Document Design 1(2), 87–97

DANIEL J. O'KEEFE

Variability of persuasive message effects

Meta-analytic evidence and implications

Keywords: message effects, persuasion, variability of message effects

This paper reports new information concerning the variability of the persuasive effects of variables across messages. Evidence from meta-analytic reviews of persuasive effects research indicates that such variability is common and substantial, even under well-specified experimental conditions. The implications of this evidence for the design, analysis, and interpretation of research on persuasive message effects are discussed.

Theorists, researchers, and message designers have a continuing interest in identifying factors influencing the success of persuasive messages. The classic form of evidence for addressing such questions has been the single-message experimental design, in which several versions of a given message are prepared, varying specifically with respect to the message property of interest.

But questions have been raised about the adequacy of this research format for providing evidence supporting cross-message generalizations about message effects (see especially Jackson & Jacobs, 1983; for some subsequent discussion, see Hunter, Hamilton, & Allen, 1989; Jackson, 1992; Jackson, O'Keefe, & Brashers, 1994; Jackson, O'Keefe, & Jacobs, 1988; Jackson, O'Keefe, Jacobs, & Brashers, 1989; Morley, 1988; Slater, 1991). These concerns might be characterized most broadly as connected with the possibility of variability in the effect of a given treatment across different messages. Obviously, to the extent that the effect of a given treatment varies substantially from message to message, any single-message design will be correspondingly weakened as a basis for generalization.

The extent of such variability is an empirical question, however. This paper offers new evidence about the extent of variability in persuasive message effects, drawn from meta-analytic research reviews. It then discusses the implications of this evidence for the design, analysis, and interpretation of research on persuasive message effects.

Evidence of variability in persuasive message effects

The idea of message-to-message variability in persuasive effects can be expressed straightforwardly. For investigating the question of the relative persuasive effectiveness of messages that vary with respect to some property (fear appeal level, message sidedness, etc.), the most common way of gathering empirical evidence involves an experiment comparing the effects of two (or more) versions of a given message, where the versions differ specifically with respect to the variable of interest. This experimental comparison (e.g., between a high-fear-appeal message and its low-fear-appeal counterpart) will yield some observed difference in persuasiveness between conditions; the amount and direction of this difference — the 'effect size' — can be expressed in various intertranslatable forms such as a standardized mean

difference (d) or a correlation (r). Replications of this investigation (other studies of that same variable's persuasive effects) will provide additional effect sizes obtained using different pairs of messages. This collection of effect sizes will have some mean (some average observed effect size); it will also almost certainly have some variance, that is, there will be some variability in the observed effect from one message (that is, one message pair) to another.

The amount of variability in a collection of effect sizes can be expressed in several ways. One familiar index is the standard deviation, computed and understood in the usual way (see, e.g., Shadish & Haddock, 1994, p. 274). Such an index is plainly informative, even if not perhaps quite as transparent as might be wanted.

Birge's (1932) ratio R provides another index of variability. Birge's R has been commonly used in the physical sciences for assessing the consistency of observations. In any set of observations of a given property (e.g., a set of effect sizes), some variability is expected if only because of sampling variation (that is, some variability is expected because each effect size is based on a sample of respondents). Birge's R (which has a lower limit of zero, with no upper limit) is one when a set of observations (effect sizes) is homogeneous except for sampling variability (that is, when the observations display as much variability as might be expected given sampling error). Values above one indicate more variability (and values below one less variability) than might be expected given sampling error. R thus provides a direct index of the magnitude of heterogeneity (variability) in a collection of effect sizes; as the variability increases, R increases.

As pointed out by Hedges (1987), Birge's ratio is related to Hedges and Olkin's (1986, pp. 123, 235) Q, a common meta-analytic statistical test of the homogeneity in a collection of effect sizes. Q provides a test value suitable for assessing the statistical significance of the amount of observed variability; R provides an index of the amount of variability. Thus R and Q are related in much the same fashion as an effect-size index (such as d or r) is related to the value of a statistical test (such as t or F); for instance, a highly-significant Q does not mean that a large amount of heterogeneity is present (any more than a highly-significant t means that a large effect size is present).

The present paper reports both the standard deviation and the Birge's ratio for effects associated with a number of different social-influence variables. These variability indices were computed over the effect sizes reported in extant meta-analytic reviews of social-influence communication effects research. Meta-analysis is a family of procedures for producing a systematic quantitative summary of a set of research studies (for a general introduction, see Rosenthal, 1991). In a meta-analysis, an effect size — a measure of the magnitude of the effect of the variable under investigation — is obtained from each relevant study, and these are combined to yield an average effect (with an affiliated confidence interval). Where a potential moderating factor varies between studies, the set of effect sizes can be subdivided based on levels of the moderator; the mean effect sizes within these subgroups can then be computed and compared.

Meta-analytic reviews, by virtue of collecting effect sizes associated with a given variable, provide a natural source of information about effect-size variability. Such information is not commonly extracted or reported in meta-analytic reviews, because the focus of attention is the mean effect size (either overall, or at different levels of potential moderator variables). But plainly the collected effect sizes can also be examined for the light they can shed on the question of variability in effect.

However, in three ways, the data in these meta-analytic reviews offer a rather conservative basis for the assessment of message-to-message variability. First, it occasionally happens that experimental materials are used in more than one reported study; for example, in research on the door-in-the-face strategy, the same messages were used by Patch (1986) and Dillard and Hale (1992). Many meta-analytic reviews use the study (not the message) as the unit of analysis, and when experimental messages are re-used such a procedure probably underestimates the amount of message-to-message variability: ceteris paribus, within a given collection of effect sizes, as the number of studies using the same message increases, the study-to-study variability likely decreases.

Second, some studies use more than one message pair but fail to report separate results for each pair (e.g., Struckman-Johnson & Struckman-Johnson, 1996). Meta-analytic treatments of such a study commonly simply obtain one summary effect size, even though the effects involve several distinct messages. In such a circumstance, the amount of message-to-message variability will be underestimated; the computed effect size from such a study will conceal any message-to-message variability within the single cross-message composite effect.

Third, when a primary research report indicates an effect is statistically 'nonsignificant' (and no further effect-size information is available), sometimes meta-analysts treat the effect as equivalent to 'no effect' (e.g., $r = .00$), which leads to an underestimation of the variability in effect sizes (Pigott, 1994, pp. 167–168; Ray & Shadish, 1996, p. 1318). This practice is uncommon, but appears to have been followed in at least some of the meta-analyses reviewed here (certainly by Kuhberger, 1998, p. 35; and perhaps by, e.g., Gayle, Preiss, & Allen, 1998, and by Witte & Allen, 1996).

Method

Meta-analytic reviews of social-influence communication effects research were located; included were reviews of studies of particular variables (e.g., message sidedness) and reviews of studies of particular influence domains (e.g., messages aimed at inducing drug abuse resistance). To be included, a review had to report individual effect sizes and study *n*s for investigations examining persuasive-outcome effects (attitude change, behavioral compliance, and the like). Excluded were reviews reporting only summary meta-analytic results without specifying the individual effect sizes and *n*s on which the review was based (e.g., Bauman, 1997; Cox, Wogalter, Stokes, & Murff, 1997; Grewal, Kavanoor, Fern, Costley, & Barnes, 1997; Mullen et al., 1997) and reviews or results concerning other effects (effects on credibility perceptions were excluded, for instance). When multiple suitable meta-analyses were available for a given topic, more comprehensive and recent publications were preferred.

For each review, the reported individual effect sizes were converted to *r* (if not already given in that form). A mean *r* (specifically, the *n*-weighted mean *r*, computed using the *r*-*z*-*r* transformation procedure), the standard deviation (specifically, the ordinary unweighted sample estimate of the standard deviation of the effect sizes; see Shadish & Haddock, 1994, p. 274, eq. 18–20), and the Birge's R (obtained from computation of Hedges and Olkin's Q; see Hedges, 1987, p. 446; Hedges & Olkin, 1986, p. 235; Shadish & Haddock, 1994, p. 266) were then computed for each review's set of effect sizes.

Results

Initial results

A total of 23 suitable meta-analyses were located, covering a variety of topics (message sidedness, language intensity, the use of statistical vs. narrative evidence, the foot-in-the-door influence strategy, the induction of resistance-to-persuasion concerning drug use, and so on). Table 1 provides summary information concerning each meta-analysis. It is plain from Table 1 that substantial variability in treatment effects is quite common. Indeed, it is rare for the standard deviation to be smaller than .12, rare for the standard deviation to be smaller than the mean, and rare for R to be less than 2. The *k*-weighted average of the standard deviations in Table 1 is .171, and of the Rs is 3.83; the corresponding simple (unweighted) averages are .148 and 3.04.

Subsets of effects

One possible reaction to these initial results might be to think that the apparent variability could be made to disappear if finer subdivisions of effect sizes were to be employed. In fact, some meta-analysts have pursued the policy that, when a set of significantly heterogeneous effect sizes is found, one aims to divide the cases into subcategories that will contain homogeneous groups of effect sizes (e.g., Dillard et al., 1984). Thus it might be thought that still-finer (and more) subdivisions of the effect sizes would eventually eliminate the variability. It is certainly true that one could

produce such subdivisions, and thereby find apparently-homogeneous (i.e., not significantly heterogeneous) subsets of effect sizes. But in the present context, this reaction would be misguided, for two reasons.

First, subdividing a collection of effect sizes until the variability becomes nonsignificant can be misleading, because the absence of apparent variability can reflect the small number of cases within a given subcategory. A small number of cases provides low power for detecting heterogeneity. In principle it will always be possible to make the variability seem to disappear (that is, become nonsignificant) through appropriate subdivision, but this would not be good evidence for a lack of intrinsic variability within subcategories.

Second, the fact that a collection of effect sizes can be divided in a fashion that minimizes variability within subcategories does not erase the fact of variability in the larger collection of effect sizes. The present results make it quite plain that, across a number of persuasion-effects factors, there is substantial variability in effect from implementation to implementation. Any lack of variability within even finer subcategories could not refute the existence of variability within superordinate categories. And the contrasts embodied in these superordinate categories — one-sided versus two-sided messages, statistical versus narrative evidence, and so forth — are of substantive interest to students of persuasion (analysts, theorists, message designers, and so forth), and hence the existence of variability in effect within such categories is itself of some importance.

Table 1 Variability in Persuasion Effects: Overall Analyses

Research domain	k	mean r	SD	R
Communication medium				
audio-visual vs. audio	7	.031	.120	1.06
audio-visual vs. print	8	−.005	.129	2.56
audio vs. print	8	−.011	.071	0.86
(Boster & Levine, 1997)				
Conclusion explicitness	32	.138	.153	6.37
(O'Keefe, 1997)				
Delayed communicator identification	10	.040	.391	5.44
(O'Keefe, 1987)				
Door-in-the-face strategy	88	.083	.220	3.58
(O'Keefe & Hale, 1998)				
Drug abuse resistance education				
effects on drug attitudes	8	.051	.053	2.17
effects on drug use	5	.022	.023	0.62
(Ennett, Tobler, Ringwalt, & Flewelling, 1994)				
Fear appeals				
effects on attitude	34	.146	.161	2.71
effects on intention	41	.131	.167	6.34
effects on behavior	27	.164	.206	7.09
(Witte & Allen, 1996)				
Foot-in-the-door strategy	33	.106	.149	2.56
(Dillard, Hunter, & Burgoon, 1984)				
Forewarning	12	.174	.089	0.79
(Benoit, 1998)				
Guilt appeals	5	−.256	.152	2.20
(O'Keefe, in press)				
Language intensity	15	.017	.129	3.25
(Hamilton & Hunter, 1998)				
Message framing	13	.018	.094	1.03
(Kuhberger, 1998)				
One-sided vs. two-sided messages	107	−.000	.192	4.66
(O'Keefe, 1999)				
Powerful vs. powerless language	5	.230	.143	1.99
(Burrell & Koper, 1998)				
Rhetorical questions	18	.054	.138	1.54
(Gayle, Preiss, & Allen, 1998)				
Statistical vs. narrative evidence	16	.073	.204	3.89
(Allen & Preiss, 1997)				
Support explicitness				
information-source citation	23	.072	.129	2.74
argument completeness	27	.113	.166	2.89
quantitative specificity	8	.066	.118	3.49
(O'Keefe, 1998)				

Note: k = number of effect-size cases; mean r = average effect size (mean correlation); SD = standard deviation of the effect sizes; R = Birge's ratio.

Even so, it may be worth examining the variability to be found within subsets of the effect sizes reviewed in these meta-analyses. Several of the meta-analyses report appropriate information for examining the variability to be found within levels of proposed moderators. As these meta-analyses sometimes consider a large number of potential moderators, reproduction of all the relevant analyses is not possible here. But Table 2 provides a sampling of some of the results obtained when effect sizes are segregated by levels of potential moderator variables. (Some meta-analyses examined moderator-variable effects but did not report how individual studies were coded with respect to the moderator, thus preventing such analyses from being included here; e.g., Dillard et al., 1984.)

Table 2 Variability in Persuasion Effects: Moderator-Variable Analyses

Research domain	k	mean r	SD	R
Conclusion explicitness				
conclusion explicitness (overall)	32	.138	.153	6.37
conclusion omission	14	.102	.162	4.16
conclusion specificity	18	.147	.149	8.16
(O'Keefe, 1997)				
Delayed communicator identification				
all sources	10	.040	.391	5.44
low-credibility sources	5	.166	.453	5.92
high-credibility sources	5	−.098	.316	3.84
(O'Keefe, 1987)				
Door-in-the-face strategy				
overall	88	.083	.220	3.58
optimal moderator values	45	.156	.240	3.73
suboptimal moderator values	43	.027	.184	2.75
(O'Keefe & Hale, 1998)				
Forewarning				
all kinds	12	.174	.089	0.79
topic-position warnings	8	.161	.093	0.72
persuasive-intent warnings	4	.193	.077	1.11
(Benoit, 1998)				
Message framing				
all cases	13	.018	.094	1.03
single risky event	10	−.005	.091	0.59
multiple risky events	3	.101	.103	1.67
(Kuhberger, 1998)				
One-sided vs. two-sided messages				
one-sided vs. two-sided (overall)	107	−.000	.192	4.66
one-sided vs. refutational two-sided	42	.067	.170	2.50
one-sided vs. nonrefutational two-sided	65	−.030	.191	5.50
(O'Keefe, 1999)				
Rhetorical questions				
overall	18	.054	.138	1.54
indirect	7	.129	.118	1.30
direct	8	−.035	.094	0.59
mixed	3	.088	.126	2.17
(Gayle, Preiss, & Allen, 1998)				

Note: k = number of effect-size cases; mean r = average effect size (mean correlation); SD = standard deviation of the effect sizes; R = Birge's ratio.

As indicated in Table 2, even when effect sizes are broken out into subsets on the basis of some moderator variable, substantial variability is still possible. In the 15 subsets of effect sizes reported in Table 2, the k-weighted average of the standard deviations is .129, and of the Rs is 3.73; the corresponding simple (unweighted) averages are .153 and 2.98. These means are quite similar to those observed in the unpartitioned sets of effect sizes.

The variability observed in the subcategories of the door-in-the-face (DITF) influence strategy is especially noteworthy. O'Keefe and Hale (1998) identified five variables that appear to moderate the size of DITF effects, and then classified cases by whether the cases were ones in which all five moderator variables had optimal values (i.e., the conditions were such as to maximize DITF effects) as opposed to cases in which at least one of the five moderators had a suboptimal value. As indicated in Table 2, even under the narrowly-specified set of optimal conditions (involving the nature of the topic and communicator, the medium of communication, and so forth), there was considerable variability in effects (SD = .240, R = 3.73).

In short, there is commonly substantial variability in

persuasive effects, even when moderator variables are used to create subclasses of effect sizes. Moreover, as mentioned previously, all these estimates of variability (both the overall estimates and the within-subcategory estimates) are conservative. It seems plain that message-to-message variability in persuasive effect is genuine and common.

Implications of variability in effects

Variability and generalization about effects

In considering the implications of this observed variability in effects, it may be useful to say at the outset that such variability does not mean that one cannot draw sound general conclusions about the message factor under investigation. In this connection, there are two notable possible errors in reasoning about the meaning of apparent heterogeneity (or homogeneity) within a set of effect sizes.

First, it is a mistake to think that if a set of effect sizes is significantly heterogeneous (i.e., contains statistically significant variability), then no legitimate sound generalization about the mean effect may be drawn. These two matters are not intrinsically related, and are appropriately addressed through two different statistical tests — a test of the significance of the observed variability (heterogeneity) and a test of the significance of the observed mean effect. It is possible for a significantly-heterogeneous set of effect sizes to have a mean effect size that is dependably different from zero. For example, refutational two-sided messages are dependably more persuasive than their one-sided counterparts although the relevant set of effect sizes is significantly heterogeneous (O'Keefe, in press-b). In such a circumstance, a useful generalization is plainly appropriate even given the evidence of heterogeneity.

Second, it is a mistake to think that if a set of effect sizes is apparently homogeneous (i.e., is not significantly heterogeneous), then the observed mean effect size is the population effect. Such reasoning can lead analysts to incorrectly treat mean effects as dependable in the absence of a significance test's having been performed (that is, even though there is no evidence that the mean effect is in fact dependably different from zero). For example, having found homogenous sets of effect sizes, Allen (1991, p. 396) treated the corresponding means as indicating genuine effects even though no significance-test information was reported.

But, again, such reasoning fails to appreciate the difference between the results of a test for the significance of the observed variability and the results of a test for the significance of the observed mean effect. It is possible for a set of effect sizes to be apparently homogeneous (that is, not significantly heterogeneous) but for the observed mean effect size to not be significantly different from zero. For example, messages with direct rhetorical questions are not significantly more persuasive than messages lacking such questions, even though the set of effect sizes is homogeneous (that is, not significantly heterogeneous; see Gayle, Preiss, and Allen, 1998).[1]

In short, there is a distinction to be appreciated between the variability in a set of effect sizes and the significance of the mean effect size: a collection of effect sizes could have substantial variability with or without a dependable (statistically significant) mean effect, or could have little variability with or without a dependable mean effect. In the present context, the point to be noticed is that substantial variability in effect, as observed in the present results, does not necessarily imply any impaired ability to reach reliable generalizations about mean effects.

The importance of replications

These results plainly indicate the importance of examining replications (multiple messages) when one is interested in generalizing across messages about persuasive effects. As Jackson and Jacobs (1983) noted, if message effects are uniform from implementation to implementation (message to message), then — so far as generalizing across implementations is concerned — one implementation is as good as another for estimating the mean effect. But if there is substantial variability across implementations, then a good

estimate of the mean effect will require evidence from multiple implementations.

The meta-analytic investigations reviewed above indicate that substantial variability is the norm, not the exception, in persuasion effects research. Given such evidence, results from single-message designs should presumptively be treated as a dubious basis for generalization across messages. Multiple-message designs ought to be the preferred form of primary research evidence; when a single-message design is used, the interpretation of results should be tempered accordingly.

An investigator planning a replicated (multiple-message) design will need to decide how many message replications to include so as to have reasonable statistical power (the likelihood of finding a statistically significant effect given that some genuine effect exists). The power of a replicated design depends upon (inter alia) the variability of effect in the population, the size of the population effect, the number of participants, and the number of replications in the design. It might be feared that the variability observed here is so substantial as to make multiple-message designs unrealistic, because of the potentially large number of replications required for adequate power.

The average variability in the meta-analyses reviewed here (as given by the k-weighted average of the standard deviations from Table 1) is .17. Information provided by Jackson and Brashers (1994) suggests that variability of this magnitude can permit rather good power (in excess of .70 with .05 alpha), even with as few as 10 message replications (across 400 respondents), so long as the population effect size is sufficiently large (in the neighborhood of $r = .20$). Where the population effect size is small, however, it can be difficult to obtain substantial power, even with much smaller variability than observed here. For example, with a population mean r of .08 (the k-weighted average of the absolute values of the mean rs in Table 1) and a population standard deviation of only .07, using 25 replications with 400 participants will yield power of approximately .40 (at .05 alpha); with the same population mean r and a population standard deviation of .17, the power is no better than .32.[2]

In short, the variability observed here is sufficiently large to make one skeptical of single-message designs as a basis for generalization, but not so large as to suggest intrinsically unrealistic requirements for multiple-message designs. The central barrier to adequate power in experimental research on persuasive effects appears not to be the existence of message-to-message variability in effect, but potentially small population effects.

Of course, multiple-message designs can sometimes face practical challenges beyond considerations of statistical power. In some research settings, it may not be feasible or appropriate to implement a replicated (multiple-message) design. In particular, research undertaken in applied settings is often focussed on the effects of particular messages not as representatives of any broader message category but simply as objects of interest in their own right. For example, an advertiser may wish to know which of two specific advertisements is more persuasive, or a manufacturer may wish to learn which of several product warning labels is most effective; such research questions obviously do not require replicated designs. Applied settings can be attractive research venues (e.g., because of their realism), but may provide evidence of limited utility insofar as generalization is concerned. Researchers interested in generalization should not mindlessly forego research opportunities in such settings, but neither should they ignore the potential tradeoffs involved.

Analyzing replications

These results also have implications for the analysis of data based on replications. In particular, the results underscore the appropriateness of random-effects (as opposed to fixed-effects) analysis of replications, both in primary research and in meta-analytic research. These two research applications are usefully discussed separately.

There has been substantial discussion of the choice between fixed-effects and random-effects (or 'mixed-model') analyses of replicated designs in primary research on communication effects (e.g., Burgoon, Hall, & Pfau, 1991; Hunt-

er, Hamilton, & Allen, 1989; Jackson, 1992; Jackson, Brashers, & Massey, 1992; Jackson, O'Keefe, Jacobs, & Brashers, 1989). These alternative statistical analyses differ in how they handle the presence of message replications in an experimental design. In a random-effects analysis (one that treats the message replications as a random factor), the particular messages studied are, like respondents (subjects), seen to be a source of potential error in estimating the treatment effect (the effect of the independent variable of interest); thus in assessing the treatment effect, the statistical computations in a random-effects analysis take into account message-to-message variability. In a fixed-effects analysis (one that treats the message replications as fixed), the assumption is that replications are not a source of error in the estimation of the treatment effect. These different assumptions lead to differences in the statistical assessment of hypotheses concerning the treatment effect (for extensive and careful discussion, see Jackson, 1992).

Perhaps the most simple and compelling argument underwriting the use of random-effects analyses is that only random-effects analyses test the hypotheses that usually are of interest to investigators. It is only occasionally the case that investigators are interested in the specific messages under study; more commonly, the messages under investigation are simply representatives of a larger class of messages to which generalization is desired. But a fixed-effect analysis does not underwrite such generalization, because its analyses do not take into account message-to-message variability;

> such an analysis establishes that two concrete groups of messages each differ from one another, but since any differences between the two concrete groups may reflect nothing more than case-to-case differences occurring even within categories, the observation that the two concrete groups of messages differ does not justify the conclusion that the categories differ (Jackson, 1992, p. 95).

By contrast, because a random-effects analysis takes into account the observed message-to-message variability in estimating the size of the treatment effect, it underwrites generalizations about the message categories involved (as opposed to merely the specific messages studied). As a number of commentators have suggested, such interest in generalization beyond the cases at hand can be an entirely sufficient justification for the use of random-effects analyses (e.g., Jackson, 1992; Raudenbush, 1994).

But the current results further underwrite such a preference, by showing that message-to-message variability is not rare or trivial. If there were actually little message-to-message variability, then single-message designs might suffice as bases for generalization, and any multiple-message designs might be analyzed in ways that ignore such variability. Given the presence of common and substantial variability, however, analyses that take such variability into account — as random-effects analyses do — seem preferable.

The observed variability in effect also provides a rationale for believing that in meta-analytic reviews of persuasion-effects research, random-effects meta-analytic procedures should be the default choice (as opposed to the more common fixed-effects analyses). As in the case of primary-research replicated designs, a preference for random-effects meta-analytic procedures can be justified simply on the basis that only such analyses test the hypothesis of interest (which involves generalizing beyond the cases at hand; see Hedges & Vevea, 1998). But the present results provide further indication of the appropriateness of random-effects procedures. This can be illustrated concretely by considering three alternative procedures for constructing a confidence interval around some mean effect size (some mean correlation, for instance).

In a procedure described by Hunter and Schmidt (1990, p. 208), the width of the confidence interval around an observed mean correlation is influenced by the total number of participants (the total N), but not by the number of different studies (effect sizes). That is, given some N, the confidence interval is the same whether the meta-analysis is based on two studies or 200. This surely is prima facie an implausible general procedure, if only because it fails to reflect the presumably greater confidence to which one is entitled as (ceteris paribus) the number of studies grows.

In standard fixed-effects meta-analytic procedures (as described, e.g., by Shadish & Haddock, 1994, pp. 265–273), the width of the confidence interval around a mean effect is affected by the number of effect sizes but not by the variability among the observed effect sizes. The consequence may be seen by envisioning two circumstances, one in which the observed effects all cluster tightly around the mean and another in which the observed effects vary greatly. If one's interest is in estimating the location of the population mean effect — that is, including cases (messages or implementations) not yet studied — then presumably the existence of substantial variability among the observed cases should make one less secure in one's estimate of the location of the mean, and hence should yield a larger confidence interval.

In random-effects meta-analytic procedures (e.g., as described by Shadish and Haddock, 1994, pp. 273–278), the width of the confidence interval is influenced by, inter alia, the observed variability among effect sizes: with greater variability, the confidence interval widens. Given an interest in generalizing beyond the cases at hand, this last procedure surely seems the most appropriate.

Obviously, if there were no substantial variability in effect sizes in persuasive-effects research, then it would be inconsequential (to the width of the confidence interval) whether one's analytic procedures bothered to take into account such variability. But the present results suggest that effect-size variability in this research domain is, in fact, common and substantial — thus indicating the importance of acknowledging such variability in one's analyses.

Message design implications

Message designers might reasonably expect that research on factors influencing persuasive effects will provide some guidance about principles of effective message design. But the results reported here suggest two cautions that might be kept in mind about such principles.

First, a principle of effective persuasive message design, to be well-supported, requires evidence obtained across a number of message replications. The classic single-message design would be a good source of dependable generalizations if message effects were entirely consistent from one message to the next. But message-to-message variability in effects is plainly quite common — which suggests that, as tempting as it might be to draw a generalization from a study of just one message, message designers (like theorists and researchers) should await better evidence.

Second, even with an appropriately supported general principle, message designers should be prepared for variability in effect. That is, even a well-evidenced general principle of message design provides no guarantee about the effects to be found in any specific case. For example, even if the overall mean effect of a given treatment is dependably positive, a substantial number of individual implementations may nevertheless produce negative effects. Well supported generalizations about persuasive message effects can be useful guides to effective message construction, but do not provide ironclad assurances.

Conclusion

There is substantial message-to-message variability in effects concerning persuasive outcomes. It is an empirical question to what extent similar variability obtains in other domains of message effects. But there is no obvious reason to suppose that persuasive message effects will be dramatically different in variability than will effects of educational or informative messages (e.g., effects on learning outcomes or comprehension), social support messages (e.g., effects on feelings of well-being or on health), managerial messages (e.g., effects on subordinate job satisfaction), self-disclosure messages (e.g., effects on liking or perceived trust), and so forth. At a minimum, then, the evidence reviewed here cautions against any easy presumption of uniformity of effects across messages.

Notes

1. Relatedly, some have apparently thought that if a set of effect sizes is homogeneous (i.e., not significantly heterogeneous), then no significant moderator can be at work within the set of effect sizes (e.g., Benoit, 1998, p. 145). But in fact a nonsignificant heterogeneity test does not guarantee the absence of significant moderators (see Cook et al., 1992, pp. 313–314; Hall & Rosenthal, 1991, p. 440).

2. Jackson and Brashers (1994) reported their results in terms of the effect-size index *d* [specifically, Hedges and Olkin's (1986) *d*], not *r*. To facilitate the use of Jackson and Brashers's results, the present data were also analyzed using *d* (by converting the individual study *r*s to *d*s, then conducting the parallel analyses across meta-analyses as were done with *r*). For the cases in Table 1, the simple (unweighted) mean *d* (using absolute values) is .177; the corresponding k-weighted mean is .160. For those cases, the simple (unweighted) standard deviation is .318; the corresponding k-weighted value is .371. Jackson and Brashers's results indicate that for a population effect size corresponding to $d = .18$ and variance of .1 (i.e., a standard deviation of approximately .32), a design with 25 replications and 400 participants will have power of approximately .32 (with .05 alpha); for a population effect size of $d = .18$ and variance of .02 (a standard deviation of approximately .14), a design with 25 replications and 400 participants will have power of approximately .40 (with .05 alpha).

References

Allen, M. (1991). Meta-analysis comparing the persuasiveness of one-sided and two-sided messages. *Western Journal of Speech Communication, 55*, 390–404.

Allen, M., & Preiss, R. W. (1997). Comparing the persuasiveness of narrative and statistical evidence using meta-analysis. *Communication Research Reports, 14*, 125–131.

Bauman, K. E. (1997). The effectiveness of family planning programs evaluated with true experimental designs. *American Journal of Public Health, 87*, 666–669.

Benoit, W. L. (1998). Forewarning and persuasion. In M. Allen & R. W. Preiss (Eds.), *Persuasion: Advances through meta-analysis* (pp. 139–154). Cresskill, NJ: Hampton Press.

Birge, R. T. (1932). The calculation of errors by the method of least squares. *Physical Review, 40* (2nd series), 207–227.

Boster, F. J., & Levine, K. J. (1997, May). The impact of the channel variable on persuasive messages: A meta-analytic review. Paper presented at the annual convention of the International Communication Association, Montreal.

Burgoon, M., Hall, J., & Pfau, M. (1991). A test of the "messages-as-fixed-effect fallacy" argument: Empirical and theoretical implications of design choices. *Communication Quarterly, 39*, 18–34.

Burrell, N. A., & Koper, R. J. (1998). The efficacy of powerful/powerless language on attitudes and source credibility. In M. Allen & R. W. Preiss (Eds.), *Persuasion: Advances through meta-analysis* (pp. 203–215). Cresskill, NJ: Hampton Press.

Cook, T. D., Cooper, H., Cordray, D. S., Hartmann, H., Hedges, L. V., Light, R. J., Louis, T. A., & Mosteller, F. (1992). *Meta-analysis for explanation: A casebook*. New York: Russell Sage Foundation.

Cox, E. P., III, Wogalter, M. S., Stokes, S. L., & Murff, E. J. T. (1997). Do product warnings increase safe behavior? A meta-analysis. *Journal of Public Policy and Marketing, 16*, 195–204.

Dillard, J. P., & Hale, J. L. (1992). Prosocialness and sequential request compliance techniques: Limits to the foot-in-the-door and the door-in-the-face? *Communication Studies, 43*, 220–232.

Dillard, J. P., Hunter, J. E., & Burgoon, M. (1984). Sequential-request persuasive strategies: Meta-analysis of foot-in-the-door and door-in-the-face. *Human Communication Research, 10*, 461–488.

Ennett, S. T., Tobler, N. S., Ringwalt, C. L., & Flewelling, R. L. (1994). How effective is drug abuse resistance education? A meta-analysis of project DARE outcome evaluations. *American Journal of Public Health, 84*, 1394–1401.

Gayle, B. N., Preiss, R. W., & Allen, M. (1998). Another look at the use of rhetorical questions. In M. Allen & R. W. Preiss (Eds.), *Persuasion: Advances through meta-analysis* (pp. 189–201). Cresskill, NJ: Hampton Press.

Grewal, D., Kavanoor, S., Fern, E. F., Costley, C., & Barnes, J. (1997). Comparative versus noncomparative advertising: A meta-analysis. *Journal of Marketing, 61*(4), 1–15.

Hall, J. A., & Rosenthal, R. (1991). Testing for moderator variables in meta-analysis: Issues and methods. *Communication Monographs, 58*, 437–448.

Hamilton, M. A., & Hunter, J. E. (1998). The effect of language intensity on receiver evaluations of message, source, and topic. In M. Allen & R. W. Preiss (Eds.), *Persuasion: Advances through meta-analysis* (pp. 99–138). Cresskill, NJ: Hampton Press.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist, 42*, 443–455.

Hedges, L. V., & Olkin, I. (1986). *Statistical methods for meta-analysis*. New York: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486–504.

Hunter, J. E., Hamilton, M. A., & Allen, M. (1989). The design and analysis of language experiments in communication. *Communication Monographs, 56*, 341–363.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Beverly Hills, CA: Sage.

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

Jackson, S., & Brashers, D. E. (1994). M > 1: Analysis of treatment x replication designs. *Human Communication Research, 20*, 356–389.

Jackson, S., Brashers, D. E., & Massey, J. E. (1992). Statistical testing in treatment by replication designs: Three options reconsidered. *Communication Quarterly, 40*, 211–227.

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

Jackson, S., O'Keefe, D. J., & Brashers, D. (1994). The messages replication factor: Methods tailored to messages as objects of study. *Journalism Quarterly, 71*, 984–996.

Jackson, S., O'Keefe, D. J., & Jacobs, S. (1988). The search for reliable generalizations about messages: A comparison of research strategies. *Human Communication Research, 15*, 127–142.

Jackson, S., O'Keefe, D. J., Jacobs, S., & Brashers, D. E. (1989). Messages as replications: Toward a message-centered design strategy. *Communication Monographs, 56*, 364–384.

Kuhberger, A. (1998). The influence of framing on risky decisions: A meta-analysis. *Organizational Behavior and Human Decision Processes, 75*, 23–55.

Morley, D. D. (1988). Meta-analytic techniques: When generalizing to message populations is not possible. *Human Communication Research, 15*, 112–126.

Mullen, P. D., Simons-Morton, D. G., Ramirez, G., Frankowski, R. F., Green, L. W., & Mains, D. A. (1997). A meta-analysis of trials evaluating patient education and counseling for three groups of preventive health behaviors. *Patient Education and Counseling, 32*, 157–173.

O'Keefe, D. J. (1987). The persuasive effects of delaying identification of high- and low-credibility communicators: A meta-analytic review. *Central States Speech Journal, 38*, 63–72.

O'Keefe, D. J. (1997). Standpoint explicitness and persuasive effect: A meta-analytic review of the effects of varying conclusion articulation in persuasive messages. *Argumentation and Advocacy, 34*, 1–12.

O'Keefe, D. J. (1998). Justification explicitness and persuasive effect: A meta-analytic review of the effects of varying support articulation in persuasive messages. *Argumentation and Advocacy, 35*, 61–75.

O'Keefe, D. J. (in press). Guilt and social influence. In M. E. Roloff (Ed.), *Communication yearbook 23*. Thousand Oaks, CA: Sage.

O'Keefe, D. J. (1999). How to handle opposing arguments in persuasive messages: A meta-analytic review of the effects of one-sided and two-sided messages. In M. E. Roloff (Ed.), *Communication yearbook 22* (pp. 209–249). Thousand Oaks, CA: Sage.

O'Keefe, D. J., & Hale, S. L. (1998). The door-in-the-face influence strategy: A random-effects meta-analytic review. In M. E. Roloff (Ed.), *Communication yearbook 21* (pp. 1–33). Thousand Oaks, CA: Sage.

Patch, M. E. (1986). The role of source legitimacy in sequential request strategies of compliance. *Personality and Social Psychology Bulletin, 12*, 199–205.

Pigott, T. D. (1994). Methods for handling missing data in research synthesis. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 163–175). New York: Russell Sage Foundation.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 301–321). New York: Russell Sage Foundation.

Ray, J. W., & Shadish, W. R. (1996). How interchangeable are different estimators of effect size? *Journal of Consulting and Clinical Psychology, 64*, 1316–1325.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). Newbury Park, CA: Sage.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis* (pp. 261–281). New York: Russell Sage Foundation.

Slater, M. D. (1991). Use of message stimuli in mass communication experiments: A methodological assessment and discussion. *Journalism Quarterly, 68*, 412–421.

Struckman-Johnson, D., & Struckman-Johnson, C. (1996). Can you say condom? It makes a difference in fear-arousing AIDS prevention public service announcements. *Journal of Applied Social Psychology, 26*, 1068–1083.

Witte, K., & Allen, M. (1996, November). When do scare tactics work? A meta-analysis of fear appeals. Paper presented at the Speech Communication Association annual convention, San Diego, CA.

ABOUT THE AUTHOR

Daniel J. O'Keefe is an Associate Professor in the Department of Speech Communication at the University of Illinois at Urbana-Champaign. His research concerns persuasion and argument. He is the author of *Persuasion: Theory and research* (Sage).