Journal of Applied Psychology
1985, Vol. 70, No. 1, 101-107

Generalizing Across Stimuli as Well as Subjects: A Neglected Aspect of External Validity

Gail A. Fontenelle, Amanda Peek Phillips, and David M. Lane
Rice University

In order to generalize the results of an experiment beyond the specific stimuli employed, it is necessary to consider variance due to stimulus sampling. This can be accomplished by treating stimuli as a random effect rather than the traditional procedure of treating stimuli as a fixed effect. The serious consequences of the traditional approach are illustrated using examples from applied psychology. Statistical and design considerations for generalizing the results of experiments are discussed.

In the design of psychological research, it is of utmost importance to be able to separate the effects of experimental manipulations from the effects of extraneous variables. It is inevitably the case that subjects possess attributes that are unmeasured, uncontrolled, and have unpredictable effects. However, it is well known that randomization allows the researcher to estimate the magnitude of the effects of these extraneous variables and, through the use of standard techniques of statistical inference, to determine the probability that differences between conditions as large as (or larger than) those obtained would occur if these extraneous variables were operating alone.

Just as in the case of subjects, it is likely that unmeasured and uncontrolled attributes of stimuli affect the experimental outcome in unpredictable ways. The potential for this problem exists in many areas of applied psychology where extraneous aspects of stimuli are difficult to assess. For example, consider a hypothetical study of sex discrimination in which four males and four females play the role of an applicant interviewing for a managerial position. Although applicants read the same script, there are (unavoidably) differences in poise, physical attractiveness, and presentation style. If the male applicants were rated as being significantly more qualified than the female applicants, would the conclusion that the difference was due to sex bias be warranted? Not necessarily. It is possible that, by chance, the four male applicants possessed extraneous attributes, irrespective of sex, that accounted for their higher ratings; thus, it would be a fallacy to generalize these results to the entire population of males and females.

Although the problem of generalizing from a stimulus sample to a stimulus population has been discussed in other contexts (Clark, 1973, 1976; Cohen, 1976; Coleman, 1964; Keppel, 1976; Smith, 1976; Wike & Church, 1976), it has been essentially ignored in industrial/organizational psychology and other applied fields. The purposes of this article, therefore, are to call attention to the problem of generalizing across stimuli and to discuss some possible courses of action that can be taken.

It is frequently the case that an experimenter has no choice other than to nest stimuli within treatment conditions. This is particularly true in studies of sex and race bias, for in these studies it is generally not possible for a person to be in both the male ratee and female ratee conditions (for example). As an example of this kind of research, consider a study by Bigoness (1976) in which the effects of ratee sex, ratee race, and level of performance on the evaluation of ratee's suitability for a job were examined. These variables were manipulated in a 2 × 2 × 2 factorial design in which a total of eight ratees (one representing each of the eight cells in the design) and 60 subjects were employed. Although low-performing males

Requests for reprints should be addressed to David M. Lane, Department of Psychology, Rice University, P.O. Box 1892, Houston, Texas 77251.

and low-performing females were rated nearly identically, high-performing females were rated higher than high-performing males. Despite the finding that this effect was highly significant in a standard analysis of variance (ANOVA), the conclusion that high-performing females are rated higher than high-performing males is not justified. As will be shown subsequently, not only does a significance test provide little protection against making a Type I error, but the probability of making a Type I error approaches 1.0 as the number of subjects employed in the design increases. Moreover, the problem is very serious even with small sample sizes.

For purposes of explication, assume that the population treatment effect in Bigoness' experiment was actually zero. That is, if all the ratees in the ratee population were rated by all the raters in the rater population, then there would be no difference between the mean rating of high-performing males and the mean rating of high-performing females. How, other than by a fluke, might Bigoness have found a significant effect under these conditions? It is not implausible that some characteristics or attributes of the two particular high-performing female ratees other than their sex may have caused raters to inflate the female performance ratings. That is, it may have been simply that the high-performing females were, by chance, inherently better (with respect to these irrelevant features) than were the high-performing males. The significance test properly rejects only the null hypothesis that there is no difference between the *specific* ratees used in the experiment. It says little about the *population* of ratees.

Are studies that use many subjects less vulnerable to this problem than studies that use few subjects? Interestingly, using more subjects (raters in this case) only increases the probability of incorrectly rejecting the null hypothesis. In situations of this sort, the mean value of the irrelevant attributes for one set of stimuli is bound to be different from the mean value of the irrelevant attributes for the other set. If enough subjects are employed, these random (but very real) differences due to stimulus sampling are virtually certain to lead to a rejection of the null hypothesis.

The problem is serious even for experiments with small sample sizes. For example, Forster and Dickinson (1976) performed a Monte Carlo simulation to estimate Type I error rates and found that the Type I error rates were grossly overestimated: with ten raters, five ratees, and a nominal Type I error of .05, the actual error rate was .24. When the number of raters was increased to 20, the actual error rate rose to .31.

Would increasing the number of ratees lessen the problem? With more ratees, the "true" mean of the sample stimuli (the mean rating of the sample stimuli if rated by all raters in the population) would be closer to the population mean (the mean rating of all stimuli if rated by all raters in the population). But it is not clear how many ratees is enough. Even studies that go to great lengths to use many stimuli are not immune from this problem. For example, Schmitt and Lappin (1980) found, among other things, an effect of ratee's race on judged performance. Although these investigators used 60 ratees, far more than is typical, they were still not justified in assuming that all irrelevant ratee attributes were completely equated. Some differences, although probably only small ones, undoubtedly remained. Because there is no way of knowing how much these differences biased the statistical analysis, these results are not as conclusive as they might have been.

How, then, is an experimenter to control for these irrelevant attributes? One solution is essentially the same as that used to control for differences between conditions resulting from the random assignment of subjects: The variance between conditions is assessed relative to the variation within conditions. Exactly how this is done is shown in the example given below.

In the hypothetical study of sex discrimination discussed previously, four females and four males each played the role of an applicant interviewing for a managerial position. Male and female stimulus persons read the same script of an applicant being interviewed and were each rated by 20 male subjects on their qualifications for the position. Table 1 presents ratings of all applicants as a function of subjects (raters) and sex of ratee.

Table 1
Qualifications Ratings of Male and Female Applicants

	Applicants							
	Males				Females			
Subject	1	2	3	4	5	6	7	8
1	8	6	5	4	7	5	4	3
2	9	7	7	5	8	6	4	4
3	7	7	6	3	7	5	3	2
4	8	5	5	4	6	4	3	2
5	9	7	6	5	7	5	3	3
6	9	8	7	6	8	6	4	3
7	7	7	5	5	7	5	3	2
8	8	6	5	5	6	5	4	2
9	7	5	4	4	6	4	3	1
10	9	6	6	4	7	4	4	2
11	8	7	6	5	7	6	5	5
12	9	7	6	4	7	5	3	2
13	8	6	5	4	6	4	3	3
14	7	7	6	3	7	5	3	2
15	8	6	5	4	7	5	4	3
16	9	7	7	5	8	6	4	4
17	7	5	5	4	7	5	2	1
18	8	6	5	4	6	4	3	1
19	9	7	6	4	8	5	3	2
20	9	6	6	5	7	5	4	3

The typical way of testing for sex bias would be to compute the mean rating each subject gave male and female ratees and conduct a Subjects (20) × Ratee Sex (2) ANOVA. For the hypothetical data shown in Table 1, this method of analysis results in a highly significant effect of ratee sex, $F(1, 38) = 67.83$, $p < .0001$. From this point on, this type of F ratio will be called $F1$.

The problem with this approach can be seen by examining the expected mean squares associated with the design. As shown in Table 2, the expected mean square of ratee sex differs from the expected mean square of the Ratee Sex × Subjects interaction by two terms: $r\sigma_R^2$ and $qr\sigma_A^2$, the former representing the effect of ratee and the latter representing the effect of ratee sex. Thus, if the mean square for ratee sex is significantly larger than the mean square for the Ratee Sex × Subjects interaction, as it is in this case, any of three possible states of nature are possible:

1. $\sigma_A^2 > 0$ and $\sigma_R^2 = 0$,
2. $\sigma_A^2 = 0$ and $\sigma_R^2 > 0$,
3. $\sigma_A^2 > 0$ and $\sigma_R^2 > 0$,

where σ_A^2 is the variance due to ratee sex and σ_R^2 is the variance due to ratees. Clearly, the important possibility (Number 2, above) that there was no real effect of ratee sex cannot be ruled out by a significant $F1$.

What then might be properly concluded from this analysis? Statistically speaking, the subjects factor is treated as a random effect and the ratees factor is treated as a fixed effect. As such, the significant F allows the generalization of the ratee sex effect to the population of subjects but not to the population of ratees. An effect is usually not of much value if it is specific to the sample of stimuli employed because the goal of most research is to generalize results beyond both the sample of subjects and the sample of stimuli.

In order to generalize the results to the population of ratees, a second F ratio can be constructed. A mean rating across subjects is computed for each ratee, and an ANOVA is conducted on these means. For the present example, this results in a simple one factor design (ratee sex) with four ratees per condi-

Table 2
Sources of Variance and Expected Mean Squares; Within-Subjects Three-Factor Design With One Fixed Effect and Two Random Effects

Label	Sources of variance	*df*	Expected value of mean square
A	Ratee sex (p)	$p - 1$	$\sigma_e^2 + \sigma_{RS}^2 + q\sigma_{AS}^2 + r\sigma_R^2 + qr\sigma_A^2$
RwA	Ratees (q) within ratee sex	$p(q - 1)$	$\sigma_e^2 + \sigma_{RS}^2 + r\sigma_R^2$
S	Subjects (r)	$r - 1$	$\sigma_e^2 + \sigma_{RS}^2 + pq\sigma_S^2$
A × S	Ratee Sex × Subjects	$(p - 1)(r - 1)$	$\sigma_e^2 + \sigma_{RS}^2 + q\sigma_{AS}^2$
S × RwA	Subjects × Ratees Within Ratee Sex	$p(p - 1)(r - 1)$	$\sigma_e^2 + \sigma_{RS}^2$

tion. For the hypothetical data, this analysis results in an $F(1, 6) = 1.79$, which is not significant, $p = .23$. An F ratio from this type of analysis will be referred to as $F2$.

In this analysis, the subjects factor is treated as a fixed effect and the ratees factor is treated as a random effect. As such, the failure to find a significant $F2$ means that there is no statistical basis on which to generalize the effect of ratee sex to the population of ratees. In order to conclude that there is a true effect of ratee sex (or, that $\sigma_A^2 > 0$) then $F2$ as well as $F1$ should be significant. If only $F2$ is significant, then it cannot be safely inferred that ratee sex has a real effect. Examination of Table 2 shows that a significant $F2$ allows three possibilities:

1. $\sigma_A^2 > 0$ and $\sigma_{AS}^2 = 0$,
2. $\sigma_A^2 = 0$ and $\sigma_{AS}^2 > 0$,
3. $\sigma_A^2 > 0$ and $\sigma_{AS}^2 > 0$,

where σ_A^2 is the variance due to ratee sex and σ_{AS}^2 is the variance due to the Ratee Sex × Subjects interaction. Because the variance due to ratee sex is zero in Possibility 2, a significant $F2$ by itself does not allow the conclusion that there is sex bias.

In summary, a significant $F1$ indicates that the results generalize to the population of raters for the sample of ratees, whereas a significant $F2$ indicates that the results generalize to the population of ratees for the sample of subjects. If both $F1$ and $F2$ are significant, one can be reasonably confident that the effect is not due to sampling error. However, if one wishes to make a valid statistical generalization to the population of ratees as judged by the population of subjects, both subjects and ratees should be treated as random effects simultaneously. The most common method of doing this employs Quasi-F ratios.

As can be seen in Equation 1, the Quasi-F ratio, F', is computed from four mean squares.

$$F' = (MS_A + MS_{S\times R_wA})/(MS_{AS} + MS_{R_wA}) \quad (1)$$

Algebraic manipulation of the mean squares from Table 2 demonstrates that the numerator of Equation 1 exceeds the denominator by only the desired term, $qr\sigma_A^2$, which is proportional to the variance due to ratee sex. Thus, if there is no effect of ratee sex, the expected value of the numerator equals the expected value of the denominator. Although the variance ratio is not exactly distributed as F, it is well approximated by the F distribution (Clark, 1973; 1976; Santa, Miller, & Shaw, 1979).

To return to the example, when the Quasi-F ratio is computed and its significance tested, the effect of ratee sex is not significant, $F'(1, 6) = 1.74$, $p = .25$. The method for calculating degrees of freedom for the Quasi-F test is somewhat cumbersome and is shown in the Appendix.

It is not the case that the problem of generalizing across stimuli is only serious when stimuli are nested within treatments. For example, consider the study by Imada and Hakel (1977) in which raters assessed a ratee's qualifications after observing an employment interview. The same female ratee was used in two conditions differing in the degree of interpersonal distance maintained. Because it is not unreasonable to suppose that some ratees are rated higher if they maintain little interpersonal distance, whereas other ratees are rated higher if they maintain a lot of interpersonal distance, the results could depend primarily on the particular ratee who, by chance, happened to be selected for the study. Therefore, the results from the ANOVA on this ratee could not be generalized validly beyond this specific ratee. If several ratees had been used, it would have been possible to compute a Quasi F and generalize to both the population of raters and the population of ratees.

It is important to note that the Quasi F has little power unless a relatively large number of stimuli is employed. Because it is generally the case that the power of the Quasi-F test will be slightly lower than the power of the less powerful of $F1$ and $F2$, both $F1$ and $F2$ must have adequate power in order for the Quasi F to have adequate power. Because there are usually fewer ratees than raters, power is typically controlled by $F2$, not $F1$. Although the power of the Quasi-F test is difficult to compute, a practical solution for the researcher is to compute the power of $F1$ and $F2$. If both are adequate, the power of the F' will probably be adequate. In general, the same considerations and/or rules of thumb used to determine the number of raters (or subjects) should also be used when determining the number of ratees.

Treatment of stimuli as a random effect in the statistical model has met with some controversy. One viewpoint holds that because stimuli are rarely sampled randomly from some specified population, the treatment of stimuli as a random effect is inappropriate (Wike & Church, 1976). Although this is certainly a legitimate concern, we do not feel that the resulting problems are serious enough to preclude the use of the random effects model. First, it is possible for a researcher to sample stimuli in a manner that approximates random sampling to a reasonable degree. This is particularly true in applied psychology. For example, in the study of performance appraisal there is no reason that the sampling of ratees (stimuli) cannot be just as random as the sampling of raters (subjects). Because (a) neither subjects nor stimuli are typically sampled randomly from a specifiable population, and (b) it is generally accepted that the sampling of subjects is close enough to random to justify the use of statistical models that assume random sampling, it seems unreasonable to accept one approximation and reject the other. Second, it is important to weigh the cost of not satisfying the random-sampling assumption completely against the cost of ignoring the stimulus-sampling problem altogether. We believe the latter cost to be far greater, because it entails the acceptance of a seriously inflated Type I error rate.

Another argument against the treatment of stimuli as a random effect reflects uncertainty about the use of the Quasi-*F* test (Cohen, 1976; Wike & Church, 1976). First, the Quasi-*F* test is only approximate: Although the numerator and the denominator have the same expected value when the null hypothesis is true, the ratio is not distributed exactly as *F* (Winer, 1971). Second, the statistical properties of the Quasi *F* have not yet been fully investigated. Specifically, Wike and Church (1976) have questioned the robustness of the Quasi *F* in the face of violations of its assumptions. However, preliminary evidence has indicated the Quasi *F* is quite robust (Forster & Dickinson, 1976; Santa et al., 1979).

It is important to note that Quasi *F* is not the only approach to analyzing mixed designs with more than one random effect. For example, tests based on the maximum likelihood approach have been developed, and a computer program for calculating them is available (Dixon & Brown, 1979). However, the statistical properties of these tests have received less attention and are thus less well known than those of the Quasi *F*.

Although this discussion has focused on controlling for extraneous variance statistically, it is important to recognize that the problem can sometimes be avoided in the design of the experiment. For example, if it were possible to pair each subject with a unique stimulus, extraneous variance due to subjects and extraneous variance due to stimuli would be combined in each data point. Variance due to stimuli would be totally confounded with variance due to subjects and, therefore, the error term in the typically-done analysis of variance would, properly, include variance due to stimuli, variance due to subjects, and variance due to the Subjects × Stimuli interaction. Maudlin and Laughery (1981) provide one example of an experimental design that used this type of solution. In a study of facial recognition, these authors examined the effects of constructing an Identikit composite of a target face on subsequent facial recognition. Because each subject viewed a different target face, this experimental design allowed the researchers to generalize to the population of all subjects as well as the population of all stimuli.

It should be noted that generalization can sometimes be based on logical rather than statistical considerations (Keppel, 1982). Logical generalization requires the representative selection of a subset of a stimulus population. As Wike and Church (1976) note, representative selection may often be the preferred method of stimulus selection because it is more economical. However, when representative selection is used, the generalization of results is based on a researcher's judgment and therefore is always somewhat subjective. As a result, one is generally on firmer ground with statistical generalization.

In addition to logical generalization, results can sometimes be generalized if certain assumptions are made. For example, although the statistical model for regression analysis treats the predictor variables as fixed, the results can be validly generalized to values of the predictor variables not included in the study if linearity is assumed. As Cramer and Appelbaum (1978) point out, the only reason that more than two values of the predictor variable(s) are needed is so that the assumption of linearity can be tested.

In an effort to determine the extent to which this stimulus-generalization problem exists in the applied psychological literature, a review of the *Journal of Applied Psychology* was conducted; the review covered articles published from January 1977 through May 1983. Only one study was found (Harris, 1977) in which both subjects and stimuli were treated appropriately as random factors in the analyses. On the other hand, 40 studies were found in which the results were incautiously generalized beyond the sample of stimuli employed.[1] Thirty-four of these did not provide a valid basis to generalize beyond the sample of stimuli due to the way in which they were designed. For the most part, these studies employed too few stimuli to allow statistical generalization. Six studies employed an adequate number of stimuli but failed to compute the appropriate statistics.

The problem of generalizing to the population of stimuli applies to many paradigms other than the performance rating paradigm discussed in previous examples. For example, in an investigation of the accuracy of eyewitness testimony, Clifford and Hollin (1981) found that the testimony of witnesses to a violent incident was significantly less accurate than the testimony of witnesses to a nonviolent incident. Each subject viewed a videotape of one of six incidents, three violent and three nonviolent. In order to generalize beyond the six incidents employed in the experiment, the Quasi-F ratio should have been employed. Although it could be argued that the Quasi-F test is not practical with so few stimuli, this does not excuse inappropriate generalization beyond the sample. If the effect does not vary (to a nontrivial degree) as a function of the stimuli, then the error term for $F2$ would be very small and the Quasi F would have reasonable power even with only six stimuli. If the effect does vary (to a nontrivial degree) as a function of the stimuli employed, then it is clear that too few stimuli were used to allow generalization by statistical or other means. Thus, there is a problem with the conclusions of the study, although it is an open question as to whether the problem is in the design or in the analysis. We categorized it, somewhat arbitrarily, as a problem in design.

In conclusion, generalizing research findings is a major aim of all experimentation. Although experimenters as a rule are very careful to make sure that their results generalize to the population of subjects, the problem of generalizing to the population of stimuli had been neglected. Within applied psychology this problem has been addressed only in the context of theories of measurement (Cronbach, Gleser, Nanda, & Rajaratnam, 1971) and not in the context of hypothesis testing. When it is the intention of an experimenter to generalize results beyond the particular sample of stimuli employed, the statistical treatment of stimuli as a fixed effect is generally inappropriate. If stimulus effects are treated inappropriately as fixed effects, the Type I error rate is severely inflated. Thus, unless a researcher is willing to limit the generalizability of his or her findings severely, the effect of stimulus sampling must be considered both in the design of the experiment and in the analysis of the results.

[1] A list of these studies is available from the third author upon request.

References

Bigoness, W. J. (1976). Effect of applicant's sex, race, and performance on employers' performance ratings: Some additional findings. *Journal of Applied Psychology, 61,* 80–84.

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.

Clark, H. H. (1976). Reply to Wike and Church. *Journal of Verbal Learning and Verbal Behavior, 15,* 257–261.

Clifford, B. R., & Hollin, C. R. (1981). Effects of the type of incident and the number of perpetrators on eyewitness memory. *Journal of Applied Psychology, 66,* 364–370.

Cohen, J. (1976). Random means random. *Journal of Verbal Learning and Verbal Behavior, 15,* 261–262.

Coleman, E. B. (1964). Generalizing to a language population. *Psychological Reports, 14,* 219–226.

Cramer, E. M., & Appelbaum, M. I. (1978). The validity of polynomial regression in the random regression model. *Review of Educational Research, 48,* 511–515.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1971). *The dependability of behavioral measurements.* New York: Wiley.

Dixon, W. J., & Brown, M. B. (Eds.). (1979). *BMDP-79: Biomedical Computer Programs.* Berkeley: University of California Press.

Forster, K. I., & Dickinson, R. G. (1976). More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for *F*1, *F*2, *F*', and min *F*'. *Journal of Verbal Learning and Verbal Behavior, 15,* 135–142.

Harris, R. J. (1977). Comprehension of pragmatic implications in advertising. *Journal of Applied Psychology, 62,* 603–608.

Imada, A. S., & Hakel, M. D. (1977). Influence of nonverbal communication and rater proximity on impressions and decisions in simulated employment interviews. *Journal of Applied Psychology, 62,* 295–300.

Keppel, G. (1976). Words as random variables. *Journal of Verbal Learning and Verbal Behavior, 15,* 257–266.

Keppel, G. (1982). *Design and analysis: A researcher's handbook.* Englewood Cliffs, NJ: Prentice-Hall.

Maudlin, M. A., & Laughery, K. R. (1981). Composite production effects of subsequent facial recognition. *Journal of Applied Psychology, 66,* 351–357.

Santa, J. L., Miller, J. J., & Shaw, M. L. (1979). Using Quasi *F* to prevent alpha inflation due to stimulus variation. *Psychological Bulletin, 86,* 37–46.

Schmitt, N., & Lappin, M. (1980). Race and sex as determinants of the mean and variance of performance ratings. *Journal of Applied Psychology, 65,* 428–435.

Smith, J. E. K. (1976). The assuming-will-make-it-so fallacy. *Journal of Verbal Learning and Verbal Behavior, 15,* 257–266.

Wike, E. L., & Church J. D. (1976). Comments on Clark's "The language-as-fixed-effect fallacy." *Journal of Verbal Learning and Verbal Behavior, 15,* 249–255.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Appendix

Calculation of Degrees of Freedom for the Quasi-*F* Test

Assume $F'(i, j) = (MS_T + MS_{S\times MT})/(MS_{T\times S} + MS_{MT})$ where MS_T = mean square for the treatment effect (T); MS_{MT} = mean square for the effect of stimuli (M) within treatment condition; $MS_{S\times MT}$ = mean square for the interaction of subjects within treatment condition, and $MS_{T\times S}$ = mean square for the treatment by subjects interaction.

The two mean squares in the numerator of F' will be referred to as MS_1 and MS_2. The two mean squares in the denominator will be referred to as MS_3 and MS_4. Let n_1, n_2, n_3, n_4 be the respective degrees of freedom for the four mean squares.

i and j are then computed as follows:

$$i = (MS_1 + MS_2)^2/(MS_1^2/n_1 + MS_2^2/n_2),$$

$$j = (MS_3 + MS_4)^2/(MS_3^2/n_3 + MS_4^2/n_4).$$

Received January 23, 1984
Revision received April 6, 1984 ■