ology: Learning, Memory, & Cognition, 9,
ssage: Theories of influence on mass media
85). Human experimental psychology. New
ith factorial experiments. Journal of Mathe-
York: Oxford University Press.

9 Designing Experiments That Assess Psychological Responses to Media Messages

Byron Reeves
Stanford University

Seth Geiger
University of California, Santa Barbara

The apparent simplicity of experiments is deceptive. There are numerou diffi-
cult decisions to be made in an experiment as stimuli, treatments, measure s, and
subjects all compete for resources. This chapter is about design decisi ons in
studies that examine cognitive responses to media messages. The disc ission
emphasizes solutions that are guided by theories of message processing, b it also
acknowledges that research is planned with limited resources. Conseq ently,
decision making not only involves knowing which options are best theoret cally,
but knowing which decisions have priority, and when to make comprom ses.

We have framed key experimental decisions as practical issues. Each de cision
could easily (and usefully) be discussed in the language of statistics and ormal
experimental methods; however, we have chosen terms and examples that relate
specifically to media studies. For example, we refer to a randomized block
factorial design as one where different people see different messages. In so
doing, we hoped to translate the logic of experimentation into practical o tions
for media research. But even more important than a taxonomy of designs is the
explicit question, "Why and when would you want to do that?" and the im plicit
question, "Why didn't you do something else?" Both questions are relev nt to
choices in any design involving media messages.

The chapter is organized around three decisions that are part of all e periments with media messages: (a) decisions about message samples, (b) dec sions
about creating variance with messages, and (c) options for assigning pe le to
experimental conditions. Our discussion about each of these decisions ref cts a
particular philosphy, albeit an informal one, about experimental procedu s (as
opposed to theories). This philosophy recognizes the importance of qui ness
and ease in laboratory setup and measurement, and emphasizes a program natic

progression of tests, none of which dominates resources, and each of which tells part of a larger story. This allows changes in research plans as the program progresses, and is in lieu of a single grand design that requires a major time commitment to one set of messages and outcomes.

THREE CRITICAL DESIGN DECISIONS

In designing research, every decision seems as if it should come first. This is quite true for considerations about messages and subjects. However, it may be important to consider first how messages will be used. The primary reason is an assumption that an interest in media is the entree to research rather than a general interest in cognitive processing or other independent variables. This may be the most important difference between a psychological and communication approach to message processing. And because the first decision will constrain subsequent ones, a consideration of media is primary.

Media messages are never an example of one thing and nothing else. Consequently, when attempting to theorize about a particular message attribute, it is quite difficult (if not impossible) to isolate this single feature to use as a stimulus. This presents serious problems for creating variance between messages, because messages are rarely, if ever, distinguishable on a single attribute. Consequently, if we begin with an interest in a natural and complex stimulus, we have, by definition, accepted the problem of message confounds.

Message generalizability is not as common a concern in media experiments as whether results will generalize to other samples of subjects (e.g., are college sophomores representative of real people?), and to other situations (e.g., do people watch TV in a laboratory the same way they do in their home?). These issues are difficult, but it is important to note that they are part of any experimental research, and it is doubtful that there are significant caveats for experiments about media. Instead, we concentrate on a different external validity question: Will research results generalize to other samples of media? This question is especially critical because the complexity of media makes any single message example unique—perhaps even more unique than any single person that processes it.

Because there is such a large amount of variance between messages, the method of selecting messages for an experiment is a crucial decision. We trace three important decisions related to experiments using messages:

1. How many messages at each level of the treatment are necessary to represent a population of messages, and how will they be chosen?
2. Should message variance be created by altering the same presentation or by sampling different messages within each level of the treatment?

How these two message questions are decided has implications for the third question:

3. How should people be assigned to the various treatment conditions?

Decision 1: Creating Message Variance

It is difficult to imagine a question about message processing that would not benefit from the presentation of multiple messages in an experiment. Even when the specified message unit is quite small and theoretically well defined, messages are so complex that one example will rarely capture the natural variation in a single feature. Any given message is likely to represent one of a number of possible levels of other message factors, as well as the feature of interest. Also, single messages are likely to be confounded with a number of factors that could influence subject responses. Consequently, a sample of messages should be used, rather than one exemplary segment. Each message added to the sample further reduces systematic between-message differences to less pernicious random error. In fact, if we put our time and money where the variance is, we would probably worry about message samples even more than subject samples.[1]

How Many Messages? How many messages are enough? The theoretical answer that runs counter to most experimentation in mass communication is almost always, "More than one." In many cases, it would be reasonable to have as many, or even more messages than subjects. It is more common to have 20 people watch one or two messages, than to have a small sample of people watch 20 messages. It is important to remember in considering the second option that the error corrected by message sampling cannot be remedied with greater measurement reliability or a better subject sample. The choice between these extremes should depend primarily on whether it is more important to worry about the representativeness of the people you observed or to worry about generalizations to a larger category of messages.

In designs where a number of messages are used to represent each level of a treatment, the determination of the size of a message sample depends on two factors: (a) how much variance there is between messages at different levels of a treatment; and (b) how much variance there is between messages chosen to represent the same treatment level. Exposing subjects to large message samples

[1]An exception to the need for message samples may be evaluations of prominent single messages; for example, special newscasts or political debates (Bradac, 1983). In these cases, there is no larger a group of messages to which the research need generalize. The research questions in these evaluations, however, will be limited to evaluations of single messages rather than generalizations about processing relevant to a message category.

will be most beneficial when there is relatively high variance between messages at a single level of a factor. For example, consider a memory test between television commercials that use either informative or persuasive strategies. Informativeness or persuasiveness usually applies to an entire commercial, even though only one part, and maybe a small part, of each message distinguishes it as informative or persuasive. Even precise definitions of these two message categories are certain to allow substantial variance within the categories. Advertisements also vary on product type, number of people, complexity of visuals, type of music, and innumerable other characteristics. And several of these features could be reasonably associated with memory. A large message sample will tend to cancel these errors.

A second situation calling for larger message samples occurs when small differences between treatment levels are expected. As an experimental strategy, we attempt to maximize variance, but there are situations (often associated with compelling policy considerations) where it is important to compare treatment levels that we know are not substantially different. Larger samples would help stabilize the responses in different treatment conditions, and insure that the question was not dismissed due to null results attributable only to high variance between messages; variance that exists more probably in small stimulus samples. Alternatively, an experimenter could make the opposite error of overestimating differences by picking one of only a few message examples that produces differences, thus jeopardizing generalizability to other messages. Either of these errors could occur with small message samples even if many people responded identically to each message. Therefore, it is message variance rather than subject variance that should determine message sample sizes.

Now a more practical response to the question, "How many messages are enough?" In spite of the previous discussion, you still have to pick a number. How many messages should you choose? An honest answer to this question is as follows: You would want to use as many suitable messages as you could locate and that you could reasonably expect people to attend to. Although it is difficult to imagine a study where you would reach an ideal limit for a message sample before you ran out of time and money to find additional examples, people may begin to respond unreliably to television after about 60 minutes of laboratory viewing.[2] You can then select the number of message units that fit that time block, assuming they are available.

[2]One of the obvious constraints on the length of experiments is the measure used. Passive measures, such as recording eye gaze during viewing, pose the fewest problems. With this measure, there is no additional burden on subjects other than watching the screen, and sessions can last as long as people are willing to view, probably about 1 hour. This could be even longer if the purpose of the research is to allow other distractions (talking, eating, other media) to compete with the television. The 1-hour guideline applies also to physiological measures. For reaction-time tasks, and measures that require active participation by subjects, 1 hour may be too long for a single session.

Choosing the Messages. After deciding on the number of messages, particular ones must be selected. Ideally, the entire population of messages at one treatment level would be defined, with messages randomly selected from that group. This is rarely (if ever) done, and for good reason. First, it requires an exact definition of the population of messages, and it assumes that we have equal access to each message selected. More commonly, access is limited (unless we are producing our own messages), and consequently, we tend to work in the other direction beginning with those messages that are possible to use. Once these are identified, we then judge (but rarely report) the larger group of messages that are represented. There is a potential for systematic bias when researchers use their own judgments about which messages to include (imagine the criticism that would ensue if we picked subjects by availability or by our preferences for them as people).

The system used to select messages also applies to the selection of messages within different treatment levels. When discussing treatment levels, there are two unique samples to consider, one associated with the determination of the treatment levels, and the second with the messages that reflect each level. If the treatment levels are defined along a continuum, then specific levels must be "sampled" before examples of messages at each chosen level can be selected.[3] For example, if the dimension of interest is visual complexity, defined by the number of cuts within messages, the first sampling decision involves selecting treatment levels (assuming that examples at all possible levels cannot be used). Typically, this selection is anything but random, and usually includes two values at or near the ends of a continuum (Geiger & Reeves, 1991). Consequently, any notion of a randomized sample of stimuli is compromised even before messages are selected.

The next step involves the selection of messages at each treatment level. Although some form of systematic selection is necessitated by practical constraints, there are certain strategies that can minimize the problems associated with this form of sampling. The single best aid in systematic selection is knowledge of the specific confounds that jeopardize a particular study. For example, if persuasive and informative advertisements are compared for memorability, the selection of ads in each category should be guided by knowledge of other message features that could influence memory, and reflect typical aspects of advertisements. For example, if music enhances memory, then either all the advertisements should have music, or the same number in each treatment should.

This sounds straightforward, but it is actually quite hard to accomplish. Be-

[3]This issue is often treated statistically, but the conceptual implications are equally important. The issue of fixed versus random effects in experiments has been addressed in the psychological literature (Clark, 1973; Coleman, 1979), and in the communication literature (Burgoon, Hall, & Pfau, 1991; Jackson & Jacobs, 1983; Jackson, O'Keefe, & Jacobs, 1988; Slater, 1991). In the communication literature, the discussion is centered on the interpersonal and language areas, and has been largely concerned with statistical assumptions and analysis strategies. Each of these issues, however, is quite relevant to the present discussion about media messages.

cause there are a large number of possible confounds between messages, there should be several different constraints on message selection. Because message repetitions are hard to come by, it is difficult to meet all of these constraints in controlling for confounds. Many media studies use only a single example of each treatment level, and even studies that include several repetitions are typically limited to a few messages. Any systematic selection of two groups of messages, no matter how competently selected, and no matter how obviously they differ on a feature of interest, will still result in two groups of messages that systematically vary on a long list of attributes—possibly an infinitely long list. Consequently, we should not allow our ability to think of possible confounds to be the final criterion by which the research is planned or evaluated.

There are two views about message confounds that are more optimistic. The most favorable of the two is that confounds are often mistaken for characteristics of media that are part of, not competitive with, the message feature under study. Consider a comparison of television messages that are negative or positive in emotional tone. A pretest could confirm this difference by asking people to describe the primary emotion present in each message. Once messages were selected, the possible confounds in the selection could then be identified, and plans could be made to control for each one. Perhaps the negative messages are slower paced or have more close-ups than the positive messages. An important question is whether these are confounds that should be controlled or features associated with emotional valence of messages. If negative means slower pacing and close-ups, then controlling for these elements will steal variance that should be featured. The second view is based on an admission of failure, but failure with a known remedy. Perhaps we merely know less than we have acknowledged about the forms of media to which results are applicable. This is no small problem, but like external validity issues with subjects, the problem of generalizability is preferable to the problem of inaccurate findings, for any group of stimuli. Optimism comes from the knowledge that external validity can be increased with time and money, rather than significant changes in theory or measures.

What Is Being Sampled? A final sampling issue has to do with what is being selected. Cognitive responses typically deal with message units that do not exactly overlap traditional message boundaries. This means there are often two potential units to sample: one that pertains to traditional units (e.g., television commercials), and one for theory units (i.e., the specific feature within the traditional unit that affects responses). For a study about cuts between scenes, for example, there could be two different samples: first, a television program or programs, and second, the cuts themselves. The exact units to be sampled are best defined by theory (cognitive theories are rarely about television programs); however, it is possible to think of these two samples in sequence. Programs may be selected intentionally, yet the actual cuts where responses are measured may be randomly selected from within the programs.

It is easy to think that sampling should pertain only to the larger message package because those messages are often what motivates research. The people responsible for traditional messages often fund research, and even if they do not, it is nice if results apply to a popularly recognized message category. But we should not take on the unnecessary burden of perpetuating these categories in research. If, for example, there was an interest in television messages that create negative emotion, it would be quite a bit easier to find a good sample of negative segments if the research was not constrained to commercials, news stories, or television programs. And we might learn something about television that would not be apparent if the study were limited to one category.

Decision 2: How to Create Treatment Differences

A second question about messages concerns creating variance: How should the differences between messages be manipulated? The answer to this question cannot be divorced from message sampling issues (previously discussed as Decision 1), or from design strategies involving subjects (Decision 3). Consequently, it may be important to use an iterative planning strategy; one that cycles through the same questions several times rather than a strategy that finalizes a single decision before proceeding. There are two ways to create variance between messages: (a) the same message can be altered to produce two or more versions; or (b) different messages can be sampled to represent unique treatment levels.

Much of the prior discussion of confounds in message selection assumes that messages are sampled within treatment levels. The concern about unequal message groups could be eliminated, however, if variance was created by changing a single feature of interest in two versions of the same message. This insures that all other message characteristics are equivalent between the two manipulations. Consequently, given a choice, the likely first option would usually be to alter messages. And rather than alter existing messages, we would prefer to alter different versions of original productions to gain maximum control over extraneous message features. Although this may be easy with print messages, it is still a time consuming and costly task with video and film. In spite of the expense, however, it is worth considering because it is the only sure way to have precise control over the vast array of extraneous message features that could doom an experiment. For example, if you are interested in the effects of camera movement (e.g., zooms and dollies) on judgments about people on television, an ideal experiment would be to compare a dolly and a zoom for the same characters and materials in the same setting. This guarantees that all other message features would be identical, a major accomplishment given the number and importance of message features that would be discrepant if two different messages were used.

A more feasible strategy for altering messages is to consider message variance as a postproduction problem. This has many advantages over original production, but avoids time and expenses because available material is simply edited. This strategy also insures that the overall production quality of the messages is

not compromised, as it might be in the case of original production. The most simple version of this strategy is to vary the presence or absence of the features of interest. For example, if you were interested in how graphic violence in news stories influenced memory for news, you could edit two versions of a news story: one with a graphic segment, and one without.

Unfortunately, this technique is not possible for all research questions, especially those that involve language or other features that occur within rather than between scenes. It's easier to insert and delete segments that can be manipulated as a whole or those that are audio-only or video-only. If two different leads are needed for the news stories, for example, it is unlikely that the exact language differences can be found by recording material from the networks or using other available material. It would be best for the same anchor to read the same story twice.

Varying the presence or absence of segments, however, does not eliminate all problems with incomparable versions. The messages may be different lengths (and subsequently affect processing differently), they may become confusing when material is omitted, or they may be recognized as altered. A similar technique that solves these problems involves substituting material. The presence or absence of a featured segment is edited out of one version, and other material, irrelevant to the manipulation, is substituted.[4] This allows message length to be controlled, and with skill, plausible sequences can be created. For example, in a study about children's responses to toy advertising, the same toy commercials were edited to include or not include video segments that showed the toy moving by itself (Reeves, Thorson, & Lometti, 1987). In the version without movement, material from a similar commercial was substituted. In addition, short segments from other parts of the commercial were repeated with no apparent change in the quality of the ad.

There are two other considerations when altering messages: believability and cost. It is not difficult to produce realistic messages by editing video. In fact, most people are not aware of editing changes, even severe mistakes like jump cuts, that would make professionals cringe (Drew & Cadwell, 1985). Consequently, this may be less of a problem than it seems. A second potential disadvantage of message altering is cost, usually in dollars for production, and time in which to create the messages.

The other choice in creating treatment variance is to sample messages in each treatment condition. Here, sampling applies to a different problem than in the prior section where sampling referred to the strategies used to find any messages, regardless of whether they would be altered or not. Even when messages are

[4]The three altering options discussed in this chapter were original production, editing material to vary the presence or absence of specific segments and editing to substitute segments. Altering digitized versions of messages represents a fourth option that is readily available for print messages, and will soon be more feasible for video.

altered, the need for replication within treatment levels is obviously no less, and different examples must be "sampled" from some larger population. Sampling can also refer to an alternate strategy for creating treatment variance. If you cannot create or alter matched pairs of messages to represent each treatment, then different messages must be sampled for each level.

The most substantial problem with this strategy is the previously discussed issue of confounds. Although this strategy may cause more interpretive problems than message altering, there are two obvious practical advantages that may often make sampling the favored strategy. First, messages do not have to be produced or edited, they need only be found. If the research question is about message units that are relatively easy to collect but hard to alter, then sampling messages to create treatment variance will obviously work better. Second, there are times when editing will simply not work because the message feature of interest cannot be inserted or deleted as a unit. These include manipulations of audio (it is difficult to insert audio if speech is synched with the visuals), and the need to vary the presence and absence of visuals that do not exist in the original material.

In summary, the major purpose of altering techniques is to retain control over other message features that may become confounded with the featured attribute. This is substantially more difficult if messages are sampled within treatment levels. Sampling is the strategy of choice when altering is not practical, when a lot of material is available, and when the message feature of interest cannot be manipulated as a unit.

Decision 3: What to Show People in an Experiment

The last decision refers to the assignment of people to different message conditions. Decisions about subjects have traditionally been limited to concerns about who they are and how many are needed; however, these may not be the most important issues for experiments about mental processing. A more critical question concerning subjects is, "Who will see what material?" Will different groups see different messages or will everyone see all of the messages? The most common strategy in communication research is the between-subjects design, where different groups of subjects are randomly assigned to be exposed to a single level of a message factor. The within-subjects design is a repeated-measures design where all subjects are exposed to all levels of all of the factors.

Between-Subjects Designs. Between-subjects designs have been used most often in communication experiments. Each group is exposed to a different treatment level, with the mean scores on some criterion compared relative to within-group variation. The most significant advantages of this design are its simplicity and the absence of the influence of message presentation order or other treatment levels on responses. Watching negative commercials, for example, cannot bias how you process positive ones if all you are exposed to is one type of message.

Another advantage is that there is less opportunity for time-based influences on results such as an increase in performance with experience, or for fatigue.

The choice of a design is not independent from decisions regarding message variance (Decision 2). The best case for between-subject designs can be made for studies where the same message has been altered to create message variance. In these cases, a within-subject design would require people to see different versions of the same messages—messages that were identical except for a single feature. For this reason, altered messages potentially cause more contamination between treatment levels than situations where the messages comprising each level are completely different.

These obvious advantages probably account for the popularity of between-subject designs in media research. But the advantages of grouping subjects are not absolute. There are many problems that between-subject designs solve—at greater cost than within-subject alternatives—that may not be critical to processing studies. And there are also compelling positive arguments for within-subject designs, arguments that are especially relevant to psychological experiments.

The most fundamental objection to between-subjects factors is the amount of error associated with individual differences, differences that are not likely to be of particular import. For the measures discussed in this volume, this may include motor coordination differences (for reaction time tasks), concentration ability (for selective looking), and a range of individual differences for each physiological measure (e.g., handedness, gender, physical ailments, and a host of unknown explanations). Unfortunately, these individual differences result in the confounding of subjects and treatments. Random assignment of subjects to conditions helps with these problems, but there is no way to separate treatment effects and between-subject variance, a particularly pernicious problem for cognitive measures.

Within-Subjects Designs. Within-subject factors generally give a clearer picture of treatment effects because the variance within treatment levels is dramatically reduced by having each subject in each condition. Because treatment effects are determined by each subject's averaged response within a treatment, divided by the same subject's response averaged over all the treatments, each participant serves as his or her own control group. This is particularly helpful in message processing experiments because individual differences in the levels of many measures will often be large. Within-subject designs may consequently reduce error variance as much as one half to one fifth that of comparable between-subject factors (Calfee, 1985).

There are three other advantages, one related to ecological validity, and two practical. First, the validity argument. When messages representing two treatment levels are shown to the same subject, we usually worry about contamination or carry over. This concern, however, should not be confused with an opportunity for comparison. Differences between treatment levels may not be

apparent unless messages can be compared. A negative news story, for example, may not be influential unless it can be compared with a positive one. This would be especially true for media examples that are not expected to produce substantially different responses. The ecological validity argument centers on this question of comparison: Are people likely to see messages from each condition naturally, or not? The more ecologically valid laboratory situation may be one that replicates the range of messages available in the real world; that is, one that allows subjects to compare messages in each treatment level (Greenwald, 1976). This argument is quite relevant to media experiments because, for many treatments in communication research, all levels of the treatment do in fact exist within the context of a single individual's experiences.

The most popular objection to within-subject factors is that the juxtaposition of treatments will sensitize subjects to the intent of an experiment and interfere with responses. This is not equally true for all research questions, however. If the test is between two methods for teaching math, it would not be difficult for someone to guess what the study is about and adjust responses accordingly. But for many questions about cognitive responses to media, it is just not that easy to figure out the research question, even if the treatments are all presented to the same person. If messages are sampled within each treatment level, and message presentation is randomized across treatment levels, most adults will not be able to determine that the research question is about informative versus persuasive advertisements, visually complex versus visually simple news stories, or many other features that are typically studied. In these cases, message complexity is an advantage. There are so many things happening in messages that it is hard for a subject to pinpoint the highlighted feature.

The practical rationale for within-subject designs is quite compelling. First, the same power in experiments can be achieved with substantially fewer subjects. A group of twenty subjects participating in a within-subject study has the equivalent power of 40 subjects if the treatment factor has two levels, and 60 subjects if there are three levels. And it is likely that the advantage is even greater, because the variance between groups in a single condition will be larger than the variance within groups across conditions. This will be particularly true for measures that vary substantially between subjects.

A second practical convenience is that the set-up time for each subject is substantially reduced. When there are complicated instructions or other complex arrangements necessary for data collection (e.g., hooking up physiological recorders), it is desirable to have the subject provide as much data as possible relevant to the entire study. In a within-subject design, each person accounts for multiple observations, substantially reducing the burden of measuring a number of individuals a single time.

Sequence and Order Problems. The choice of within-subjects factors does produce additional problems with message sequence. Note, however, that some

sequence problems are present in any experiment. If more than one message example is used for each condition in a between-subject design, order must be accounted for because effects of practice and fatigue are still possible. However, the increased number of messages in a within-subject design, and the possibilities of sensitization and carry-over effects, makes the control of presentation order more important.

There are three strategies for dealing with order effects. The first concerns skill and facility in responding to criterion measures and is the simplest: Give people time to acclimate to the experimental situation and let them practice the tasks. Many of the measures reviewed in this volume require tasks that are unusual. If people are given a chance to use response apparatus, get used to computers and video gear in the room, acclimate to recording devices, and practice tasks, their response are likely to be a better indicator of their true responses.

A second strategy is also relatively easy, and often overlooked. Order effects are diminished to the extent that stimuli can be separated psychologically. The most obvious and simplest strategies are to maximize the time between presentations, or to include distraction tasks that disrupt stimulus sequences, and increase psychological distance between repetitions. Order effects in processing studies are often the result of stimuli presented close together in a sequence. The greatest sensitivity to prior material occurs the first several seconds after one presentation ends and another begins and becomes less of a factor after about thirty seconds (Geiger & Reeves, 1993a). Simply allowing time between presentations will diminish the mental activity carried over from prior trials. Better still, subjects can be required to perform a task relevant to the study (e.g., an evaluation of the prior message) or an irrelevant task (e.g., counting backwards by intervals of 13) that will allow more time between presentations, and will interfere with the thought processes that might promote contamination between trials and treatments.

The third and most difficult strategy is to counterbalance presentation order in the actual stimulus sequence. This requires the construction of a number of different presentation orders and assignment of different people to different sequences, each of which represents the same treatment level. This insures that responses are not confounded with sequence effects, because the same message is located in different positions across the presentation orders. If there is an influence of sequence, it will be equally distributed across all levels of the treatment (interactions between sequence order and treatment are not controlled, however).

The need for counterbalancing is certainly not a radical suggestion. Almost all experiments with multiple presentations attempt this control. There is less agreement on how to achieve counterbalancing, however, and the method chosen can have important practical and statistical implications. The simplest case is a two-level treatment with a single message at each level. There are two possible

orders, and each can be inexpensively created through videotape editing or collating, and included in the design. Systematic error is evenly distributed between the treatment levels, and order can even be examined as a separate factor if it represents interesting variance. Counterbalancing becomes considerably more complicated (and time consuming) if the treatment has three or more levels, or many examples within each treatment level, because the number of possible orders increases exponentially. Decisions at this point should balance the need to control order with the practical problem of creating stimuli, a time consuming process with videotape.[5]

The most common solution for insuring the independence of treatment by presentation sequence is the latin square method. This procedure efficiently limits the number of orders, and it has been used effectively in media experiments (Slater, 1990; Geiger & Reeves, 1993b), yet it is hardly a standard technique in communication. The latin square method is recommended more as a way of restricting the number of orders rather than as a formal design. If the columns and rows of a latin square are used as separate factors (factors that represent order and time of presentation), it is difficult to examine interactions between other factors in the experiment (Calfee, 1985). Consequently, responses to a particular message are usually averaged across their locations in the different orders, and not examined as a separate factor.

It is critical to note that the latin square strategy only distributes error associated with the order in which treatment levels are presented. Based on previous arguments, it is likely that there also will be (or should be) several messages within each level. This presents yet another sequence problem: the time related to the effects of contiguous messages within the same condition. There are now two sequencing problems: one for the order of treatment levels, and one for the sequence of materials within each condition. A Greco-Latin square strategy addresses the need to randomize sequences and messages, but this design totally precludes examination of interactions. It is still possible, however, to use the Greco-Latin square strategy to sample sequences and materials, but then aggregate across the different presentation versions to obtain values for each message. In an experiment with positive, negative, and neutral messages, for example, this randomizes the sequence of positive messages as well as the serial position of positive ones relative to the other two conditions.

When treatment levels and message examples within each treatment are balanced, it is probably best to mix the two within a single presentation. If there are

[5]When video or film is used to present full motion sequences in experiments, it is generally feasible to compile only a few different sequences of those that are possible. New technology (e.g., videodisc, compact disc) offers the ability to totally randomize presentation of several segments, because final sequences need not be compiled. The sequences are played in random order by allowing the videodisc player or computer to quickly cue up each segment as the material is presented.

four commercials in each of three emotion categories, for example, a sample of the different possible sequences of 12 messages (a sample of two or three will probably suffice) can be edited without maintaining a break between the conditions. This will also make it more difficult to "figure out" experiments, because the treatment levels are not preserved as a group of messages. This strategy is difficult, however, if there are manipulations (in addition to message variance) that require all repetitions to be run as a group; for example, viewing several messages at three different distances from a television set. For a within-subject design in this case, it would be most convenient to have people view all the different messages at one viewing distance, and then move the chair to another location.

Message Subset Designs. A final option combines elements of within- and between-subject strategies. In a typical randomization of treatment orders and message sequences, the different versions of stimuli within a treatment level represent a sample of the possible sequences for a given number of messages. It is also possible to consider these versions to be different samples of a larger group of available messages. This situation could occur, for example, for the researcher who is fortunate enough to have six good examples of messages in each of three conditions, but time enough to show only three per condition to each viewer. The question is whether it is better to select just three of the six messages in each condition for presentation in a number of different sequences, or to use different three-message samples for each subject. For many questions, the second option represents an additional opportunity to insure that conclusions are representative of a greater range of messages.

The option of using subsets of messages within each condition can be combined with message altering to achieve a design that is particularly sensitive to between-message variance. Treatment effects in this design would have higher external validity with respect to a larger message population. When message subsets are used, treatment variance is usually achieved by sampling messages within each condition. For example, six commercials could be sampled from a larger population in each of two emotion categories, negative and positive. In this case, the messages in each condition are different, precluding any message by treatment confounds.

A more extreme example of this strategy would include altering each message to represent all treatment levels. For example, Group 1 would see Message 1 in a positive version and Message 2 in a negative version, with Group 2 exposed to a negative Message 1 and a positive Message 2. If resources allow (i.e., it is a tough editing job), the message subsets could be altered for each subject or at least for two or three different groups of subjects. This design has the advantage of retaining some benefits of a within-subject factor (each person is in each condition) and of reducing the variance between messages in the two conditions (altered versions rather than sampled messages represent each treatment level). It

is important to note that in adopting this mixed design, the statistical analyses can become quite complicated.

SUMMARY

Three general strategies for assigning subjects to messages have been discussed: between-subject assignments, within-subject assignments, and assignments to message subsets. The last strategy, message subsets, could include both assignments to different sequences of messages within a condition, or assignment to different subsamples of messages from a larger pool. When these three strategies are crossed with the two message options discussed (message altering and message sampling), six basic experimental strategies are formed. Combinations of these designs in multivariate studies give many more possibilities, but we believe these summarize the most useful single-factor media experiments with multiple messages.

It is impossible not to wonder with great fascination about the outcome of six research projects, each aimed at the same theoretical question, and armed with the same cognitive measures and budget, but each pursuing only one of the six strategies. For all but the most obvious questions, we think there would be some disagreements. We doubt, however, that the different results would be theoretically competitive. More likely, some designs would show significance and others would not. We think that the within-subject designs have the greatest potential to uncover differences that will hold up across the greatest range of material.

We close with a comment about replication, a cherished concept in the philosophy of science, but a neglected feature of communication research programs. The single most important competitive difference in designs will likely be between those with single (or small) message samples and those with well planned message samples. We advocate designs that allow for replication within a single research effort, a far easier prospect than replicating entire studies. It is this replication that offers the best prospect to account for the complexity of media stimuli and to continue refinement of psychological definitions of media.

REFERENCES

Bradac, J. J. (1983). On generalizing cabbages, messages, kings, and several other things: The virtues of multiplicity. *Human Communication Research, 9,* 181–186.

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.

Calfee, R. (1985). *Experimental methods in psychology.* New York: Holt, Rinehart & Winston.

Clark, H. H. (1973). The language as fixed-effects fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12,* 335–359.