Changing Conceptions of “Message Effects”
A 24-Year Overview

DALE E. BRASHERS
University of Illinois, Urbana-Champaign

SALLY JACKSON
University of Arizona

Jackson and Jacobs (1983) argued for three changes in the conduct of message effects research: inclusion of multiple message replications as instances of a treatment, recognition of message replications as a source of random variation in the estimation of treatment effects, and attention to issues of message sampling. This review updates their argument and examines 24 years of research published in Human Communication Research for evidence of attention to these recommendations. The review shows the following: the prevalence of studies failing to replicate has declined, replications are still rarely recognized as random factors, and researchers who use replications appear to do so for purposes of generalizability and control over confounding but without carefully analyzing the burden of proof associated with those purposes. An explicit framework for discussion of treatment effects in communication is proposed as an advance over the original reasoning of Jackson and Jacobs.

Asking whether Clinton’s “apology” to the American people was effective is a very different question from asking which message strategies best serve the communicator’s aims in making an apology. Experimental communication research occasionally focuses on the former type of question, but mostly it is concerned with the latter type; what is wanted is some sort of abstract description of what works and what does not in some class of communication situations. Questions posed at this more abstract level require a conceptualization of message effects that makes a distinction between a strategic choice understood abstractly and a particular concrete embodiment of that strategic choice in

Dale Brashers (Ph.D., University of Arizona, 1994) is an assistant professor in the Department of Speech Communication at the University of Illinois at Urbana-Champaign, 244 Lincoln Hall, 702 S. Wright Street, University of Illinois, Urbana, IL 61801; e-mail: dbrasher@uiuc.edu. Sally Jackson (Ph.D., University of Illinois at Urbana-Champaign, 1980) is a professor in the Department of Communication at the University of Arizona, Speech Building, Tucson AZ 85721; e-mail: sjackson@u.arizona.edu.

Human Communication Research, Vol. 25 No. 4, June 1999 457-477
© 1999 International Communication Association
a message. Research practices tailored to our subject matter should respect this distinction and, if possible, should promote increasing sophistication over time in how we think about and study message effects.

Increased sophistication in our methods of study is one aspect of scientific progress that commonly is overlooked in retrospective commentary on the field. In this article, we examine progress in scientific practice as reflected in methods for evaluation of message effects.

**HISTORICAL CONTEXT**

In 1974, when *Human Communication Research* (HCR) began publication, a completely standard experimental protocol was in broad use by communication researchers, assumed to be shared by other social science fields. This experimental protocol could be considered a "substitution instance" of one or another of Campbell and Stanley's (1963) recommended experimental designs, such as the randomized posttest-only design adapted below for comparison between two "treatments" ($X_1$ and $X_2$) presented to randomly assigned groups of respondents. The design is adapted to communication research through the substitution of contrasting messages or other stimuli for the treatments.

\[
\begin{array}{ccc}
R & X_1 & O_1 \\
R & X_2 & O_2 \\
\end{array}
\]

Although there is no easy way to document classroom teaching practices, we feel confident that most methods curricula, even today, present the Campbell and Stanley (1963) designs as canonical and as a good fit with our disciplinary subject matter. In this representation of the structure of an experiment, it is natural (although not quite correct) to think of the stimuli presented to experimental participants as the treatments. Jackson (1992) pointed out that the substitution for *treatment* in this design is not actually the concrete stimulus but a feature assumed to be the only point of variation between the contrasting stimuli. This subtlety is missed in the Campbell and Stanley (1963) diagram (where the treatment appears as $X$ whether $X$ is a concrete stimulus or an abstract procedure for generating stimuli) and in much recent and current research practice.

A typical substitution instance of the randomized posttest-only design involves presentation of an experimental message in two or more versions to independent groups of respondents. For example, two audiences might read two partially matched messages on drunk driving with difference in the intensity of the language used. Such a protocol makes sense if we think we can isolate the effect of a variable such as intensity and if we
think that intensity has one effect that is more or less uniform from setting to setting. There is a strong appearance of control in this protocol, and that appearance can be heightened by elaborate attention to the creation of the experimental stimulus.

Jackson and Jacobs (1983) attacked this unquestioned standard for research on message effects in a highly controversial article on generalizations in communication research. Jackson and Jacobs argued against the reductionist presumptions behind this protocol and attempted to show that serious threats to internal and external validity (to abstraction and to generalization) followed from experimenting on messages in this way. Specifically, they argued that use of single instances of message categories and message treatments left experiments vulnerable to confounding and made it impossible to evaluate the uniformity or variability of treatment effects across settings.

Jackson and Jacobs (1983) made three recommendations for change in research practices connected to a distinction between the treatment of interest and the concrete message materials used to embody it: first, use of multiple messages in experiments using concrete messages to represent abstract message categories or to carry abstract message treatments; second, treatment of individual messages as a source of uncontrolled variation in estimates of the treatment effect; and third, attention to the quality of the individual messages as a sample from the category of interest. These themes have been elaborated and refined through vigorous debate in HCR (Jackson, O'Keefe, & Jacobs, 1988; Morley, 1988) and other journals of the field (M. Burgoon, Hall, & Pfau, 1991; Hunter, Hamilton, & Allen, 1989; Jackson, Brashers, & Massey, 1992; Jackson, O'Keefe, & Brashers, 1994; Jackson, O'Keefe, Jacobs, & Brashers, 1989; Slater, 1991), spawning additional reflections on design and analysis options (Jackson, 1991, 1992, 1993; Jackson & Brashers, 1993, 1994a, 1994b). The subsequent writings have made clear that the arguments made in 1983 apply equally to other experimental components (such as the use of confederates).

The purpose of this article is to evaluate change in research practices, focusing on the 1983 Jackson and Jacobs proposals, to reflect on what these changes in research practices say about our changing conceptions of message effects and to speculate on future directions in both research procedures and analytic perspectives.

METHOD

To develop a historical overview of research practices, every article published in HCR from 1974 to 1998 was examined and coded for several features. Using procedures developed by Brashers (1994), articles were screened for relevance to the Jackson and Jacobs (1983) proposals,
eliminating all studies except those in which stimuli (e.g., messages, confederates, situations, or tasks) were used to instantiate a treatment difference. This screening eliminated from further analysis theoretical essays, methodological articles, articles reporting observational studies of naturally occurring conversations (e.g., discourse or conversation analysis), and articles reporting correlational studies of individual differences or other naturally varying phenomena, none of which incorporate manipulated stimuli.

Next, studies with potential need for replication were classified as to use of replications: as having appropriate replications, as needing but not having replications, or as not needing replications. A study was judged to need replications if vulnerable to plausible rival hypotheses connected with either specific message content or variability in the treatment. Independent coding of two complete volumes of HCR yielded 89% exact agreement: 96% exact agreement on whether studies needed replications, and for those judged as needing replications, 82% exact agreement on whether replications were present. All classifications used in the main analyses were done by the first author.

Studies containing replications were further classified according to statistical analysis: as treating replications as random, as treating replications as fixed, as ignoring the replications in analysis, as treating the treatment comparison for each replication as a separate study, or as adopting some other analytic strategy such as excluding replications from analysis after testing their effects for significance. Again, all classifications used in the analysis were done by the first author. Finally, studies were examined for attention to message sampling issues. No formal coding was done, but discussion of message selection, representativeness, generalizability, and other issues were noted.

Analysis centered on three questions related to the three recommendations of Jackson and Jacobs (1983): (a) How pervasive is the use of replications, and how has this changed over time? (b) What is current practice in analysis of replicated designs, and how has this changed over time? (c) How are researchers reasoning about the relationships between their experimental stimuli and their abstract claims?

RESULTS AND DISCUSSION

The Use of Replications

The issue of whether to include replications affects research in a very wide range of subfields. By our coding, 35% of all articles ever published in HCR reported studies in which replications were (arguably) needed. Of these, 58% actually incorporated replications into the design. However,
TABLE 1
Relevance and Use of Replications Across Volumes of Human Communication Research

<table>
<thead>
<tr>
<th>Volumes</th>
<th>Total Articles in Volumes</th>
<th>Total Needing Replications</th>
<th>Replications Included Where Needed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Yes (%)</td>
</tr>
<tr>
<td>1-3</td>
<td>86</td>
<td>33</td>
<td>13 (39)</td>
</tr>
<tr>
<td>4-6</td>
<td>87</td>
<td>26</td>
<td>12 (46)</td>
</tr>
<tr>
<td>7-9</td>
<td>74</td>
<td>19</td>
<td>13 (68)</td>
</tr>
<tr>
<td>10-12</td>
<td>80</td>
<td>33</td>
<td>17 (52)</td>
</tr>
<tr>
<td>13-15</td>
<td>79</td>
<td>31</td>
<td>20 (65)</td>
</tr>
<tr>
<td>16-18</td>
<td>67</td>
<td>28</td>
<td>22 (79)</td>
</tr>
<tr>
<td>19-21</td>
<td>63</td>
<td>23</td>
<td>14 (61)</td>
</tr>
<tr>
<td>22-24</td>
<td>67</td>
<td>18</td>
<td>11 (61)</td>
</tr>
<tr>
<td>Total</td>
<td>603</td>
<td>211</td>
<td>122 (58)</td>
</tr>
</tbody>
</table>

this proportion is not homogeneous throughout the 24-year publishing history. Instead, among studies published before Volume 9, only 46% used replications, whereas for those published after Volume 9, 63% did so ($\chi^2 = 5.23, p < .05, \phi = .16$). These proportions fluctuate considerably from year to year; Table 1 and Figure 1 provide a summary of proportions for each of the eight completed editorships.

Because studies were coded as needing but not having replications only when we could identify plausible threats to validity connected with the experimental stimuli, it may be helpful to review the nature of these threats. Jackson (1992) described a number of specific threats to internal validity, including case-category confounding (in which there is no way to tell whether a difference is between two categories or between two individual cases chosen to exemplify the two categories), superfluous variation (in which the experimental treatment contains potential important variations other than that intended by the manipulation), gestalt effects (in which the experimental treatment "leaks" into the supposedly controlled context), and unexamined effect variability (in which the treatment has a wide range of possible effects, only one of which is measured in the experiment).

In studies using single cases to represent categories defined by the explanatory factors, case-category confounding threatens conclusions about the category to the extent that it might be the cases themselves, and not the categories they belong to, that lead to differences in response. For example, J. K. Burgoon, Buller, Hale, and deTurck (1984) used one male and one female confederate to represent male and female speakers in a study of the impressions created by nonverbal behaviors, resulting in an unreplicated categorical comparison. Confounding of category and case
is evident from the fact that respondents asked to evaluate the speakers might evaluate them differently because of their genders or because of their unique qualities as individuals apart from gender. As Burgoon et al. (1984) noted, claims about presence or absence of gender effects are not justified here.

In studies using a single stimulus "template" (e.g., a message or confederate that can be altered to represent both levels of the treatment contrast) as the context for a manipulation, confounding of the simpler sort appears to be controlled, but only if (a) the treatment segments used to vary the stimulus are "pure" treatment unconfounded with unwanted additional variations, and (b) the treatment segments leave the controlled context constant in meaning and quality. In designs of this type, the threats to look for are surplus variation in the treatment segments, gestalt effects, and unexamined variability in the treatment effect. For example, Trees and Manusov (1998) attempted to evaluate a model of politeness by
varying message qualities (bald-on-record vs. facework) within otherwise matched complaints. However, the use of a single passage for each level of the treatment resulted in two messages that vary not only in presence of facework but also in other conceptually distinct ways (e.g., indicting a target's romantic partner vs. complaining that the target's romantic relationship is interfering with the target's friendship with the speaker).

Many studies published in HCR did include replications of stimuli such as messages (e.g., Bauchner, Kaplan, & Miller, 1980; Chaudhuri & Buck, 1995; Danes, Hunter, & Woelfel, 1978; Dillard, Palmer, & Kinney, 1995; Fiedler & Walka, 1993; O'Keefe & McCormack, 1987; Putnam & Sorensen, 1982; Roloff & Greenberg, 1980), confederates or speakers (e.g., J. K. Burgoon, Coker, & Coker, 1986; J. K. Burgoon & Le Poire, 1993; deTurck & Miller, 1990; Dillard et al., 1995; Freimuth, 1976; Greene, O'Hair, Cody, & Yen, 1985; Hocking, Bauchner, Kaminski, & Miller, 1979; Knowlton & Berger, 1997; Le Poire & J. K. Burgoon, 1994; Thompson & Seibold, 1978), situations or scenarios (e.g., Clark & Delia, 1977; Dillard & Kinney, 1994; Dillard et al., 1995; Doelger, Hewes, & Graham, 1986; Hewes, Graham, Doelger, & Pavitt, 1985; Neuliep & Mattson, 1990; Reeves, Lang, Thorson, & Rothschild, 1989), and tasks (e.g., Knowlton & Berger, 1997; Meyers, 1989). These studies show a long-standing if uneven attention to the contributions of replications to communication research.

Our examination of the entire body of literature published in HCR suggests that there are indeed cases in which replications are not needed. However, these cases are ones in which plausible rival hypotheses cannot be built from the idiosyncrasies of the materials. Of course, the plausibility of rival hypotheses is highly dependent on the claim being made as well as on the structure of the experiment. What we see in the change over time is not simple compliance with new design standards but rather increased attentiveness to substantive threats to validity related to messages and other stimuli: an increasing willingness to evaluate claims about communication effects against alternatives tied to the particulars of the experimental materials.

Statistical Analysis of Replicated Studies

When replications have been included as a strategy for permitting generalization, one way of thinking about what it means to generalize is to assume that what is wanted is a conclusion about an abstract message quality whose possible implementations are merely sampled by the specific stimuli included in the experiment. This line of thinking has suggested to many that replications should be treated not as fixed levels of an explanatory factor but as levels of a random factor that contributes sample-related variance to the size of the effects of theoretical interest (see Bonge, Schuld, & Harper, 1992; Clark, 1973; Coleman, 1964, 1979;
Crits-Christoph & Mintz, 1991; Fontenelle, Phillips, & Lane, 1985; Jackson, 1992; Jackson & Brashers, 1994b; Jackson et al., 1992; Maxwell & Bray, 1986; Richter & Seay, 1987; Santa, Miller, & Shaw, 1979; Wickens & Kep- pel, 1983; Zucker, 1990). Jackson and Brashers (1994a) noted that respondents are treated as random because they are a source of unpredictable variation in an experiment’s outcome and argued that “analogous reasoning applied to replications would suggest that replications be treated as random if they are recognized as a source of unsystematic variation in experimental outcomes and if they are replaceable in principle with other replications of like kind” (p. 359).

However, the mere inclusion of replications in an experiment does not constrain the experimenter to treatment of replications as random. Other choices open to experimenters include ignoring replications altogether, treating replications as a fixed factor, correcting in some fashion for the effect of replications, and analyzing each replication as a separate case.

Among these choices, the least attractive is to ignore the replications factor. Ignoring replications can lead to biased tests of treatment effects due to nonindependence of observations (Jackson & Brashers, 1993; see also Kenny & Judd, 1986). If replications are nested within categories (i.e., replicated categorical comparisons), observations within categories are nonindependent. Any nonzero association (nonindependence) between observations within categories will result in inflated Type I error rates (see Barcikowski, 1981; Kenny & Judd, 1986). If replications are crossed with treatments (i.e., replicated treatment comparisons), nonindependence can be both within treatments (for observations grouped within treatment levels by replication) and across treatments (for each level of the replication). Thus, the bias attributable to ignoring replications that are crossed with treatments is a complicated combination of influences. If the nonindependence takes the form of main effect differences across the levels of the replications factor, the test will be more conservative than expected. But if the nonindependence takes the form of interaction effects (a Treatment $\times$ Replication interaction), the bias is positive and the test of the treatment effect is more liberal than intended, resulting in an inflated Type I error rate. If both forms of nonindependence are present, the relative sizes of the intraclass correlations of nonindependent scores will determine the effect on the Type I error rate.

Nevertheless, among replicated experiments published in HCR, the most common analytic choice remains to ignore replications ($n = 45$). To protect against the relatively well-known biases due to nonindependence, this strategy is sometimes bolstered with preliminary testing of replication effects, dropping replications from analysis if pretests of replication main effects or Treatment $\times$ Replications interactions are nonsignificant. This may dampen but does not eliminate the bias
documented by Kenny and Judd (1986) (see Jackson & Brashers, 1993),
because any protection against bias depends on the power of the statistic-
tical test in the first step (i.e., the pretests of replication main effects or Treat-
ment \times Replications interaction effects), which typically is not carefully
controlled.

Treating replications as fixed is an advance over ignoring them out-
right because it at least handles nonindependence among observations.
However, Type I error rates can be alarmingly high for designs treating
replications as fixed (see Forster & Dickinson, 1976; Jackson & Brashers,
1994a; Santa et al., 1979; Zucker, 1990). This can be attributed to the failure
to account for the contribution of the replications-related variance to the
treatment variance. In a treatment by replication design, a significant F
test for a treatment effect that contains variability due to replications that
is not accounted for in the ratio of mean squares (e.g., a fixed effects test)
can mean that (a) the treatment effect is not equal to zero; (b) the treatment
effect is zero, but the Treatment \times Replication interaction is not zero; or (c)
both are not zero (Fontenelle et al., 1985; Jackson & Brashers, 1994a; Jack-
son et al., 1992). Jackson and Brashers (1994a) argued that the treatment of
replications as fixed rested on a picture of the treatment effect as invariant
from one implementation of the treatment to another.

Despite this problem, most studies including the replications factor in
the analysis treated the replications factor as fixed (n = 15). The potential
for alpha inflation is worst in studies that use few replications and many
participants (Jackson & Brashers, 1994a). Most of the studies examined
here included very few replications, which exacerbates the problem of
inflated Type I error rates (see Jackson & Brashers, 1994a). Brashers (1994)
estimated that the Type I error rate for designs such as that used by
McCormack, Levine, Solowczuk, Torres, and Campbell (1992) could be as
high as 95%, given the reported size of the Treatment \times Replication inter-
action and the study size (number of respondents = 1,074, number of
situation replications = 2). In the current review, the number of partici-
pants in each study ranged from a low of 16 to a high of 1,159 (mean = 179,
median = 127), and the number of replications ranged from a low of 2 to a
high of 120 (mean = 10, median = 3). The number of replications included
is generally very low: only two replications in 40% of all studies using rep-
lications and three or four in another 36%. With a large number of partici-
pants and a few replications, it is quite likely that a significant finding will
be obtained even in the absence of a true effect when replications are
treated incorrectly as fixed effects.

Similar results occur if replications are treated as separate cases (n =
11). A general problem with separate analyses is that they can fail to
account for the fact that the average of the distribution of treatment effects
may be zero even though the effect of any one replication can vary from
zero, sometimes far enough away to be counted as a significant effect. With two or more replications treated separately, the opportunity for at least one to turn up significant is the familiar "experimentwise" alpha (when the size of Treatment × Replication interaction is zero). When the per comparison Type I error rate inflates, as it will if replication-specific effects fluctuate around a true average of 0, then the experimentwise alpha magnifies this effect.

Studies that fail to account for replications-related variability in treatment effects provide tenuous support for claims of differences due to a treatment. Treating replications as random is a straightforward way of dealing with that problem. Only two studies in this review treated replications as an explicit random factor in their designs (Tracy, 1983; Villaume, Jackson, & Schouten, 1989). In several other studies in which replications were confounded with participants (confederates in Andersen, Guerrero, Buller, & Jorgensen, 1998; Manusov, 1995; and Palmer & Simmons, 1995; messages in O'Keefe & McCormack, 1987), replications are essentially a random factor in which the replications- and participant-related vari- ances cannot be distinguished (see Jackson, 1992, for a discussion of this strategy).

In none of the research published in HCR to date have we found any effort to explicitly model treatment effects as potentially variable from one implementation of the treatment to another, except in the statistical arguments of Jackson and Brashers (1994a). We will return to this point shortly.

Attention to the Role of Replications in Research

Across many studies, researchers gave explicit attention to the use of replications in communication research through discussion of generalizability, replication selection, representativeness, and other issues. For example, researchers frequently argued that their use of replicated designs was intended to increase generalizability (e.g., Andrews, 1987; Armstrong & Greenberg, 1990; Arnston, Mortensen, & Lustig, 1980; Berg & Archer, 1983; Bingham & Burleson, 1989; Cantrill & Seibold, 1986; Cusella, 1982; Hample & Dallinger, 1987; Honeycutt, Cantrill, & Allen, 1992; Honeycutt, Cantrill, & Greene, 1989; Hosman, 1989; Jordan & Roloff, 1990; Kazoleas, 1993; Kim, 1994; Leets & Giles, 1997; Lorch et al., 1994; Powell, 1974; Reeves & Greenberg, 1977; Runco & Pezdek, 1984; Samter, Burleson, & Basden-Murphy, 1989; Sparks, 1991; Weiss & Wilson, 1998) and to avoid confounding (e.g., Daly, Vangelisti, & Daughton, 1987; Geiger & Reeves, 1993; Knapp, Hart, & Dennis, 1974; Roloff & Janiszewski, 1989). Even researchers who failed to use replications sometimes offered that fact as a limitation to their study (e.g., Andersen & Kibler, 1978; Giles,

In addition, many researchers noted the strategy they used for selecting replications. In hopes of approximating the range of possible values in the distribution, Jackson and Jacobs (1983) cautioned that researchers should attempt to represent naturally occurring messages in their samples of replications. Jackson (1992) noted that samples of message replications generated by a researcher can be sources of invalidity due to bias (e.g., a choice of messages partial to the hypothesis) or collection-category mismatch (e.g., failing to represent the full range of message variability). Bradac (1986) argued that "elicited, purloined, and contrived messages" can fail to generalize to a population of messages for a variety of reasons, including "idiosyncratic warping" (p. 60) (e.g., tendencies of the researcher to unconsciously choose similar style or language in constructing messages).

Although no general strategy is readily apparent for selecting replications, it is important to get a sense of how well replications might approximate the diversity of an imagined stimulus population. Jackson (1992) offered some useful hints for generating message samples, and other types of replications factors will require that researchers use imaginative techniques for generating samples.

In the studies reviewed here, some authors described a form of sampling that can help deflect charges of researcher bias. Leichty and Applegate (1991) chose their replications from interviews with resident assistants in dormitories about the persuasive tasks they encountered on the job because the participants in the study also were resident assistants in dorms. Other times, stimuli were not sampled but chosen purposively. In the Miller et al. (1992) experiment designed to test the effects of positive, negative, or boastful disclosures on perceptions of a speaker, they chose their four situations (job hunting, rushing a fraternity, attending an awards dinner, and attending a party) because "it was felt that university students would be more responsive to stimuli that involved student concerns and experiences" (p. 370). When sampling was purposive, often this led to pretesting to ensure replications adequately represented the levels of the explanatory factor. Hawkins, Pingree, Fitzpatrick, Thompson, and Bauman (1991) studied the effects of marital interaction schemata on expectations about couple interactions. To choose replications of couples that fit certain schemata, they viewed 200 videos and selected 80 scenes with couples to represent the three couple types. Then, they had participants rate the couples on the Relational Dimension Inventory (RDI). From those ratings, they selected six tapes for the experiment: An Early Frost and Tender Mercies to represent the traditional couple type, Mr. Mom and Micki and Maude to represent the independent couple type, and Irreconcilable Differences and Desperately Seeking Susan to represent the separate couple
type. They “chose scenes showing marital interactions between spouses without others present, and talking about normal conversational topics (e.g., the relationship, external issues, and other people, but not murder or kidnapping)” (p. 490).

All of these studies point to the fact that generating samples of replications can be a complicated task (sometimes requiring a balance between sampling and purposive selection), but it is certainly a task worthy of attention. Questions of generalization to a larger “population” of stimuli will be answered in part by how well the stimuli represent the range of possible choices of stimuli. Although descriptions of the sampling of replications offer some indication that a variety of potential stimuli are being considered, often that sampling is purposive (e.g., selecting confederates who can enact a manipulation).

RECONCEPTUALIZING EFFECTS

How should we think about treatments and about treatment effects in communication research? We have argued here and elsewhere for drawing a conceptual distinction between abstract strategic choices made by communicators and the specific messages or other stimuli that result somehow from the making of these choices. When we think about treatment variables in communication, we should be thinking about the abstract construct and not the concrete stimuli. On the other hand, we must never forget that the abstract construct does not occur anywhere as an object but only as part of a mostly unanalyzed package deal known as a message or as a confederate or as some other aggregate. This creates a tension in our thinking about treatment effects, which we now propose to address in a preliminary way.

From standard textbook treatments of statistical models, we grow accustomed to thinking of effects as specific numerical values that combine with other specific numerical values to generate the predictable variation in a measurement. In thinking about message variables and similar sources of variation, this standard textbook treatment invites us to think about an individual stimulus as having an effect of its own that is supplemented by the addition of an effect related to a treatment applied to the stimulus. However, this picture collapses on any detailed examination, because the treatment is not any one thing but an abstract potential that gets realized in indefinitely many different ways. When we speak of “the treatment,” what we really mean is some indefinitely large class of related things that embody, in diverse ways, a common production strategy applied in diverse circumstances. We never, for example, observe the effects of two-sidedness in argumentation; we only observe the advantage (or disadvantage) of certain two-sided messages over their one-sided
alternatives. Our interest is in the production strategy: to answer or ignore arguments against one's own position. Our observations are for arbitrary instances of the result of applying this production strategy to otherwise one-sided arguments.

Suppose we were to explicitly theorize that communication effects for abstract strategic choices come not as fixed increments added to or subtracted from a message's base value but rather as distributions of advantage and disadvantage requiring description over their entire range. This requires an attitude adjustment: a willingness to entertain alternatives to the pictures invited by the sort of statistical training most of us share. It also requires change in research practices: a turn away from simply testing hypotheses about the overall treatment effect and toward modeling of the effect as variable and to some extent unpredictable.

Formally, it is possible for a treatment effect to be understood as a random variable $\Delta$, taking on different "true" values, $\delta_1, \delta_2, \ldots, \delta_n$ for different objects (e.g., confederates or messages) to which it might be applied. To describe such a treatment effect, what would be wanted is not some single value but a description of a whole range of values, such as a distribution with a mean $\delta$ and a variance $\sigma^2(\Delta)$. A constant treatment effect would have every replication-specific effect $\delta_j = \Delta$ and a variance $\sigma^2(\Delta) = 0$. Any nonzero $\sigma^2(\Delta)$ would represent an effect that varied randomly from instance to instance (which appears in a standard analysis of variance as a Treatment x Replication interaction). If the variance $\sigma^2(\Delta)$ is nonzero, even if the mean of that distribution is zero (so that the null hypothesis $\Delta = 0$ is true), nearly all finite samples of replications will have an average treatment effect $\bar{\delta}$ that is nonzero. The null hypothesis $\bar{\delta} = 0$ (that the average treatment effect across a sample of instances is zero) will not be true even when the null hypothesis of interest $\Delta = 0$ is true. (As noted earlier, this fact leads to alpha inflation for studies that treat sampled materials as fixed effects or that analyze replications as separate studies.)

Reichert (1998) incorporated this picture of treatment effects as intrinsically variable into an experimental analysis of the potential benefits of using sexual appeals in social marketing. The benefits and risks of using sexual content in commercial ads are well known. Reasoning that sexual appeals might be differentially advantageous depending on the resources presented by any given topic, Reichert created experimental contrasts between sexual content and neutral content for 13 different topics (including diverse topics such as support for the arts, safe sex, sun safety, and diet). In Reichert's results, the use of sexual content always enhanced attention to the ad, with attention effects ranging from a standardized mean difference of .22 to a standardized mean difference of 1.55. By contrast, sexual content was neither uniformly advantageous nor uniformly disadvantageous for recall of the content of the ad, ranging from a
standardized mean difference of -.49 to a standardized mean difference of .37. Notice the difference this picture of effects makes for how we discuss the effect of sexual content: Instead of simply noting that the effect on recall is small (averaging .04 across the 13 messages), we note that the effect is variable enough that, although the average effect is positive, the effect in any given instance has a high probability of being negative.

There would be no reason to adopt this more complex modeling strategy if treatment effects could usefully be understood in terms of fixed increments added to or subtracted from the base effect for any occurrence of the treatment (i.e., if treatment effects could be thought of as a single value). However, with the change in research practices documented in this review (i.e., incorporating replications of stimuli within experiments) has come a flood of evidence that treatment effects in communication cannot be usefully understood this way. Considerable evidence now exists that, across a variety of domains of communication research, treatment effects vary from replication to replication. In a study of “advertisorials” designed specifically to try to debunk the “messages-as-fixed-effects fallacy,” M. Burgoon et al. (1991) found that differences in treatment effects depended on which message was used even though the treatment procedure was identical for all six replications. They admitted that, “in fact, the messages operated so differently in the inoculation situation that had a single message been chosen for the study, support, null findings, or the obverse of the hypothesized relationships could have been obtained” (p. 31).

As noted earlier, variation in the size or direction of a treatment effect is manifest in the Treatment $\times$ Replication interaction (Jackson, 1991, 1992). In the current review, significant interactions of replications and treatments were found in many studies (e.g., Bingham & Burleson, 1989; Canary & Spitzberg, 1987; Doeiger et al., 1986; Hample & Dallinger, 1987; Hewes et al., 1985; Hocking et al., 1979; Le Poire & J. K. Burgoon, 1994), including the only two studies in which replications were treated as an explicit random factor (Tracy, 1983; Villaume et al., 1989). Even studies that ignored the replications factor sometimes found interactions of treatment effects and replications in auxiliary analyses (e.g., Weiss & Wilson, 1998). In addition, evidence that results differed from replication to replication existed in many of the studies that chose separate analyses or non-parametric tests (e.g., Honeycutt et al., 1989; Neuliep & Mattson, 1990). Other evidence of variability in treatment effect sizes is summarized by Jackson et al. (1989) and Jackson and Brashers (1994a).

In the face of this accumulated evidence of effect variability, a well-justified conclusion about the overall effect of a treatment tells us only its average over many cases, when we could have not only knowledge of the overall effect, but also a detailed picture of the uniformity of this effect
over its many diverse instances. In common practice, we do not interpret main effects without reference to interactions that may affect them; what we propose here is extension of this commonly accepted notion to qualify our conclusions about treatment effects by explicitly measuring not only their averages but also their tendency to vary from one instance of the treatment to another, as in Reichert’s (1998) study.

A highly unified basis for the Jackson and Jacobs (1983) recommendations emerges from this picture of communication effects. Thinking of any treatment effect as having a distribution of true values means we would never think of trying to describe the distribution by examining a single point. Nor would we try to describe the distribution without giving thought to the selection of points to observe. The design and analysis practices critiqued in this review—failing to replicate treatments or failing to take into account random variation associated with replications—reflect deep, unexamined conceptualizations of treatments and their effects, grounded not unreasonably but neither felicitously in our methodological training.

CONCLUSION

The trend of studies published in HCR toward incorporation of replications into experiments on communication effects is pronounced, and it represents a move toward more rigorous standards of proof. Methods for the study of messages and other stimuli necessarily advance as we learn more about our objects of study—and what we learn about our objects of study may take unexpected turns as we improve our methods of study. Here is one such case. We have begun to be very attentive as a field to problems of confounding that result from failure to replicate contrasting categories or treatments, and in responding with replicated designs, we have generated evidence that pricks our taken-for-granted pictures of treatment effects. Our next move should be toward a better class of claims—claims not just about whether a treatment makes a difference but about the shape of the difference that treatment makes across the diverse domain in which it might be applied.

NOTES

1. The decision about whether a study used replications often turned on how the replications were labeled. Two television programs included for greater generality are replications, unless the two programs get repurposed as levels of a program type factor used to make claims about the differences between, for example, comedies and dramas. The relative difficulty of deciding whether stimuli were intended simply as replications or as levels of
theoretical classifications is partially a reflection of a transitional period in the field's thinking about how stimuli stand for theoretical contrasts.

2. The assumptions of the $\chi^2$ contingency test, in general, are plausible. The observations within levels of the "pre- vs. post-Volume 9" are relatively independent, and the expected variable values are all greater than 5.

3. The confederates were crossed with nonverbal behaviors (e.g., they each enacted a high degree of eye contact in one condition and a lower degree of eye contact in the other condition); thus, this design can be considered a replicated treatments design for the theoretical variables crossed with the specific confederates.

4. Some studies had replications of several stimuli (e.g., scenarios, messages, and confederates in Dillard, Palmer, & Kinney, 1995; confederates and tasks in Knowlton & Berger, 1997; conversational segments and conversational extensions in Tracy, 1983, and Villaume, Jackson, & Schouten, 1989). Other studies needed replications of several stimuli but did not include them (e.g., the use of a single scenario, confederate, and message in Trees & Manusov, 1998).

5. Some authors offered justifications for ignoring replications (e.g., Hawkins, Pingree, Fitzpatrick, Thompson, & Bauman, 1991). J. K. Burgoon, Coker, and Coker (1986) noted that "where inclusion of the confederate variable (which was a random variable) was nonsignificant or failed to improve the efficiency of the design, effects were tested with a three way, fixed model ANOVA" (p. 513). Despite finding differences among the observer ratings of their confederates, J. K. Burgoon, Walther, and Baesler (1992) defended ignoring the replications in their design because "the confederate variable failed to increase the power of most analyses, frequently reduced the power, and in no cases produced significant results different from those found in fixed four factor models (in which the random confederate factor was omitted)" (p. 249). However, J. K. Burgoon and colleagues (1992) fail to notice that the "increased efficiency" achieved by ignoring replications (a) comes at the expense of control of Type I error and (b) results in a test of a hypothesis that is probably not the true hypothesis of interest.

REFERENCES


Coleman, E. B. (1979). Generalization effects vs. random effects: Is $\sigma^2_r$ a source of Type I or Type II error? *Journal of Verbal Learning & Verbal Behavior, 18,* 243-256.


