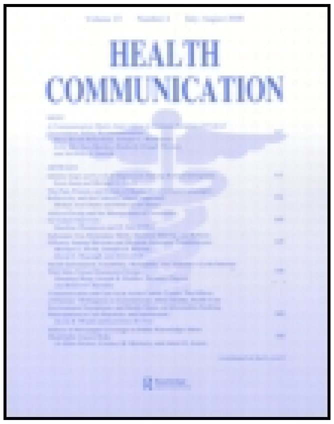

Health Communication
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/hhth20

Message Generalizations That Support Evidence-Based Persuasive Message Design: Specifying the Evidentiary Requirements

Daniel J. O'Keefe[a]

[a] Department of Communication Studies, Northwestern University

Published online: 03 Dec 2015.

Health Communication, 30: 106–113, 2015

ISSN: 1041-0236 print / 1532-7027 online
DOI: 10.1080/10410236.2014.974123

Message Generalizations That Support Evidence-Based Persuasive Message Design: Specifying the Evidentiary Requirements

Daniel J. O'Keefe
Department of Communication Studies
Northwestern University

Evidence-based persuasive message design can be informed by dependable research-based generalizations about the relative persuasiveness of alternative message-design options. Five propositions are offered as specifying what constitutes the best evidence to underwrite such generalizations: (1) The evidence should take the form of replicated randomized trials in which message features are varied. (2) Results should be described in terms of effect sizes and confidence intervals, not statistical significance. (3) The results should be synthesized using random-effects meta-analytic procedures. (4) The analysis should treat attitudinal, intention, and behavioral assessments as yielding equivalent indices of relative persuasiveness. (5) The replications included in research syntheses should not be limited to published studies or to English-language studies.

Designing effective persuasive messages is widely recognized as an important element in addressing a great many health challenges. Encouraging regular exercise, disease screening, vaccination, medication adherence, sunscreen use, safer sex practices, health care worker hand hygiene, and so forth—all these may require persuasion. But health message designers should not have to grope around hoping to stumble across some way of making their messages more persuasive. Ideally, their message design choices should be evidence based, that is, informed by good research evidence.

Persuasive message design might be guided by research evidence in any number of ways. For example, designers might use well-established general theoretical frameworks such as reasoned action theory (Fishbein & Ajzen, 2010; Yzer, 2013) or the extended parallel process model (Basil & Witte, 2012; Witte, 1992) to inform message design; these frameworks identify recurring issues that persuaders may need to address (e.g., the audience's perceived self-efficacy). Or designers might use application-specific research data, that is, information collected specifically to inform the particular project at hand; for example, health campaign planners might collect focus-group data to guide message design, might pretest specific possible messages to see which is most effective, and so forth.

Correspondence should be addressed to Daniel J. O'Keefe, Department of Communication Studies, Northwestern University, 2240 Campus Drive, Evanston, IL 60208. E-mail: d-okeefe@northwestern.edu

The focus of this essay is another basis for evidence-based message design choices, namely, empirical generalizations about the relative persuasiveness of alternative message varieties. If, for example, strong threat-appeal messages are in general (or in specific kinds of circumstances) typically more persuasive than weak threat-appeal messages, then message design choices can be guided appropriately. Such message design practice is "evidence based" in ways parallel to evidence-based medicine: If the relevant research evidence underwrites a generalization that drug A is typically more effective than drug B in treating a given condition (in general or for particular kinds of circumstances, e.g., certain types of patients), then medication choices can be guided correspondingly.

Indeed, the desire for—and to all appearances a belief in the existence of—such evidence-based persuasive communication principles is apparently widespread, judging from the titles of popular books such as *Yes!: 50 Scientifically Proven Ways to Be Persuasive* (Goldstein, Martin, & Cialdini, 2008) or *The Science of Influence* (Hogan, 2011) and from the titles of academic books such as *Writing Health Communication: An Evidence-Based Guide* (Abraham & Kools, 2012), *Behavioural Change: An Evidence-Based Handbook for Social and Public Health* (Browning & Thomas, 2005), or *Persuasive Advertising: Evidence-Based Principles* (Armstrong, 2010).

However, if persuasive message design is to be guided by evidence-based principles, evidence will be needed.

The question this essay takes up is this: What are the evidentiary requirements for generalizations that might support evidence-based persuasive message design? In what follows, five propositions are offered as providing the specifications for what constitutes the best evidence to underwrite generalizations about effective design of persuasive messages.

PROPOSITIONS

Replicated Trials

The first proposition: The evidence should take the form of replicated randomized trials in which message features are varied.

Randomized trials. For familiar reasons, randomized trials are to be preferred over other forms of evidence concerning claims about the relative persuasiveness of different forms. Only randomized trials provide suitable evidence about the causal relationships for which generalizations are desired. In certain health applications, traditional randomized trials can be modified in response to the ongoing results of the trial (e.g., Brown et al., 2009), and studies of eHealth interventions can raise distinctive issues about the choice of comparison conditions and the characteristics of participants (e.g., Glasgow, 2007)—but randomization of participants into condition remains a key methodological feature.

Variation in message features. The trials of interest involve systematic variation in the features of persuasive messages, precisely because it is such variations that are of interest to message designers. To use familiar examples: Researchers may vary whether the message states its overall conclusion explicitly or leaves it unstated; whether it mentions, refutes, or ignores opposing arguments; whether it describes the advantages of doing the advocated action or the disadvantages of not doing that action; and so forth. The point of such studies is to see what difference it makes to the persuasiveness of a message when one or more of its characteristics is experimentally varied. Results from these studies can thus be the basis for straightforward advice to message designers about how to construct messages so as to maximize persuasiveness.

Replications. A randomized trial is a good thing—but one is not enough. Replications are essential. Indeed, the importance of replications has for some time been very widely acknowledged in a number of social-scientific fields (see, e.g., Evanschitzky, Baumgarth, Hubbard, & Armstrong, 2007; Rosenthal, 1991; Tsang & Kwan, 1999), to the point that recent discussion has turned to practical questions about how replications can be encouraged (e.g., Asendorpf et al., 2013; Koole & Lakens, 2012).

There is a distinctive aspect to replications in the context of persuasion research: *message* replications. For decades, the most common research design in studies of persuasive message effects has been a "single-message design," in which each abstract message category is represented by a single concrete example. For example, Sponberg (1946) varied whether the most important arguments came first or last in a persuasive message but had only one concrete message for each argument order. Similarly—but more than 60 years later—Igou and Bless (2007) varied whether arguments were presented in a pro/con or con/pro order but had only one concrete message for each argument order.

However, as has been recognized for quite some time, single-message designs do not provide a good basis for generalization (see especially Jackson & Jacobs, 1983). Indeed, it is all too easy to find instances in which an initial single-message design study found substantial effects of a given message variation, only to have subsequent replications fail to reproduce those effects. As just one illustration: Meyerowitz and Chaiken's (1987) classic study of message framing and breast cancer screening found a loss-framed appeal to be significantly more persuasive than a gain-framed appeal, but subsequent reviews found no general advantage for loss-framed messages (e.g., Gallagher & Updegraff, 2012; O'Keefe, 2011b; O'Keefe & Jensen, 2006).

Of course, if a large number of single-message-design studies accumulate, then the needed message replications will be in hand. This way of achieving replication evidence is subject to the proviso that new studies of a given message variation employ new messages, however—which unfortunately is not always the case (for an example of reuse of experimental messages concerning flossing, see Mann, Sherman, & Updegraff, 2004; Sherman, Mann, & Updegraff, 2006; Updegraff, Sherman, Luyster, & Mann, 2007). Still, it is possible to obtain a body of message replications by amassing single-message-design studies.

But another, more efficient, way to obtain replications is to employ multiple-message designs. Rather than representing each message category by only one concrete message, instead each message type would be represented by multiple concrete messages. That is, message replications can be built into primary research designs (for some examples, see Goodall, Slater, & Myers, 2013; Jensen, 2008; Kim, Bigman, Leader, Lerman, & Cappella, 2012; Lee, Cappella, Lerman, & Strasser, 2011; Slater, Goodall, & Hayes, 2009).

To be sure, a multiple-message design cannot address all possible replication-related concerns. For example, in multiple-message designs, the same specific dependent measures will characteristically be used across the set of messages, which leaves open the question of whether results would replicate when other measures (of the same constructs) were employed.

Even so, multiple-message designs ought to be preferred, precisely because they address the issue of generalizability across messages. Any number of commentators have previously urged the use of multiple-message designs (e.g., Jackson, 1992; Reeves & Geiger, 1994; Siegel et al., 2008;

Thorson, Wicks, & Leshner, 2012; see, relatedly, Kay & Richter, 1977; Murayama, Pekrun, & Fiedler, 2014; Uncles & Kwok, 2013; Wells & Windschitl, 1999)—in some cases quite pointedly so: "It cannot be emphasized enough that using multiple messages is key to rigorous experimental investigations of mass media effects" (Grabe & Westley, 2003, p. 283). And the frequency of use of multiple-message designs has increased (Brashers & Jackson, 1999). But a glance through recent journal issues will confirm that the use of single-message designs is still all too common.

Given that (a) multiple-message designs provide much better evidence concerning the research questions of interest and (b) the report of a multiple-message design will not require that much more space than the report of a single-message design, it surely is time for reviewers and editors to begin to demand more from researchers. The suggestion here is not that reports of single-message designs never see print, but rather that the publication bar be raised where single-message designs are concerned, especially in the best journals. The presumption (on the part of reviewers and editors) should be that, absent a compelling reason for failing to include replications in the study design, results from single-message designs do not merit publication space. After all, if the results from a single-message design are sufficiently tantalizing to lead one to hope that the observed effects are in fact general, then researchers can do what Kim et al. (2012) did: report, in a single article, results both from an initial single-message design and from a follow-up multiple-message design.

Effects Sizes, Not Statistical Significance

The second proposition: Results should be described in terms of effect sizes (ESs) and confidence intervals (CIs), not statistical significance.

The potential pitfalls and misunderstandings associated with null hypothesis significance testing (NHST) have been well publicized (for general discussion, see Harlow, Mulaik, & Steiger, 1997; for discussion focused specifically on communication research, see Levine, Weber, Hullett, Park, & Lindsey, 2008; Levine, Weber, Park, & Hullett, 2008). As now seems widely recognized, simple NHST should be replaced by information that gives ESs and CIs. And yet, for all that the virtues of ESs and CIs have been repeatedly articulated (see, e.g., Cohen, 1994; Cumming, 2014), still research practices have not entirely changed. ESs and (especially) CIs are commonly not routinely reported—and even when reported they are not necessarily discussed (see, e.g., Cumming et al., 2007; Faulkner, Fidler, & Cumming, 2008; Fritz, Scherndl, & Kühberger, 2013; Sun, Pan, & Wang, 2010).

There are at least two good reasons for shifting to ESs and CIs as ways of understanding results. First, focusing on statistical significance can too easily mislead. For example, just because a given effect is statistically significant in one study and not statistically significant in a second study does not necessarily mean that the study results are inconsistent (and specifically does not necessarily mean that the two effects are significantly different from each other; for an illuminating discussion, see Nieuwenhuis, Forstmann, & Wagenmakers, 2011). Similarly, just because an effect is statistically significant does not guarantee that it is a practically significant (large) effect.

Second, using ESs and CIs can achieve all the functions of using NHST—but also provides additional useful information, namely, the magnitude of effect and the range of plausible population values given the sample effect (O'Keefe, 2011a). Knowing not only the direction of effect (whether message type A is more persuasive than message type B or vice versa) but also the size of the effect (the size of the difference in persuasiveness between the two messages) can obviously be useful to message designers. (Such information can also be used by researchers who want to ensure adequate statistical power for subsequent studies.)

To encourage this shift (from NHST to ESs and CIs), it may be helpful to explicitly formulate the research problem as one of estimation: The goal is to estimate the size of a given effect. (For a nice treatment of this idea, see Cumming [2014].) In the context of persuasion effects research, the task is estimating the size of the effect associated with a given message variation, that is, the size of the difference in persuasiveness between two message forms.

Random-Effects Meta-Analyses

The third proposition: The results should be synthesized using random-effects meta-analytic procedures.

Random-effects meta-analysis. Given the desirability of describing results in terms of effect sizes, perhaps it follows naturally that meta-analytic methods are to be preferred as a means of synthesizing research findings.[1] In such meta-analytic treatments, replication should be treated as a random factor; that is, random-effects meta-analytic procedures should be preferred over fixed-effect procedures.

The choice between random-effects meta-analyses and fixed-effect meta-analyses has been much discussed, with various considerations adduced as potentially bearing on that choice. But arguably the crucial consideration is what research question the investigator wants to answer, because the two procedures answer different questions. Each procedure provides an estimate of a mean effect and an associated confidence interval, but these figures represent answers to substantively different questions (Borenstein, Hedges, Higgins, & Rothstein, 2009; Card, 2012).

[1] There are meta-analytic methods that do not involve synthesizing effect sizes, such as vote-counting procedures (Bushman & Wang, 2009). But the most familiar meta-analytic methods—and the ones of natural interest in the present context—are ones that synthesize effect sizes.

Expressed briefly: In fixed-effect analyses, the effect size from each message pair (case, study, implementation) is taken to reflect the one general (fixed) population effect and hence only human sampling variation is responsible for variability among effect sizes. For estimating the population effect size, a fixed-effect analysis provides both a mean and a 95% CI. The 95% CI is in effect the answer to the question, "If there were more participants in the studies that have been done, where could the mean plausibly be?"[2] But this question is not really of very much interest, especially where the larger purpose is that of informing future message design. Our interest in the particular messages that have already been studied is not for those specific messages themselves, but rather for what those messages can tell us about other—future—messages.

In a random-effects analysis, the assumption is that each message pair (implementation, study, case) has its own population effect, which is estimated by data from the human sample for that study. Thus, the observed variability in a collection of effect sizes reflects not only human sampling variation but also variation in the underlying effects associated with the different implementations. For estimating the location of the average across the universe of those population effects, a random-effects analysis provides both a mean and a 95% CI.[3] The 95% CI is in effect the answer to the question, "If there were more message pairs (studies) like these, with more participants, where could the mean plausibly be?" That is, the random-effects analysis answers the question of presumable interest (involving generalization beyond the cases in hand), whereas the fixed-effect analysis answers a question typically of little or no interest (involving conclusions limited to the studies already completed).[4]

Hence "meta-analysts using fixed-effects models are only justified in drawing conclusions about the specific set of studies included in their meta-analysis" whereas "the use of random-effects models justifies inferences that generalize beyond the particular set of studies included in the meta-analysis to a population of potential studies" (Card, 2012, p. 233). Because our interest here is in dependable generalizations about message effects—generalizations that extend beyond the messages already studied—random-effects analyses should be employed. The guidance message designers need is not provided by results from fixed-effect analyses.

Similarity of results? One potential source of confusion is that sometimes the numerical results from random-effects and fixed-effect analyses can look rather similar. Some meta-analysts appear to have been misled by such similarities into supposing that the choice between these two analyses doesn't make much difference. But this is a mistake, for two reasons.

First: Fixed-effect and random-effects analyses may sometimes give *numerically* similar results, but they never give *substantively* similar results. Because the two analyses address substantively different questions, the answers are necessarily different.

A parallel case: The questions "How many states are there in the United States?" and "What is the atomic number of tin?" have the same answer, namely, "50." But what "50" means—what "50" represents—is utterly different in those two answers. The two questions have the "same" answer only in a silly or trivial sense. Substantively, those two questions have different answers.

And similarly with fixed- and random-effects meta-analytic results: The two results may sometimes be "the same" (or similar) numerically, but the numbers do not mean the same things because they do not represent the same property. Fixed- and random-effects analyses never yield the same substantive result, because "the fixed-effect model and the random-effects model address different hypotheses" (Borenstein et al., 2009, p. 272). Of course, this first point would be mere pedantry if fixed- and random-effects analyses always gave numerically similar results. After all, if the numbers were always the same, then it wouldn't really matter that the numbers had substantively different meanings. But . . .

Second: Fixed-effect and random-effects analyses sometimes give numerically divergent results. In particular, the widths of the CIs can differ in consequential ways.

Consider, as an example, Scott-Sheldon, DeMartini, Carey, and Carey's (2009) meta-analysis of the effect of alcohol interventions on (inter alia) college students' consumption intentions. Their conclusion, on the basis of fixed-effect results, was that such interventions were successful in improving intentions (p. 818)—but their random-effects analysis did not yield a statistically significant effect (p. 814, Table 3). That is, the evidence did not in fact support a general conclusion that such interventions improve consumption intentions. As another example: Hart et al.'s (2009) meta-analysis of selective exposure (congeniality) research examined attitudinal confidence as one potential moderator of the general preference for congenial information. On the basis of the results of a fixed-effect analysis, Hart et al. concluded that "congeniality is weaker at high (vs. low or moderate)

---

[2]Thus the width of the confidence interval in a fixed-effect analysis is a function of the total human sample size across the various studies. It is not affected by the number of different studies (message pairs) or by the distribution of participants over those studies.

[3]Because in random-effects analyses, each different message pair (implementation) is seen to have its own population effect, the mean effect in a random-effects analysis is sometimes described as the estimate of the mean across the universe of those various population effects, rather than itself being a population effect (see, e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009, p. 79). "Random-effects models conceptualize a population distribution of effect sizes, rather than a single effect size as in the fixed-effects model" (Card, 2012, p. 230).

[4]This brief exposition necessarily passes over a number of complexities; for fuller discussions, see Borenstein et al. (2009, pp. 77–86) and Card (2012, pp. 229–256).

levels of confidence" (p. 581). But a random-effects analysis found no significant differences in congeniality as a function of variations in confidence (p. 576, Table 3). That is, the evidence did not underwrite a general conclusion that congeniality varies as a function of attitudinal confidence.

Given that (a) fixed-effect and random-effects analyses can yield numerically divergent results and (b) only random-effects analyses underwrite the desired sorts of generalizations, random-effects analyses should be required. Indeed, in most cases it is arguably a bad idea to even report results from fixed-effect analyses: Usually "we do want to make inferences about a wider population," and if fixed-effect results are reported, "readers will make these inferences even if they are not warranted" (Borenstein et al., 2009, p. 84). Reporting fixed-effect results invites avoidable misunderstandings.

Differences in power? Another source of confusion is that fixed-effect analyses appear to have greater statistical power than do random-effects analyses. But appearances are once again deceiving, because the seemingly greater statistical power arises from answering a different—and less demanding—research question: "It is not meaningful to compare power for fixed- and random-effects analyses since the two values of power are not addressing the same question" (Borenstein et al., 2009, p. 272). Yes, for a given data set (collection of effect sizes), the fixed-effect analysis has greater power for answering its research question than the random-effects analysis has for answering its research question. But in choosing between the two analyses, the relevant criterion is not "Which of these research questions is easier to answer?" but rather "Which question do we want to answer?" If the goal is dependable generalizations about persuasive message effects, the research question of interest involves conclusions that go beyond the cases in hand—and thus random-effects analyses are required.

Summary. The temptation to report fixed-effect analyses can be quite strong, given the seemingly greater power and the seemingly similar results. Consider, for example, Hart et al.'s (2009) characterization of their selective exposure meta-analysis: "Generally, the results from fixed- and random-effects models converged. Thus, we focus on the fixed-effects models, which are more powerful" (p. 575). But this reasoning is mistaken. The results did not "converge" (the numbers might have looked similar, but substantively the results were by definition different), and the fixed-effect analyses were not "more powerful" in any straightforward sense (comparing the power of the two analyses is inappropriate because the two analyses answer different questions). Both the apparent greater power of fixed-effect analyses and the apparent similarity of results between fixed- and random-effects analyses are illusions.

Collapsing Persuasive Outcomes

The fourth proposition: The analysis should treat attitudinal, intention, and behavioral assessments as yielding equivalent indices of relative persuasiveness.

The rationale for this proposition is simply the observed equivalency of these different outcomes when used as assessments of relative persuasiveness. O'Keefe (2013) reanalyzed data from 29 meta-analyses of 13 message variations, including gain–loss framing, message sidedness, conclusion explicitness, and several threat–appeal variations. The relative persuasiveness of message types was largely invariant across attitude, intention, and behavior outcomes, in the sense that for a given message variation, the observed mean effect sizes did not significantly differ across those outcomes. That is, the three different outcome measures yielded statistically indistinguishable estimates of the relative persuasiveness of alternative message kinds.

As one example: In Witte and Allen's (2000) meta-analysis concerning the relative persuasiveness of strong and weak fear appeals, the mean effect sizes (expressed as correlations) were .14 for attitudinal outcomes, .15 for intention outcomes, and .16 for behavioral outcomes; these mean effect sizes were not significantly different from each other. Thus, these three outcomes were functionally interchangeable indices of relative persuasiveness; no matter which outcome was examined, the same conclusion was reached about both the direction and the size of the difference in persuasiveness between the two message varieties.

Just to be clear: Attitude, intention, and behavior outcomes are not interchangeable indices of *persuasiveness*; on the contrary, a given message might be very persuasive when assessed in terms of intention outcomes but relatively unpersuasive when behavioral effects are examined. However, these different outcomes do provide equivalent (interchangeable) indices of *relative persuasiveness*: "If message type A is more persuasive than message type B with attitudinal outcomes, it is also—and equally—more persuasive with intention and behavioral outcomes" (O'Keefe, 2013, p. 221).

Hence for building generalizations about the relative persuasiveness of message types, these three outcome variables can and should be treated as functionally equivalent. Rather than analyzing these three outcomes separately (or restricting one's meta-analysis to one kind of outcome), meta-analysts should combine effect sizes across these outcomes—which will provide a better estimate of the effect size of interest (and correspondingly enhance statistical power and reduce vulnerability to false positives).[5]

[5]The observed functional equivalence of attitude, intention, and behavior outcomes as indices of relative persuasiveness has implications beyond meta-analytic methodological choices. In particular, where formative persuasive campaign research compares two or more possible messages with the purpose of identifying the one most effective in influencing behavioral outcomes, message designers need not collect behavioral data (because, e.g., intention data will yield the same conclusion about relative persuasiveness).

This proposition does not underwrite collapsing any and all "persuasive outcome" variables when the research question concerns the relative persuasiveness of two message forms. Nor does it underwrite collapsing these outcomes where other research questions are pursued. But specifically where research questions about the relative persuasiveness of message forms are involved, the evidence in hand plainly indicates the functional equivalence of attitude, intention, and behavior outcomes.

Broad Literature Retrieval

The fifth proposition: The replications included in research syntheses should not be limited to published studies or to English-language studies.

Including unpublished studies. Research synthesis should be based on both published and unpublished studies, because well-known processes distort the research results that appear in published form. In particular, the editorial process is biased in favor of publishing statistically significant effects, and (perhaps unsurprisingly) researchers correspondingly are inclined to engage in practices such as selective reporting—and this naturally leads to a published research literature that can be misleading (Chan, Hrobjartsson, Haahr, Gøtzsche, & Altman, 2004; Dwan, Gamble, Williamson, Kirkham, & the Reporting Bias Group, 2013; Gerber & Malhotra, 2008; Ioannidis, 2005, 2008; Simmons, Nelson, & Simonsohn, 2011).

To address this concern, sometimes meta-analysts examine only published studies and then test the collection of studies for evidence of publication bias. However, this is less desirable than obtaining unpublished results, for two reasons. First, publication-bias detection procedures are imperfect in various ways (e.g., characteristically low power) and often misapplied (for discussion, see Ioannidis & Trikalinos, 2007; Kromrey & Rendina-Gobioff, 2006). Second, having additional cases permits better estimation of the effect of interest. The larger the number of cases (studies), the narrower the confidence interval (ceteris paribus)—and correspondingly the better the estimate of the underlying effect. So even absent publication bias, including unpublished studies is still important.

Sometimes it is supposed that restricting one's review to published peer-reviewed studies provides a measure of confidence about study quality. But there is actually little evidence that peer review guarantees quality (see, e.g., Jefferson, Rudin, Folse, & Davidoff, 2007)—and there is good reason to believe that reviewers' assessments are influenced not simply by a study's methods (an appropriate basis for judging study quality) but also by irrelevant considerations such as whether the results were statistically significant or consistent with the reviewer's own hypotheses (Ernst & Resch, 1994; Mahoney, 1977). When a plausible case can be made that "most published research findings are false" (Ioannidis, 2005), it is difficult to credit hand-waving assertions that excluding unpublished studies assures quality.

Including non-English studies. It is common to see meta-analyses explicitly exclude studies that have been reported in some language other than English. Some examples: "We applied search limiters to exclude studies . . . written in a language other than English" (Cushing & Steele, 2010, p. 939); "The search was restricted to English peer-reviewed journal articles, books, and book chapters to minimize the risk of admitting studies of poor quality" (Porath-Waller, Beasley, & Beirness, 2010, p. 712); and "only English papers were included" (Wanyonyi, Themessl-Huber, Humphris, & Freeman, 2011, p. 349).

It is difficult to discern a rationale for this practice. Perhaps some have been dissuaded by the apparent challenges of translation. But these challenges are not as daunting as one might think, for two reasons. First, one need not translate the entire research report. The methods and results sections usually contain all the information a meta-analyst needs in order to extract effect sizes, sample sizes, and information relevant to coding study characteristics. Second, free online translation sites (such as Google Translate) make the task dead simple—even if the research report is in Afrikaans or Croatian or Urdu or . . . .

SUMMARY

There are many ways in which the abstract idea of evidence-based persuasive message design might be realized other than the application of empirically based generalizations about the relative persuasiveness of different message kinds. For example, in a specific campaign, formative research might directly pretest the relative effectiveness of particular alternative messages (thus informing an evidence-based choice between them). And even when empirical generalizations are used to guide message design, the challenges of moving from such generalizations to their creative and effective application can be considerable (Jackson & Aakus, 2014).

But when dependable generalizations about the comparative persuasiveness of alternative message types are wanted, the evidence for such generalizations should take the form of data about relative persuasiveness (treating attitudinal, intention, and behavioral outcomes as functionally equivalent) from replicated randomized trials (no matter whether published or unpublished, and no matter in what language), with the results described in terms of effect sizes and synthesized using random-effects meta-analysis.

ACKNOWLEDGMENTS

A version of this article was presented at the Kentucky Conference on Health Communication Preconference on Message Design in Health Communication, April 2014.

REFERENCES

Abraham, C., & Kools, M. (Eds.). (2012). *Writing health communication: An evidence-based guide*. Los Angeles, CA: Sage.

Armstrong, J. S. (2010). *Persuasive advertising: Evidence-based principles*. New York, NY: Palgrave Macmillan.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119.

Basil, M., & Witte, K. (2012). Health risk message design using the extended parallel process model. In H. Cho (Ed.), *Health communication message design: Theory and practice* (pp. 41–58). Los Angeles, CA: Sage.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chicester, West Sussex, UK: Wiley.

Brashers, D. E., & Jackson, S. (1999). Changing conceptions of "message effects": A 24-year overview. *Human Communication Research*, *25*, 457–477.

Brown, C. H., Ten Have, T. R., Jo, B., Dagne, G., Wyman, P. A., Muthen, B., & Gibbons, R. D. (2009). Adaptive designs for randomized trials in public health. *Annual Review of Public Health*, *30*, 1–25.

Browning, C. J., & Thomas, S. A. (Eds.). (2005). *Behavioural change: An evidence-based handbook for social and public health*. Edinburgh, UK: Elsevier Churchill Livingstone.

Bushman, B. J., & Wang, M. C. (2009). Vote-counting procedures in meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 207–220). New York, NY: Russell Sage Foundation.

Card, N. A. (2012). *Applied meta-analysis for social science research*. New York, NY: Guilford.

Chan, A. W., Hrobjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Journal of the American Medical Association*, *291*, 2457–2465.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1002.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., . . . Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science*, *18*, 230–232.

Cushing, C. C., & Steele, R. G. (2010). A meta-analytic review of eHealth interventions for pediatric health promoting and maintaining behaviors. *Journal of Pediatric Psychology*, *35*, 937–949.

Dwan, K., Gamble, C., Williamson, P. R., Kirkham, J. J., & the Reporting Bias Group. (2013). Systematic review of the empirical evidence of study publication bias and outcome reporting bias: An updated review. *PLoS ONE*, *8*(7), e66844.

Ernst, E., & Resch, K. L. (1994). Reviewer bias: A blinded experimental study. *Journal of Laboratory and Clinical Medicine*, *124*, 178–182.

Evanschitzky, H., Baumgarth, C., Hubbard, R., & Armstrong, J. S. (2007). Replication research's disturbing trend. *Journal of Business Research*, *60*, 411–415.

Faulkner, C., Fidler, F., & Cumming, G. (2008). The value of RCT evidence depends on the quality of statistical analysis. *Behaviour Research and Therapy*, *46*, 270–281.

Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action approach*. New York, NY: Psychology Press.

Fritz, A., Scherndl, T., & Kühberger, A. (2013). A comprehensive review of reporting practices in psychological journals: Are effect sizes really enough? *Theory and Psychology*, *23*, 98–122.

Gallagher, K. M., & Updegraff, J. A. (2012). Health message framing effects on attitudes, intentions, and behavior: A meta-analytic review. *Annals of Behavioral Medicine*, *43*, 101–116.

Gerber, A., & Malhotra, N. (2008). Do statistical reporting standards affect what is published? Publication bias in two leading political science journals. *Quarterly Journal of Political Science*, *3*, 313–326.

Glasgow, R. E. (2007). eHealth evaluation and dissemination research. *American Journal of Preventive Medicine*, *32*, S119–S126.

Goldstein, N. J., Martin, S. J., & Cialdini, R. B. (2008). *Yes!: 50 Scientifically proven ways to be persuasive*. New York, NY: Free Press.

Goodall, C. E., Slater, M. D., & Myers, T. A. (2013). Fear and anger responses to local news coverage of alcohol-related crimes, accidents, and injuries: Explaining news effects on policy support using a representative sample of messages and people. *Journal of Communication*, *63*, 373–392.

Grabe, M. E., & Westley, B. H. (2003). The controlled experiment. In G. H. Stempel III, D. H. Weaver, & G. C. Wilhoit (Eds.), *Mass communication research and theory* (pp. 267–298). Boston, MA: Allyn and Bacon.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.) (1997). *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, *135*, 555–588.

Hogan, K. (2011). *The science of influence: How to get anyone to say yes in 8 minutes or less!* (2nd ed.). Hoboken, NJ: Wiley.

Igou, E. R., & Bless, H. (2007). Conversational expectations as a basis for order effects in persuasion. *Journal of Language and Social Psychology*, *26*, 260–273.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 696–701.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–648.

Ioannidis, J. P. A., & Trikalinos, T. A. (2007). The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *Canadian Medical Association Journal*, *176*, 1091–1096.

Jackson, S. (1992). *Message effects research: Principles of design and analysis*. New York, NY: Guilford Press.

Jackson, S., & Aakhus, M. (2014). Becoming more reflective about the role of design in communication. *Journal of Applied Communication Research*, *42*, 125–134.

Jackson, S., & Jacobs, S. (1983). Generalizing about messages: Suggestions for design and analysis of experiments. *Human Communication Research*, *9*, 169–181.

Jefferson, T., Rudin, M., Folse, S. B., & Davidoff, F. (2007). Editorial peer review for improving the quality of reports of biomedical studies. *Cochrane Database of Systematic Reviews 2007*, issue 2, article MR000016.

Jensen, J. D. (2008). Scientific uncertainty in news coverage of cancer research: Effects of hedging on scientists' and journalists' credibility. *Human Communication Research*, *34*, 347–369.

Kay, E. J., & Richter, M. L. (1977). The category-confound: A design error. *Journal of Social Psychology*, *103*, 57–63.

Kim, H. S., Bigman, C. A., Leader, A. E., Lerman, C., & Cappella, J. N. (2012). Narrative health communication and behavior change: The influence of exemplars in the news on intention to quit smoking. *Journal of Communication*, *62*, 473–492.

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, *7*, 608–614.

Kromrey, J. D., & Rendina-Gobioff, G. (2006). On knowing what we do not know: An empirical comparison of methods to detect publication bias in meta-analysis. *Educational and Psychological Measurement*, *66*, 357–373.

Lee, S., Cappella, J. N., Lerman, C., & Strasser, A. A. (2011). Smoking cues, argument strength, and perceived effectiveness of antismoking PSAs. *Nicotine and Tobacco Research*, *13*, 282–290.

Levine, T. R., Weber, R., Hullett, C., Park, H. S., & Lindsey, L. L. (2008). A critical assessment of null hypothesis significance testing in quantitative communication research. *Human Communication Research*, *34*, 171–187.

Levine, T. R., Weber, R., Park, H. S., & Hullett, C. (2008). A communication researchers' guide to null hypothesis significance testing and alternatives. *Human Communication Research*, *34*, 188–209.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *1*, 161–175.

Mann, T., Sherman, D., & Updegraff, J. (2004). Dispositional motivations and message framing: A test of the congruency hypothesis in college students. *Health Psychology*, *23*, 330–334.

Meyerowitz, B. E., & Chaiken, S. (1987). The effect of message framing on breast self-examination attitudes, intentions, and behavior. *Journal of Personality and Social Psychology*, *52*, 500–510.

Murayama, K., Pekrun, R., & Fiedler, K. (2014). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review*, *18*, 107–118.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*, 1105–1107.

O'Keefe, D. J. (2011a). The asymmetry of predictive and descriptive capabilities in quantitative communication research: Implications for hypothesis development and testing. *Communication Methods and Measures*, *5*, 113–125.

O'Keefe, D. J. (2011b). Generalizing about the persuasive effects of message variations: The case of gain-framed and loss-framed appeals. In T. van Haaften, H. Jansen, J. de Jong, & W. Koetsenruijter (Eds.), *Bending opinion: Essays on persuasion in the public domain* (pp. 117–131). Leiden, The Netherlands: Leiden University Press.

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. In E. L. Cohen (Ed.), *Communication yearbook 37* (pp. 221–249). New York, NY: Routledge.

O'Keefe, D. J., & Jensen, J. D. (2006). The advantages of compliance or the disadvantages of noncompliance? A meta-analytic review of the relative persuasive effectiveness of gain-framed and loss-framed messages. In C. S. Beck (Ed.), *Communication yearbook 30* (pp. 1–43). Mahwah, NJ: Lawrence Erlbaum Associates.

Porath-Waller, A. J., Beasley, E., & Beirness, D. J. (2010). A meta-analytic review of school-based prevention for cannabis use. *Health Education and Behavior*, *37*, 709–723.

Reeves, B., & Geiger, S. (1994). Designing experiments that assess psychological responses to media messages. In A. Lang (Ed.), *Measuring psychological responses to media messages* (pp. 165–180). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rosenthal, R. (1991). Replication in behavioral research. In J. W. Neuliep (Ed.), *Replication research in the social sciences* (pp. 1–30). Newbury Park, CA: Sage.

Scott-Sheldon, L. A. J., DeMartini, K. S., Carey, K. B., & Carey, M. P. (2009). Alcohol interventions for college students improves antecedents of behavioral change: Results from a meta-analysis of 34 randomized controlled trials. *Journal of Social and Clinical Psychology*, *28*, 799–823.

Sherman, D. K., Mann, T., & Updegraff, J. A. (2006). Approach/avoidance orientation, message framing, and health behavior: Understanding the congruency effect. *Motivation and Emotion*, *30*, 165–169.

Siegel, J. T., Alvaro, E. M., Crano, W. D, Lac, A., Ting, S., & Jones, S. P. (2008). A quasi-experimental investigation of message appeal variations on organ donor registration rates. *Health Psychology*, *27*, 170–178.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Slater, M. D., Goodall, C. E., & Hayes, A. F. (2009). Self-reported news attention does assess differential processing of media content: An experiment on risk perceptions utilizing a random sample of US local crime and accident news. *Journal of Communication*, *59*, 117–134.

Sponberg, H. (1946). A study of the relative effectiveness of climax and anti-climax order in an argumentative speech. *Speech Monographs*, *13*, 35–44.

Sun, S., Pan, W., & Wang, L. L. (2010). A comprehensive review of effect size reporting and interpreting practices in academic journals in education and psychology. *Journal of Educational Psychology*, *102*, 989–1004.

Thorson, E., Wicks, R., & Leshner, G. (2012). Experimental methodology in journalism and mass communication research. *Journalism & Mass Communication Quarterly*, *89*, 112–124.

Tsang, E. W. K., & Kwan, K.-M. (1999). Replication and theory development in organizational science: A critical realist perspective. *Academy of Management Review*, *24*, 759–780.

Uncles, M. D., & Kwok, S. (2013). Designing research with in-built differentiated replication. *Journal of Business Research*, *66*, 1398–1405.

Updegraff, J. A., Sherman, D. K., Luyster, F. S., & Mann, T. L. (2007). The effects of message quality and congruency on perceptions of tailored health communications. *Journal of Experimental Social Psychology*, *43*, 249–257.

Wanyonyi, K. L., Themessl-Huber, M., Humphris, G., & Freeman, R. (2011). A systematic review and meta-analysis of face-to-face communication of tailored health messages: Implications for practice. *Patient Education and Counseling*, *85*, 348–355.

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin*, *25*, 1115–1125.

Witte, K. (1992). Putting the fear back into fear appeals: The extended parallel process model. *Communication Monographs*, *59*, 329–349.

Witte, K., & Allen, M. (2000). A meta-analysis of fear appeals: Implications for effective public health programs. *Health Education and Behavior*, *27*, 591–615.

Yzer, M. (2013). Reasoned action theory: Persuasion as belief-based behavior change. In J. P. Dillard & L. Shen (Eds.), *The* Sage *handbook of persuasion: Developments in theory and practice* (2nd ed., pp. 120–136). Thousand Oaks, CA: Sage.