Stimulus Sampling and Social Psychological Experimentation

Gary L. Wells
Iowa State University
Paul D. Windschitl
University of Iowa

*The authors discuss the problem with failing to sample stimuli in social psychological experimentation. Although commonly construed as an issue for external validity, the authors emphasize how failure to sample stimuli also can threaten construct validity. They note some circumstances where the need for stimulus sampling is less obvious and more obvious, and they discuss some well-known cognitive biases that can contribute to the failure of researchers to see the need for stimulus sampling. Data are presented from undergraduate students (*N = *106), graduate students (*N = *72), and psychology faculty (*N = *48) showing insensitivity to the need for stimulus sampling except when the problem is made rather obvious. Finally, some of the statistical implications of stimulus sampling with particular concern for power, effect size estimates, and data analysis strategies are noted.*

Suppose that you read an article that stressed the importance of the gender of the experimenter, such as the following:

> The results support our chivalry hypothesis that men are more courteous to women than to men. Male participants were much more likely to return their money to the female experimenter than they were to the male experimenter.

Or, suppose that you read an article on eyewitness identification that stated the following:

> White participants witnessed a staged robbery in which the perpetrator was either Black or White. Confirming our prediction, participants tended to misidentify the Black actor more than they did the White actor. We conclude that White eyewitnesses have more difficulty identifying Blacks than they do identifying Whites.

Or, suppose that you reviewed a manuscript on gender and persuasion and read the following:

> The participants in this experiment viewed a videotape of either the male or the female giving the scripted speech. Because the only difference in the two conditions was the gender of the speaker, and participants were randomly assigned to speaker, we can conclude that any significant differences in attitude change were due to the gender of the speaker.

These three examples represent a serious problem that plagues a surprising number of experiments in the social psychological literature. What is surprising is that the authors of studies such as these do not recognize the problem and, perhaps more surprising, the reviewers and editors allowed conclusions based on designs of this sort.[1] These are cases in which we would argue that the functional sample size in the experiment is $n = 1$, regardless of the number of participants that the experimenter runs in the experiment. In effect, both the obtained means and the error variance estimates are of no real value in these three examples. The critical missing feature of these experiments is stimulus sampling.

Stimulus sampling refers to the use of multiple instances of a stimulus category in research. The need for stimulus sampling exists whenever individual instances in the category potentially vary from one

Authors' Note: We thank Brad Bushman, Susan Cross, Veronica Dark, and David Kenny for comments on an earlier version of this article. Correspondence concerning this article should be addressed to Gary L. Wells, Department of Psychology, Iowa State University, Ames, IA, 50011; e-mail: glwells@iastate.edu.

PSPB, Vol. 25 No. 9, September 1999 1115-1125

another in ways that might be relevant to the dependent measure.

Commonly, stimulus sampling is treated as an issue of external validity in which the question is whether the results can be generalized across other participants, stimuli, times, settings, and so on. Here, we emphasize how failure to sample stimuli can threaten construct validity. Construct validity is threatened when "the operations which are meant to represent a cause or effect can be construed in terms of more than one construct" (Cook & Campbell, 1979, p. 59). We note that such a situation usually applies whenever a single stimulus instance from one category is used to represent one condition of an experiment and a single stimulus instance from some other category is used to represent another condition of the experiment. The use of only one stimulus to represent a category can confound the unique characteristics of the selected stimulus with the category. What might be portrayed as a category effect could in fact be due to the unique characteristics of the stimulus selected to represent that category. In the earlier example, the use of an individual male experimenter and an individual female experimenter confounds the unique properties of the individual experimenter with the gender of the experimenter.

Many of the best thinkers regarding statistical issues in psychological science have discussed stimulus sampling (e.g., Brunswik, 1947; Campbell, 1960; Campbell & Stanley, 1966; Clark, 1973; Cook & Campbell, 1979; Cronbach & Meehl, 1955; Kenny, 1985; Maher, 1978; Rosenthal & Rosnow, 1991). Our analysis, however, is unique from these other treatments in several ways. First, previous treatments have cast the problem as being primarily a problem for stimulus generalization. It is indeed a problem for generalization, but, as we describe, the issue is also a fundamental one for construct validity. Second, previous treatments have focused almost exclusively on the issue of how to analyze data when there has been appropriate sampling of stimuli (see Kay & Richter, 1977, for an exception). We think that the statistical analysis problem has been fairly well resolved (e.g., see Kenny, 1985; Richter & Seay, 1987). Our concern, on the other hand, is with the question of when stimulus sampling is needed, why it often goes as an unrecognized problem in research, and how it serves as a threat to construct validity.

Although the problem of failing to sample stimuli can occur in any area of research, we see the problem as particularly likely to occur when people are used as stimuli, as most of our examples illustrate.[2] One or two individuals are sometimes used as the stimuli to represent their gender category, their racial group, or a category of attractive people, for instance.

It is important to note that we are not concerned here with experiments in which a particular instance from a category is used as a stimulus but the stimulus itself is not the manipulated variable. For instance, a researcher might design a levels-of-processing study in which participants process a single stimulus face by making either self-relevant judgments about the face or other-relevant judgments about the face. In this case, the manipulation is the self-relevant versus other-relevant judgments, not the particular instance of the face. Although there might be concerns about external validity in the sense that we might want to know how well this effect would generalize across a broader sample of faces, the particular face selected in this study is not confounded with the experimental conditions because the same face(s) appear in all conditions. Our concern about stimulus sampling is better exemplified by a study in which a characteristic of the stimulus face (e.g., gender or race) is treated as the manipulation of interest.

Shades of Obviousness

Like many problems, there are occasions in which the existence of the problem is obvious even to the casual observer. Consider one hypothetical example in which the stimulus sampling problem is obvious. Suppose that a researcher hypothesizes that people with polysyllabic given names are perceived to be more intelligent than those with monosyllabic given names. The researcher has participants evaluate an essay and tells them that the author's name is Fred (monosyllabic) or tells them the author's name is Mirajoul (polysyllabic). Clearly, we would not even consider the idea that we could test this hypothesis by having participants evaluate the intelligence of the author based on only one example of a particular monosyllabic name versus a particular polysyllabic name. In this case, the specific instance (Fred vs. Mirajoul) manipulates the focal variable (number of syllables) in a way that is potentially confounded with ethnicity or other unknown properties of the particular names chosen. And yet, this is precisely what the researchers are doing in the three examples that we gave in the opening of this article. In those cases, the researchers were testing whether a particular male is more or less persuasive than a particular female, or whether a particular Black person is easier to recognize than a particular White person, or whether a particular male experimenter has a different impact than a particular female experimenter. In each of these examples, the broader construct (e.g., gender, race, number of syllables in given names) is operationalized by use of a single instance that may or may not represent the central tendencies of the population of instances that are contained in the construct. It would be absurd to assume, for instance, that the name Mirajoul represents the central

tendency of all polysyllabic names in perceived intelligence. It seems less obvious to people that a particular male may be unrepresentative of the persuasive impact of males in general or that a particular Black actor in a staged crime eyewitness identification experiment might be unrepresentative of the recognizability of Black people in general.

Consider some less obvious examples. A researcher proposes that people are more likely to be influenced by misleading questions for peripheral information than for central information regarding a scene that they witnessed on slides. After viewing the scene, some participants are asked a misleading question about a central feature, such as the actor, or about a peripheral feature, such as a pop can. Others are asked neutrally worded questions. Suppose that the effect of the misleading question occurs only for the peripheral feature and not for the central feature. The author concludes that central information is more resistant to the misleading question effect than is peripheral information. Consider another example in which a researcher proposes that people give more personal space to males than to females. To test this, people at a mall had to walk by either a male or a female who was standing in the middle of an aisle. Hundreds of mall patrons were observed in terms of how much room they gave the male versus the female as they passed the stimulus person in the aisle. On average, they gave the male an additional 12 centimeters of distance, an effect that was both highly significant and had a large effect size.

The failure of stimulus sampling and the extent to which it seems to be a problem in each of these two examples might seem less obvious than the hypothetical Fred versus Mirajoul experiment. And yet, they can be made equally obvious by restating the findings. For instance, we could say that it was easier to mislead the participant about a pop can than it was about a person. We could say that people gave more room to Stan as they walked by him than they gave to Mary. Stated this way, it becomes more obvious that we should be concerned about stimulus sampling in these cases.[3] Perhaps one way to increase the obviousness of the problem in experiments that report experimenter gender effects would be to require the authors to refer to the experimenters by name rather than refer to the gender category. Imagine an article, for instance, in which the author says, "Participants were more reactive to the bad news from our experimenter Albert than they were from our experimenter Trixie." We suspect that the researcher who cast such a sentence would then be reluctant to say, "Hence, participants react more to bad news from males than from females.

Typicality judgments. One stimulus characteristic that is likely to make the need for stimulus sampling less obvious is the appearance of typicality or class resemblance for the stimulus selected. For instance, if one were testing the hypothesis that background rock music interferes with learning more than does background classical music, it might seem acceptable to compare the Rolling Stones with Beethoven because they resemble or seemingly typify these categories of music. However, this use of the representativeness heuristic to make judgments about the lack of a need for stimulus sampling may be just as questionable as making use of the representativeness heuristic for making judgments of probability (as demonstrated by Tversky & Kahneman, 1971). In this case, neither the Rolling Stones nor Beethoven may be particularly representative of their respective categories because both are closer perhaps to their ideals than they are to the central tendencies of their categories. Representativeness judgments of this sort are insensitive to base rates and other properties of statistical distributions.

Ease of imagining within-category variation. The extent to which the need for stimulus sampling is more obvious or less obvious seems to be related to the ease with which people can imagine variation across instances within a category. When variation across instances within a category is naturally salient or is made salient, the need for stimulus sampling seems obvious. When variation across instances within a category is not salient, the need for stimulus sampling is less obvious. Of course, when the actual variance across instances within a category is low, there is in fact less need for stimulus sampling. But the perception of variation is based on heuristic psychological processes such as availability, representativeness, stereotyping, and related processes that can often bear little resemblance to actual variance (Kahneman, Slovic, & Tversky, 1982). We know, for instance, that people often overlook differences between individuals within groups and exaggerate differences between the groups themselves (Taylor, 1981; Wilder, 1978). Hence, we might expect that groups or categories that are well formed or stereotyped would be especially likely to "hide" the need for stimulus sampling. Perhaps this is why the use of one male and one female, or one Black person and one White person, is so common in studies purported to show effects of stimulus gender and race. We do not think it heretic to suggest that researchers, ourselves included, are subject to the same biased cognitive processes that psychological researchers attribute to other people. The fact that these psychological biases affect even the psychologists who have written about them is somewhat ironic testimony to the very strength of the biases.

One difference between the "names" experiment and the "gender persuasion" experiment is that people do not have a natural, a priori grouping for polysyllabic ver-

sus monosyllabic names but they do have a natural, a priori grouping for males versus females. Hence, it readily occurs to someone to question whether you can test the polysyllabic names hypothesis with a comparison between only the names Fred and Mirajoul, but it seems less likely to occur to someone that you cannot test the gender question using only one instance from each category.

If a researcher explicitly considers the possibility that there is considerable variance across stimuli within a category, then the need for stimulus sampling becomes more obvious. If we consider, for instance, that mall patrons would walk closely to some males and give other males more distance, then we would have to confront the question of how to select the particular male who would represent the category of males. Lacking any particular way of knowing how to select the average male or how to select the male who can represent the category of males, it becomes clear that one should sample a number of males to use as stimuli for the mall patrons.

A few years ago, one of the current authors noticed that there were some glaring inconsistencies in an emerging literature on the credibility of children as trial witnesses. Some studies were showing that young children were perceived by participant-jurors to be less credible eyewitnesses than were adult eyewitnesses, other researchers were showing no differences, and still others were showing children as young as 6 to 8 years old to be perceived as more credible than their adult counterparts. An examination of the methods used in these studies proved informative. The standard paradigm involved having either a child or an adult give scripted testimony, which was then videotaped and shown to juror-participants whose task was to evaluate the credibility of the eyewitness. Because the content of the testimony was controlled by the script, the authors of these studies reasoned that any differences in perceived credibility must be due to the age of the eyewitness. Although previous research had already demonstrated vast individual differences in the perceived credibility of eyewitnesses of the same age (e.g., Wells, Lindsay, & Ferguson, 1979), these researchers used only one instance (in some cases two) of an adult or child eyewitness to serve as the stimulus representative for the entire age category. In an attempt to see what happens when one uses samples of children and adults as testimony stimuli, Wells, Turtle, and Luus (1989) found that the variation across stimulus-persons within an age category greatly exceeded differences across age categories.

Figure 1, taken from Wells et al. (1989), is particularly instructive. Figure 1 is a scatter plot of the perceived testimony credibility of 14 8-year-old, 14 12-year-old, and 14 adult eyewitnesses who served as stimuli. Each of the 42 data points in Figure 1 is a mean that is based on 7 different participant-jurors who evaluated a particular eyewitness (yielding a total of 294 participant jurors). An important characteristic of these data is that there is considerable variance from one 8-year-old to another, one 12-year-old to another, and one adult to another. Individual attributes of the witnesses clearly have more effect on perceived credibility than do the ages of the witnesses. These individual attributes affecting perceived credibility undoubtedly include the perceived confidence of the witness (e.g., Wells et al., 1979), as well as other variables such as tone of voice, apparent verbal fluency, and apparent genuineness. As a result, the distributions of means overlap across age such that any randomly selected 8-year-old stimulus person has about a 50/50 chance of being perceived as more credible or less credible than any randomly selected adult. Little wonder that some researchers found their child stimulus to be more credible than their adult stimulus and other researchers found the reverse. All that these studies were really testing was whether the particular child that was used in their study was more or less credible than the particular adult that was used.

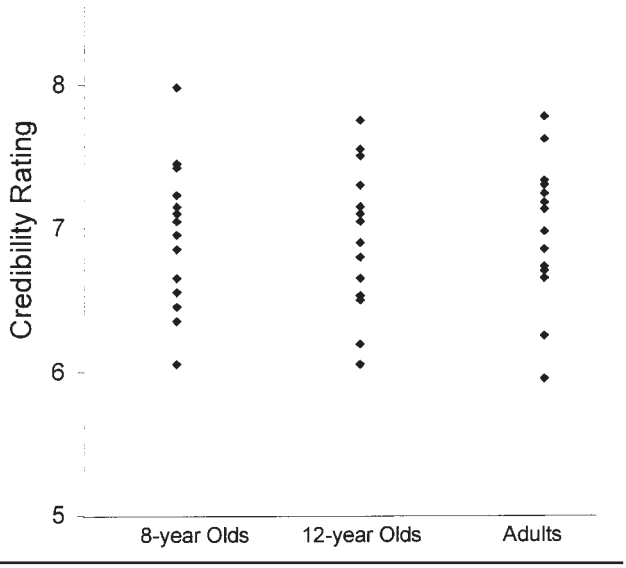

Figure 1 Scatterplot of means for perceived credibility of samples of 8-year-old, 12-year-old, and adult eyewitnesses.
SOURCE: Wells, Turtle, and Luus (1989).

Apparently, age seems to be one of the grouping variables that can lead a researcher to not think much about differences among people within ages and to instead treat age groupings as though there were homogeneity within the group. Somewhat ironically, neglecting heterogeneity across instances for a stimulus category (such as gender, age, and race) may be an especially strong propensity for the researchers themselves because their hypotheses are focused on the differences between stimulus categories rather than on the variation of

instances within categories. We have yet to see a study on experimenter gender effects, for instance, that discusses differences between experimenters within gender.

Evidence That Researchers Can Fail to Detect the Stimulus Sampling Problem

We believe that most psychological researchers understand the concept and need for stimulus sampling but commonly fail to recognize its need in a given study except under relatively obvious conditions. One way to make the problem more obvious is to use the exemplar label rather than the category label to describe the condition. We conducted a simple e-mail experiment to test this hypothesis.

We e-mailed 80 psychology faculty, 110 psychology graduate students, and 150 undergraduate psychology students asking them to evaluate the validity of the conclusion reached in a simple psychology study. Three large public universities were used, none of which were affiliated with the current authors, each of which have graduate programs in psychology. Participants were sent one of two versions of the fictional (but allegedly real) study. The category-label version is printed below and the portions in brackets were substituted for the italicized portions in the exemplar-label version.

> Researcher Tim Stang, *along with his male accomplice and female accomplice* [along with his male accomplice Frank Miller and female accomplice Judy Tendore], took their research on the road across the United States to find out if men and women get the same help when asked for directions. Large cities, such as New York, Los Angeles, and Chicago, and small towns, such as Newton (Kansas), Clinton (Pennsylvania), and Johnston (California), produced the same findings time and again. At the flip of a coin, people were asked directions by either *the male accomplice* [Frank Miller] or *the female accomplice* [Judy Tendore]. When people on the street were asked for directions by *the male accomplice* [Frank Miller], 88% gave directions. When *the female accomplice* [Judy Tendore] asked them for directions, only 57% gave directions, the remainder usually indicated that they should ask someone else. "We were surprised to find that this failure to give directions to a woman was just as strong for the females we approached as it was for the males." In total, the researchers approached 1,020 people, half approached by the female accomplice, half approached by the male accomplice. After each approach, Stang then reapproached each person and ascertained whether they actually knew the location of the place they were asked about. "Our direction questions were simple and 90% knew the answer, regardless of whether they were approached by *the male or the female* [Frank or Judy]," noted Stang, "For some reason, people withhold directions from females more than they do from males." Stang would not speculate on the underlying reasons for the results except to note that "it clearly indicates that the gender of the person asking for directions has an effect on the likelihood that someone who knows the answer will actually give them the information."

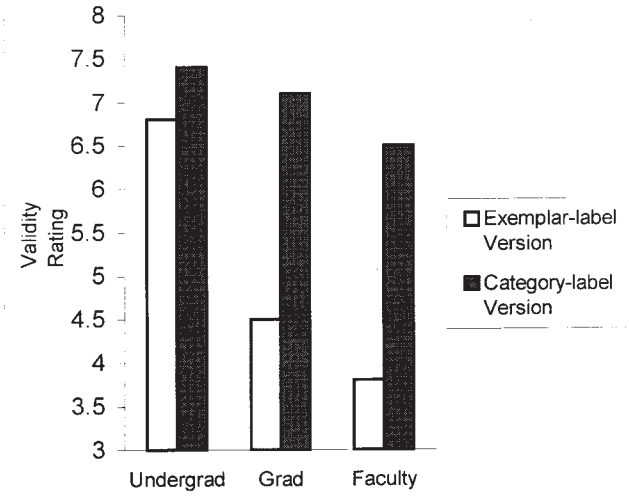

Figure 2 Ratings of the validity of the conclusions as functions of category-label versus exemplar-label conditions by participant population.

Participants were then asked to rate the following on a scale from 1 (*not at all valid*) to 11 (*completely valid*), "To what extent can the conclusion by Stang, as stated in the last sentence, be considered a valid conclusion based on this study?" In addition, participants were asked, "Are there any apparent flaws in the research methods used by Stang that could prevent the conclusion he stated from being valid? (Please list briefly)."

Validity ratings. We received 48 responses from faculty members, 72 from graduate students, and 106 from undergraduate students. Return rates did not differ as a function of whether the category or exemplar label was used. Responses to the 11-point validity measure are shown in Figure 2. The 3 (undergraduate, graduate, faculty) × 2 (category vs. exemplar label) interaction was significant, $F(2, 220) = 6.33$, $p < .01$. By inserting the names of the specific male and female who asked for help (exemplar label), both the graduate students and the faculty reduced their ratings of the validity of the conclusion, Fs(1, 220) = 5.11 and 7.93, respectively, ps < .01, but the undergraduates did not display a significant decrease in perceived validity of the conclusion when the names were used $F(1, 220) = 1.43$, *ns*.

Identification of stimulus sampling flaw. Analyses of the listing of flaws supports the idea that the problem with stimulus sampling became obvious to the graduate students and the psychology faculty when the exemplar labels were used.[4] The percentages of participants

mentioning the failure to use other males and females in addition to Frank and Judy is shown in Figure 3. None of the undergraduates, fewer than 10% of the graduate students, and fewer than 15% of the faculty mentioned this flaw in the category-label version. When the exemplar-label version was used, however, 8%, 46%, and 63% of these participants mentioned the problem of failing to use other males and females in addition to Frank and Judy in the study. The interaction is especially important to these results because it indicates that making the stimulus sampling problem "obvious" is not itself sufficient to yield recognition of the problem; recognition of the problem, even in the exemplar-based version, was largely confined to faculty and graduate students. There is reasonable evidence to indicate that graduate training in psychology leads to enhanced reasoning abilities (see Lehman, Lempert, & Nisbett, 1988), suggesting that this effect might be due to formal training rather than a selection process regarding who goes to graduate school in psychology. However, if faculty and graduate students were trained to be sensitive to stimulus sampling, why didn't they notice it in the category-label version of the problem?

In certain respects, the results of this study are a bit disturbing. Although we are pleased that psychology faculty can recognize a stimulus sampling problem when undergraduate students could not, 92% of the faculty did not recognize the problem in the category-label condition even though it was quite clear from the description that there was only one male and one female. Furthermore, their task was to look for problems. This suggests to us that the failure to sample stimuli is not a problem that psychologists are usually looking for. Furthermore, we are not consoled by the fact that the majority of faculty detected the problem in the exemplar-label version because we see this version as much more obvious than anything we see in the published literature. The published literature virtually always uses the category-label style of writing, which is the version in which the faculty performed about as poorly as the undergraduates.

Using Two or Three Stimuli

In the exemplar-label condition of our study, those who mentioned the need for additional males and females usually did not mention how many more females or males should be selected or how they should be selected. When a specific number was mentioned, however, it was almost always one or two. For instance, one faculty member said, "I'd like to see the effect replicated with another male and female." Hence, even when researchers recognize the need to have more than one stimulus represent the category, they somehow decide that only two or perhaps three such stimuli would be sufficient. Consistent with our speculation that well-known cognitive biases may be operating on researchers just as they operate on the person on the street, we suspect that this strategy seems sufficient because of the representativeness heuristic. One of the apparent consequences of the representativeness heuristic is a belief in the law of small numbers (Tversky & Kahneman, 1971). In general terms, this is a tendency to think that relatively small samples are representative of the parent population from which they are drawn. Clearly, research psychologists know not to make this error in reasoning when sample size refers to the number of participants. When sample size refers to the number of stimuli within a stimulus category, however, the idea that one might need only two or three exemplars seems especially appealing.

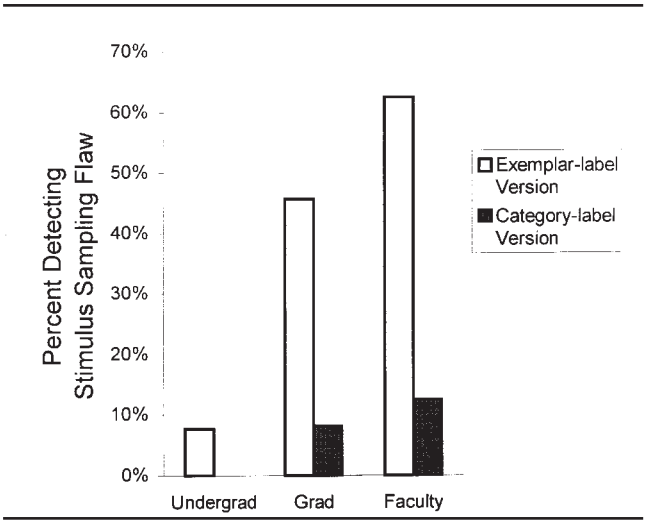

Figure 3 Percentage who detected the stimulus sampling flaw as functions of category-label versus exemplar-label conditions by participant population.

Studies of the cross-race recognition effect, gender and persuasion, or age and credibility, for instance, might use two or three stimuli to represent the category. We concur that this is a better approach than using only one stimulus, especially if there is some type of strategy articulated by which we might assume these two or three cases to be typical of the category. But the selection of only two or three stimuli hardly satisfies the need for broader stimulus sampling. Consider the age and credibility data in Figure 1. When sampling only two 8-year-old children and two adults under the null hypothesis of no true population differences between adult and child eyewitnesses, there are six possible patterns (i.e., both children more credible than both adults selected, both children less credible than both adults, or one of four possible mixed cases), each being equally likely. Two of these patterns involve a clear direction favoring either the children or the adults. Hence, there is a one third chance that a true null difference in the population of

stimuli will yield an ordering of the means favoring both children over both adults or vice versa. As long as one uses enough participant-jurors to detect these differences, the differences will be statistically reliable even though there are no actual differences in the populations.

Three Types of Error

The child versus adult credibility study makes it clear that Type I error (rejecting the null when there is not true difference) can balloon when one or two stimuli are used to represent a category. Somewhat less obvious perhaps is the fact that Type II error also can balloon. Even less obvious is a type of error that we call diametrical error.

Type II error. Figure 4 uses the same data as Figure 1 except that we have arbitrarily shifted up the distribution for the adult participants and we have dropped the 12-year-olds' data. Assume for current purposes that the samples closely approximate the true distribution for the populations from which they were drawn. In this case, there are true differences between the populations, but, of course, the distributions overlap. If we use only one stimulus person from each age group, we have no basis for knowing whether the selected stimulus is close to the central tendency of its parent population. Under these conditions, there is some chance, which cannot be calculated if we use only one or two exemplars, that we could select stimuli that fail to show any difference between the categories even though there are true differences in the two populations. This situation is indicated by selection of Child A and Adult C (or by Child B and Adult D) in Figure 4. This illustrates how the failure to sample stimuli can lead to Type II error (i.e., concluding that there are no differences when in fact the null is false). The usual way to avoid Type II error is to increase the power of the experiment through such things as increasing the number of participants in each condition. Notice, however, that Type II error in this case is not attributable to lack of power in the traditional sense. The cause of Type II error in this case is attributable to the stimuli that were sampled. If Child A and Adult C are selected as the stimuli, even a sample of 1,000 participants per condition is unlikely to produce a significant difference.

Diametrical error. Perhaps more interesting yet is the type of error noted in Figure 4 by the selection of Child A and Adult D. Here, there is a true difference between the population of stimuli in the two categories, but the stimulus that is sampled from the lower distribution (8-year-olds) is drawn from the upper portion of its parent population, whereas the stimulus that is sampled from the higher distribution (adults) is drawn from its

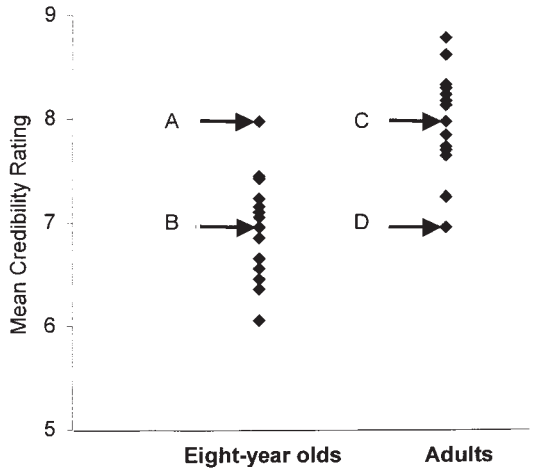

Figure 4 Scatterplot of means for perceived credibility of samples of 8-year-old and adult eyewitnesses with main effect favoring adults.

lower portion. We are unaware of any term for this type of problem, so we call it diametrical error. Diametrical error is different from Type I error or Type II error. With Type I error, the null hypothesis is true but the null is falsely rejected. With Type II error, the null is false but is not rejected. With diametrical error, the null hypothesis is false and the null is rejected but the direction of the true population difference is opposite to the obtained data. Diametrical error is highly unlikely under conditions where there is no stimulus sampling issue because inferential statistics and alpha levels control reasonably well for chance. However, under conditions where there is a clear need for stimulus sampling and a researcher selects only one stimulus per category, there is no control for chance in selecting the stimuli and diametrical error might not be uncommon. The greater the overlap between the two stimulus distributions, the greater the chance that diametrical error will occur.

In the case of Figure 4, there is an approximate 50% overlap in the distributions of stimuli. For purposes of mathematical convenience, assume that this is a uniform distribution and that a single stimulus person was randomly drawn from the 14 8-year-olds and another randomly from the adults. Under these conditions, there would be a 10.7% chance of a diametrical error in which the 8-year-old stimulus person selected is more credible than the adult selected.[5] If the overlap were 71%, which is less overlap than the overlap of gender height distributions, then the chances of diametrical error would rise to nearly 23%.[6] These are rather large chances for finding a significant difference in a direction that is precisely opposite to the actual population difference. Notice how this can render the traditional interpretation of a *p* value

or an effect size estimate meaningless. Diametrical error can produce huge effect size estimates owing to overlap of the stimulus distributions. In fact, it is almost ironic to note that increasing the power of the experiment (e.g., by increasing the number of participants) actually increases the chances of diametrical error when only one stimulus is used to represent a category.

Interactions

One of the circumstances that seems to lessen researchers' concerns about the absence of stimulus sampling is when the researcher predicts and then obtains a statistical interaction. Consider, for instance, the personal space / mall study we described earlier. Suppose that the researcher obtained an interaction such that the male participants in the mall gave more distance to the male stimulus person than to the female stimulus person, whereas female participants in the mall gave equal distance to both stimulus persons. Many researchers seem to treat this interaction as though it nullifies any concerns about stimulus sampling. The reasoning seems to be that any unique characteristics of the particular male or particular female who was chosen as the stimulus person should have affected the male and female participants equally.

We suspect that the lure of the interaction argument is particularly strong because it resembles a more valid argument that is commonly used with regard to interactions. In many cases, theoretical concerns about how a variable is manipulated can be diminished by predicting and obtaining an interaction between that variable and some other variable. For instance, participants who are instructed to make personality judgments about a face tend to be better able to later recognize the face than are participants instructed to make physical feature judgments while viewing the face. Because the instruction manipulation fails to control for a number of possible confounding differences (such as effort in viewing the face or amount of time spent processing the face), it is difficult to interpret the instruction main effect. However, this instruction manipulation has the reverse effect on the accuracy of verbal descriptions of the faces (i.e., physical feature judgments are superior to personality trait judgments for the description test). The interaction between instruction and test diminishes our concern that the instruction manipulation merely increased effort or attention paid to the faces (Wells & Turtle, 1988). Hence, the presence of an interaction can sometimes lessen or even nullify one's concerns about how a variable is manipulated, depending on the nature of the hypothesis being tested.

This tends not to be the case with regard to failures to sample stimuli. Returning to the personal space / stimulus gender study, we agree that it is harder to think of why male participants gave this particular male stimulus greater space than they gave this particular female stimulus, whereas female participants did not. Nevertheless, in principle, we don't see why explaining the interaction with respect to the gender of the stimulus person should be any different from explaining the interaction in terms of the unique qualities of the particular person selected to represent that gender. Of course, in experiments such as these, the reader is not armed with information about the unique qualities of the individual stimulus but are simply told of the gender of the stimulus. Hence, we have no basis for focusing on anything other than the gender explanation for the interaction. It remains the case that we don't know how male participants versus female participants would react to a broad sample of male and female stimuli. All we know is that males gave Stan more space than they gave to Mary, whereas females gave Stan and Mary equal space. Unique features of Stan and Mary remain confounded with the stimulus gender variable regardless of whether an interaction is observed involving Stan and Mary.

Effect Size Estimates

Researchers might assume that there was little need for stimulus sampling because of the large effect that they were able to obtain. The reasoning seems to be that some observed effect, *X*, must be attributable to the construct invoked because the effect was so large. The onus of explanation is then shifted to the critic who cannot quite come up with a coherent or persuasive account of such a large effect based merely on the unique properties of the stimulus that was used. Once again, however, the critic has little or no information about the individual stimulus that was used (other than, for example, that the stimulus was male, or Black, or 8-years-old) because the researcher is describing the stimulus only in terms of the hypothesized general category. Returning to Figure 1, we could note that a chance selection of the most credible 8-year-old as the stimulus and the least credible adult as the stimulus would yield a whopping effect size estimate favoring the 8-year-old under conditions in which there are no true population differences in credibility.

A single stimulus or exemplar of a category might be considered an acceptable operationalization for the category to the extent that it represents the central tendency of the population of stimuli in that category. Assuming that the sample data in Figure 4 approximate the distribution of eyewitness credibility ratings for a population of 8-year-old and adult eyewitnesses, for instance, we might consider the stimulus person closest to the mean or the median within each age to be an acceptable exemplar who might in some sense represent the age group as a whole. Selecting Child B and Adult C (see Figure 4) would satisfy the central-tendency crite-

rion for selecting a single stimulus. Unfortunately, it is difficult to know where any particular stimulus falls in the stimulus distribution for a category unless one samples stimuli and defines the relevant dimension(s) to be measured. This is likely to require as much or more effort and cost than simply sampling stimuli within the experiment itself.

Data Analysis Issues

It is not our purpose to delve into the statistical methods for analyzing data in which stimulus sampling is a part of the design. We believe that the technical procedures for analyzing such data are well described in the literature (e.g., see Kenny, 1985). We do think it is important, however, to understand the basic idea from a general statistical perspective.

The data in Figures 1 and 4 help to make clear our earlier claim that the use of one stimulus to represent a category can be construed as functionally equivalent to conducting an experiment with a sample size of $n = 1$. The typical way for many researchers to think about sample size is to consider sample size to be equivalent to the number of participants. Similarly, they think of error variance only as the variance across participants within a condition. In the case of the data in Figure 1, sample size is better construed as the number of children and adults who served as stimuli, and error variance is the variance across these stimuli within age groups. Hence, the data in Figure 1 could be analyzed with the sample stimuli as the unit of analysis ($n = 14$ per category, 42 total) rather than participant-jurors ($n = 7$ per stimulus, 98 per age category, 294 total) as the unit of analysis (see Wells et al., 1989).

A more powerful approach to analyzing data in which stimulus sampling is used was suggested by Kenny (1985). Kenny recommends treating stimuli as a random factor in the analysis and calculating quasi-*F*-ratios (by taking linear combinations of the mean squares). This is a good way to analyze data in which there is stimulus sampling, but there is no statistical requirement that stimuli be sampled randomly to treat stimuli as a random factor. The analysis is appropriate regardless of how the stimuli were sampled. In effect, this test using quasi-*F*-ratios is analogous to one in which the effect is replicated across different stimuli within a single experiment.

Back to Reality

It is not our intent to suggest that all studies that fail to include stimulus sampling should be rejected for publication or that interpretations of the results of such studies should never be attempted.[7] Our point is that there are some common examples that are particularly egregious violations of the need for stimulus sampling. A researcher who uses two experimenters and finds that participants in an experiment react to the male experimenter differently than to the female experimenter, for instance, ought not interpret this as an effect of experimenter gender. The chances of a Type I error, Type II error, or a diametrical error are not controllable or even estimable under these conditions. As noted earlier, even the presence of an interaction (e.g., between gender of experimenter and gender of participant) does not circumvent the problem that the gender of the experimenter is confounded with the individual characteristics of the selected experimenter.

Even if money, time, and effort were not barriers, it is often difficult or impossible to define the stimulus population and develop a truly random sampling strategy. Furthermore, there are no clear rules or formulas for defining the appropriate size for these stimulus samples. Neither are there clear criteria for knowing when the sample is an adequate representation of the variances and means of the stimuli in the general category. Therefore, one could always argue that stimulus sampling was not carried out to the maximum and hence resist acceptance of most any experimental conclusion for which stimulus sampling could be an issue. It is for these reasons that we do not think that failure to sample a large number of stimuli nor failure to sample randomly automatically prevents researchers from reaching conclusions or renders a study unpublishable. As Kenny (1985) notes, experimenters rarely even sample participants randomly.[8]

Although we are dubious of the use of shallow heuristics for selecting stimuli, we believe that a researcher's good judgment is a necessary part of minimizing the stimulus sampling problem. In some cases, it is not plausible to believe that there is meaningful variance across stimuli within category, and empirical proof of this seems unnecessary. In cases where it is plausible to believe that stimuli vary within category, however, researchers ought to provide evidence or reasoned argument that the particular stimulus is likely to be at or around the central tendency of the category. In still other cases, a researcher might use the strategy of matching the stimuli on some relevant variables so as to argue that these variables are not confounded with condition.[9] In any case, it is not our purpose to argue for unattainable or overly costly research designs.[10] Our purpose instead is to note the speciousness of some arguments that are used to dismiss the need for stimulus sampling (e.g., the interaction argument), note the ways in which heuristic thinking (e.g., representativeness heuristic) can lead to decreased sensitivity to the problem, note how the failure to sample stimuli can in many cases be considered a problem of confounding rather than merely an issue of generalization, and to remind

researchers that stimulus sampling, when feasible, is the best strategy for many problems.

NOTES

1. We chose to not single out individual publications as examples for which there were stimulus sampling problems because there are enough examples in the literature that it would be arbitrary to single out only an unlucky few for criticism. Also, we are not suggesting that all instances of inadequate stimulus sampling are slipping through the editorial review process. It is quite possible that most instances are detected and rejected in the review process. If so, however, then our concern is even greater because it indicates that the published literature is the tip of an iceberg for this problem. There might be a considerable amount of hidden research effort being spent on designs that are flawed by the failure to sample stimuli.

2. The stimulus sampling problem also occurs with other stimuli and in areas outside of social psychology. Clark (1973), for example, describes stimulus sampling problems in verbal learning, human learning, and psycholinguistics. As with other prior treatments of stimulus sampling, however, Clark focuses primarily on the question of what is the appropriate statistical test when stimuli have been properly sampled, whereas we are concerned with why the problem continues to exist, when stimulus sampling is and is not needed, and how failure to sample stimuli can threaten construct validity.

3. Stimulus sampling is not the only solution in the misleading question study example, and probably not the best solution. A counterbalancing might be possible in which the pop can is used as the central stimulus and the person is used as the peripheral stimulus for half of the participants. Issues in generalization might still plague such a study, but the need for stimulus sampling for purposes of construct validity is diminished.

4. Every respondent listed at least one flaw and often several flaws, indicating that respondents were in fact attempting to find flaws. Not counting the stimulus sampling problem, the most common flaw listed by all three populations was along the lines of suggesting that the person who was asked the question might not have known the answer. It is not a true flaw of the study, in our opinion, because the person approached was randomly assigned to be approached by either the male or female and, hence, fails to explain the male-female difference. The second most common flaw mentioned was that the person should have asked another question (e.g., the time of day) rather than just a question about directions. This is probably a valid criticism because Stang's conclusion is stated very generally rather than restricting itself to questions about directions. The third most commonly mentioned flaw had to do with the sample and / or the failure to report differences between samples. People said that Stang failed to note possible differences between large and small cities or failed to control for which people from those cities ended up in the sample. Other than the stimulus sampling problem, these were the only significant flaws listed, the others being uncodable or only mentioned by a couple of participants.

5. There are 196 possible pairs to draw (14×14) and 21 of these ($6 + 5 + 4 + 3 + 2 + 1$) are pairings of an 8-year-old with a less credible adult.

6. Of the 196 possible pairs to draw, 45 of these ($9 + 8 + 7 + 6 + 5 + 4 + 3 + 2 + 1$) are pairings of an 8-year-old with a less credible adult.

7. Even when only one stimulus instance was used to represent a category, an experiment has some value for meta-analyses. Note, however, that our observations about stimulus sampling have implications for meta-analytic techniques when there is a stimulus sampling issue. We are not aware of any meta-analyses that have used the number of sampled stimuli as a weighting variable in combining the results of experiments. Consider, for instance, the meta-analytic pooling of two experiments that tested the hypothesis that a male speaker would produce more attitude change than a female speaker. Experiment A used 300 participants and 1 male and 1 female speaker, whereas Experiment B used 150 participants and 15 different male and 15 different female speakers. A standard meta-analytic approach would be to weight Experiment A twice as strongly as Experiment B for purposes of estimating effect size. Although the precise statistical solutions are beyond the scope of the current article, Experiment B ought to have as much, or perhaps more, weight than Experiment A in estimating the reliability and size of the gender effect.

8. Nevertheless, participants can be randomly assigned to condition, whereas, in the problems we have identified here, stimuli cannot be randomly assigned to condition (e.g., a male experimenter cannot be assigned randomly to be the female experimenter).

9. Suppose, for instance, that a researcher used a particular movie scene to be the violent media stimulus and some other particular movie scene to be the nonviolent media stimulus. This is a stimulus sampling problem because these two scenes differ in ways other than violent content (such as interestingness, dialogue, production quality, excitability). An alternative to stimulus sampling in this case might be to find scenes that are matched on these potential confounds. It is then up to the judgment of reviewers to decide whether the use of a single stimulus is problematic. An excellent example of this approach can be seen in the work of Bushman (1995), who matched scenes on numerous relevant dimensions to examine the effects of media violence. Our point is that the absence of stimulus sampling itself should not automatically render results uninterpretable if appropriate other measures are taken to rule out specific alternative explanations.

10. Another alternative to the strategy of sampling stimuli is to manipulate the variable at the level of the construct itself rather than manipulate the variable via the use of particular instances. For example, the gender, age, or race of a stimulus person might be manipulated via an assertion to the participants that the stimulus is male or female, young or old, White or Black. This generally requires, of course, that the participants never actually see the stimulus person. Researchers might find in many cases that this solution is not particularly satisfactory because such a manipulation lacks salience or because it is not the typical way that people encounter this variable in everyday life.

REFERENCES

Brunswik, E. (1947). *Systematic and representative design of psychological experiments*. Berkeley: University of California Press.

Bushman, B. J. (1995). The moderating role of trait aggressiveness in the effects of violent media on aggression. *Journal of Personality and Social Psychology, 69*, 950-960.

Campbell, D. T. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist, 15*, 546-553.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

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

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation*. Boston: Houghton Mifflin.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. New York: Cambridge University Press.

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

Kenny, D. A. (1985). Quantitative methods for social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 1, pp. 487-508). New York: Random House.

Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effects of graduate training on reasoning: Formal discipline and thinking about everyday-life events. *American Psychologist, 43*, 431-442.

Maher, B. A. (1978). Stimulus sampling in clinical research: Representative design reviewed. *Journal of Consulting and Clinical Psychology, 46*, 643-647.

Richter, M. L., & Seay, M. B. (1987). ANOVA designs with subjects and stimuli as random effects: Applications to prototype effect on recognition memory. *Journal of Personality and Social Psychology, 53*, 470-480.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis*. New York: McGraw-Hill.

Taylor, S. E. (1981). A categorization approach to stereotyping. In D. Hamilton (Ed.), *Cognitive processes in stereotyping and ingroup behavior.* Hillsdale, NJ: Lawrence Erlbaum.

Tversky, A., & Kahneman, D. (1971). The belief in the "law of small numbers." *Psychological Bulletin*, *76*, 105-110.

Wells, G. L., Lindsay, R.C.L., & Ferguson, T. J. (1979). Accuracy, confidence, and juror perceptions in eyewitness identification. *Journal of Applied Psychology*, *64*, 440-448.

Wells, G. L., & Turtle, J. W. (1988). What is the best way to encode faces? In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 163-168). New York: John Wiley.

Wells, G. L., Turtle, J. W., & Luus, C. A. (1989). The perceived credibility of child eyewitnesses: What happens when they use their own words? In S. J. Ceci, D. F. Ross, & M. P. Toglia (Eds.), *Children take the stand: Adult perceptions of children's testimony* (pp. 23-39). New York: Springer-Verlag.

Wilder, D. A. (1978). Perceiving persons as a group: Effect on attributions of causality and beliefs. *Social Psychology*, *41*, 13-23.

Received February 11, 1998
Revision accepted August 17, 1998