Journal of Communication ISSN 0021-9916

ORIGINAL ARTICLE

Message Pretesting Using Assessments of Expected or Perceived Persuasiveness: Evidence About Diagnosticity of Relative Actual Persuasiveness

Daniel J. O'Keefe

Department of Communication Studies, Northwestern University, Evanston, IL 60208, USA

Do formative assessments of the expected or perceived effectiveness of persuasive messages provide a good guide to the relative actual effectiveness of those messages? The correlational evidence usually invoked on this question is less than ideal. The most relevant evidence compares two messages' relative standing on perceived message effectiveness (PME) and actual message effectiveness (AME) as assessed in separate samples. Across 151 such comparisons, the direction of difference in PME matched that of AME in 58% of cases (ns). Diagnostic accuracy did not differ significantly depending on the size or significance of the PME difference, the size of the PME sample, whether PME assessments came from experts or target-audience representatives, the referent of the PME assessment, or whether the PME assessment involved comparing messages.

Keywords: Formative Research, Perceived Effectiveness (PE), Perceived Message Effectiveness (PME), Persuasion, Message Pretesting.

doi:10.1093/joc/jqx009

Formative research—research designed to shape subsequent campaign or intervention messages—is widely recognized as valuable. Such research can identify relevant target audiences and channels, specify the campaign focus, pretest messages or interventions, and so forth (for a general discussion, see Atkin & Freimuth, 2013).

Across a variety of communication contexts and formats, formative research has itself become the focus of recent work aimed at assessing and improving formative practices (e.g., Truong, Hayes, & Abowd, 2006; Willoughby & Furberg, 2015; Yardley, Morrison, Bradbury, & Muller, 2015). For example, research on human–computer interaction (HCI) has taken up such questions as how to effectively provide crowdsourced critique of designs (Easterday, Rees Lewis, & Gerber, 2017), the utility of different think-aloud protocols in usability testing (Alhadreti & Mayhew, 2017),

Corresponding author: Daniel J. O'Keefe; e-mail: d-okeefe@northwestern.edu

and how different forms of iterative feedback can improve web designers' effectiveness (Dow et al., 2010).

A common practice in formative research is that of eliciting participant preferences for alternative messages or formats. For example, in the context of risk communication, pretest participants may be asked for assessments of alternative ways of presenting risk information, with these preferences then shaping subsequent message design (e.g., Morgan, Fischhoff, Bostrom, & Atman, 2002, pp. 111–120).

This article focuses on the realization of this kind of practice in the specific context of persuasive communication design: pretesting messages or message concepts by asking respondents how persuasive or convincing a message will be, or how important various factors are in influencing the behavior of interest. Such data are used to inform message design decisions; designers use the messages thought to be more effective, or create messages focused on the factors described as having the largest influence.

The purpose of this article is to assess the research evidence bearing on this practice. The central question is whether such pretest data are dependably diagnostic of differences in actual message persuasiveness. In what follows, the current focus is initially concretized by considering examples of formative assessments of perceived or expected persuasiveness. The discussion then addresses how to identify the best evidence bearing on the diagnosticity of those assessments. The article then turns to locating and analyzing such evidence.

Predicting relative persuasiveness

Assessments of expected or perceived persuasiveness

Examples of assessments

Formative research often gathers information concerning the likely effects of persuasive messages by asking respondents about expected or perceived persuasiveness. The purpose of these assessments is to provide guidance about future actual message effectiveness (AME), understood in this context as a message's producing the intended persuasive effects on attitudes, intentions, or behaviors.

Some investigators have used a straightforward one- or two-item assessment. For example, in a study of messages to promote stair climbing, Webb and Eves (2007, p. 51) asked respondents "to rate how much each message would encourage them to use the stairs." In a study of antismoking advertisements, Pechmann, Zhao, Goldberg, and Reibling (2003, p. 6) asked seventh- and tenth-graders to assess ads using a single question: "Overall, I think this ad is effective for kids my age" (with a 5-point response scale anchored by "strongly agree" and "strongly disagree"). Byrne, Katz, Mathios, and Niederdeppe (2015) had participants evaluate cigarette package warnings by responding to two items: "The images I just viewed are convincing" and "The images I just viewed would have the intended effect." (Similarly, see Dillard & Ha, 2016; Ganzach, Weber, & Ben-Or, 1997, Study 2; Malo, Gilkey,

Hall, Shah, & Brewer, 2016; Noar, Palmgreen, Zimmerman, Lustria, & Li, 2010; Volk et al., 2015.)

Some researchers have obtained message evaluations using multi-item scales with a variety of items, with some (but not all) items specifically focused on persuasiveness (or convincingness, effectiveness, etc.), reporting that such effect-focused items were sufficiently highly correlated with others to warrant combining the items. For example, Popova, Neilands, and Ling (2014) used five scales with end-anchors of convincing–unconvincing, effective–ineffective, believable–unbelievable, realistic–unrealistic, and memorable–not memorable ($\alpha = .95$). (Similarly, see Bogale, Boer, & Seydel, 2010; Jasek et al., 2015; McLean et al., 2016; Santa & Cochran, 2008; cf. Dillard, 2013.)

In some studies, participants have been asked to rank-order messages in terms of effectiveness (e.g., Mendez et al., 2012) or to indicate the most effective message from a set (e.g., Healey & Hoek, 2016; Hernandez et al., 2014). In related procedures, message design has been guided by pretest respondents' assessments of the importance of various factors in influencing the behavior of interest. For example, Latimer et al. (2012) had participants rate, on a 10-point scale, how much each of nine different factors affected their desire to quit smoking (factors such as short-term health risks, long-term health risks, and financial costs, etc.); this information then shaped the selection of content for subsequent antismoking videos. (Similarly, see Bartlett, Webb, & Hawley, 2017; Glynn et al., 2003; Nolan, Schultz, Cialdini, Goldstein, & Griskevicius, 2008, Study 1.)[1]

Formative-research participants are commonly representatives of the eventual target audience of interest. For example, in developing a condom-use campaign aimed at African American women, Hood, Shook, and Belgrave (2017) undertook formative research with African American women as participants. But in some studies, participants are relevant experts—professionals working in the substantive domain of interest, marketing experts, and the like. For example, Taylor (2015, p. 1169) had 86 infection-control professionals assess six message strategies aimed at encouraging handwashing, asking them to rate whether "the statement will lead to health care workers washing their hands more often."

"Perceived effectiveness"

At least some of the assessments of interest here have previously been studied under the label of "perceived effectiveness" (PE) (e.g., Dillard, Shen, & Vail, 2007) or "perceived message effectiveness" (PME) (Yzer, LoRusso, & Nagler, 2015). As Yzer et al. (2015) have pointed out, there is considerable diversity among the measures that have been collected under these labels. But one element common to many such measures, though not all, has been obtaining people's perceptions of whether the message influenced *them*.

The current project is both broader and narrower than an interest in PE thusly understood, that is, perceived effects on oneself. It is broader, because the interest here is with any sort of assessment that bears on likely (future) persuasiveness, not

just those concerning whether respondents believed that the message was effective in influencing them. For example, a question such as "which of these two messages would be more persuasive?" does not ask respondents whether a given message persuaded them, and so by some definitions might not be taken to be a measure of PE.

The current project is also narrower, because PE-for-oneself represents a phenomenon worthy of study in its own right. For example, researchers might pursue questions concerning the causes and effects of people's believing they were persuaded by a message. Such questions are valuable, but are not the current interest.

The specific focus of the present project is the practice of asking formative-research participants for assessments of expected or perceived persuasiveness, either for the self or for others. For reasons of convenience and familiarity, the assessments of interest are given the acronym PME—with the understanding that this includes measures that under some definitions would not be included under that label.[2]

The commonality of this practice bespeaks a belief that such assessments are useful for message design purposes—and specifically a belief that such assessments are diagnostic of differences in AME. The next section considers what evidence might underwrite such a belief.

Identifying the best evidence

Correlational evidence

Because the question at hand concerns the predictability of AME from PME, PME–AME correlations would seem to be a natural source of evidence. For example, Dillard, Weber, and Vail's (2007) meta-analysis reviewed 40 cases, reporting that the mean correlation between PME and AME was .41. Their conclusion was that "overall, the results empirically demonstrate the value of PE judgments in formative research" (p. 613).

Such correlational findings are commonly invoked as underwriting the formative use of PME data. As Yzer et al. (2015, p. 125) put it, "Clearly, if PE measures can predict the likely effects of a health message with sufficient precision, then PE can at the very least help filter out ineffective messages before allocating resources to message implementation." (For other examples, see Brennan, Durkin, Wakefield, & Kashima, 2014; Choi & Cho, 2016; Davis, Nonnemaker, Duke, & Farrelly, 2013; Davis, Uhrig, Bann, Rupert, & Fraze, 2011; Noar et al., 2016.) Similarly, weak or negative PME–AME correlations have been offered as a reason for thinking that PME data will not be diagnostic of AME (e.g., O'Keefe, 2002, p. 28).

But the utility of PME–AME correlations for assessing the diagnosticity of PME data varies depending on how the PME–AME correlation is computed. PME–AME correlations might be computed either within the data for a given message or across the data for a set of messages. (For an earlier treatment of this distinction, see Dillard & Ha, 2016.)

Within-message PME–AME correlations

When a PME–AME correlation is computed within the data for a single message, that correlation does not provide evidence relevant to the question of the

diagnosticity of PME data for formative decisions. Even if messages' PME ratings are individually (within-message) very strongly positively correlated with their AME ratings, that does not necessarily mean that the relative PME standing of two messages will match the relative AME standing of those messages.

Abstractly put, the reason is that such correlations do not contain information about the means of the variables involved. To see this, imagine a small data set in which PME and AME data (with each variable scored from 0 to 100) are available for two messages, with $n = 5$ for each message. For message A, the participants have the following (PME, AME) pairs of scores: (77, 47), (76, 46), (75, 45), (74, 44), and (73, 43). For message B, the participants have the following (PME, AME) pairs of scores: (52, 62), (51, 61), (50, 60), (49, 59), and (48, 58). For message A, the PME–AME correlation is +1.00; for message B, the PME–AME correlation is also +1.00. Message A has a better mean PME score (75.0) than message B (50.0)—but message B has a better mean AME score (60.0) than message A (45.0).

Even though in this hypothetical data set PME and AME are perfectly positively correlated within each message, the messages' relative standing on PME is the opposite of their relative standing on AME. As this illustrates, positive within-message PME–AME correlations do not, and cannot, show that relative PME standing will match relative AME standing.

For similar reasons, weak within-message PME–AME correlations also cannot possibly provide good evidence. Imagine a second small data set (with PME and AME again scored from 0 to 100). For message C, the five participants have the following (PME, AME) pairs of scores: (82, 66), (81, 68), (80, 70), (79, 72), and (78, 74). For message D, the five participants have the following (PME, AME) pairs of scores: (62, 20), (61, 25), (60, 30), (59, 35), and (58, 40). For message C, the PME–AME correlation is −1.00; for message D, the PME–AME correlation is also −1.00. Even so, the relative standing of the two messages on PME matches their relative standing on AME: message C has a better mean PME score (80.0) than message D (60.0), and message C also has a better mean AME score (70.0) than message D (30.0). As this illustrates, it is possible for within-message PME–AME correlations to be strongly negative and yet for PME data to give the right answer about which of two messages will actually be more effective.

In short, within-message PME–AME correlations are not relevant to the question of whether messages' relative standing on PME is diagnostic of their relative standing on AME.

Across-message PME–AME correlations

When PME and AME data are collected concerning two messages, the PME–AME correlation computed for data combined across messages does provide evidence relevant to the question of the diagnosticity of PME data. Strong positive across-message PME–AME correlations are an indication that messages' relative standing on PME will be diagnostic of their relative standing on AME; weak or negative correlations are a sign of poor diagnosticity. The two hypothetical data sets above

illustrate this: Across the data for messages A and B, the PME–AME correlation is -.96, correctly indicating poor diagnosticity; across the data for messages C and D, the PME–AME correlation is .92, correctly suggesting good diagnosticity.[3]

Correlational evidence reconsidered

Whether PME–AME correlations are relevant to the diagnosticity of PME assessments depends on whether those correlations are within-message or across-message correlations. But the distinction between these two kinds of correlation does not appear to have been sufficiently appreciated. For example, Dillard, Weber, et al.'s (2007) meta-analysis of PME–AME correlations included both irrelevant within-message correlations (e.g., Hullett, 2004) and relevant across-message correlations (e.g., Hullett, 2002).

Some PME–AME correlations that have been offered as relevant to the formative use of PME assessments for message selection have been irrelevant within-message correlations. For example, Davis et al.'s (2011) study of an HIV testing campaign found that perceived ad effectiveness predicted subsequent intentions, and so concluded that their PME measures "may be useful for quantitatively pretesting messages in future campaigns" (p. 58). But in this research, all participants were exposed to the same campaign materials. That is, the PME–AME correlations were within-message correlations—and hence are not relevant to the use of PME data for pretesting alternative messages. But other proffered PME–AME correlations have been the relevant sort, across-message correlations (e.g., Davis et al., 2013; Popova et al., 2014).

However, even though across-message PME–AME correlations are relevant to the assessment of the utility of PME measures in formative research, such correlations are an imperfect source of evidence. To compute across-message PME–AME correlations, the same participants must supply both PME and AME data, as when participants are exposed to a message and then complete both PME measures and AME measures (e.g., Chen, McGlone, & Bell, 2015). This might upwardly bias across-message correlations. If a participant reports that the sunscreen message she saw was very effective and then is asked "do you intend to wear sunscreen in the future?" one might reasonably expect some rough consistency.

Comparison of relative PME and AME standing

Against this backdrop, an alternative source of evidence naturally recommends itself. Because the concrete formative task uses the relative standing of two messages with respect to PME as a guide to the relative standing of those messages with respect to AME, more suitable evidence will consist simply of data that permit one to see whether relative PME standing matches relative AME standing.

Most formative studies that have collected PME data do not provide relevant evidence, because—understandably—those studies do not collect appropriate AME data. A message designer who collects PME data during formative research will choose the message that seems most likely to be effective, meaning that no comparative AME data would be available.

Thus the kind of study that will provide the most suitable data is one in which PME data are collected on at least two messages from one set of participants, and AME data are collected on those same messages but from a different set of participants. Such studies maximize the realism (external validity) of the research design, at least in the sense of paralleling the circumstances faced in formative research.

A simple metric

Given a set of such studies, one can compute what amounts to a batting average: the percentage of cases in which two messages' relative PME standing matches their relative AME standing. If relative PME standing correctly predicts relative AME standing in (say) 90% of cases, then PME data will look to be a very good guide for message selection; on the other hand, if relative PME standing were to match relative AME standing only 50% of the time, then message designers would be just as well served by randomly choosing a message as by collecting PME data.

This metric speaks directly to the needs of formative researchers in ways that across-message correlational data do not. Formative decisions are commonly based on a simple comparison of PME means, with or without a significance test (see, e.g., Maddock et al., 2008; Malo et al., 2016; Mendez et al., 2012; Pechmann et al., 2003; Webb & Eves, 2007). It is entirely reasonable for a message designer to want to have an answer to the question "If I follow this decision procedure, how often will I be choosing the more effective message?" Across-message correlations obscure, rather than clarify, the degree of diagnosticity of PME data. It is not obvious just how good an indicator PME assessments are if the average across-message PME–AME correlation is .63 or .09 or .41—but the diagnosticity is apparent if correct predictions are found to occur in 68% or 85% or 51% of cases.

Methods

Identification of relevant cases

Literature search

Relevant research reports were located by searching Google, PsycINFO, ProQuest Dissertations and Theses Global, and the Web of Science databases for "perceived message effectiveness" (and the latter for "perceived effectiveness"); through personal knowledge of the literature; and by examining review discussions of formative research (e.g., Abraham & Kools, 2012; Atkin & Freimuth, 2013), citations to key articles (Dillard, Weber, et al., 2007; Yzer et al., 2015; Zhao, Strasser, Cappella, Lerman, & Fishbein, 2011), and reference lists in relevant reports.

Inclusion criteria

To be included, a study had to meet three criteria. First, the study had to provide quantitative PME and AME data on each of two (or more) messages, such that it was possible to compare messages' relative rankings on PME and AME. Excluded by this criterion were studies in which PME was assessed through focus groups or

other non-quantitative methods (e.g., Booth-Butterfield et al., 2007); focus groups, although useful in formative research, might be thought to provide insufficiently systematic assessments of PME. Also excluded by this criterion were studies in which the effects of individual messages or message types could not be distinguished (Bigsby, Cappella, & Seitz, 2013) and studies that assessed messages' PME but not AME (e.g., Andsager, Austin, & Pinkleton, 2001).

To be included as a measure of PME, a measure had to provide some manifest assessment of a message's expected or perceived persuasive effectiveness. This criterion thus included indices composed entirely of effect-oriented questions, whether a single item (e.g., Pechmann et al., 2003) or a multi-item index (e.g., Byrne et al., 2015). A multi-item index containing both effect-oriented items and other items (e.g., focused on other attributes such as "logical" or "biased") was included only if there was appropriate evidence of inter-item consistency such as Cronbach's alpha (e.g., Hullett, 2000). This criterion also included assessments in which participants rank-ordered or selected messages or message contents based on expected persuasiveness (e.g., Paul, Redman, & Sanson-Fisher, 1997). This criterion excluded measures that did not directly assess expected or perceived persuasiveness (e.g., measures of message liking, memorability, bias, clarity, and so forth; e.g., Latimer et al., 2012). To determine relative PME standing when more than one measure was available, an effect size (r) was computed for each and then these were averaged (all such averages used the r-to-z-to-r transformation procedure, weighted by n).

To be included as a measure of AME, a measure had to assess one or more of three common persuasion outcomes: attitude, intention, and behavior. To determine relative AME standing when more than one measure was available, an effect size (r) was computed for each and then these were averaged. These three measures were treated as interchangeable because relative persuasiveness appears to be invariant across these three outcomes (O'Keefe, 2013).

Second, the PME data and AME data had to come from different sets of participants, those providing PME data had to be either plausible representatives of a potential target audience or putative experts in the relevant domain, and those providing AME data had to represent a corresponding potential target audience. This criterion maximized the similarity of the included cases to the circumstance of formative research. Excluded by this criterion were studies in which PME and AME data came from the same participants (e.g., Jasek et al., 2015) and studies with PME data from participants who were neither experts nor representatives of a potential target audience (Dillard & Ha, 2016).

Third, the messages being compared had to represent plausible formative message comparisons, that is, ones that might reasonably arise in formative research. This criterion excluded cases in which researchers purposefully set out to create relatively ineffective messages (e.g., Druckman, Peterson, & Slothuus, 2013), such as research on elaboration likelihood model hypotheses about argument quality (e.g., Cacioppo, Petty, & Morris, 1983).

Main analysis

Unit of analysis

The main unit of analysis was the message pair. A study that reported data for only two messages thus provided one case (one message pair). When a design had three or more messages not arising from a factorial design, all possible message pairs were included. For example, Byrne et al. (2015) studied five substantively different designs for antismoking messaging on cigarette packages, which yielded 10 message pairs. When a design had three or more messages because two or more message variables were manipulated in a factorial design, only the contrasts associated with the message factors were included. For example, Piccolino's (1966) study of safety messages had 12 experimental messages generated by a 3 (threat: high, medium, low) × 2 (realism: high, low) × 2 (specificity: high, low) design. Rather than examining all 66 pairs, only the comparisons associated with each message factor were included. This choice was meant to reflect the likely formative interest in learning about the message factors (as opposed to any particular message).

Metric

The metric of interest was whether, for each message pair, the direction of effect on PME (i.e., which message had the higher PME mean) matched the direction of effect on AME (which message had the higher AME mean).[4] If two messages differed on PME but had identical AME means, that case was scored as having the same direction of effect; in such circumstances, the PME data would not have led to choosing a demonstrably inferior message (and so the PME data would not have led to a poor message design decision). If two messages had identical PME means but differed on AME, that case was scored as having different directions of effect for PME and AME; in such a circumstance, the PME data would not have identified the more effective message.

Other properties

For each case, six other properties were also recorded (when relevant information was available, either in the research report or through correspondence with authors), because each represented a potential moderator of PME diagnosticity. Several of these concerned the PME effect size, that is, the difference between the PME mean for one message and the PME mean for the other message; the effect size was computed as r and then converted to d for reporting. (When more than one PME measure was available, an average effect size was computed.)

First, whether the PME effect size was statistically significant: Diagnosticity might be greater where significant differences are found between the two messages' PME values.

Second, the magnitude of the PME effect size: Larger PME differences between the two messages, whether statistically significant or not, might be expected to be more diagnostic.

Third, the PME sample size, that is, the number of participants contributing to the PME effect size: Independent of whether a statistically significant PME effect is

observed, diagnosticity might be expected to be greater as the size of the pretest sample increases.

Fourth, whether the participants providing PME data were representatives of the target audience or were experts: Data from one kind of participant might be more diagnostic.

Fifth, the referent for the PME assessment: As Yzer et al. (2015) noted, PME assessments vary in the specification of the referent. Some assessments ask about the respondent (e.g., "would this motivate you to do X?"). Some assessments ask about convincing other people (sometimes specifying a particular sort of other: "kids your age," "green consumers," et al.). Some leave the referent unspecified, as when respondents are asked to rate how "persuasive" or "convincing" a message is, without specifying for whom. And some combine two or more referents, as when respondents are asked both how convincing a message is and whether it persuaded them.

Sixth, whether the PME assessment was comparative or non-comparative: It might be that PME assessments will be more diagnostic when respondents make what amount to comparative judgments about relative persuasiveness, as opposed to rating a single message. A PME assessment was classified as comparative if a PME respondent assessed both messages (or message kinds) that were compared on AME. So, for example, cases in which the PME assessment involved rank-ordering a set of messages or rating a number of different kinds of message were classified as comparative. Cases in which a PME respondent assessed only one message or message type were classified as non-comparative.

Additional analysis

The collected cases were also analyzed in a second way to address potential issues of statistical independence. Such issues can arise when a given study examines more than two messages. For example, a study with three messages (A, B, and C) yields three message pairs (A vs. B, A vs. C, and B vs. C), but the same participants would contribute to more than one comparison. As long as message pair is the unit of analysis, this problem can be avoided only by discarding cases, and there is no obviously defensible way of choosing cases to be put aside.

Hence the data were also analyzed using study as the unit of analysis, with the rank-order correlation between the PME standings and the AME standings computed for each study. A meta-analytic mean correlation was then computed across studies. This study-based metric is not quite as transparent as that based on message pairs, because the mean rank-order correlation does not convey the diagnosticity of relative PME standing in a straightforward manner. But this additional way of analyzing the data is useful for addressing issues of statistical independence.

Results

Overall effects

A total of 151 relevant comparisons (message pairs) were located. The list of included cases, with codings and reference citations, is archived at the Open Science

Table 1 Results: Message-Pair Analysis

	k	proportion of correct predictions	95% CI
All cases	151	.576	.496, .652
Moderator variables			
PME statistical significance			
d significant ($p < .05$)	72	.667	.551, .765
d non-significant	53	.528	.397, .656
PME difference (effect size)			
$d \geq .326$	63	.635	.511, .743
$d < .326$	62	.581	.457, .695
PME sample size			
$N \geq 87$	74	.608	.494, .712
$N < 87$	77	.545	.435, .652
PME sample composition			
target audience	133	.571	.486, .652
experts	18	.611	.385, .798
PME referent			
self	18	.556	.337, .755
other	59	.525	.400, .647
multiple	5	.600	.229, .884
unspecified	51	.627	.490, .747
PME comparativeness			
comparative	92	.554	.453, .652
non-comparative	59	.610	.482, .724

Note: The confidence interval (CI) is the 95% adjusted Wald confidence interval.

Framework: osf.io/rh2bf. The included studies were: Barnett, Klassen, McMinimy, & Schwarz, 1987; Bhattacharjee, Berger, & Menon, 2014; Byrne et al., 2015; Falk, Berkman, & Lieberman, 2012; Falk et al., 2016; Faulkner & Kennedy, 2008; Ganzach et al., 1997; Goldsmith & Dhar, 2013; Gollust, Niederdeppe, & Barry, 2013; Hansmann, Loukopoulos, & Scholz, 2009; Hornikx, 2008; Hullett, 2000; Leoniak & Maj , 2016; Lockwood, Marshall, & Sadler, 2005; McIntyre et al., 1987; Nolan, Kenefick, & Schultz, 2011; Nolan, Schultz, Cialdini, Goldstein, & Griskevicius, 2008; Paul, Redman, & Sanson-Fisher, 1997, 2003; Pechmann et al., 2003; Pettigrew et al., 2014, 2016; Piccolino, 1966; Tal-Or, Shilo, & Meister, 2009; Thornton, Kirchner, & Jacobs,1991; Trawalter, Driskell, & Davidson, 2016; Turner et al., 2010; Turner, Banas, Rains, Moore, & Jang, 2005; Weaver, Hock, & Garcia, 2016; Wogalter, Begley, Scancorelli, & Brelsford, 1997. The median PME effect size was $d = .326$; the median PME sample size was 87.

Across all 151 message pairs, the direction of difference in PME matched the direction of difference in AME in 58% of the cases (87/151, .576); see Table 1. This proportion is not significantly different from .50 ($z = 1.87$, $p = .061$). With 151

cases, power for a population effect of .75 exceeded .99, and for a population effect of .65 was .96.[5]

Moderating variables

Significance of PME effects

In the 72 cases in which the observed PME effect was statistically significantly different from zero, the direction of difference in PME matched that of AME in 67% of the cases (48/72, .667). This proportion is significantly different from .50 ($z = 2.83$, $p = .005$).

In the 53 cases in which the observed PME effect was not statistically significantly different from zero, the direction of difference in PME matched that of AME in 53% of the cases (28/53, .528). This proportion is not significantly different from .50 ($z = .41$, $p = .680$). With 53 cases, power for a population effect of .75 was .97, and for a population effect of .65 was .59.

These two proportions (.667 and .528) are not significantly different: $\chi^2(1) = 2.43$, $p = .119$. For a medium-sized difference between proportions (per Cohen, 1988), power was .78.

Size of PME effects

A median split distinguished cases with a relatively large ($d \geq .326$, $k = 63$) or small ($d < .326$, $k = 62$) PME effect size. In the 63 cases with a relatively large effect size, the direction of difference in PME matched that of AME in 63% of the cases (40/63, .635). This proportion is significantly different from .50 ($z = 2.14$, $p = .032$).

In the 62 cases with a relatively small effect size, the direction of difference in PME matched that of AME in 58% of the cases (36/62, .581). This proportion is not significantly different from .50 ($z = 1.27$, $p = .204$). With 62 cases, power for a population effect of .75 was .99, and for a population effect of .65 was .67.

These two proportions (.635 and .581) are not significantly different: $\chi^2(1) = .38$, $p = .536$. For a medium-sized difference between proportions (per Cohen, 1988), power was .78.

Size of PME sample

A median split distinguished cases with a relatively large ($N \geq 87$, $k = 74$) or small ($N < 87$, $k = 77$) PME sample size. In the 74 cases with a relatively large sample, the direction of difference in PME matched that of AME in 61% of the cases (45/74, .608). This proportion is not significantly different from .50 ($z = 1.86$, $p = .063$). With 74 cases, power for a population effect of .75 exceeded .99, and for a population effect of .65 was .74.

In the 77 cases with a relatively small sample, the direction of difference in PME matched that of AME in 55% of the cases (42/77, .545). This proportion is not significantly different from .50 ($z = .80$, $p = .425$). With 77 cases, power for a population effect of .75 exceeded .99, and for a population effect of .65 was .76.

These two proportions (.608 and .545) are not significantly different: $\chi^2(1) = .603$, $p = .438$. For a medium-sized difference between proportions (per Cohen, 1988), power was .85.

Composition of PME sample

In the 133 cases in which PME data were obtained from representatives of the target audience, the direction of difference in PME matched that of AME in 57% of the cases (76/133, .571). This proportion is not significantly different from .50 ($z = 1.65$, $p = .100$). With 133 cases, power for a population effect of .75 exceeded .99, and for a population effect of .65 was .94.

In the 18 cases in which PME data were obtained from experts, the direction of difference in PME matched that of AME in 61% of the cases (11/18, .611). This proportion is not significantly different from .50 ($z = .94$, $p = .346$). With 18 cases, power for a population effect of .75 was .57, and for a population effect of .65 was .24.

These two proportions (.571 and .611) are not significantly different: $\chi^2(1) = .10$, $p = .750$. For a medium-sized difference between proportions (per Cohen, 1988), power was .57.

Referent of PME assessment

When the formative-research participants were experts, the PME referent was, understandably, never the self; experts were not asked how they themselves would react. To remove this confound, the analysis of the PME referent moderator examined only cases in which the participants were representatives of the target audience.

In the 18 cases in which the referent of the PME assessment was the respondent, the direction of difference in PME matched that of AME in 56% of the cases (10/18, .556). This proportion is not significantly different from .50 ($z = .47$, $p = .637$). With 18 cases, power for a population effect of .75 was .57, and for a population effect of .65 was .24.

In the 59 cases in which the referent of the PME assessment was some other, the direction of difference in PME matched that of AME in 53% of the cases (31/59, .525). This proportion is not significantly different from .50 ($z = .391$, $p = .696$). With 59 cases, power for a population effect of .75 was .99, and for a population effect of .65 was .64.

In the five cases in which the PME assessment had multiple referents, the direction of difference in PME matched that of AME in 60% of the cases (3/5, .600). This proportion is not significantly different from .50 ($z = .447$, $p = .655$). With five cases, power for a population effect of .75 was .17, and for a population effect of .65 was .09.

In the 51 cases in which the referent of the PME assessment was unspecified, the direction of difference in PME matched that of AME in 63% of the cases (32/51, .627). This proportion is not significantly different from .50 ($z = 1.82$, $p = .069$).

With 51 cases, power for a population effect of .75 was .97, and for a population effect of .65 was .58.

No two of these proportions were significantly different from each other. The proportion for the respondent as referent (.556) was not significantly different from that for some other (.525; $\chi^2(1) = .05$, $p = .824$), for multiple referents (.600; $\chi^2(1) = .03$, $p = .862$), or for unspecified referents (.627; $\chi^2(1) = .71$, $p = .399$). The proportion for others as referent (.525) was not significantly different from that for multiple referents (.600; $\chi^2(1) = .10$, $p = .750$) or for unspecified referents (.627; $\chi^2(1) = 1.15$, $p = .283$). The proportion for multiple referents (.600) was not significantly different from that for unspecified referents (.627; $\chi^2(1) = .01$, $p = .905$). For a medium-sized difference between proportions (per Cohen, 1988), power for these comparisons ranged from .11 to .73.

Comparative vs. non-comparative PME assessments

In the 92 cases in which PME assessments were comparative, the direction of difference in PME matched that of AME in 55% of the cases (51/92, .554). This proportion is not significantly different from .50 ($z = 1.04$, $p = .297$). With 92 cases, power for a population effect of .75 exceeded .99, and for a population effect of .65 was .83.

In the 59 cases in which PME assessments were non-comparative, the direction of difference in PME matched that of AME in 61% of the cases (36/59, .610). This proportion is not significantly different from .50 ($z = 1.69$, $p = .091$). With 59 cases, power for a population effect of .75 was .99, and for a population effect of .65 was .64.

These two proportions (.554 and .610) are not significantly different: $\chi^2(1) = .46$, $p = .500$. For a medium-sized difference between proportions (Cohen, 1988), power was .84.

Additional analysis

A random-effects meta-analysis was undertaken examining the rank-order correlation of messages on PME and AME, using study as the unit of analysis and weighting cases by the number of participants providing PME data (Borenstein & Rothstein, 2005); the list of cases, with codings and reference citations, is archived at the Open Science Framework: osf.io/rh2bf. Across 35 cases, the mean rank-order correlation was -.053, not significantly different from zero ($p = .916$); the 95% confidence interval is [−.776, .730].

Discussion

The present results

These results offer at best a mixed picture concerning the formative use of assessments of perceived or expected persuasiveness. Across all cases, pretesting messages by asking respondents about perceived or expected persuasiveness was no more informative about relative actual persuasiveness than flipping a coin: Such measures matched the observed direction of actual difference only 58% of the time, a value

statistically indistinguishable from 50% (despite excellent statistical power for detecting a population diagnosticity of 65%). And the mean correlation between PME rank and AME rank was almost zero (−.05); how messages ranked on PME was unrelated to how they ranked on AME.

There is a hint in these results that message designers might have more confidence in relying on PME assessments when the PME effect size is statistically significant or relatively large; under those conditions, the diagnosticity of PME measures was statistically significantly different from 50%. But even this conclusion must be tempered, because the relevant moderator tests did not yield significant effects: Studies with statistically significant PME effects were no more diagnostic than those with nonsignificant effects, and studies with studies with larger PME differences were no more diagnostic than those with smaller effects. And, similarly, studies with relatively large PME differences were not more (or less) diagnostic than those with relatively small effects, studies with participants representing the target audience were no more (or less) diagnostic than those with experts, studies with comparative PME assessments were no more (or less) diagnostic than those with noncomparative assessments, and diagnosticity did not vary as a function of the referent of the PME assessment.

Moving forward

Read optimistically, these results might at least suggest some potential usefulness of PME assessments. All of the observed mean diagnosticity values exceeded 50%, even if not always statistically significantly different from 50%. And the absence of dependable moderator effects might in some cases be ascribed to weak statistical power. Even so, the present results suggest that there is considerable room for improvement in the use of PME assessments for diagnosing relative message persuasiveness.

As a starting point for improving PME diagnosticity, consider that when pretest respondents assess how "persuasive" a message will be, they are presumably relying on their naive (perhaps nonconscious) conceptions of what makes messages persuasive. When a given message has the properties they associate with persuasiveness, respondents judge it as persuasive.

Booth-Butterfield et al.'s (2007) pretesting of messages about risks to firefighters may provide a useful illustration. "Our message pretesting with focus groups (…) conclusively demonstrated that virtually every participant strongly preferred executions that featured more color, graphics, and design qualities. The typical government documents were almost uniformly declared to be less useful, attention getting, and memorable" (p. 87). But two subsequent randomized field experiments "found better reception and processing results with the standard format than with the high-design format" (p. 87). Plainly, pretest respondents had erroneous conceptions of what would make these messages effective.

Indeed, whenever PME pretesting is not diagnostic of actual differences in effectiveness, it might be that the pretest respondents had inaccurate lay theories of

persuasiveness—theories that misled them. PME pretests can be expected to be diagnostic only when the relevant lay beliefs (on which the PME judgments are based) are accurate.

This reasoning suggests that the nature of the messages being pretested is a crucial influence on PME diagnosticity. The very same PME measure (e.g., "will this message be persuasive?") used in two different pretesting circumstances might vary dramatically in diagnosticity—not because of some shortcoming of the measure itself, but because of variations in respondents' accuracy for judging different kinds of messages.

Hence one general approach to improving the diagnosticity of PME measures may be to give more attention both to the lay beliefs that underlie PME judgments and to the messages being pretested. This will eventually require not only an articulated account of lay conceptions, but also an appropriate abstract framework for describing message variations—a framework that distinguishes different message variations on the basis of their susceptibility to accurate lay assessment. Improving the diagnosticity of PME assessments cannot simply be a matter of making adjustments to PME measurement procedures; researchers will want also to consider whether the variations in the messages being pretested are variations about which respondents are likely to have sound lay theories.

However, there may be limits to the improvement of PME diagnosticity. If message designers are sufficiently good at devising initial candidate messages so that there are only small differences between them in effectiveness, then a pretesting procedure would need to be especially sensitive to detect such differences. It may not be possible for any pretesting procedure to be highly diagnostic under such circumstances.[6]

Perhaps the best way to pretest message effectiveness is to do just that: pretest message effectiveness. Over 25 years ago, the Advertising Research Foundation undertook to examine the predictive validity of a number of different advertising pretesting ("copy-testing") procedures such as ad likeability and recall (Haley & Baldinger, 1991). Among all these, the best single general-purpose copy-testing measure appeared to be assessments of persuasion—brand preference, purchase intention, and the like (Rossiter & Eagleson, 1994).

Thus, in formative research, message designers might dispense with questions about expected or perceived persuasiveness, and instead pretest messages for actual effectiveness (e.g., Whittingham, Ruiter, Zimbile, & Kok, 2008). That solution will not always be feasible; for instance, sometimes it will be difficult to recruit many participants from the target audience (for an example, see Siegel, Lienemann, & Rosenberg, 2017), and sometimes the number of candidate messages may be so large as to prevent efficient AME assessment (e.g., Bigsby et al., 2013). But where pretesting using AME assessments is possible, it surely should be considered.

The larger picture

The present discussion of pretesting persuasive messages can be seen as part of an emerging broad discussion of best practices for formative research. Indeed, the

kinds of concerns raised here have also arisen in other formative contexts. For example, Barnes, Hanoch, Miron-Shatz, and Ozanne (2016) found that women preferring graphical risk formats had lower risk comprehension when information was presented in that format than when it was presented in a numeric format. This was particularly striking, given that less numerate women were more likely to prefer graphical rather than numeric risk formats. As Barnes et al. concluded, such findings "point to the potential perils of tailoring risk communication formats to patient preferences" (p. 1007).

Results such as these underscore the need for continuing attention to the practices of formative research. Designing effective messages is too important to be left to chance.

Acknowledgment

Thanks to Mark Barnett, Sahara Byrne, Jos Hornikx, Sherri Jean Katz, Christine Paul, and Connie Pechmann for additional information, and to Nancy Grant Harrington, Andy King, Associate Editor Robin Nabi, and three anonymous reviewers for manuscript comments.

Notes

1 All of the preceding examples provide some sort of quantitative basis for assessing expected or perceived persuasive effectiveness, but formative research can instead, or also, use qualitative assessments such as derived from focus-group discussions (e.g., Booth-Butterfield, Welbourne, Williams, & Lewis, 2007; Maddock, Silbanuz, & Reger-Nash, 2008; Mowbray, Marcu, Godinho, Michie, & Yardley, 2016; Pollard et al., 2016; Riker et al., 2015).
2 Thus in two ways, the label "perceived message effectiveness" is perhaps a bit misleading for the present purposes. First, any sort of assessment of perceived likely future persuasiveness is relevant to the present undertaking (not just respondents' reports about whether a message was persuasive to them). Second, the interest here is specifically with *persuasive* effectiveness (as opposed to, say, effectiveness in informing). But "perceived message effectiveness" (PME) is a familiar term and hence is used here.
3 Some will recall Simpson's paradox: A relationship between two variables that obtains in every subset of a data set can be reversed when data are aggregated.
4 Comparing AME means gives the best estimate of AME differences, but, like all estimates, contains error.
5 All reported power analyses are based on two-tailed tests with .05 alpha.
6 As a reader noted, PME measures might be more diagnostic at early stages of formative research, when the candidate messages may vary more in effectiveness, than at later stages when the field has been narrowed to a smaller set of contenders.

References

*References marked with an asterisk indicate studies included in the analyses.
Abraham, C., & Kools, M. (Eds.). (2012). *Writing health communication: An evidence-based guide*. Los Angeles, CA: Sage.

Alhadreti, O., & Mayhew, P. (2017). To intervene or not to intervene: An investigation of three think-aloud protocols in usability testing. *Journal of Usability Studies*, **12**, 111–132.

Andsager, J. L., Austin, E. W., & Pinkleton, B. E. (2001). Questioning the value of realism: Young adults' processing of messages in alcohol-related public service announcements and advertising. *Journal of Communication*, **51**, 121–142. doi:10.1111/j.1460-2466.2001.tb02875.x

Atkin, C. K., & Freimuth, V. (2013). Guidelines for formative evaluation research in campaign design. In R. E. Rice & C. K. Atkin (Eds.), *Public communication campaigns* (4th ed., pp. 53–68). Los Angeles, CA: Sage.

Barnes, A. J., Hanoch, Y., Miron-Shatz, T., & Ozanne, E. M. (2016). Tailoring risk communication to improve comprehension: Do patient preferences help or hurt? *Health Psychology*, **35**, 1007–1016. doi:10.1037/hea0000367

*Barnett, M. A., Klassen, M., McMinimy, V., & Schwarz, L. (1987). The role of self- and other-oriented motivation in the organ donation decision. *Advances in Consumer Research*, **14**, 335–337.

Bartlett, Y. K., Webb, T. L., & Hawley, M. S. (2017). Using persuasive technology to increase physical activity in people with chronic obstructive pulmonary disease by encouraging regular walking: A mixed-methods study exploring opinions and preferences. *Journal of Medical Internet Research*, **19**, e124. doi:10.2196/jmir.6616

*Bhattacharjee, A., Berger, J., & Menon, G. (2014). When identity marketing backfires: Consumer agency in identity expression. *Journal of Consumer Research*, **41**, 294–309. doi:10.1086/676125

Bigsby, E., Cappella, J. N., & Seitz, H. H. (2013). Efficiently and effectively evaluating public service announcements: Additional evidence for the utility of perceived effectiveness. *Communication Monographs*, **80**, 1–23. doi:10.1080/03637751.2012.739706

Bogale, G. W., Boer, H., & Seydel, E. R. (2010). Reaching the hearts and minds of illiterate women in the Amhara highland of Ethiopia: Development and pre-testing of oral HIV/AIDS prevention messages. *SAHARA-J: Journal of Social Aspects of HIV/AIDS*, **7**(1), 2–9. doi:10.1080/17290376.2010.9724949

Booth-Butterfield, S., Welbourne, J., Williams, C., & Lewis, V. (2007). Formative field experiments of a NIOSH alert to reduce the risks to firefighters from structural collapse: Applying the cascade framework. *Health Communication*, **22**, 79–88. doi:10.1080/10410230701310331

Borenstein, M., & Rothstein, H. (2005). Comprehensive meta-analysis (Version 2.2.023) [Computer software]. Englewood, NJ: Biostat.

Brennan, E., Durkin, S. J., Wakefield, M. A., & Kashima, Y. (2014). Assessing the effectiveness of antismoking television advertisements: Do audience ratings of perceived effectiveness predict changes in quitting intentions and smoking behaviours? *Tobacco Control*, **23**, 412–418. doi:10.1136/tobaccocontrol-2012-050949

*Byrne, S., Katz, S. J., Mathios, A., & Niederdeppe, J. (2015). Do the ends justify the means? A test of alternatives to the FDA proposed cigarette warning labels. *Health Communication*, **30**, 680–693. doi:10.1080/10410236.2014.895282

Cacioppo, J. T., Petty, R. E., & Morris, K. J. (1983). Effects of need for cognition on message evaluation, recall, and persuasion. *Journal of Personality and Social Psychology*, **45**, 805–818. doi:10.1037/0022-3514.45.4.805

Chen, M., McGlone, M. S., & Bell, R. A. (2015). Persuasive effects of linguistic agency assignments and point of view in narrative health messages about colon cancer. *Journal of Health Communication*, **20**, 977–988. doi:10.1080/10810730.2015.1018625

Choi, J., & Cho, H. (2016). Perceived message effectiveness, attitude toward messages, and perceived realism. In D. K. Kim & J. W. Dearing (Eds.), *Health communication research measures* (pp. 139–152). New York: Peter Lang.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Davis, K. C., Nonnemaker, J., Duke, J., & Farrelly, M. C. (2013). Perceived effectiveness of cessation advertisements: The importance of audience reactions and practical implications for media campaign planning. *Health Communication*, **28**, 461–472. doi:10.1080/10410236.2012.696535

Davis, K. C., Uhrig, J., Bann, C., Rupert, D., & Fraze, J. (2011). Exploring African American women's perceptions of a social marketing campaign to promote HIV testing. *Social Marketing Quarterly*, **17**(3), 39–60. doi:10.1080/15245004.2011.595536

Dillard, J. P. (2013). The effects of prior behavior on judgments of perceived message effectiveness: Evaluating HPV vaccine messages. *Journal of Vaccines & Vaccination*, **4**, 193. doi:10.4172/2157-7560.1000193

Dillard, J. P., & Ha, Y. (2016). Interpreting perceived effectiveness: Understanding and addressing the problem of mean validity. *Journal of Health Communication*, **21**, 1016–1022. doi:10.1080/10810730.2016.1204379

Dillard, J. P., Shen, L., & Vail, R. G. (2007). Does perceived message effectiveness cause persuasion or vice versa? 17 consistent answers. *Human Communication Research*, **33**, 467–488. doi:10.1111/j.1468-2958.2007.00308.x

Dillard, J. P., Weber, K. M., & Vail, R. G. (2007). The relationship between the perceived and actual effectiveness of persuasive messages: A meta-analysis with implications for formative campaign research. *Journal of Communication*, **57**, 613–631. doi:10.1111/j.1460-2466.2007.00360.x

Dow, S. P., Glassco, A., Kass, J., Schwarz, M., Schwartz, D. L., & Klemmer, S. R. (2010). Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Transactions on Computer-Human Interaction*, **17**, article 18. doi:10.1145/1879831.1879836

Druckman, J. N., Peterson, E., & Slothuus, R. (2013). How elite partisan polarization affects public opinion formation. *American Political Science Review*, **107**, 57–79. doi:10.1017/S0003055412000500

Easterday, M. W., Rees Lewis, D., & Gerber, E. M. (2017). Designing crowdcritique systems for formative feedback. *International Journal of Artificial Intelligence in Education*, **27**, 623–663. doi:10.1007/s40593-016-0125-9

*Falk, E. B., Berkman, E. T., & Lieberman, M. D. (2012). From neural responses to population behavior: Neural focus group predicts population-level media effects. *Psychological Science*, **23**, 439–445. doi:10.1177/0956797611434964

*Falk, E. B., O'Donnell, M. B., Tompson, S., Gonzalez, R., Dal Cin, S., Strecher, V., … An, L. (2016). Functional brain imaging predicts public health campaign success. *Social Cognitive and Affective Neuroscience*, **11**, 204–214. doi:10.1093/scan/nsv108

*Faulkner, M., & Kennedy, R. (2008). A new tool for pre-testing direct mail. *International Journal of Market Research*, **50**, 469–490.

*Ganzach, Y., Weber, Y., & Ben-Or, P. (1997). Message framing and buying behavior: On the difference between artificial and natural environment. *Journal of Business Research*, **40**, 91–95. doi:10.1016/S0148-2963(96)00208-1

Glynn, S. A., Williams, A. E., Nass, C. C., Bethel, J., Kessler, D., Scott, E. P., … Schreiber, G. B. (2003). Attitudes toward blood donation incentives in the United States: Implications for donor recruitment. *Transfusion*, **43**, 7–16. doi:10.1046/j.1537-2995.2003.00252.x

*Goldsmith, K., & Dhar, R. (2013). Negativity bias and task motivation: Testing the effectiveness of positively versus negatively framed incentives. *Journal of Experimental Psychology: Applied*, **19**, 358–366. doi:10.1037/a0034415

*Gollust, S. E., Niederdeppe, J., & Barry, C. L. (2013). Framing the consequences of childhood obesity to increase public support for obesity prevention policy. *American Journal of Public Health*, **103**, e96–e102. doi:10.2105/AJPH.2013.301271

Haley, R. I., & Baldinger, A. L. (1991). The ARF Copy Research Validity Project. *Journal of Advertising Research*, **31**(2), 11–32.

*Hansmann, R., Loukopoulos, P., & Scholz, R. W. (2009). Characteristics of effective battery recycling slogans: A Swiss field study. *Resources, Conservation and Recycling*, **53**, 218–230. doi:10.1016/j.resconrec.2008.12.003

Healey, B., & Hoek, J. (2016). Young adult smokers' and prior-smokers' evaluations of novel tobacco warning images. *Nicotine & Tobacco Research*, **18**, 93–97. doi:10.1093/ntr/ntv041

Hernandez, C., Wang, S., Abraham, I., Angulo, M. I., Kim, H., Meza, J. R., … Uddin, S. (2014). Evaluation of educational videos to increase skin cancer risk awareness and sun-safe behaviors among adult Hispanics. *Journal of Cancer Education*, **29**, 563–569. doi:10.1007/s13187-014-0624-z

Hood, K. B., Shook, N. J., & Belgrave, F. Z. (2017). "Jimmy cap before you tap": Developing condom use messages for African American women. *The Journal of Sex Research*, **54**, 651–664. doi:10.1080/00224499.2016.1168351

*Hornikx, J. (2008). Comparing the actual and expected persuasiveness of evidence types: How good are lay people at selecting persuasive evidence? *Argumentation*, **22**, 555–569. doi:10.1007/s10503-007-9067-6

*Hullett, C. R. (2000). *The mediating role of values in value-expressive attitudes*. Doctoral dissertation, ProQuest 9985400, Michigan State University, East Lansing, MI.

Hullett, C. R. (2002). Charting the process underlying the change of value-expressive attitudes: The importance of value-relevance in predicting the matching effect. *Communication Monographs*, **69**, 158–178. doi:10.1080/714041711

Hullett, C. R. (2004). Using functional theory to promote sexually transmitted disease (STD) testing: The impact of value-expressive messages and guilt. *Communication Research*, **31**, 363–396. doi:10.1177/0093650204266103

Jasek, J. P., Johns, M., Mbamalu, I., Auer, K., Kilgore, E. A., & Kansagra, S. M. (2015). One cigarette is one too many: Evaluating a light smoker-targeted media campaign. *Tobacco Control*, **24**, 362–368. doi:10.1136/tobaccocontrol-2013-051348

Latimer, A. E., Krishnan-Sarin, S., Cavallo, D. A., Duhig, A., Salovey, P., & O'Malley, S. A. (2012). Targeted smoking cessation messages for adolescents. *Journal of Adolescent Health*, **50**, 47–53. doi:10.1016/j.jadohealth.2011.04.013

*Leoniak, K. J., & Maj, K. (2016). A slice of hygiene: Justification and consequence in the persuasiveness of prescriptive and proscriptive signs. *Social Influence*, **11**, 271–283. doi:10.1080/15534510.2016.1267663

*Lockwood, P., Marshall, T. C., & Sadler, P. (2005). Promoting success or preventing failure: Cultural differences in motivation by positive and negative role models. *Personality and Social Psychology Bulletin*, **31**, 379–392. doi:10.1177/0146167204271598

Maddock, J. E., Silbanuz, A., & Reger-Nash, B. (2008). Formative research to develop a mass media campaign to increase physical activity and nutrition in a multiethnic state. *Journal of Health Communication*, **13**, 208–216. doi:10.1080/10810730701807225

Malo, T. L., Gilkey, M. B., Hall, M. E., Shah, P. D., & Brewer, N. T. (2016). Messages to motivate human papillomavirus vaccination: National studies of parents and physicians. *Cancer Epidemiology, Biomarkers & Prevention*, **25**, 1383–1391. doi:10.1158/1055-9965.EPI-16-0224

*McIntyre, P., Barnett, M. A., Harris, R. J., Shanteau, J., Skowronski, J. J., & Klassen, M. (1987). Psychological factors influencing decisions to donate organs. *Advances in Consumer Research*, **14**, 331–334.

McLean, S. A., Paxton, S. J., Massey, R., Hay, P. J., Mond, J. M., & Rodgers, B. (2016). Identifying persuasive public health messages to change community knowledge and attitudes about bulimia nervosa. *Journal of Health Communication*, **21**, 178–187. doi:10.1080/10810730.2015.1049309

Mendez, R. D. R., Rodrigues, R. C., Spana, T. M., Cornélio, M. E., Gallani, M. C., & Pérez-Nebra, A. R. (2012). Validation of persuasive messages for the promotion of physical activity among people with coronary heart disease. *Revista Latino-Americana de Enfermagem*, **20**, 1015–1023. doi:10.1590/S0104-11692012000600002

Morgan, M. G., Fischhoff, B., Bostrom, A., & Atman, C. J. (2002). *Risk communication: A mental models approach*. New York: Cambridge University Press.

Mowbray, F., Marcu, A., Godinho, C. A., Michie, S., & Yardley, L. (2016). Communicating to increase public uptake of pandemic flu vaccination in the UK: Which messages work? *Vaccine*, **34**, 3268–3274. doi:10.1016/j.vaccine.2016.05.006

Noar, S. M., Hall, M. G., Francis, D. B., Ribisl, K. M., Pepper, J. K., & Brewer, N. T. (2016). Pictorial cigarette pack warnings: A meta-analysis of experimental studies. *Tobacco Control*, **25**, 341–354. doi:10.1136/tobaccocontrol-2014-051978

Noar, S. M., Palmgreen, P., Zimmerman, R. S., Lustria, M. L. A., & Li, H.-Y. (2010). Assessing the relationship between perceived message sensation value and perceived message effectiveness: Analysis of PSAs from an effective campaign. *Communication Studies*, **61**, 21–45. doi:10.1080/10510970903396477

*Nolan, J. M., Kenefick, J., & Schultz, P. W. (2011). Normative messages promoting energy conservation will be underestimated by experts … unless you show them the data. *Social Influence*, **6**, 169–180. doi:10.1080/15534510.2011.584786

*Nolan, J. M., Schultz, P. W., Cialdini, R. B., Goldstein, N. J., & Griskevicius, V. (2008). Normative social influence is underdetected. *Personality and Social Psychology Bulletin*, **34**, 913–923. doi:10.1177/0146167208316691

O'Keefe, D. J. (2002). *Persuasion: Theory and research* (2nd ed.). Thousand Oaks, CA: Sage.

O'Keefe, D. J. (2013). The relative persuasiveness of different message types does not vary as a function of the persuasive outcome assessed: Evidence from 29 meta-analyses of 2,062 effect sizes for 13 message variations. *Annals of the International Communication Association*, **37**, 221–249. doi:10.1080/23808985.2013.11679151

*Paul, C. L., Redman, S., & Sanson-Fisher, R. W. (1997). The development of a checklist of content and design characteristics for printed health education materials. *Health Promotion Journal of Australia*, **7**, 153–159.

*Paul, C. L., Redman, S., & Sanson-Fisher, R. W. (2003). Print material content and design: Is it relevant to effectiveness? *Health Education Research*, **18**, 181–190. doi:10.1093/her/18.2.181

*Pechmann, C., Zhao, G. Z., Goldberg, M. E., & Reibling, E. T. (2003). What to convey in antismoking advertisements for adolescents: The use of protection motivation theory to identify effective message themes. *Journal of Marketing*, **67**(2), 1–18. doi:10.1509/jmkg.67.2.1.18607

*Pettigrew, S., Jongenelis, M., Chikritzhs, T., Slevin, T., Pratt, I. S., Glance, D., & Liang, W. B. (2014). Developing cancer warning statements for alcoholic beverages. *BMC Public Health*, **14**, article no. 786. doi:10.1186/1471-2458-14-786

*Pettigrew, S., Jongenelis, M. I., Glance, D., Chikritzhs, T., Pratt, I. S., Slevin, T., … Wakefield, M. (2016). The effect of cancer warning statements on alcohol consumption intentions. *Health Education Research*, **31**, 60–69. doi:10.1093/her/cyv067

*Piccolino, E. B. (1966). *Depicted threat, realism, and specificity: Variables governing safety poster effectiveness*. Doctoral dissertation, UMI no. 68-4473, Illinois Institute of Technology, Chicago, IL.

Pollard, C. M., Howat, P. A., Pratt, I. S., Boushey, C. J., Delp, E. J., & Kerr, D. A. (2016). Preferred tone of nutrition text messages for young adults: Focus group testing. *JMIR mHealth and uHealth*, **4**, e1. doi:10.2196/mhealth.4764

Popova, L., Neilands, T. B., & Ling, P. M. (2014). Testing messages to reduce smokers' openness to using novel smokeless tobacco products. *Tobacco Control*, **23**, 313–321. doi:10.1136/tobaccocontrol-2012-050723

Riker, C. A., Butler, K. M., Ricks, J. M., Record, R. A., Begley, K., Anderson, D. G., & Hahn, E. J. (2015). Creating effective media messaging for rural smoke-free policy. *Public Health Nursing*, **32**, 613–624. doi:10.1111/phn.12188

Rossiter, J. R., & Eagleson, G. (1994). Conclusions from the ARF's copy research validity project. *Journal of Advertising Research*, **34**(3), 19–32.

Santa, A. F., & Cochran, B. N. (2008). Does the impact of anti-drinking and driving public service announcements differ based on message type and viewer characteristics? *Journal of Drug Education*, **38**, 109–129. doi:10.2190/DE.38.2.b

Siegel, J. T., Lienemann, B. A., & Rosenberg, B. D. (2017). Resistance, reactance, and misinterpretation: Highlighting the challenge of persuading people with depression to seek help. *Social and Personality Psychology Compass*, **11**, e12322. doi:10.1111/spc3.12322

*Tal-Or, N., Shilo, S., & Meister, T. (2009). Third-person perception and purchase behavior in response to various selling methods. *Psychology and Marketing*, **26**, 1091–1107. doi:10.1002/mar.20314

Taylor, R. E. (2015). The role of message strategy in improving hand hygiene compliance rates. *American Journal of Infection Control*, **43**, 1166–1170. doi:10.1016/j.ajic.2015.06.015

*Thornton, B., Kirchner, G., & Jacobs, J. (1991). Influence of a photograph on a charitable appeal: A picture may be worth a thousand words when it has to speak for itself. *Journal of Applied Social Psychology*, **21**, 433–445. doi:10.1111/j.1559-1816.1991.tb00529.x

*Trawalter, S., Driskell, S., & Davidson, M. N. (2016). What is good isn't always fair: On the unintended effects of framing diversity as good. *Analyses of Social Issues and Public Policy*, **16**, 69–99. doi:10.1111/asap.12103

Truong, K. N., Hayes, G. R., & Abowd, G. D. (2006). Storyboarding: An empirical determination of best practices and effective guidelines. In *DIS '06: Proceedings of the 6th conference on designing interactive systems* (pp. 12–21).

*Turner, M. M., Banas, J. A., Rains, S. A., Jang, S. A., Moore, J. L., & Morrison, D. (2010). The effects of altercasting and counterattitudinal behavior on compliance: A lost letter technique investigation. *Communication Reports*, **23**, 1–13. doi:10.1080/08934211003598759

*Turner, M. M., Banas, J. A., Rains, S., Moore, J., & Jang, S. A. (2005, November). *Testing the effects of altercasting, source status, and target liking on compliance using the lost letter technique.* Paper presented at the annual conference of the National Communication Association, Boston, MA.

Volk, R. J., Kinsman, G. T., Le, Y.-C. L., Swank, P., Blumenthal-Barby, J., McFall, S. L., … Cantor, S. B. (2015). Designing normative messages about active surveillance for men with localized prostate cancer. *Journal of Health Communication*, **20**, 1014–1020. doi:10.1080/10810730.2015.1018618

*Weaver, K., Hock, S. J., & Garcia, S. M. (2016). "Top 10" reasons: When adding persuasive arguments reduces persuasion. *Marketing Letters*, **27**, 27–38. doi:10.1007/s11002-014-9286-1

Webb, O. J., & Eves, F. F. (2007). Promoting stair climbing: Effects of message specificity and validation. *Health Education Research*, **22**, 49–57. doi:10.1093/her/cyl045

Whittingham, J., Ruiter, R. A. C., Zimbile, F., & Kok, G. (2008). Experimental pretesting of public health campaigns: A case study. *Journal of Health Communication*, **13**, 216–230. doi:10.1080/10810730701854045

Willoughby, J. F., & Furberg, R. (2015). Underdeveloped or underreported? Coverage of pretesting practices and recommendations for design of text message-based health behavior change interventions. *Journal of Health Communication*, **20**, 472–478. doi:10.1080/10810730.2014.977468

*Wogalter, M. S., Begley, P. B., Scancorelli, L. F., & Brelsford, J. W. (1997). Effectiveness of elevator service signs: Measurement of perceived understandability, willingness to comply and behaviour. *Applied Ergonomics*, **28**, 181–187. doi:10.1016/S0003-6870(96)00063-4

Yardley, L., Morrison, L., Bradbury, K., & Muller, I. (2015). The person-based approach to intervention development: Application to digital health-related behavior change interventions. *Journal of Medical Internet Research*, **17**, e30. doi:10.2196/jmir.4055

Yzer, M., LoRusso, S., & Nagler, R. H. (2015). On the conceptual ambiguity surrounding perceived message effectiveness. *Health Communication*, **30**, 125–134. doi:10.1080/10410236.2014.974131

Zhao, X., Strasser, A., Cappella, J. N., Lerman, C., & Fishbein, M. (2011). A measure of perceived argument strength: Reliability and validity. *Communication Methods and Measures*, **5**, 48–75. doi:10.1080/19312458.2010.547822