Research Report

CONTEMPORARY ISSUES IN THE ANALYSIS OF DATA: A Survey of 551 Psychologists

Miron Zuckerman,[1] Holley S. Hodgins,[1] Adam Zuckerman,[1] and Robert Rosenthal[2]
[1]*University of Rochester and* [2]*Harvard University*

Abstract—*We asked active psychological researchers to answer a survey regarding the following data-analytic issues: (a) the effect of reliability on Type I and Type II errors, (b) the interpretation of interaction, (c) contrast analysis, and (d) the role of power and effect size in successful replications. Our 551 participants (a 60% response rate) answered 59% of the questions correctly; 46% accuracy would be expected according to participants' response preferences alone. Accuracy was higher for respondents with higher academic ranks and for questions with "no" as the right answer. It is suggested that although experienced researchers are able to answer difficult but basic data-analytic questions at better than chance levels, there is also a high degree of misunderstanding of some fundamental issues of data analysis.*

In this article, we examine how well psychologists apply statistics to the analysis of data. Recent commentaries on this issue have found fault with a number of statistical practices in our field—for example, the prevalence of yes-no decisions at the magic .05 level and the failure to consider the power of statistical tests (Cohen, 1990), misinterpretations of the meaning of interactions (Rosnow & Rosenthal, 1989a, 1989b), and overreliance on single studies in comparison to meta-analyses of research domains (Rosenthal, 1991). Following the footsteps of these observations, the goal of the present work was to provide yet another look at the statistical scene.

For the most part, commentaries on statistical analyses have used two sources of data: surveys of published articles (Cohen, 1962; Rosnow & Rosenthal, 1989b) and personal knowledge or impressions (Cohen, 1990; Rosenthal, 1991). In contrast, we conducted a survey of active psychological researchers, asking each one to respond to a number of statistical problems. Surveys of researchers' views have been conducted in the past (Rosenthal & Gaito, 1963; Minturn, Lansky & Dember, 1972; Nelson, Rosenthal, & Rosnow, 1986). However, in these earlier investigations, both the number of questions and the number of investigators surveyed were relatively small. We asked more questions and attempted to reach more psychologists.

The survey covered basic issues in statistical analysis, including (a) the practical distinction between Type I and Type II errors, (b) the interpretation of interaction, (c) the question of omnibus versus focused tests, and (d) the role of power and effect size as criteria for successful replications. Although not new, these topics have been brought to the fore by recent advances in statistical analysis. For example, the development of meta-analytic techniques requires the use of focused (as opposed to omnibus) tests, the calculation of effect size, and perhaps a new definition of what replication means (Rosenthal, 1991; Rosenthal & Rosnow, 1991). The "new" emphasis on power analysis (new in the sense that in spite of all the advocacy, one still does not see "any mention of power in the journals"; Cohen, 1990, p. 1311) requires an understanding of the distinction between Type I and Type II errors as well as calculation of effect size (Cohen, 1988). Consideration of effect size (slowly gaining in acceptance as an indicator to be reported along with significance tests) requires that we understand clearly what an interaction effect is. Completing the circle, power analysis and calculation of effect size draw attention to focused tests (contrast analysis). Contrasts boost power and endow the effect size with a more specific meaning (Pearson *r*, a measure of effect size, can be calculated for contrasts but not for omnibus tests).

To study these concerns, we formulated five multiple-choice questions. Each could be answered by one of three alternatives: "it depends," "yes," and "no." The complete questions, as seen by each respondent, are presented in Appendix 1; the solutions for the questions are presented in Appendix 2. Briefly, the survey contained the following items: The first question asked whether low reliability can cause spuriously significant results. The second question asked whether a test of simple effects is the correct approach to the interpretation of interactions. The third question asked whether a multivariate analysis of variance (MANOVA) followed by univariate analyses of variance (ANOVAs) appropriately tests a predicted pattern among means. The fourth question asked whether a small *n* can lead to spuriously significant results. The fifth question asked whether a nonsignificant result based on a smaller *n* than the *n* associated with a previously significant result (both results going in the same direction) is a failure to replicate.

THE SAMPLE

Accompanying the survey was a request for background information (academic rank, sex, and year of Ph.D.) and a cover letter. These materials were sent to 931 authors of articles published primarily in the second half of 1989 and the first half of 1990 in the following journals of the American Psychological Association: *Developmental Psychology, Journal of Abnormal Psychology, Journal of Consulting and Clinical Psychology, Journal of Counseling Psychology, Journal of Educational Psychology, Journal of Experimental Psychology: General, Journal of Experimental Psychology: Human Perception and Performance, Journal of Experimental Psychology: Learning, Memory, and Cognition,* and *Journal of Personality and Social Psychology*. For each article appearing in one of the journals, the author contacted

Address correspondence to Miron Zuckerman, University of Rochester, Rochester, NY 14627.

was the individual listed as the one to receive correspondence about the published work. There were three successive mailings, spaced 2½ weeks apart from each other. Each mailing went to all participants since returns were anonymous and we had no way of knowing who had responded.

Of the 931 survey forms that were sent out, 10 were returned unopened, 9 with the standard message "not deliverable as addressed" and 1 with a note that the person to whom it was addressed was deceased. Of the remaining 921 potential participants, 4 persons returned uncompleted questionnaires, 1 with an original if unusual note: "I think hypothesis testing has virtually no place in the social sciences. So these questions aren't meaningful to me."

Five hundred and fifty-one persons, 371 men and 158 women (and 22 who did not report their gender), returned codable questionnaires. Our response rate (59.8%) was twice as high as that (29%) of the Nelson et al. (1986) survey. That the questionnaire had piqued the interest of many participants was also shown by various best-wishes messages, including the following very enthusiastic comment: ". . . I receive nearly 6 surveys each week. Your cover letter was wonderful and your survey is enclosed." Of course, we did not receive the comments of those who did not return the survey and whose opinions were probably much less charitable.

The respondents included 17 students, 175 assistant professors, 134 associate professors, 182 full professors, and 43 holders of nonacademic jobs. The earliest year of Ph.D. was 1943 and the median was between 1980 and 1981. As could be expected, year of Ph.D. and academic rank were highly correlated (r = .80, $p < .001$). Men tended to report an earlier year of Ph.D. (r = .15, $p < .01$) and somewhat higher ranks (r = .08) than women. Participants who sent their replies earlier (in terms of the three mailings of the survey) reported higher academic ranks (r = .15, $p < .01$), were less likely to leave questions unanswered (r = .14, $p < .01$), and were more likely to write comments elaborating on their choice of answer (r = .10, $p < .05$). Interestingly, volunteers for psychology experiments (compared with nonvolunteers) seem more involved and expect to do better in the investigations (Rosenthal & Rosnow, 1975)—attributes that are conceptually similar to those of our early respondents. Finally, respondents who wrote comments were more likely to choose "it depends" as an answer (r = .18, $p < .01$)—a strategy that reflects a Talmudic spirit (nothing is exactly what it seems to be . . .), though that spirit did not help accuracy.

ACCURACY AND ITS DETERMINANTS

The mean accuracy for the entire sample was .59; that is, our respondents provided accurate answers for 2.93 of the 5 questions. The distribution of responses for each of the questions is presented in Table 1. Response categories included the three options provided by the questionnaire and leaving a question unanswered (questions left blank were often accompanied by comments such as "I don't know the answer"). On the basis of response preference alone (presented in the right-hand column of Table 1), accuracy would have been .46, significantly lower than the .59 score that was obtained, $t(550) = 13.21$, $p < .001$, r = .49.

Note the strong evidence of a negativity bias. For each question (including questions for which "yes" was the right answer), participants preferred the "no" response over every other category. Therefore, questions with "no" as the right answer (Questions 2, 3, and 5) were more likely to elicit correct responses than questions with "yes" as the right answer (Questions 1 and 4). As seen in Table 2, the mean accuracy of the former group was far higher than the mean accuracy of the latter group, $F(1, 550) = 349.42$, $p < .001$, r = .62. The advantage of questions with "no" as the right answer would disappear, however, if obtained accuracy were compared with accuracy expected on the basis of response preference, .59 for Questions 2, 3, and 5 and .26 for Questions 1 and 4 (see right-hand column in Table 1). However, comparisons of actual accuracy with accuracy expected on the basis of response preference should be interpreted with caution. Our measure of response preference may reflect in part differences in difficulty of questions rather than only response bias of the participants. For example, if Questions 2, 3, and 5 were in fact less difficult than Questions 1 and 4, a preference for "no" answers would be shown, although not of the magnitude that was obtained in the present study.

Table 2 also shows that higher academic rank was related to higher accuracy scores, Academic Rank linear contrast $F(1, 504) = 16.52$, $p < .001$, r = .18. For the most part, this pattern was repeated for each of the five questions. A noticeable exception, however, was the higher accuracy of students on Ques-

Table 1. *Distribution of responses to each of the five questions*

Response category	Question 1	2	3	4	5	Unweighted mean[a]
It depends	97 (17.5)	112 (20.0)	78 (14.0)	37 (6.7)	40 (7.2)	71.8 (12.9)
Yes	198[b] (35.7)	88 (15.7)	128 (23.0)	235[b] (42.6)	13 (2.3)	146.4 (26.4)
No	254 (45.8)	353[b] (63.1)	334[b] (60.0)	275 (49.8)	499[b] (90.0)	329.9 (59.4)
Blank	5 (0.9)	6 (1.0)	16 (2.9)	5 (0.9)	2 (0.3)	6.5 (1.2)
Total	554	559	556	552	554	554.7

Note. The total number of responses to each question exceeds 551 because some participants provided more than one answer; numbers in parentheses are percentages of the total number of responses.

[a] Mean of the average score of the three questions with "no" as the right answer and the average score of the two questions with "yes" as the right answer.

[b] Correct answer.

Table 2. *Mean accuracy scores by question and academic rank*

Academic rank	*n*	Question 1	2	3	4	5	Mean
Student	17	.47	.41	.41	.59	.88	.55
Assistant professor	175	.30	.63	.50	.34	.89	.53
Associate professor	134	.35	.63	.66	.43	.94	.60
Full professor	182	.43	.67	.67	.51	.89	.63
Mean (unweighted)	—	.39	.59	.56	.47	.90	.58
Mean (weighted)	—	.36	.63	.60	.43	.90	.59

tions 1 and 4. We proposed earlier that overall accuracy for Questions 1 and 4 was relatively low because of a negativity bias. It is not surprising, therefore, that students' higher performance on these items was accompanied by an overall tendency to give fewer "no" responses. In a sense, students suffered less from a negativity bias. The mean proportion of "no" answers across all five questions was .48 for students versus .63 for assistant professors, .64 for associate professors, and .61 for full professors; the contrast comparing the students to the three other academic ranks was significant, $F(1, 504) = 8.79$ $p < .005$, $r = .13$. A cautionary note regarding the students' results should be added here. Because the number of students in our sample was smaller than the number of subjects with other academic ranks, the students' responses to our questionnaire are less reliable indicators of the responses we would expect from the total student population.

A series of studies by Amabile may be relevant to trying to decipher the origin of the negativity bias. Amabile and Glazerbrook (1981) found a negativity bias in the evaluation of a target's intellectual ability when the audience of the evaluation was high in status. Amabile (1983) showed that authors of negative book reviews are perceived as more intelligent and competent than authors of positive reviews. With these results in mind, an occupational hazard of our profession appears to be the tendency to overly criticize others' work. After all, this is what we train our graduate students to do, this is what is required of us as peer reviewers of papers and grant proposals, and, according to Amabile's studies, this may be a mechanism by which we bolster our self-concept as scholars. The negativity bias in the present study may be a carryover from a habit that is well practiced in academic life. That the graduate students in this investigation showed less of the bias may simply indicate that they did not yet have sufficient time to develop the habit fully.

RESPONDENTS' COMMENTS AND THEIR IMPLICATIONS

Many participants followed their choice of a response alternative with an explanation of their answer. These comments often included a rationale that was identical to our own.

Let us consider Question 2 (Should we calculate simple effects to understand an interaction?) first. A participant who answered "no" added this short but incisive comment: "Look at residuals with main effects and error extracted." It was not unusual for participants, however, to provide the right answer with the wrong rationale. For example, some "no" answers to this question were accompanied by comments such as "There could be other sources of the interaction—other pairwise comparisons" and "need to test all cells simultaneously in all possible pairs." Evidently, some respondents thought that more comparisons between means were necessary for a correct interpretation of the interaction. On a lighter side, one respondent answered "no" and then added, "But I do it anyway. . . ."

"Yes" and "it depends" answers to Question 2 were often accompanied by the familiar logic of simple effects: "Only if interaction is significant"; and "It's what I was taught." Obviously, there is a strong and perhaps difficult to change tradition of following a significant interaction with a test of simple effects. One respondent was aware of Rosnow and Rosenthal's (1989a, 1989b) critique of testing simple effects as the basis for understanding interaction but commented that "Rosenthal and Rosnow are incorrect."

According to Dawes (1969), the difficulty in interpreting interaction effects may be in part the result of "the lack of perfect correspondence between the meaning of 'interaction' in the analysis of variance model and its meaning in other discourse" (p. 57). It seems to us that matters of language are also relevant to difficulties in distinguishing between Type I and Type II errors. In the current survey, we asked (Questions 1 and 4) whether conditions (low reliability and a small *n*) that increase Type II error also increase Type I error (produce "spuriously significant results"). These questions were the most difficult of all five, in part because of the negativity bias ("yes" was the correct answer to both questions). However, language also might have contributed to the low accuracy rate. It appears that respondents often generalized from characteristics of the achievement scale in Question 1 (unreliable) and of the sample in Question 4 (unrepresentative) to characteristics of the results. Language facilitates such generalization because the same terms can characterize conditions related to both Type II and Type I error. Consequently, participants sometimes generalized from an unreliable scale to unreliable results and from an unrepresentative sample to unrepresentative results. Note the following comments: "If the predictor is not reliable, then how reliable is its relation to the criterion?"; "Low power means high probability of making a Type II error if not a Type I error"; "Small *n* could affect representativeness of sample which in turn could affect statistics"; and "With low power, spurious effect may be more likely than detection of population effect."

While language provided misleading cues for answering Questions 1 and 4, it also provided helpful hints for answering Question 5. Our original intention was to examine researchers' views of replication. Accordingly, Question 5 presented

two studies. In the first study, Lisa obtained a significant sex difference; in the second study, Karen ran fewer subjects and obtained a nonsignificant but same-size sex difference. Rather than asking whether Karen failed to replicate Lisa's finding, we asked whether Karen obtained a smaller sex difference than did Lisa. In psychology, terms like smaller or bigger imply significantly smaller or bigger. Respondents could easily calculate or guess that Karen's finding was not significantly different from that of Lisa. Accordingly, the great majority of participants (90%) provided the right answer. Most comments were right on target (e.g., "weak power, difference in effect size probably not significant"). We were left wishing we had asked whether Karen failed to replicate.

Even in this case, however, there were hints that researchers focused on Karen's failure to obtain a significant difference and did not take into account the lower power of her test. Thus, participants who rejected the claim that Karen obtained a smaller sex difference added these comments: "One must accept the conclusion of no sex difference to visual cues"; and "Karen found no evidence for visual cues, not evidence for a smaller effect."

Finally, most comments written in response to Question 3 echoed our own thinking. The question was whether a MANOVA and three univariate ANOVAs test a specific ordering of the effects of self-awareness on the self-monitoring subscales. Respondents who objected to these analyses offered comments of the following sort: "No direct comparison of effect for Scale 1 to effect of Scale 3"; and "A priori contrast would most directly test the predicted effects." Some respondents, however, criticized the MANOVA and ANOVAs as inappropriate, yet preferred less powerful post hoc tests (e.g., paired comparison or a Scheffé test) over contrast analysis.

CONCLUSIONS

Our participants did not do very well. A level of .59 accuracy is not very impressive, particularly when compared with a .46 baseline. Thus, one possible implication of the present results is that active researchers in psychology have only limited understanding of basic issues in statistics. We were warned in advance against such conclusions. Thus, one participant wrote, ". . . a disappointing set of response options. I worry that you're going to add this up across respondents and conclude that experienced people have a naive view of statistics." Other respondents criticized our use of certain words ("What do you mean by 'spurious'—very unclear"), certain questions ("question is unclear"), or the entire questionnaire ("The problems are so ill-defined I should have marked 'it depends' on all of them"). No doubt the questions could have been phrased better, but it is not certain that improvement in clarity would have resulted in higher accuracy.

We have our own doubts, however, about whether the survey can be used as a test of statistical knowledge. The questions we asked represent only a small number of statistical issues. The conditions under which the survey was filled out were less than ideal. We found in a pilot study and informed our respondents that the survey takes an average of 6.5 min to fill out. We assume that most participants declined to invest more than a very short time in an anonymous questionnaire.

The above problems are exacerbated by the lack of a control group in our sample. There are, of course, demonstrations of important effects in psychology without control groups (e.g., Milgram's, 1963, studies on obedience and Asch's, 1955, studies on conformity). In these studies, the deviation of data from expectations was so compelling that a control group did not seem necessary. That may not be the case in the present study. Given the ambiguity, selectivity, and time constraints associated with the survey instrument, expectations are more difficult to establish.

Even if the survey cannot serve as a test of knowledge, it can be used as a platform for issues that are crucial to the application of statistics to psychological research. Concepts such as Type I and Type II error, interaction, power, effect size, and replication are basic to any data analysis. If we have motivated researchers to think about these issues, then an important purpose has been accomplished.

Acknowledgments—The authors are grateful to all the researchers who volunteered to complete the survey questionnaire. This study could not have been done without their cooperation. Preparation of this article was supported in part by the Spencer Foundation. The content is, of course, solely the responsibility of the authors.

REFERENCES

Amabile, T.M. (1983). Brilliant but cruel: Perceptions of negative evaluators. *Journal of Experimental Social Psychology, 19,* 146–156.

Amabile, T.M., & Glazerbrook, A.H. (1981). A negativity bias in interpersonal evaluation. *Journal of Experimental Social Psychology, 18,* 1–22.

Asch, S. (1955). Opinions and social pressures. *Scientific American, 193,* 31–35.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145–153.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dawes, R.M. (1969). "Interaction effects" in the presence of asymmetrical transfer. *Psychological Bulletin, 71,* 55–57.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67,* 371–378.

Minturn, E.B., Lansky, L.M., & Dember, W.N. (1972, April). *The interpretation of levels of significance by psychologists: A replication and extension.* Paper presented at the meeting of the Eastern Psychological Association, Boston.

Nelson, N., Rosenthal, R., & Rosnow, R.L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist, 44,* 1276–1284.

Rosenthal, R. (1991). Cumulating psychology: An appreciation of Donald T. Campbell. *Psychological Science, 2,* 213, 217–221.

Rosenthal, R., & Gaito, J. (1963). The interpretation of level of significance by psychological researchers. *Journal of Psychology, 55,* 33–38.

Rosenthal, R., & Rosnow, R.L. (1975). *The volunteer subject.* New York: Wiley-Interscience.

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

Rosnow, R.L., & Rosenthal, R. (1989a). Definition and interpretation of interaction effects. *Psychological Bulletin, 105,* 143–146.

Rosnow, R.L., & Rosenthal, R. (1989b). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276–1284.

(RECEIVED 12/11/91; REVISION ACCEPTED 5/26/92)

APPENDIX 1: SURVEY QUESTIONS

Question 1

Miller found a significant effect ($p < .05$) of achievement motivation on responses to helplessness training. The data also showed, however, that the scale used to measure achievement motivation had low reliability (Cronbach's alpha = .35). Anticipating attempts to dismiss the results altogether, Miller argued that low reliability cannot cause spuriously significant results. Do you agree with Miller's reasoning?

____It depends ____Yes ____No

Question 2

One way to understand interaction effects is to calculate the simple effects, i.e., the effect of each factor at a constant level of the other factors. In a 2 × 2 design, for example (see table below), an interaction effect can be interpreted by conducting two t tests; one t comparing cell a to cell b and another t comparing cell c to cell d. Is this a correct approach to the interpretation of interactions?

a	b
c	d

____It depends ____Yes ____No

Question 3

Smith collected data to test whether the manipulation of public self-awareness influences the first subscale of self-monitoring the most, and the third subscale of self-monitoring the least. He started the analyses with a Manova (which was significant) and continued with three univariate Anovas, one for each subscale. Are these analyses appropriate for the purpose of the study?

____It depends ____Yes ____No

Question 4

Chris planned to run a study with 32 subjects. The design was a 2 × 4 × 2 (all factors are between-subjects). A reviewer argued that with such a complicated design a small n may lead to a spuriously significant three-way interaction ($df = 3$). Chris answered that size of n does not affect the probability of getting spuriously significant results. Do you agree with Chris's reasoning?

____It depends ____Yes ____No

Question 5

Lisa showed that females are more sensitive to auditory nonverbal cues than are males, $t = 2.31$, $df = 88$, $p < .05$. Karen attempted to replicate the same effect with visual cues but obtained only a t of 1.05, $df = 18$, $p < .15$ (the mean difference did favor the females). Karen concluded that visual cues produce smaller sex differences than do auditory cues. Do you agree with Karen's reasoning?

____It depends ____Yes ____No

APPENDIX 2: SOLUTIONS TO SURVEY QUESTIONS

Solution to Question 1

Miller is right in that low reliability increases the probability of Type II error (failing to reject a null hypothesis that is false) but does not increase the probability of Type I error (rejecting a null hypothesis that is true; i.e., obtaining "spuriously significant results"). Or, as stated by Cohen and Cohen (1983, p. 70), "Unreliability . . . is a sufficient reason for low correlations; it cannot cause correlations to be spuriously high."

Solution to Question 2

Interaction effects cannot be interpreted on the basis of comparisons between cell means (the so-called simple effects) because these means combine the effect of the interaction with the effects of rows and columns (the main effects). Stated differently, the main effects may contribute to the simple effects as much as or even more than the interaction does (Rosenthal & Rosnow, 1991; Rosnow & Rosenthal, 1989a). The meaning of the interaction is defined in terms of the interaction residuals (i.e., leftovers of the cell means after all lower order effects have been removed). Of course, comparisons between cell means may be important in their own right. However, such comparisons do not reveal the pattern of the interaction.

Solution to Question 3

No, these analyses do not test the predictions of the study. In the framework of fixed-effect ANOVAs, specific predictions can be optimally tested with appropriate contrast analyses. A contrast is a 1-df test of significance that examines whether the pattern of obtained means (or interaction residuals) matches the pattern of predicted values (the so-called contrast weights). To test the prediction that the manipulation of public self-awareness (e.g., high vs. low) influences the first self-monitoring subscale the most and the third subscale the least, the data can be examined in a 2 × 3 ANOVA (Self-Awareness × Self-Monitoring Subscale) with self-awareness as a between-subjects factor and subscale as a within-subjects (repeated-measure) factor. In case of difference in variance among the subscales, they should be standardized. The prediction itself is tested by the interaction between self-awareness and the linear contrast of the subscales (contrast weights assigned to the three subscales would be +1, 0, and −1 under high self-awareness and −1, 0, and +1 under low self-awareness).

The MANOVA tests whether self-awareness influences the subscales in some way; it does not test, however, a specific pattern of influences as described by the prediction. The univariate ANOVAs examine the effect of self-awareness on each subscale; they do not test, however, whether the first subscale was affected most and the third subscale affected least. An exact test of this latter pattern is the contrast analysis described above.

Solution to Question 4

The problem and its solution parallel those of Question 1. A small n increases the probability of Type II error (i.e., it decreases the power of the test); it does not increase Type I error (i.e., it cannot increase the probability of getting spuriously significant results).

Solution to Question 5

No, a smaller sex difference means a significantly smaller effect size. Karen failed to compare the sex difference she obtained with the one obtained by Lisa. Had she done so, she would have discovered that the effect size of this difference is $r = .24$[1] in both studies. Any difference between the levels of statistical significance obtained by the two investigators can be accounted for by differences in ns and, hence, by differences in statistical power. An "extra credit" response would have added a comment that Karen's study would have been stronger as a replication had she included an auditory condition.

---

1. The estimate of effect size, the Pearson r, was computed as

$$r = \sqrt{\frac{F_{1,-}}{F_{1,-} + df_{error}}}$$

(Rosenthal & Rosnow, 1991).