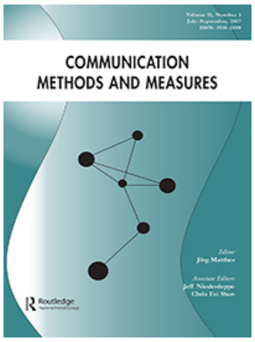

Routledge
Taylor & Francis Group

Communication Methods and Measures

ISSN: 1931-2458 (Print) 1931-2466 (Online) Journal homepage: http://www.tandfonline.com/loi/hcms20

Misunderstandings of Effect Sizes in Message Effects Research

Daniel J. O'Keefe

To cite this article: Daniel J. O'Keefe (2017) Misunderstandings of Effect Sizes in Message Effects Research, Communication Methods and Measures, 11:3, 210-219, DOI: 10.1080/19312458.2017.1343812

To link to this article: https://doi.org/10.1080/19312458.2017.1343812

Published online: 12 Jul 2017.

COMMUNICATION METHODS AND MEASURES
2017, VOL. 11, NO. 3, 210–219
https://doi.org/10.1080/19312458.2017.1343812

TEACHER'S CORNER

OPEN ACCESS Check for updates

Misunderstandings of Effect Sizes in Message Effects Research

Daniel J. O'Keefe

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

ABSTRACT

Widespread acknowledgement of the weaknesses of null hypothesis significance testing has led to correspondingly greater appreciation of the central role of effect size estimation in social-scientific research. But in the context of message effects research, it is easy to confuse an effect size—a quantitative representation of the effect of a variable on an outcome—with the size of the effect of a message on an outcome. Illustrations of this misunderstanding are offered, its unhappy consequences specified, and remedies discussed.

KEYWORDS
Effect sizes; message effects; meta-analysis

In message effects research, effect sizes have become more prominent in both primary research and in secondary treatments such as meta-analytic reviews. This increasing attention to effect sizes has brought several advantages. Instead of simply knowing that message A was more effective than message B, it is possible to say something about the size of the observed difference (the effect size) and the associated confidence interval. And familiarity with the concept of effect sizes naturally encourages greater attention to questions of statistical power, now seen as bearing on the accuracy of effect size estimates. (For some general treatments, see Cumming, 2014; Kline, 2013; Morey, Rouder, Verhagen, & Wagenmakers, 2014. Concerning communication research specifically, see Levine, 2013, and Levine, Weber, Hullett, Park, & Lindsey, 2008).

However, as with any such methodological development, misunderstandings or misapprehensions can arise. Unfortunately, in message effects research, the concept of an effect size seems particularly prone to being misunderstood. What follows identifies a core misunderstanding, offers examples of the problem, and discusses possible remedies.

As an initial clarification: The focus of the present discussion is the understanding of effect sizes in message effects research. "Message effects research" is here used as shorthand for experimental studies of the effects of different messages or interventions on outcomes of interest. Thus, it is meant to include (as examples) studies comparing the persuasiveness of strong and weak threat appeals for encouraging protective health behaviors, studies examining the relative effectiveness of different media literacy interventions, studies comparing different means of encouraging behavioral self-efficacy, studies examining the effects of different risk-information formats on risk perceptions, studies of different means of encouraging participation in online collective enterprises (e.g., Wikipedia), and so forth. What all these have in common is an interest in the relative effectiveness of different messages (interventions, treatments, strategies, formats, etc.) for influencing some outcome of interest.

The core misunderstanding

The primary point to be grasped is this: In message effects research, an "effect size" does not describe the size of a message's effect. Paradigmatically, it describes the *difference* between the sizes of the effects of *two* messages.

CONTACT Daniel J. O'Keefe d-okeefe@northwestern.edu Department of Communication Studies, Northwestern University, 2240 Campus Drive, Evanston IL 60208.

This point can be concretized by considering one common way of expressing effect sizes, the standardized mean difference (*d*). In a posttest-only randomized trial comparing the outcomes from two different conditions (e.g., narrative message vs. argument-based message, statistical information vs. case study, etc.), the effect size *d* is computed as the difference between the two means on the outcome variable, divided by the pooled standard deviation.[1]

Imagine a hypothetical message effects study, study A, comparing the effectiveness of two messages. The outcome variable ranges from zero (indicating the message was completely ineffective) to 100 (completely effective). The means in the two conditions are 25.0 and 15.0, with a pooled standard deviation of 10. The effect size in study A would be d = 1.0 (or, equivalently expressed as a correlation, r = .45). In a parallel hypothetical study B, with the same outcome variable, the means are 90.0 and 80.0 (again with *sd* = 10). The effect size in study B also would be d = 1.0 (r = .45). As these two examples illustrate, the magnitude of the effect size does not necessarily indicate the effectiveness of the messages: the messages in Study B were much more effective than those in Study A, but the effect size (the difference between the two messages in a study) was identical across the two studies.

So in message effects research, a large effect size means a large difference between the two messages being compared—not a highly effective message or a large message effect. From knowing what the difference is between two conditions, you cannot tell the value of either condition. If you know that Susie has $5 more in her saving account than Johnny has in his, you cannot tell how much money either one has; all you know is how big the difference is, not how big either of the two values is. If you know that the effect size in a study is d = .25, you cannot tell how effective the messages were in that study; all you know is how big the difference is, not how effective either message was in absolute terms.

This may seem clear enough, but as the next section indicates, misunderstandings of effect sizes are all too possible.

Examples of misunderstandings

This section offers examples of how effect sizes have been misunderstood in message effects research. These misapprehensions arise both when a single effect size is examined and when effect sizes are compared.

Examining one effect size

When considering a single effect size, whether from a primary research study or a meta-analysis, one possible misunderstanding is to think that the effect size represents the absolute level of one treatment condition on the outcome variable. As an example: Yzer, Southwell, and Stephenson (2013) discussed the meta-analytic findings of Boster and Mongeau (1984), Witte and Allen (2000), and de Hoog, Stroebe, and de Wit (2007) concerning threat appeals. Those meta-analyses reported mean effect sizes for the effect of various message variations (e.g., manipulations of depicted threat severity) on fear assessments. The mean effect sizes (expressed as *r*s) ranged from .16–.36. Yzer et al. (2013, p. 171) characterized these results as follows: “The observed correlations suggest that, on average, fear appeal research has not been able to produce very strong levels of fear… . Put differently, much of what we think we know about fear as a mechanism is based on a comparison between low and moderate levels of fear.”

But that interpretation is not correct. At best, the observed correlations might suggest that, on average, the message manipulations in fear appeal research have not been able to produce large *differences* in fear levels (differences between two message conditions in aroused fear). But one

[1] Effect sizes can be expressed in a variety of equivalent forms, including a standardized mean difference (*d*) and a correlation coefficient (*r*). For discussion and details, see Grissom and Kim (2012) and Rosnow and Rosenthal (2009).

cannot tell from these effect sizes what the *absolute* level of fear was in the studies reviewed; the reported effect sizes are consistent with any number of different absolute levels of fear arousal.

Another example of the same kind of misunderstanding is provided by Gallagher and Updegraff's (2012) message framing meta-analysis. The advantage of gain-framed messages over loss-framed messages for health prevention behavior outcomes was reported as a mean effect size (expressed as *r*) of .08. Concerning the magnitude of that effect, they wrote "Although the effect of message framing on prevention behavior might seem relatively small in magnitude, it is important to keep in mind that health behaviors are complex in nature, and health message framing is but one aspect of an intervention that can contribute to its success. Indeed, Latimer and colleagues offer that 'the small changes induced by framed messages may contribute to the additive effects of multiple intervention components'" (p. 113, quoting Latimer, Salovey, & Rothman, 2007).

But the small mean effect size (r = .08) does not necessarily mean that there were only "small changes induced by framed messages." It means that the *difference* between the changes induced by the gain-framed message and those induced by the loss-framed message was not very large. Each message might have induced very large changes—but the two message types did not differ very much in how much change they induced.

The same kind of misapprehension is reflected in Ashford, Edmunds, and French's (2010) description of the results of their meta-analysis of the effects of various intervention elements on physical activity self-efficacy. The reviewed studies included comparisons of interventions with and without a given design element—with or without (for example) exposure to vicarious success, or inclusion of a graded-mastery component, and so on. For each such comparison, a mean effect size was computed. Ashford et al. (2010, p. 277) reported that "Providing feedback by comparing participant's performance with the performance of others produced the largest effect size estimates (d = 0.44), followed by feedback on the participants' past performances (d = 0.43)." That is, of all the comparisons examined, those two feedback-related comparisons "produced the highest effect size estimates generated in this meta-analysis" (p. 279).

But compare this description of these effect sizes: "Interventions that included feedback on past or others' performance produced the highest levels of self-efficacy found in this review" (p. 265). This is not a correct portrayal of the obtained results. Taken at face value, the observed mean effect sizes suggest that, on average, variation in the presence or absence of those feedback intervention elements produced the largest *differences* in physical activity self-efficacy (that is, the largest differences in self-efficacy between two conditions). But one cannot tell from these effect sizes what the *absolute* level of self-efficacy was in any single condition.

A variant on this sort of misunderstanding is reflected in representations of individual effect sizes as reflecting the "effectiveness" of a given kind of message or treatment. As an example: In Tannenbaum et al.'s (2015) meta-analytic review of fear appeal research, the focal comparison of interest was between "messages designed to depict relatively high levels of fear compared to conditions designed to depict relatively lower levels of fear" (p. 1183). They reported that "the average weighted effect size comparing outcomes for treatment to comparison groups was d = 0.29 with a 95% CI [0.22, 0.35]" (pp. 1192–1193), glossing this result by saying "Fear appeals are effective" (p. 1196). But the reported results do not show that higher-level fear appeals are effective, not in any absolute sense. The results show only that higher-level fear appeals are more effective than lower-level fear appeals. It might be that higher-level fear appeals in fact aren't all that effective—even though they are more effective than lower-level appeals.

Tannenbaum et al. (2015) also looked for circumstances under which the observed effect reversed, that is, conditions under which lower-level fear appeals would be more effective than higher-level appeals. No such circumstances were apparent: "there was not one level of any moderator that we tested for which fear appeals backfired to produce worse outcomes relative to the comparison group" (p. 1186). This finding was glossed as indicating that "there are no identified circumstances under which they [fear appeals] backfire and lead to undesirable outcomes" (p. 1178).

But the reported results do not show that higher-level fear appeals are not ineffective in absolute terms, in the sense of leading to undesirable outcomes. It might be that both higher-level and lower-level fear appeals produce undesirable outcomes (e.g., boomerang attitude change)—but with those effects being larger for lower-level fear appeals than for higher-level fear appeals. Taken at face value, the results show that higher-level fear appeals are generally a better choice for persuaders than lower-level fear appeals, but that is a comparative judgment (about relative effectiveness), not an absolute judgment (about whether higher-level appeals are effective or ineffective, i.e., backfire).

A similar misapprehension about effectiveness was reflected in Eisend and Tarrahi's (2016) meta-meta-analysis of 44 previous meta-analyses concerning factors influencing advertising outcomes such as attitudes, credibility judgments, behavior, and the like. The individual meta-analyses examined not only variations in message properties (one-sided vs. two-sided advertising, comparative vs. noncomparative advertising, etc.) but also variation in recipient characteristics (e.g., prior attitudes), source characteristics (e.g., celebrity vs. non-celebrity), and advertising strategy (variations in exposure, repetition, etc.). Those meta-analyses yielded 324 meta-analytic effect sizes of interest, expressed as correlations. For each meta-analytic effect size, "the absolute values were coded because we were interested in the size of the effect and not its direction" (p. 524). That is, the question was (for example) "the strength of the effect of message characteristics compared to other input variables" (p. 525). Across all 324 meta-analytic effect sizes, the "meta-meta-analytic" mean *r* was .199. Eisend and Tarrahi (2016, p. 528) characterize this result as follows: "The meta-meta-analytic effect size of .2 shows that … advertising is effective."

But this misrepresents the obtained grand-mean effect size. That effect size represents the average effect size across a number of different specific contrasts: celebrity vs. non-celebrity source, comparative vs. noncomparative advertising, greater vs. lesser exposure, and so on. For each such contrast, the associated mean effect size represents the average difference between the two conditions. But from knowing the average difference between two conditions, one cannot tell what the absolute value is of either; knowing the size of the average difference between (say) celebrity and non-celebrity sources does not provide information about the absolute effectiveness of either kind of source. And computing a grand average across many such comparisons indicates the average size of the difference associated with the variations being reviewed—but that grand average does not, and cannot, indicate the absolute effectiveness of advertising.

Another confusion of effect sizes with message effectiveness is apparent in Hornik, Ofir, and Rachamim's (2016) review. This review examined effect sizes from studies of various kinds of advertising appeals, with an interest in both persuasion outcomes and attitude toward the ad (A-ad) outcomes. Two broad appeal categories were distinguished: "emotional appeals" (represented by variations in sex appeals, humor appeals, and fear appeals) and "rational appeals" (represented by variations in comparative appeals, gain-framed appeals, and two-sided appeals). Mean effect sizes were computed for each of the six specific appeal types (that is, for the sexual-ad-vs.-non-sexual-ad contrast, the humorous-ad-versus-non-humorous ad contrast, and so on) and for the two broader categories. The reported effect sizes, expressed as correlations, ranged from .09 (the effect of one-sided vs. two-sided ads on ad liking) to .46 (the effect of sexual vs. non-sexual ads on persuasion).

One overall conclusion drawn by Hornik et al. was this: "In general, the effectiveness of appeals might be substantially less than many managers tend to believe … ranging from .09 to .46" (p. 216). But this is a mistake. The reported effect sizes do not indicate the absolute effectiveness of advertising generally or of any one kind of ad. The effect sizes represent the *difference* in effectiveness between two kinds of ads. And knowing the size of the difference in effectiveness between two types of ads does not provide information about the absolute effectiveness of any one type of ad.

In short, in message effects research, individual effect sizes are prone to being misunderstood as representing the absolute level of one treatment condition on the outcome variable (e.g., the absolute effectiveness of a given message or message type).

Comparing two effect sizes

When comparing two effect sizes, whether from a primary research study or a meta-analysis, related misunderstandings can arise. These can occur when comparing two effect sizes for different independent variables with a given dependent variable, comparing two effect sizes for one independent variable but different dependent variables, or comparing two effect sizes obtained at different times.

Different independent variables

When comparing two effect sizes for different independent variables with a given dependent variable, one mistake that is invited is to suppose that the comparison will indicate the relative effectiveness of the messages. For example, Hornik et al.'s (2016) review of "emotional" and "rational" advertising appeals reported a "comparison of the effect size of the three emotional appeals with that of the three rational appeals" as being "r = .31 and r = .33 compared to r = .14 and r = .17 for persuasion and A-ad, respectively." And they interpreted this as follows: "This implies, all else being equal, that consumers respond to emotional appeals more favorably than to rational ones" (p. 208). But this interpretation is mistaken: The reported effect sizes do not show that consumers respond more favorably to emotional appeals than to rational ones. The reported effect sizes show that there is a larger difference between the two message types being contrasted in the set of emotional appeals than there is between the two message types being contrasted in the set of rational appeals—but those effect sizes do not speak to the question of whether emotional appeals are more effective than rational appeals.

Different dependent variables

When comparing two effect sizes for one independent variable but different dependent variables, one misunderstanding that is invited is to suppose that the relative sizes of the two effect sizes indicate something about the relationship between the two dependent variables. This was exemplified in Rhodes and Dickau's (2012) meta-analytic review of studies of interventions concerning physical activity behavior. The inclusion criteria for this meta-analysis required, inter alia, that (a) the study contain a comparison between a treatment condition and a control condition, (b) the treatment condition produce a meaningful difference in intention between the treatment and control conditions (specifically, d > .19), and (c) a measure of behavior be taken after the intention measure. Across the 11 included studies, the mean effect size for intention (d = .45) was significantly larger than the mean effect for behavior (d = .15).

Rhodes and Dickau (2012) concluded that "these results demonstrate a weak relationship between intention and behavior" in the domain of physical activity, one that "may be below meaningful/practical value" (p. 724). The reasoning that lies behind such conclusions is presumably that because the mean effect size for behavior outcomes was smaller than the mean effect size for intention outcomes, intentions and behaviors must not have been strongly correlated. That is, the shrinkage in the mean effect size was interpreted as reflecting a weak intention-behavior correlation.

But such reasoning is faulty. The relative size of the two mean effect sizes does not indicate whether intentions and behaviors were strongly or weakly correlated. In a given study, even if the effect size for intention is larger than the effect size for behavior, intentions and behaviors might nevertheless be very strongly positively correlated. In fact, if the effect size for intention is larger than the effect size for behavior, any number of different intention-behavior correlations are possible.

To see this abstractly, consider that in a given experiment, the effect size (d) for an outcome depends on the means and standard deviations for each condition. The means and standard deviations—and hence the effect size—are unaffected by which participant in a given condition has which score. (For example, if the intention scores for two participants in the treatment condition are accidentally swapped, the mean and standard deviation will be unaffected, as will the effect size.) But the correlation between two outcomes will vary considerably depending on which participants have which scores. The plain implication is that one cannot deduce the correlation between two outcomes by comparing their effect sizes.[2]

A similar misapprehension concerning comparison of effect sizes for different outcomes was evinced in Gallagher and Updegraff's (2012) message framing meta-analysis. For health-related prevention messages, a statistically significant framing effect was found when behavioral outcomes were assessed but not when attitudinal or intention outcomes were assessed. The comparison of these effect sizes thus led to speculation about the relationship of behavior to its supposed precursors of attitude and intention: "framing effects on the adoption of prevention behaviors may not be completely mediated by the most commonly assessed beliefs used in health message framing studies (attitudes, intentions)" (p. 111).

But, again, the relationships among outcome variables cannot be determined by comparing such effect sizes. Comparison of the effect sizes representing the effects of message variations on attitudinal, intention, and behavioral outcomes cannot speak to the question of the relationship of behaviors to attitudes or intentions. The degree of similarity or dissimilarity among those effect sizes is independent of the degree of correlation among those different outcomes.[3]

Different points in time

When comparing two effect sizes obtained at different times, one mistake that is invited is to suppose that any change in the effect size over time represents a change in message effectiveness over time. To see this in a simple case, imagine a study comparing the persuasiveness of strong and weak threat appeals at two points in time, with the following results. At time-1, there was no significant difference between the two appeals, but at time-2 the strong threat appeal was significantly more persuasive than the weak threat appeal. The temptation might be to suppose that this represents an increase in message effectiveness over time, because there was no significant effect at time-1 but there was a significant one at time-2. However, as perhaps is apparent, there need not have been any such increase. Ceteris paribus, the size of the difference (between the two message conditions) was larger at time-2 than at time-1, but this could occur without any increase in effectiveness over time for any message.

A more complex version of this misunderstanding appeared in Banks et al.'s (1995) study of the effects of message framing variations on mammography. Immediate post-exposure assessments of attitude and intentions did not find any significant effects for framing, but delayed behavioral assessments revealed a significant advantage for the loss-framed appeal. Banks et al. wrote that "It is surprising that we obtained a significant framing effect on mammography in the 12-month bivariate analysis … yet did not obtain condition differences on any of the potential mediating variables" (attitudes and intentions). One of the possible explanations they offered was this: "the loss-framed message may not have had an immediate impact on attitudes, beliefs, or intentions, but

[2] Rhodes and Dickau (2012) might have based their conclusion (that "these results demonstrate a weak relationship between intention and behavior" in the domain of physical activity; p. 724) not on the relationship between the two effect sizes, but simply on the effect size for behavior outcomes. The reported mean effect size for behavior outcomes (d = .15) corresponds to r = .07. And Rhodes and Dickau (2012, p. 726) did offer a specific numerical value for the intention-behavior correlation in the physical activity domain: "Our meta-analysis of experimental evidence places this correlation coefficient at r = .07." But on the basis of the information presented, one cannot in fact tell what the mean intention-behavior correlation coefficient was in the studies that were reviewed—much less that it was specifically .07. It may be that the mean effect size for behavior outcomes (r = .07) was interpreted as if it represented the mean correlation between intention and behavior. It does not. The mean effect size for behavior outcomes represents the average difference between treatment conditions and control conditions on one variable (behavior). That number cannot possibly describe the mean correlation between two variables (intention and behavior).

[3] Another possible misstep here is also worth noticing. Just because there was a significant framing effect for behavior outcomes but not for (for example) attitude outcomes does not necessarily mean that framing affected attitudes and behaviors differently. The fact that one effect is statistically significant and another effect is not does not necessarily mean that the two effects are significantly different from each other. To see whether two effects are significantly different, one needs to compare them directly. For related results, with similar (mis)interpretations, see Williams, Clarke, and Borland (2001). For related discussion, see Gelman and Stern (2006) and Nieuwenhuis, Forstmann, and Wagenmakers (2011).

had a delayed impact on behavior by way of information seeking or cognitive processing after the experimental session was over" (p. 182).

But this contains a confusion concerning the two key findings. The finding of an initial nonsignificant effect of framing on attitudes and intentions does not necessarily mean that the loss-framed message did not have a large initial effect on attitudes or intentions; that finding means that there was no significant *difference* between the loss-framed and gain-framed messages in those effects. Similarly, the finding of a delayed significant effect of framing on behavior does not necessarily mean that the loss-framed message had a large effect on behavior at the delayed assessment; the finding means that there was a significant *difference* between the loss-framed and gain-framed messages in those effects. Both the loss-framed and the gain-framed message could have had a large immediate impact on attitudes and intentions, with each having a smaller delayed impact on behavior—but with that effectiveness decaying more over time for the gain-framed appeal than for the loss-framed appeal.[4]

Summary

To summarize these misunderstandings briefly: In message effects research, an effect size for the effect of a given independent variable on some outcome variable does not contain information about the absolute value of either of the two message conditions being compared. When effect sizes for different independent variables are being compared for a given outcome, the relative magnitude of the effect sizes does not contain information about the relative placement of a condition mean from one effect size and a condition mean from the other effect size. When two effect sizes for a given independent variable are available for different outcome variables, the relative magnitude of the two effect sizes does not contain information about the relationship of the two outcome variables. And when comparing two effect sizes at different points in time, the relative magnitude of the effect sizes does not contain information about changes in message effectiveness over time.

Consequences, causes, and remedies

Consequences

The upshot of these confusions is probably apparent: Researchers make claims that are not supported by the evidence provided. Each of the previous examples represents a case in which the evidence offered did not underwrite the claim advanced. In message effects research, effect sizes speak to questions of relative effectiveness—questions that arise both in primary-research studies (is message A more effective than message B?) and in meta-analytic research (are messages of kind A more effective than messages of kind B?). But effect sizes cannot address questions of the absolute effectiveness of a given message or kind of message. It's one thing to say "message A is more effective than message B," and quite something else to say "message A is effective." When researchers invoke the magnitude of an effect size to support claims about the magnitude of a message's effects, they evince a misunderstanding of effect sizes.

So researchers who fall prey to these misunderstandings make claims that aren't true—or at a minimum are not supported by the evidence offered. And unsuspecting readers who do not fully understand effect sizes are not in a position to detect this slippage between evidence and claim, which permits misunderstandings to metastasize.

It's not easy to say just how widespread these misunderstandings are. There is no plausible automated way to detect such problems, no systematic search procedure to recommend. But it is worth noticing that the examples offered here are not limited to any one journal, any one academic

[4] Additionally, as noted earlier: just because there was a significant framing effect for behavior outcomes but not for (for example) attitude outcomes does not necessarily mean that framing affected attitudes and behaviors differently. To see whether the two effects were significantly different, one would need to compare them directly.

field, any one kind of message effect. On the contrary, they seem to be widespread, at least in those senses. Indeed, this misunderstanding seems sufficiently common that one hesitates to be too critical of those who have fallen prey to it. There is plainly something beguiling about the intersection of "message effects" and "effect size" that somehow invites missteps.

Causes

Why such misunderstanding of effect sizes? The finger naturally points at the language being used. In the context of message effects research, the word "effect," conjoined with the word "size," naturally invites the misunderstanding of "effect size" as "the size of the effect of a message." If the label had not been "effect size" but "difference size" (the size and direction of the *difference* in effect between two message conditions), perhaps some of this confusion might have been avoided. But the language of "effect size" is entirely too well established to imagine ever replacing it.[5]

Remedies

A design solution?

It might be thought that at least some of these confusions arise from the nature of the research designs being used—and hence that appropriate design changes might mitigate against such misunderstandings. Specifically, the absence of a no-treatment (no-message) control condition might be thought to encourage these confusions. Having a no-treatment control condition, this reasoning suggests, would allow for a comparison between a treatment condition and the no-treatment control condition, and the corresponding effect size could straightforwardly be interpreted as an index of treatment effectiveness. If every research design had a no-treatment control condition, this analysis suggests, then other effect sizes might not so easily be confused with those arising from a comparison of a treatment condition and a no-treatment condition.

But there is reason to be skeptical of the ability of such a design solution to prevent the sort of misunderstandings of interest here. It can certainly be tempting to think that, in a treatment-vs.-no-treatment comparison, the associated effect size provides an index of the absolute effectiveness of the treatment. But at least in some senses, it doesn't. That effect size provides a description of the difference between the two conditions being compared, but it does not necessarily describe the absolute effectiveness of the treatment.

To see this concretely, imagine the following hypothetical circumstance. A nonprofit theatre has a subscriber base. Some percentage of subscribers make charitable donations to the theatre each year. The theatre undertakes a randomized trial in which a letter explicitly soliciting donations is sent to some subscribers (the treatment condition) but not others (the no-treatment comparison condition).

Imagine two different scenarios for the results. Scenario #1: The treatment has an donation rate (outcome percentage) of 21%, and the no-treatment comparison has a donation rate of 5%. The effect size (Cohen's h, the difference between the arcsine-transformed proportions; Cohen, 1988) is h = .50. Scenario #2: The treatment has a donation rate of 30%, and the no-treatment comparison has a donation rate of 20%. The effect size is h = .23.

In which scenario is the treatment more effective? Well, in relative terms—relative to the comparison condition—the treatment is more effective in scenario #1, because the effect size is larger. (There's a bigger increase in the donation rate in scenario #1 than in scenario #2.) But in absolute terms, the treatment is more effective in scenario #2, because the treatment has a higher donation rate in that scenario.

The point is this: Even with a no-treatment comparison condition, the effect size based on comparing the treatment condition and the no-treatment condition does not necessarily provide

[5] Moreover, outside the context of message effects research, the terminology of "effect size" is not necessarily so problematic; see, e.g., Kelley and Preacher (2012).

an index of the absolute effectiveness of the treatment.[6] And the implication, thus, is that including no-treatment control conditions in message effects designs will not necessarily prevent confusions of the sort cataloged here.

Better interpretive practices

With the terminology of "effect size" already in place, and with design variations unlikely to mitigate the problem, the most likely avenue for minimizing misunderstanding presumably will lie in how one *interprets* effect sizes. For avoiding the sorts of confusions identified here, two mental devices recommend themselves as possible preventive measures.

The first is this: Whenever the phrase "effect size" is encountered in the context of message effects research, mentally substitute some appropriate version of the phrase "difference between conditions." Any misapprehensions of effect sizes will almost certainly become more visible.

Here's an example sentence that embodies a misunderstanding: "The comparison of the two message conditions yielded a small effect size, $d = .10$ ($r = .05$), suggesting that the messages were not very effective." The reformulated version is: "The comparison of the two message conditions yielded a small difference between the two conditions, $d = .10$ ($r = .05$), suggesting that the messages were not very effective." The reasoning mistake seems rather more apparent in the reformulation.

As another example: "The augmented intervention was more effective than the standard intervention both at time-1 and at time-2; the effect size was larger at time-2 than at time-1, indicating that the augmented intervention became more effective over time." But the misunderstanding here is perhaps more easily seen in a reformulated version: "The augmented intervention was more effective than the standard intervention both at time-1 and at time-2; the difference between the two interventions was larger at time-2 than at time-1, indicating that the augmented intervention became more effective over time."

A second possible device is this: When the phrase "effect size" is encountered in the context of message effects research, mentally substitute some appropriate version of the phrase "the effect of varying X." In message effects research, effect sizes characteristically describe the effects of variables, not individual messages, and thus mentally underscoring that may be helpful.

Consider, for example, this misleading statement: "The effect size for credibility was small, indicating that high credibility sources were not especially persuasive." The reformulated version is: "The effect of varying credibility was small, indicating that high credibility sources were not especially persuasive"—where the misstep is perhaps more apparent.

These mental transformations can be useful to consumers of research, to ensure that they are not entangled in confusions unwittingly sown by researchers. But these devices can also be of value to researchers who want to avoid mischaracterizing their results. In fact, as appropriate, research reports might sometimes want to avoid the language of "effect size" in favor of more transparent descriptions.

These suggested remedies are imperfect, if only because they require greater attentiveness both from researchers describing research results and from research consumers reading research reports. But because the language of effect size can so easily lead to confused thinking about message effects, only greater awareness of the potential problems is likely to prevent them from arising.

Conclusion

Increased attention to issues of effect size estimation has been a welcome development in many social-scientific domains. But in the context of message effects research, an effect size is easily confused with the size of the effect of a message. Avoiding this misunderstanding will require vigilance from both those who conduct research and those who consume it.

[6] Perhaps the potential mistake here is to think that the value on the outcome variable in the no-treatment condition will inevitably always be literally zero, and hence the effect size—the difference between the value on the outcome variable for the treatment condition and the value on the outcome variable for the no-treatment condition—will reflect the absolute effectiveness of the treatment condition. But the value on the outcome variable in the no-treatment condition is not inevitably zero.

References

Ashford, S., Edmunds, J., & French, D. P. (2010). What is the best way to change self-efficacy to promote lifestyle and recreational physical activity? A systematic review with meta-analysis. *British Journal of Health Psychology*, *15*, 265–288. doi:10.1348/135910709X461752

Banks, S. M., Salovey, P., Greener, S., Rothman, A. J., Moyer, A., Beauvais, J., & Epel, E. (1995). The effects of message framing on mammography utilization. *Health Psychology*, *14*, 178–184. doi:10.1037/0278-6133.14.2.178

Boster, F. J., & Mongeau, P. (1984). Fear-arousing persuasive messages. *Annals of the International Communication Association*, *8*, 330–375. doi:10.1080/23808985.1984.11678581

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

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. doi:10.1177/0956797613504966

de Hoog, N., Stroebe, W., & de Wit, J. (2007). The impact of vulnerability to and severity of a health risk on processing and acceptance of fear-arousing communications: A meta-analysis. *Review of General Psychology*, *11*, 258–285. doi:10.1037/1089-2680.11.3.258

Eisend, M., & Tarrahi, F. (2016). The effectiveness of advertising: A meta-meta-analysis of advertising inputs and outcomes. *Journal of Advertising*, *45*, 519–531. doi:10.1080/00913367.2016.1185981

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. doi:10.1007/s12160-011-9308-7

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*, 328–331. doi:10.1198/000313006X152649

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). New York, NY: Routledge.

Hornik, J., Ofir, C., & Rachamim, M. (2016). Quantitative evaluation of persuasive appeals using comparative meta-analysis. *The Communication Review*, *19*, 192–222. doi:10.1080/10714421.2016.1195204

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*, 137–152. doi:10.1037/a0028086

Kline, R. B. (2013). *Beyond significance testing: Reforming data analysis methods in behavioral research* (2nd ed.). Washington, DC: American Psychological Association.

Latimer, A. E., Salovey, P., & Rothman, A. J. (2007). The effectiveness of gain-framed messages for encouraging disease prevention behavior: Is all hope lost? *Journal of Health Communication*, *12*, 645–649. doi:10.1080/10810730701619695

Levine, T. R. (2013). A defense of publishing nonsignificant (ns) results. *Communication Research Reports*, *30*, 270–274. doi:10.1080/08824096.2013.806261

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. doi:10.1111/j.1468-2958.2008.00317.x

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming (2014). *Psychological Science*, *25*, 1289–1290. doi:10.1177/0956797614525969

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*, 1105–1107. doi:10.1038/nn.2886

Rhodes, R. E., & Dickau, L. (2012). Experimental evidence for the intention-behavior relationship in the physical activity domain: A meta-analysis. *Health Psychology*, *31*, 724–727. doi:10.1037/a0027290

Rosnow, R. L., & Rosenthal, R. (2009). Effect sizes: Why, when, and how to use them. *Zeitschrift Fur Psychologie*, *217*, 6–14. doi:10.1027/0044-3409.217.1.6

Tannenbaum, M. B., Hepler, J., Zimmerman, R. S., Saul, L., Jacobs, S., Wilson, K., & Albarracín, D. (2015). Appealing to fear: A meta-analysis of fear appeal effectiveness and theories. *Psychological Bulletin*, *141*, 1178–1204. doi:10.1037/a0039729

Williams, T., Clarke, V., & Borland, R. (2001). Effects of message framing on breast-cancer-related beliefs and behaviors: The role of mediating factors. *Journal of Applied Social Psychology*, *31*, 925–950. doi:10.1111/j.1559-1816.2001.tb02656.x

Witte, K., & Allen, M. (2000). A meta-analysis of fear appeals: Implications for effective public health programs. *Health Education and Behavior*, *27*, 591–615. doi:10.1177/109019810002700506

Yzer, M. C., Southwell, B. G., & Stephenson, M. T. (2013). Inducing fear as a public communication campaign strategy. In R. E. Rice & C. K. Atkin (Eds.), *Public communication campaigns* (4th ed., pp. 163–176). Los Angeles, CA: Sage.