (Jongerius et al., 2013, p. 58), but it is the interpretation of the product owner. Dervin would not be satisfied by that, she would claim that an interpretation is not good enough, it should be the story of the user herself.

If, indeed, the stakeholder has to be put first, as many scholars in our field as well as in organization and management studies say, then we have to alter our approach and see our stakeholders no longer on the receiver’s end but as human beings who are actively making sense of their worlds in their current time and space. Dervin’s sense-making measurement approach can help to see that information has meaning only in the context of the human being. The sense-making approach mandates the framing of the research questions such that the respondent is free to name his or her own world. That is why she promotes the use of a dialogic interview setting (Dervin, 2003b, p. 200) as the only suitable measurement. The basic idea of agility that acts have to be tested in short cycles whether they fulfill needs of users instead of organizational objectives, can be made concrete by this sense-making dialogic interview.

Dervin (2003c) is very explicit what her approach has to offer to message design, which she prefers to call responsive design. Audience research, she claims, should be dialogue, and we should therefore not use the terms audience or campaign anymore (p. 222). “The term will be retained in this chapter with the assumption that the ‘audience’ involved is both the public and the institution, both willingly audenced and audiencing, and that the campaign involved is one consented to and dealt with in shared and interactive consent,” Dervin stated already in 1989. “Sense-making is a new way of listening, simultaneously ethnographic because it allows respondents to define and anchor themselves in their own realities, qualitative because it is built on open-ended interviewing and reports findings primarily in qualitative terms, quantitative because procedures for quantitative analysis have been developed, and systematic because a general theory guides the approach to listening, a theory that is applicable in all situations but allows specificity in any situation” (Dervin, 2003c, p. 223, italics in the original).

The fundamental elements of action research and sense-making methodology are harmonizing well with the concept of agility. They add to agility that user needs cannot be interpreted by the organization but need to be discussed with users. What agility brings to these two approaches of measurement is the idea of iteration in short speedy cycles in order to keep on improving and gradually growing to meet users’ needs.

Discussion and conclusion
This conceptual paper in form of a development debate has demonstrated that agility is not discussed in communication evaluation and measurement yet. It is obvious that agile evaluation and measurement are completely different from classical evaluation and measurement in communication management. Zerfass et al. (2018, p. 30) warn that “Becoming more agile has implications for all elements of the management process, including evaluation and performance management. The interconnected, entwined nature of agile projects and processes as well as the required flexibility, responsiveness, and speed of working tend to make traditional approaches to measurement and key performance indicators (KPIs) more difficult.”

As explained before, it also demands a fundamental change of our measurement and evaluation practice and our methodology and meta-theory. Consequently it is not only a change in communication management practice and theory but also a fundamental change in the field of evaluation and management theory. As explained in this paper, agile
measurement is essentially focused on needs rather than on objectives. As action research as well as sensemaking methodology teaches us, there is only one expert in this process and that is the user. His or her sense making of the world is the only benchmark in the research. The researcher is the “not knowing” researcher. That is why Dervin claims that the only measure that counts is the dialogic interview without hypotheses or any pre-set interview guide. Evaluation is, consequently, always formative and goal-free and input for next action or even for change of goals. Agility teaches us that iteration is important, in order to speed up to meet users’ needs.

Agile evaluation certainly implies that traditional planning and goal setting have to be changed (see also Van Ruler, 2015). But what will be the implication of goal-free evaluation? Will the only benchmarks then be user’s or societal needs or are there other challenges? Does it mean that classical measurement methods are becoming obsolete in agile organizations? Can we combine action research and sense making methodology in existing evaluation and measurement methodologies? We have to acknowledge that agility initiates a fundamental change of meta-theory, middle range theories and planning procedures, but future investigations have to be done to answer these questions.

According to the research conducted for this conceptual paper, Germany is the only country in which research has been done on how communication departments look at and feel about agility. “Many of the large mostly multi-national companies in Germany are already engaged in or are currently implementing corporate-wide cultural change programs that puts agility providers such as flat hierarchies, lateral structures, enabling and empowerment of staff to the fore,” Dühring and Zerfass (2018, p. 11) conclude. Most interviewees indeed confirmed insights that workforce agility is a key to success and that change to agility needs trainings and a cultural change program. Yet, they see all kinds of barriers to embrace this way of working completely; many adapt a hybrid approach: they still use vertical ways of organizing, but complement them with agile approaches (p. 17). Those are things against which agility experts give warning: you can never be a little bit agile, it is all or nothing (Schwaber and Sutherland, 2013). So, it makes no sense to include iterative, agile measurement without changing our measurement and evaluation meta-theory as well.

Agility neglects the stance of the organization as well as variables like organizational culture, historical practice, etc. and can, therefore, be seen as a very radical approach. Agile organizations have to rearrange their practices, including their business models and hierarchies. From internet businesses like Amazon we can learn that focusing solely on user needs can cause conflict with for example employer needs or societal needs. It is on us, scholars, to research these aspects of agility as well.

Note
1. What one calls public relations is by others called corporate communication or communication management. All these concepts are meant to describe the steering of the communications between an organization and its stakeholders and/or society at large.

References


Jongerius, P., Offermans, A., Vanhoucke, A., Sanwika, P. and Geel, J. van (2013), Get Agile, BIS Publishers, Amsterdam, NL.


Muller, F. Jzn. (1921), Latijnsch Woordenboek (Latin Dictionary), Wolters, Groningen.


Schwaber, K. (2014), Agile Project Management with Scrum, Microsoft Press, Redmond, WA.


Szyska, P. and Dürig, U.-M. (Eds) (2008), Strategische Kommunikationsplanung (Strategic Communications Planning), UVK Verlagsgesellschaft, Konstanz.


Weintraub, E.A. and Pinkleton, B.E. (2001), Strategic Public Relations Management: Planning and Managing Effective Communication Programs, Erlbaum, Mahwah, NJ.


Further reading


Corresponding author
Betteke Van Ruler can be contacted at: ruler@telfort.nl