

## The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research

HERBERT H. CLARK<sup>1</sup>

*Stanford University*

Current investigators of words, sentences, and other language materials almost never provide statistical evidence that their findings generalize beyond the specific sample of language materials they have chosen. Nevertheless, these same investigators do not hesitate to conclude that their findings are true for language in general. In so doing, it is argued, they are committing the language-as-fixed-effect fallacy, which can lead to serious error. The problem is illustrated for one well-known series of studies in semantic memory. With the appropriate statistics these studies are shown to provide no reliable evidence for most of the main conclusions drawn from them. A review of other experiments in semantic memory shows that many of them are likewise suspect. It is demonstrated how this fallacy can be avoided by doing the right statistics, selecting the appropriate design, and sampling by systematic procedures, or, alternatively, by proceeding according to the so-called method of single cases.

In 1964, Edmund B. Coleman published an important methodological paper called "Generalizing to a Language Population" in which he criticized some of the procedures psychologists were then using to deal with language samples in their study of verbal behavior. As he put it, "Many studies of verbal behavior have little scientific point if their conclusions have to be restricted to the specific language materials that were used in the experiment. It has not been customary, however, to perform significance tests that permit generalization beyond these specific materials, and thus there is little statistical evidence that such studies could be successfully

replicated if a different sample of language materials were used (p. 219)." Coleman then described available statistical procedures that would assure generality across language materials. Despite their importance, Coleman's criticisms got buried in the literature and have been all but totally ignored ever since. But if his criticisms were serious then, they are even more serious now, for there has been an increase in research in such areas as psycholinguistics, word perception, and semantic memory—areas particularly vulnerable to Coleman's criticisms. In the present paper, therefore, I would like to disinter Coleman's arguments from their premature grave, add a few arguments of my own, and then demonstrate with several specific examples how these arguments lead to serious doubts about the conclusions drawn in many well-known papers in verbal learning, human memory, and psycholinguistics.

Coleman's main point is best illustrated with a simple example. Imagine that Baker and Reader are two psychologists interested in reading. Independently, they come up with the hypothesis that people can read, that is,

<sup>1</sup> The preparation of this paper was supported in part by Public Health Service Grant MH-20021 from the National Institute of Mental Health. I am very grateful to William P. Banks, J. Merrill Carlsmith, Eve V. Clark, Douglas J. Herrmann, Peter Lucy, Lance J. Rips, and Edward J. Shoben for their helpful comments on the manuscript and to Thomas K. Landauer, David E. Meyer, and three anonymous reviewers for their detailed reviews of the paper. I am especially indebted to Edward E. Smith and Ewart A. C. Thomas for their generous and thoughtful counsel on many points in the paper.

perceive and vocalize, nouns faster than verbs. To test their hypotheses, each consults a dictionary, selects 10 nouns and 10 verbs at random, and collects reading latencies for the 20 words from each of 50 subjects. Let us assume, however, that contrary to their hypothesis nouns are in actuality *exactly* equal to verbs in reading latencies. Nevertheless, since the actual latencies for individual nouns and verbs vary from 500 to 1000 msec, the nouns in any particular sample will not be exactly equal to the verbs. So let us assume, quite plausibly, that in Baker's sample the nouns are actually 25 msec faster than the verbs, while in Reader's there is a 25 msec difference in the opposite direction. Independently, then, the two investigators tally their results, Baker finding a 30 msec difference in favor of the hypothesis, and Reader, a 35 msec difference against it. And since 42 out of 50 subjects showed the difference for Baker, and 45 out of 50 for Reader (both differences significant at  $p < .001$  by a sign test), Baker reports to the public that he has reliable support for the hypothesis, while Reader reports that he has reliable evidence against it.

But how could either investigator have come to his conclusion (barring a one in a thousand statistical freak) when in fact there is a zero difference between nouns and verbs? The answer lies in their statistics. With their sign tests they have demonstrated, consistent with the true means in their respective samples, that the differences they found would replicate if they gave these *same* two samples of 20 words to new samples of subjects. They have not demonstrated, however, that their differences would replicate if they gave *new* samples of 10 nouns and 10 verbs to new samples of subjects. Nor would they be able to demonstrate this. If Baker and Reader had examined the individual mean latencies to their 20 words, they would have found that the 10 nouns ranged approximately from 500 to 1000 msec, and so did the 10 verbs. Thus, if either investigator had compared the 10 nouns against the

10 verbs by any conventional statistical test, he would have found no significant difference between nouns and verbs. So even though Baker's and Reader's findings should replicate with new samples of subjects, they should not, necessarily, with new samples of words. And this is why it was possible for Baker and Reader to come to exactly contrary conclusions, complete with "statistical" evidence. In drawing their conclusions, therefore, Baker and Reader have committed a statistical error, one I will call the language-as-fixed-effect fallacy. In statistical jargon, they have treated Words as a fixed instead of a random effect, implicitly accepting the assumption that the 20 words they chose constitute the complete population of words they wish to generalize to. They have not presented any statistical evidence to show that their findings generalize beyond the 20 words they chose, yet they have drawn conclusions which presume that they have.

Although the errors in Baker's and Reader's studies are obvious, nearly every study in the current literature vulnerable to this fallacy exhibits the very same error. Modern investigators of language, of course, have been aware of the problem of language generality and have explicitly discussed such problems as the random sampling of words, item selection biases, and the sizes of language samples. Despite this concern, however, most of these investigators have been unaware of the statistical error they themselves have been committing. With few exceptions, they have failed to provide even the most elementary statistical evidence that their results generalize beyond their particular sample of words or sentences. Although in some instances this failure has probably done little harm, in far too many other instances it leaves the conclusions drawn by the investigator completely in doubt. As evidence for such doubts one could cite studies in verbal learning, memory, psycholinguistics, visual perception, or reading. To bring this task down to manageable size, I have therefore chosen to examine most of

the papers in the new and, as yet, relatively small field of semantic memory. My major example will be drawn from a series of studies by Rubenstein and his colleagues. For these I will demonstrate that when the appropriate statistics are computed, there is no longer any reliable support for most of the main conclusions drawn from them. The Rubenstein *et al.* studies have been singled out only because, commendably, they are accompanied by appendices containing data from which the necessary statistics can be calculated. Even though the remaining studies do not allow such analyses to be done, I shall examine the possible consequences of statistical procedures on their results as well. In the final section I will offer some remedies and one alternative approach to this unfortunate state of affairs.

#### CASE STUDIES

##### *The Rubenstein et al. Studies*

Rubenstein and his colleagues carried out a series of five experiments on the time it takes people to decide whether a letter string is a word or a nonword. For brevity's sake I will refer to the experiment by Rubenstein, Garfield, and Millikan (1970) as Study 1, that by Rubenstein, Lewis, and Rubenstein (1971a) as Study 2, and Experiments 1, 2, and 3 by Rubenstein, Lewis, and Rubenstein (1971b) as Studies 3, 4, and 5, respectively. These studies were designed to test a series of hypotheses about the search for words in semantic memory. Studies 1 and 2 examined the general thesis that to recognize a letter string as a word, the subject must locate this word in semantic memory. Under this hypothesis, recognition time should depend on various semantic properties of the word recognized, such as whether the words have one meaning or two. Studies 3, 4, and 5 examined the thesis that people put each letter string through a "phonemic recoding" before attempting to search the internal lexicon to see whether it constitutes a word or not. Under this hypothesis recognition time should depend on

various phonological properties of the letter strings, such as whether the nonword letter strings have the same pronunciation as English words. Since all five studies are essentially alike in procedure, I will first examine a simplified version of Study 5 in some detail and then, later, present the more complete evidence on all of them together.

*A statistical analysis of Study 5.* In this study the authors wished to compare homophones (words like *bear*, which is pronounced just like *bare*, but is spelled differently) against nonhomophones. They selected 25 homophones and 24 nonhomophones, mixed them in with nonword filler items, presented them one at a time to each of 44 subjects, and measured their word/nonword recognition times in milliseconds. The study, therefore, had three effects: (1) Homophony, consisting of two fixed categories; (2) Words nested within Homophony, consisting of a random sample of all possible homophones and nonhomophones; and (3) Subjects, consisting of a random sample of all possible people. Although this is a rather complicated mixed hierarchical design, the appropriate analysis of variance can be constructed on advice from, say, Winer (1971). Table 1 shows the appropriate sources of variance, degrees of freedom, and expected mean squares for the more general analysis in which there are  $p$  Treatments,  $q$  Words nested within each Treatment, and  $r$  Subjects.

The critical issue, as always, is how to construct the F-ratio that tests whether or not the Treatments effect is significant, in this case whether homophones differ significantly from nonhomophones. The information required for this decision is found in the expected value of the mean squares, abbreviated  $E(MS)$ , in the right-hand side of Table 1. The main goal is to show that the variance due to Treatments,  $\sigma_t^2$ , is greater than zero. This requires us to compare the  $MS_T$ , the Treatments mean square, against some "error" term,  $MS_{\text{error}}$ , such that  $E(MS_T)$  exceeds  $E(MS_{\text{error}})$  by exactly the variance due to Treatments,  $\sigma_t^2$ .

TABLE 1  
SOURCES OF VARIANCE AND EXPECTED MEAN SQUARES FOR MIXED HIERARCHICAL THREE FACTOR DESIGN WITH ONE FIXED EFFECT AND TWO RANDOM EFFECTS

Label	Sources of variance	Degrees of freedom	Expected value of mean square
T	Treatments ( $p$ )	$p - 1$	$\sigma_e^2 + \sigma_{ws}^2 + q\sigma_{ts}^2 + r\sigma_w^2 + qr\sigma_t^2$
WwT	Words ( $q$ ) within Treatments	$p(q - 1)$	$\sigma_e^2 + \sigma_{ws}^2 + r\sigma_w^2$
S	Subjects ( $r$ )	$r - 1$	$\sigma_e^2 + \sigma_{ws}^2 + pq\sigma_s^2$
T $\times$ S	Treatments $\times$ Subjects	$(p - 1)(r - 1)$	$\sigma_e^2 + \sigma_{ws}^2 + q\sigma_{ts}^2$
S $\times$ WwT	Subjects $\times$ Words within Treatments	$p(q - 1)(r - 1)$	$\sigma_e^2 + \sigma_{ws}^2$

The logic, then, is if the  $MS_T$  calculated from the data exceeds the  $MS_{error}$  calculated from the data by a sufficient amount, we can be confident that this has happened because the variance due to Treatments,  $\sigma_t^2$ , is greater than zero. More precisely, if  $\sigma_t^2 = 0$ , then the ratio  $MS_T/MS_{error}$  is distributed as  $F$  around a mean of  $n/(n-2)$ , where  $n$  is the degrees of freedom of  $MS_{error}$ ; therefore, if this ratio is enough greater than  $n/(n-2)$ , we can reject the hypothesis that  $\sigma_t^2 = 0$ .

Given these requirements let us consider  $F_1$ , the F-ratio in (1), in which the Treatments effect is tested against the Treatments by Subjects interaction:

$$(1) F_1(p-1, (p-1)(r-1)) = MS_T/MS_{T \times S}$$

Although appropriate in many designs,  $F_1$  clearly does not fulfill our requirements. Algebraically,  $E(MS_T)$  exceeds  $E(MS_{T \times S})$  by the sum of two variances, that is,  $r\sigma_w^2 + qr\sigma_t^2$ , not just one. So a significant  $F_1$  could lead to any one of three conclusions:

$$(2) \begin{array}{ll} \text{a. } \sigma_t^2 > 0 & \text{and } \sigma_w^2 = 0, \\ \text{b. } \sigma_t^2 = 0 & \text{and } \sigma_w^2 > 0, \\ \text{c. } \sigma_t^2 > 0 & \text{and } \sigma_w^2 > 0. \end{array}$$

In particular, (2b), in which  $\sigma_t^2 = 0$ , is a definite possibility, and so  $F_1$  could be significant even though there were no differences among the Treatments. The same considerations hold for  $F_2$ , the F-ratio in (3), in which the Treatments effect is tested against the Words-within-Treatments effect:

$$(3) F_2(p-1, p(q-1)) = MS_T/MS_{WwT}$$

In this case,  $E(MS_T)$  exceeds  $E(MS_{WwT})$  by  $q\sigma_{ts}^2 + rq\sigma_t^2$ , and so  $F_2$ , if significant, could indicate any one of three possibilities, as shown in (4):

$$(4) \begin{array}{ll} \text{a. } \sigma_t^2 > 0 & \text{and } \sigma_{ts}^2 = 0, \\ \text{b. } \sigma_t^2 = 0 & \text{and } \sigma_{ts}^2 > 0, \\ \text{c. } \sigma_t^2 > 0 & \text{and } \sigma_{ts}^2 > 0. \end{array}$$

Since (4b) is a definite possibility, a significant  $F_2$  does not guarantee that the variance due to Treatments,  $\sigma_t^2$ , is greater than zero either. An F-ratio with  $MS_{S \times WwT}$  as the error term can be shown to fail in the same way.

Because there is no single error term appropriate for this analysis, Winer (1971) and others recommend the use of  $F'$ , the so-called quasi F-ratio in (5):

$$(5) F'(i, j) = (MS_T + MS_{S \times WwT}) / (MS_{T \times S} + MS_{WwT})$$

The degrees of freedom  $i$  and  $j$  for this F-ratio are computed as follows. Let  $MS_1$  and  $MS_2$  be the two mean squares in the numerator of  $F'$ , and let  $n_1$  and  $n_2$  be their respective degrees of freedom. Then  $i$  is the nearest integer value of the following formula:

$$(6) i = (MS_1 + MS_2)^2 / \left( \frac{MS_1^2}{n_1} + \frac{MS_2^2}{n_2} \right)$$

The value of  $j$  is computed by the same formula but where  $MS_1$  and  $MS_2$  are the two mean squares from the denominator of  $F'$ . A little algebra will show that the expected value of the numerator  $E(MS_T + MS_{S \times WwT})$  exceeds the expected value of the denominator  $E(MS_{T \times S} + MS_{WwT})$  by exactly the wanted

term,  $qr\sigma_t^2$ , the variance due to Treatments. So when  $F'$  is significantly large, the variance due to Treatments can be assumed to be greater than zero.

As the name "quasi  $F$ -ratio" suggests,  $F'$  is an approximation to a true  $F$ -ratio and is not an exact test in the statistical sense. For almost all practical cases, however, the approximation is very close, so close that its use appears to be preferable to other possible procedures. For example, Winer (1971, p. 378) spells out two tests that might be used instead of  $F'$ . In the first procedure, one must first show that  $\sigma_{ts}^2 = 0$  by demonstrating that the  $F$ -ratio  $MS_{T \times S} / MS_{S \times W \times T}$  is not significant at some liberal alpha level (say,  $\alpha = .25$ ). Once this is accomplished, one can compute  $F_2$  as a test of the Treatments effect, since, with  $\sigma_{ts}^2 = 0$ , (4a) is the only possible interpretation of a significant  $F_2$ . In the second procedure, one must first show that  $\sigma_w^2 = 0$  with a non-significant  $F$ -ratio,  $MS_{W \times T} / MS_{S \times W \times T}$ , and then compute  $F_1$  as the test for the Treatments effect. In practice, however, neither of these procedures is very satisfactory. First, they will often not work, for the prerequisite  $F$ -ratios will come out to be significant. Note that the first procedure requires that  $\sigma_{ts}^2 = 0$ , and this is unlikely except with a rather homogeneous group of subjects. The second procedure requires the even more unlikely assumption that  $\sigma_w^2 = 0$ . Because individual words are sampled, they should vary considerably, and so investigators should rarely expect the variance due to words to be null. Second, these procedures are rather risky. In some instances they can lead to a judgment of "significant," while  $F'$  which takes into account all of these variances at once leads to the judgment of "not significant."<sup>2</sup> In

<sup>2</sup> Consider, for example, an instance where the  $F$ -ratio  $MS_{T \times S} / MS_{S \times W \times T}$  is not significant, and  $F_2$  is significant. According to the first procedure, this would allow us to conclude that the Treatments effect is reliable over both Subjects and Words simultaneously. Yet this pattern could very easily arise while  $F_1$  is not significant, and a nonsignificant  $F_1$  would lead us to the contradictory conclusion that the Treatments effects

short,  $F'$  is probably the safest test to use in most instances.

Returning to Study 5, we find that Rubenstein *et al.* did not use  $F'$  or either of the alternative procedures suggested by Winer. Instead, they computed  $F_1$ , finding  $F_1(1, 43) = 10.40$ ,  $p < .005$ , and concluded that there was a significant difference between homophones and nonhomophones. This test, of course, is the normal one for simple Treatments by Subjects factorial designs in which Subjects is assumed to be the only random effect. Thus, it would have been an appropriate test if Words could have been considered a fixed effect, that is, if the 25 homophones and 24 nonhomophones had depleted their respective language populations. But there are obviously many such words in English and other languages that Rubenstein *et al.* did not include in their experiment, and so Words within Treatments should have been treated as a random effect along with Subjects. Since the design in Table 1 is the appropriate one, their significant  $F_1$  allows the three interpretations shown in (2). In particular, (2b) might be correct, and the Treatments effect might actually be null. Interpretation (2b) is especially plausible since one would expect the variance due to Words,  $\sigma_w^2$ , to be considerable by itself. Thus, the significant  $F_1$  Rubenstein *et al.* cited is inconclusive as evidence for the homophone/nonhomophone effect.

Note that if there really were a Treatments effect and  $\sigma_t^2 > 0$ , then  $F_2$ , the  $F$ -ratio in (3), should also be significant, since the expected value of the numerator  $E(MS_T)$  exceeds the expected value of the denominator  $E(MS_{W \times T})$  by the quantity  $qr\sigma_t^2$  plus an additional quantity. Although Rubenstein *et al.* did not compute  $F_2$ , it can be readily calculated from the mean latencies for each word given in the

is not reliable over Subjects, even treating Words as a fixed effect. The quasi  $F$ -ratio, in contrast, is dependent on both  $F_1$  and  $F_2$ , and indeed,  $F'$  must be smaller than both  $F_1$  and  $F_2$  (see below). So  $F'$  would rarely if ever lead to such contradictory conclusions as these.

appendix to Study 5, as I will show later. This F-ratio,  $F_2(1, 45) = 2.00$ , is not significant, leading one to suspect that the Homophony effect is probably not reliable.

For a better test of the Homophony effect, one should calculate  $F'$ , the quasi F-ratio in (5). Note that its computation requires four mean squares,  $MS_T$ ,  $MS_{T \times S}$ ,  $MS_{wWT}$  and  $MS_{S \times wWT}$ . The first three can readily be calculated from Study 5 and its appendix, but the fourth,  $MS_{S \times wWT}$ , cannot be calculated without all of the data. It is therefore impossible to calculate the actual  $F'$  for Rubenstein *et al.*'s data. We can, however, calculate the maximum and minimum values of  $F'$  given the values of  $MS_T$ ,  $MS_{T \times S}$ , and  $MS_{wWT}$  for Rubenstein *et al.*'s data and given certain assumptions. And with  $\max F'$  and  $\min F'$ , we would be able to draw the following conclusions. If  $\max F'$  is not significant, then the actual  $F'$  for Rubenstein *et al.*'s data cannot be significant either, for  $F'$  must be smaller than  $\max F'$ . Or, if  $\min F'$  is significant, then the actual  $F'$  for their data must be significant, since  $F'$  must be larger than  $\min F'$ . Only in the case where  $\max F'$  is significant and  $\min F'$  is not would we not be able to draw any conclusions, for then the actual  $F'$  might or might not be significant.

The calculation of  $\max F'$  and  $\min F'$  follows a straightforward line of reasoning. For given values of  $MS_T$ ,  $MS_{T \times S}$ , and  $MS_{wWT}$ ,  $F'$  will be at a maximum when  $MS_{S \times wWT}$  is at a maximum and at a minimum when  $MS_{S \times wWT}$  is at a minimum. The maximum value of  $MS_{S \times wWT}$ , in turn, can be reckoned as follows. In Table 1, it can be seen that  $E(MS_{S \times wWT})$  cannot be larger than the smallest expected value of the other mean squares in the table since all of the other mean squares in the table contain extra variances. In particular,  $MS_{S \times wWT}$  cannot, on the average, exceed  $MS_{T \times S}$ , the smallest of the remaining mean squares in the Rubenstein *et al.* data. Because of sampling variation, however, both  $MS_{S \times wWT}$  and  $MS_{T \times S}$  are imperfect estimates of their expected values, and so  $MS_{S \times wWT}$  could, by chance, be somewhat larger than  $MS_{T \times S}$ . So under the rather

unlikely condition that  $\sigma_{ts}^2 = 0$ ,  $MS_{S \times wWT}$  will be significantly larger (with  $\alpha = .05$ ) than  $MS_{T \times S}$  only 2.5% of the time (in a two-tailed test). Thus, the practical limit to be placed on the size of  $MS_{S \times wWT}$  is that it not be significantly larger than  $MS_{T \times S}$ .<sup>3</sup> That is, we are interested in the critical value of the following F-ratio:

$$(7) F_3[p(q-1)(r-1), (p-1)(r-1)] \\ = MS_{S \times wWT} / MS_{T \times S}$$

Let the critical value of  $F_3$  at the .05 level be denoted by  $F_3^*$ . Then, by simple algebra the maximum allowable value of  $MS_{S \times wWT}$  is  $F_3^* MS_{T \times S}$ , and therefore  $\max F'$  is given by the following formula for ( $MS_{T \times S} < MS_{wWT}$ ):

$$(8) \max F'(i, j) = (MS_T + F_3^* MS_{T \times S}) / \\ (MS_{T \times S} + MS_{wWT})$$

where  $i$  and  $j$  are defined as in (6), but with the value of  $F_3^* MS_{T \times S}$  replacing  $MS_{S \times wWT}$  in (6). It can be shown that the actual  $F'$  will always be less significant than  $\max F'$  so long as  $MS_T / (p-1) > F_3^* MS_{T \times S} / p(q-1)(r-1)$ ; this condition, of course, will invariably hold for any interesting Treatments effects since for them  $MS_T$  will have to be larger than  $F_3^* MS_{T \times S}$  and in any case  $(p-1)$  will be smaller, typically much smaller, than  $p(q-1)(r-1)$ . The minimum value of  $MS_{S \times wWT}$ , obviously, is zero, and so  $\min F'$  is given by the formula:

$$(9) \min F'(i, j) = MS_T / (MS_{T \times S} + MS_{wWT})$$

where  $i = (p-1)$  and  $j$  is as defined in (6). It can easily be shown that the actual  $F'$  will always be more significant than  $\min F'$ .

<sup>3</sup> Note that if  $MS_{S \times wWT}$  were significantly larger than  $MS_{T \times S}$ , we would have reason to conclude (assuming that  $\sigma_{ts}^2 = 0$ ) that  $MS_{S \times wWT}$  was an overestimate of  $\sigma_e^2 + \sigma_{ws}^2$ , or that  $MS_{T \times S}$  was an underestimate of this quantity, or both. In this case  $F'$  probably has a positive bias, and consequently, the probability of a Type I error in testing the Treatments effect is higher than the stated  $\alpha$  level. When this happens, there may be something amiss in the methodology of the experiment. So when  $MS_{S \times wWT} / MS_{T \times S}$  is significant, it is best to use the more conservative  $\max F'$  instead of the actual  $F'$ , thereby treating the quasi F-ratio as having a distribution truncated at the value of  $\max F'$ .

