JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR 12, 335–359 (1973)

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

HERBERT H. CLARK[1]

*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,

[1] 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, $MS_{error}$, such that $E(MS_T)$ exceeds $E(MS_{error})$ 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 × S | Treatments × Subjects | $(p-1)(r-1)$ | $\sigma_e^2+\sigma_{ws}^2+q\sigma_{ts}^2$ |
| S × WwT | Subjects × 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:

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

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:

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

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:

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

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):

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

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):

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

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:

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

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 $MS_{T\times S}/MS_{S\times WwT}$ 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 nonsignificant F-ratio, $MS_{WwT}/MS_{S\times WwT}$, 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."[2] In 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 $E(MS_{WwT})$ 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

[2] Consider, for example, an instance where the F-ratio $MS_{T\times S}/MS_{S\times WwT}$ 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 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 $MS_{T\times S}$.[3] That is, we are interested in the critical value of the following F-ratio:

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

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 $(MS_{T\times S} < MS_{WwT})$:

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

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:

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

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* $F'$.

[3] 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* $F'$.

Now we are in a position to compute *max* $F'$ for Study 5. $F_3^*$, the critical value for $F_3(1935, 43)$ at the .05 level (two-tailed), is 1.64. Then, by the formula in (8), *max* $F'$ turns out to be 1.94 with 1 and 62 degrees of freedom. Since this value is not significant, the actual $F'$ for Study 5 is not significant either, and so there is no reliable support, that is, no statistical justification, in the data for the conclusion that homophones are recognized more slowly than nonhomophones.

In the course of this argument, I have presented three F-ratios: $F_1$, $F_2$, and $F'$. What exactly does each of them tell us? Roughly speaking, $F_1$ indicates what should happen if the same 25 homophones and 24 nonhomophones were given to a new sample of 44 subjects. Because $F_1$ was significant, we can be fairly certain that the Homophony effect will replicate on this new sample of subjects. On the other hand, $F_2$ indicates what should happen if the same 44 subjects were given a new random sample of 25 homophones and 24 nonhomophones. The fact that $F_2$ was not significant implies that the Homophony effect should not necessarily replicate on the new sample of words. Finally, $F'$ tells us what should happen both with a new sample of 44 subjects and with a new sample of 25 homophones and 24 nonhomophones. Because it was not significant, there is no assurance that the Homophony effect would replicate in this case. From these rough descriptions, we should also expect, in general, that if either $F_1$ or $F_2$ is not significant, then $F'$ will not be significant either, and this will almost always be the case (see below).

Statistically, Study 5 was not quite as simple as I have presented it so far. While $MS_T$ and $MS_{T\times S}$ were calculated directly from the text of Study 5, $MS_{WwT}$ required the use of the geometric mean latencies for each of the 25 homophones and 24 nonhomophones as reported in the appendix to Study 5. As it happened, Rubenstein *et al.* had also divided the homophones and nonhomophones into high and low frequency ranges such that the 49 words actually constituted a Homophony by Frequency design. So to calculate $MS_{WwT}$ I took logarithmic transforms of the 49 latencies, just as Rubenstein *et al.* had done, and submitted them to the appropriate $2 \times 2$ factorial design, with Words nested within Homophony and Frequency. This design is completely analogous to the typical between-subjects design except that here the sampling factor is Words, not Subjects. Since there were unequal numbers of words within the various conditions, I also had to make use of Winer's (1971) method of unweighted means for the analysis of variance. The mean square for Words within Homophony and Frequency for this design can be shown to be identical to the mean square (times 44, the number of subjects) that is required for the more complete analysis indicated above. It was this mean square that I used as the value of $MS_{WwT}$ in the above calculations. In general, $MS_T$ and $MS_{WwT}$—hence $F_2$—can be computed simply by collapsing across subjects and by applying the appropriate analysis of variance as if Words was the only random effect.

*Statistical evidence for Studies 1 through 5.* Following exactly the same procedures as I used for Study 5, I have computed $F_2$, *max* F', and *min* $F'$ for each effect originally reported as significant in all five studies by Rubenstein *et al.* and have listed these values in Table 2 opposite the values of $F_1$ reported in the original studies.[4] As Table 2 makes plain, there are large discrepancies between the values of $F_1$ reported by Rubenstein *et al.* and the values of *max* $F'$ calculated for the

[4] According to H. Rubenstein, the authors of Study 1 inadvertently listed the arithmetic, rather than the geometric, means for each letter string in the appendix to the study; H. Rubenstein has kindly sent me the geometric means, and the calculations I have done are based on them. In addition, the geometric means calculated from the appendix of Study 3 (namely, 880, 896, and 995) do not jibe exactly with the reported geometric means (859, 874, and 966, respectively), apparently because of Rubenstein *et al.*'s procedure for replacing missing data.

TABLE 2

F-RATIOS BY SUBJECT ($F_1$), F-RATIOS BY WORD ($F_2$), AND MAXIMUM AND MINIMUM QUASI F-RATIOS (*MAX F'* AND *MIN F'*) FOR STUDIES 1 THROUGH 5 BY RUBENSTEIN ET AL.

| Study | Source of variance | $F_1$ | Signif-icance | $F_2$ | Signif-icance | *Min F'* and *Max F'*[b] | Signif-icance[c] |
|---|---|---|---|---|---|---|---|
| 1 | Frequency | $F(2, 76) = 45.53$[a] | .001 | $F(2, 168) = 53.02$ | .001 | $F(2, 197) = 24.43, 25.10$ | .001 |
| | Homography (H) | $F(1, 38) = 10.72$ | .005 | $F(1, 168) = 4.26$ | .05 | $F(1, 193) = 3.05, 3.52$ | n.s. |
| | Concreteness × H | $F(1, 38) = 17.77$ | .001 | $F(1, 168) = 1.17$ | n.s. | $F(1, 187) = 1.10, 1.20$ | n.s. |
| 2 | Systematicity (S) | $F(1, 44) = 34.80$ | .001 | $F(1, 104) = 4.14$ | .05 | $F(1, 126) = 3.70, 3.87$ | n.s. |
| | Equiprobability (E) | $F(1, 44) = 19.93$ | .001 | $F(1, 104) = 2.79$ | n.s. | $F(1, 129) = 2.45, 2.65$ | n.s. |
| | S × E | $F(1, 44) = 11.59$ | .005 | $F(1, 104) = 0.98$ | n.s. | $F(1, 120) = 0.90, 1.03$ | n.s. |
| | Frequency (F) | $F(2, 88) = 27.28$[a] | .001 | $F(2, 104) = 4.62$ | .025 | $F(2, 138) = 3.95, 4.16$ | .025 |
| | S × E × F | $F(2, 88) = 7.17$[a] | .001 | $F(2, 104) = 0.18$ | n.s. | $F(3, 109) = 0.18, 0.22$ | n.s. |
| 3 | Legality | $F(1, 44) = 108.75$ | .001 | $F(1, 185) = 86.18$ | .001 | $F(1, 163) = 48.08, 48.81$ | .001 |
| | Pronounceability within illegality | $F(1, 44) = 9.83$ | .005 | $F(1, 185) = 0.81$ | n.s. | $F(1, 212) = 0.75, 0.88$ | n.s. |
| 4 | Homophony | $F(1, 43) = 34.97$ | .001 | $F(1, 53) = 7.46$ | .01 | $F(1, 74) = 6.30, 6.60$ | .025 |
| 5 | Homophony | $F(1, 43) = 10.40$ | .005 | $F(1, 45) = 2.00$ | n.s. | $F(1, 62) = 1.68, 1.94$ | n.s. |
| | Frequency | $F(1, 43) = 83.41$ | .001 | $F(1, 45) = 45.01$ | .001 | $F(1, 82) = 29.23, 29.80$ | .001 |

[a] The degrees of freedom for $F(2, 76)$ and $F(2, 88)$ in Studies 1 and 2 were incorrectly reported in the original studies as $F(2, 38)$ and $F(2, 44)$, respectively.

[b] The degrees of freedom for *min F'* and *max F'* were the same in all cases but one, the S × E × F interaction in Study 2, where *min F'* had 2 and 109 degrees of freedom and *max F'* had 3 and 109 degrees of freedom.

[c] The level of significance was the same for *min F'* and *max F'* for all 13 pairs of values.

same data. While all 13 values of $F_1$ were significant at the .005 level, only five values of *max F'* are significant, two at only the .025 level. Thus, when Words is treated as a random effect along with Subjects, a number of the effects originally reported as significant turn out to be statistically unreliable.

In addition, Table 2 illustrates two general points about quasi F-ratios. First, consider those instances where both $F_1$ and $F_2$ are larger than $F_3^*$ (1.64 for most of Table 2). These instances turn out to be the only interesting ones, since all other instances can be shown to result automatically in a non-significant $F'$. In these cases, it can be shown (see the Appendix) that *max F'*, hence the actual $F'$, will never be larger than $F_1$ or $F_2$, whichever is smaller. This agrees with our intuitions about $F'$. If $F_1$ and $F_2$ indicate what would happen with new samples of subjects and words, respectively, then $F'$ should be smaller than either, since it indicates what should happen both with new subjects *and* with new words.[5] Second, in most instances *max F'* is not much larger than *min F'*. It can be shown (see the Appendix) that when $F_1 \geqslant F_2$ (as is true for most of Table 2), *max F'* is algebraically equivalent to $(1 + F_3^*/F_1)$ times *min F'*. Consider the Systematicity effect in Study 2 (where $F_1 = 34.80$ and $F_3^* = 1.64$). There *max F'* is only 5% larger than *min F'*. For the Homophony effect in Study 5 where $F_1$ is only 10.40, *max F'* is still only 16% larger than *min F'*. All this indicates that in many instances *min F'* will not be much smaller than the actual $F'$ and could therefore be used as a convenient substitute for the actual $F'$ when the latter is too cumbersome to calculate

[5] Strictly speaking, although $F'$ must be smaller than both $F_1$ and $F_2$, it is possible for $F'$ to be significant (because of the possibility of increased degrees of freedom) even though the smaller of $F_1$ and $F_2$ is not. This possibility, however, is very remote and has never occurred in my experience.

easily (see below). In addition, it should be noted that *max F'* does not change much when the $\alpha$ level is raised from .05 ($F_3^* = 1.64$) to .01 ($F_3^* = 1.93$). In no case in Table 2 does a nonsignificant *max F'* now become significant.

How do all these statistics affect the conclusions drawn in Studies 1 through 5? The main hypothesis tested in Studies 1 and 2 required that at least some of the effects labeled Homography, Concreteness, Systematicity, and Equiprobability, or their interactions, be significant. The quasi F-ratios show no reliable evidence for any of these effects in either study. The main hypothesis tested in Studies 3, 4, and 5 required that at least some of the effects labeled Pronounceability within Illegality and Homophony be significant. There was support for this hypothesis in the Homophony effect of Study 4, where the letter strings used were nonwords, but not in the Homophony effect of Study 5, where the letter strings were actual words. One cannot conclude, of course, that those effects lacking significance in Studies 1 through 5 are not real. Other more sensitive experiments, or even more powerful analyses of these same studies with frequency handled in a more detailed way, might well show any one of these nonsignificant effects to be real. The argument is, simply, that the data in Studies 1 through 5 as analyzed provide no statistical justification for the conclusion that these effects are real.

*The Meyer Study*

A rather different example of a study committing the language-as-fixed-effect fallacy is Meyer's (1970) detailed investigation of the representation and retrieval of stored semantic information. He reported two experiments. In one, 56 subjects were timed as they made true–false judgments of 192 test items like *All chairs are furniture* (that is, *All* S *are* P). In the second, 32 subjects went through the identical procedure with a similar set of 384 sentences in which *all* had been replaced by *some*, as in *Some chairs are furniture*. Meyer classified the test sentences of each experiment into 16 categories according to the subject–predicate (S–P) relation they exhibited and then carried out a variety of comparisons among the categories in order to distinguish among a number of competing theories of semantic retrieval. For purpose of illustration I will examine only one of these comparisons in detail.

In the case under examination sentences such as *All stones are rubies* were compared with ones such as *All solids are rubies*. *Stones* is said to be a "small" superset of *rubies* and *solids*, a "large" superset of *rubies*, because *stones* is itself a subset of *solids*. This particular pair of sentences, then, can be thought of as having been constructed from the Word-triple *rubies-stones-solids*, in which *rubies* is a subset of *stones* which in turn is a subset of *solids*. Meyer composed eight such Word-triples (implicitly taking them from the population of all Word-triples with this nesting property), constructed one pair of sentences from each Word-triple, and then examined the latencies of 56 subjects to all 16 of the resulting sentences. Since each sentence in the "small" superset category was paired with one in the "large" superset category, these sentences fit into a simple factorial design with three crossed factors: Treatments (that is, Size of superset relation of S to P) Word-triples, and Subjects.

The analysis of variance for such a factorial design is indicated for the general case in Table 3 (see Coleman, 1964; Winer, 1971; and others). This design contains $p$ fixed Treatments, $q$ random Word-triples, and $r$ random Subjects. The problem here again is how to choose the correct F-ratio for testing the reliability of the Treatments effect. As in the previous design, it is not correct to use $F_1$, which tests the Treatments effect against the Treatments by Subjects interaction:

$$(10)\quad F_1(p-1,(p-1)(r-1)) = MS_T/MS_{T\times S}$$

A little algebra shows that $F_1$, if significant, guarantees only that $q\sigma_t^2 + \sigma_{tw}^2 > 0$, and this

TABLE 3

SOURCES OF VARIANCE AND EXPECTED MEAN SQUARES FOR MIXED FACTORIAL DESIGN WITH ONE FIXED EFFECT AND TWO RANDOM EFFECTS

| Label | Source of variance | Degrees of freedom | Expected value of mean square |
|---|---|---|---|
| T | Treatments ($p$) | $p-1$ | $\sigma_e^2+\sigma_{tws}^2+q\sigma_{ts}^2+r\sigma_{tw}^2+qr\sigma_t^2$ |
| W | Words ($q$) | $q-1$ | $\sigma_e^2+p\sigma_{ws}^2+pr\sigma_w^2$ |
| S | Subjects ($r$) | $r-1$ | $\sigma_e^2+p\sigma_{ws}^2+pq\sigma_s^2$ |
| T × W | Treatments × Words | $(p-1)(q-1)$ | $\sigma_e^2+\sigma_{tws}^2+r\sigma_{tw}^2$ |
| T × S | Treatments × Subjects | $(p-1)(r-1)$ | $\sigma_e^2+\sigma_{tws}^2+q\sigma_{ts}^2$ |
| W × S | Words × Subjects | $(q-1)(r-1)$ | $\sigma_e^2+p\sigma_{ws}^2$ |
| T × W × S | Treatments × Words × Subjects | $(p-1)(q-1)(r-1)$ | $\sigma_e^2+\sigma_{tws}^2$ |

gives no assurance that $\sigma_t^2 > 0$, that is, that the variance due to Treatments is greater than zero. The F-ratio in (11), which I will denote as $F_2$ since it is analogous to $F_2$ in the previous design, is not the right one either:

$$(11)\quad F_2(p-1,(p-1)(q-1)) = MS_T / MS_{T\times W}$$

It suffers from the same fault, for its significance guarantees that $r\sigma_t^2+\sigma_{ts}^2>0$, but not that $\sigma_t^2>0$. Winer (1971) therefore recommends the following quasi F-ratio, which I will again denote as $F'$:

$$(12)\quad F'(i,j) = (MS_T + MS_{T\times S\times W}) / (MS_{T\times S} + MS_{T\times W})$$

As before, the degrees of freedom $i$ is the nearest integer calculated by the formula in (13):

$$(13)\quad i = (MS_1+MS_2)^2 \Big/ \left(\frac{MS_1^2}{n_1}+\frac{MS_2^2}{n_2}\right)$$

where $MS_1$ and $MS_2$ are the two mean squares in the numerator, and $n_1$ and $n_2$ are their respective degrees of freedom. The degrees of freedom $j$ is calculated by the same formula except with the two mean squares from the denominator. Although Winer (1971, p. 378) again discusses alternatives to this method, their assumptions will be difficult to satisfy in most cases.

Like Rubenstein *et al.*, Meyer committed the language-as-fixed-effect fallacy and calculated $F_1$ instead of $F'$. Since he found $F_1(1, 48) = 5.5$, $p < .05$ for this comparison, he concluded that "large" supersets take reliably longer than "small" supersets. But to see if this 44 msec difference really is reliable, we must compute $F'$. Unfortunately, this is impossible to do from the information presented in Meyer's paper, for there is no way to calculate $MS_{T\times W}$ or $MS_{T\times S\times W}$. To illustrate what might happen, therefore, I will make reference to some data recently collected by Lance Rips in a partial replication of Meyer's two experiments with 24 subjects and nine newly sampled word-triples.[6] First, one can attempt a crude estimate of *max* $F'$ for Meyer's data. In Rips' data $MS_{T\times W}$ was more than twice as large as $MS_{T\times S}$. Since Meyer used many more subjects than Rips did (56 to 24), in Meyer's data $MS_{T\times W}$ is likely to be even more than twice the size of $MS_{T\times S}$. If we assume that $MS_{T\times W} \geq 2MS_{T\times S}$, or equivalently, $F_1 \geq 2F_2$, in Meyer's data, then *max* $F'$ turns out to be 2.34, which with 2 and 15 degrees of freedom is not significant. Second, one can turn the argument around and compute the standard deviation for the Treatments by Word-triples interaction effects that would be required for *max* $F'$ to be significant (at the .05 level) given Meyer's own $F_1$. This standard deviation turns out to be 39 msec, which is less than half the size of Rips' standard deviation of 91 msec. Thus, for Meyer's

[6] Rips, Lance J. Quantification and semantic memory. In preparation. I am deeply indebted to Rips for the use of his data.

44-msec difference to be significant, his experiment would have to have been very much more precise than Rips', since Meyer used about the same number of word-triples as Rips did. Third, Rips' data can be used in quite a different way. Note that if Meyer's 44-msec difference is really reliable across subjects and word-triples simultaneously, then it ought to replicate on a new sample of subjects *and* word-triples, as in Rips' study. But it did not. The mean difference in Rips' data went 20 msec in the opposite direction to Meyer's.

It is instructive to look at the $F_1$s and $F'$s in Rips' data for six treatment comparisons (similar to and including the one just presented) common to the Rips and Meyer studies. The $F_1$s for these six comparisons in Meyer's data (see his Tables 8 and 16) were 8.4, 19.0, 5.5, 17.4, 176.4, and 202.4, all of which were significant at at least the .05 level.[7] The respective $F_1$s in Rips data were .66, 2.32, 2.73, 13.21, 30.09, and 18.56, only the last three of which were significant (with 1 and 23 degrees of freedom). So Meyer was clearly able to produce more significant $F_1$s than Rips, and this could have happened for a number of reasons, including the greater number of subjects in Meyer's experiment. Rips' data, however, show dramatically how much lower the significance levels are for the $F'$s than for the $F_1$s calculated from the same set of data. The first three $F'$s turn out to be .57, 1.17, and 1.24, respectively, none of which approaches significance. The $F'$s for the last three comparisons are $F'(1, 13) = 3.36$, $F'(1, 14) = 8.68$, and $F'(1, 16) = 6.06$, respectively, where only the latter two are significant (at the .025 and .05 levels, respectively). Note that these last three $F'$s are each less than a third the magnitude of the respective $F_1$s, and they have fewer degrees of freedom as well. If this observation is also generally true for the 11 significant $F_1$s of similar treatment comparisons in Meyer's paper (see his Tables 8 and 16), it also suggests that those $F_1$s overestimate the actual reliability of Meyer's findings by considerable amounts, and, indeed, some of these findings may turn out to be statistically unreliable.

The Meyer study also illustrates a further consequence of the language-as-fixed-effect fallacy, and this is in the reporting of standard errors. Throughout his paper, Meyer followed the commendable practice of reporting both the mean and standard error for each type of sentence, as in $1252 \pm 12$ msec. The standard errors actually reported, however, are misleading. This particular 12-msec estimate, for example, was apparently calculated by the formula $[MS_{T\times S}/qr]^{\frac{1}{2}}$, which assumes that Subjects is the only random effect. That is, it indicates how much the mean should vary with new samples of 56 subjects judging the same eight sentences. To be appropriate, however, it should indicate how much the mean should vary when both subjects and sentences are sampled anew. This standard error can be shown to be:

$$(14)\quad \mathrm{SE} = [(MS_{T\times S} + MS_{T\times W} - MS_{T\times S\times W})/qr]^{1/2}$$

In Rips' data, the standard error treating only Subjects as a random effect was 21 msec. When calculated according to (14), however, it increased to 30 msec. It is likely that the standard errors in Meyer's study are underestimates of their more appropriate values by similar amounts.

### *Other Studies in Semantic Memory*

The studies of Rubenstein *et al.* and Meyer, of course, are not the only ones within the field of semantic memory to have committed the language-as-fixed-effect fallacy. Indeed, most of the remaining ones have too, one

[7] These six F-ratios correspond, respectively, to the following six conditions in Tables 8 and 16 in Meyer (1970): (1) Variations in P-size for Disjoint PA sentences; (2) P-size, Subset, UA sentences; (3) S-size, Superset, UA sentences; (4) S-size, Disjoint, PA sentences; (5) P-size, Subset, PA sentences; and (6) S-size, Superset, PA sentences.

way or another. For example, the studies of Collins and Quillian (1969, 1970a, b, 1971), Conrad (1972), Landauer and Freedman (1968), Landauer and Meyer (1972), Loftus (in press), Meyer (1971, 1973 in press), Meyer and Ellis (1970), Meyer and Schvaneveldt (1971), Schaeffer and Wallace (1969, 1970a, b), Schvaneveldt and Meyer (1973), Smith (1967), and Wilkins (1971) all based their conclusions on statistics in which Words was considered a fixed, instead of a random, effect. None of these studies reports statistical evidence that their results generalize to the language population sampled. Because of this oversight many of these studies rest on very precarious findings indeed, for many of the "significant" differences reported are rather small (as low as 17 msec) and many are based on small samples of words (as few as 16 words total). If the Rubenstein *et al.* studies are at all representative—and they contained a relatively large sample of words for this type of study—then many of the smaller "significant" differences in the cited studies may well turn out to be unreliable with the proper statistical tests.

Several investigators in this area, however, have recognized the language-as-fixed-effect fallacy, at least in part, and have endeavored to demonstrate that their findings do generalize beyond their language sample. For example, Freedman and Loftus (1971), Loftus (1972, 1973), Loftus and Freedman (1972), Loftus, Freedman, and Loftus (1971), and Rips, Shoben, and Smith (1973) listed their stimuli in full and computed at least some statistics appropriate to their language sample. The latter five studies, in fact, reversed the usual failing in such statistical tests. Each of them, in effect, reported $F_2$, but not $F_1$, thereby treating Words as a random effect, as was appropriate, but incorrectly treating Subjects as a fixed effect. Smith *et al.* (1972) reported that their effects "were consistent across individual subjects, individual nouns, and different levels of practice," which is equivalent to reporting both $F_1$ and $F_2$ separately.[8] In one experiment, Rosch (1973) reported the reliability of her findings over Words and Subjects separately, and this is equivalent to giving $F_1$ and $F_2$ separately; in another experiment, however, she reported the reliability only over Subjects. All of these practices, of course, are inadequate, since $F_1$ and $F_2$ can both be significant while $F'$ remains nonsignificant. Nevertheless, these studies have at least taken a step in the right direction.

There are also several studies that have reported the reliability of their findings across the language sample, but have then failed to note that this reliability did not meet the conventional levels of statistical significance. In their Experiment I, for example, Landauer and Freedman (1968) noted that the finding of interest held for six of their eight Word-pair comparisons, or for three of four such comparisons that consider only the negative responses. By a sign test neither of these results is significant, and the $F_2$s computed from their presented data are .69 and 1.26, respectively, neither of which is significant. It should be noted, however, Landauer and Freedman did run a second study to replicate their findings on a different sample of words, although in the latter experiment they did not report any language statistics. In the area of sentence memory, Bransford, Barclay, and Franks (1972) reported that the main finding in their Experiments 1 and 2 held for only 10 of the 14 sentence-pairs sampled. This is not significant by a sign test either ($p < .09$), although it could be by a more sensitive $t$-test. These examples demonstrate just how blinding the language-as-fixed-effect fallacy can be. Results

[8] Smith and Haviland (personal communication) note that the original version of the Smith *et al.* paper contained a much fuller discussion of the reliability of their findings across words. The reviewer for their paper, however, instructed them to leave out all mention of these statistics. I have had similar struggles with editors and reviewers on this issue myself and others probably have too. This may be one reason why there have been so few papers reporting such statistics.

that are unreliable for the language population by the usual criteria are nevertheless reported as though they were reliable because the statistical tests by subjects indicate significance.

## Remedies

The remedies for the language-as-fixed-effect fallacy are for the most part obvious. They include doing the right statistics, choosing the appropriate experimental design, and selecting a random or representative sample of language. In some instances, another available remedy is to argue from single cases, a method that requires rather careful thought about the purposes of psychological experimentation.

### *Do the Right Statistics*

For a variety of reasons, fledgling psychologists are not normally taught how to generalize a finding to two populations at the same time, nor even why they should. Because of the emphasis on subjects in statistics texts, most psychologists simply learn to observe the current statistical shibboleth and treat subjects as their only random effect. But as Coleman (1964) has argued, language should also be treated as a random effect in many studies, and this will require more complicated statistics, often some form of quasi F-ratio. Although I have laid out the appropriate statistics for two typical designs, one hierarchical and one factorial, there are many other possible experimental designs and they will require other statistical considerations. One should consult such reference texts as Winer (1971) and Snedecor and Cochran (1967) for further advice.

In some experiments, however, investigators may wish to use an alternative to $F'$. Often, while it is relatively easy to compute $F_1$ and $F_2$ (the F-ratios by subjects and words, respectively), the additional mean square required for $F'$, namely $MS_{\mathrm{S\times WwT}}$ or $MS_{\mathrm{T\times S\times W}}$, raises special problems. For example, this term may be very complicated to compute because of missing data that would have to be estimated; or, the term may be suspect since it is based on 0s and 1s; or, it may simply be too costly to calculate because the experiment is large. However, as long as the investigator can legitimately compute $F_1$ and $F_2$, he can eschew $F'$ and use the computationally simpler *min* $F'$. One formula for *min* $F'$ is given in (9), but an equivalent (see Appendix) and often handier formula is given in (15):

$$(15)\quad \mathit{min}\, F'(i,j) = F_1 F_2/(F_1 + F_2)$$

If $F_1$ has $n$ and $n_1$ degrees of freedom, and $F_2$ has $n$ and $n_2$ degrees of freedom, then $i = n$, and $j$ is the nearest integer given by the following expression (see Appendix for derivation):

$$(16)\quad j = (F_1 + F_2)^2 \Big/ \left(\frac{F_1^2}{n_2} + \frac{F_2^2}{n_1}\right)$$

Since *min* $F'$ will always be smaller than the actual $F'$, *min* $F'$ will never be significant unless the actual $F'$ is too. But this also means that *min* $F'$ will be more conservative than $F'$ by some amount. This amount should not be very large when $n_1$ and $n_2$ are of at least moderate size (see, for example, Table 2). Indeed, the investigator can see just how conservative *min* $F'$ could possibly be by computing *max* $F'$ according to one of the two formulae in (17):

$$(17)\ \text{a.}\quad \mathit{max}\, F' = (1 + F_3^*/F_1)\, \mathit{min}\, F', \quad \text{for} \quad F_1 \geqslant F_2$$

$$\text{b.}\quad \mathit{max}\, F' = (1 + F_3^*/F_2)\, \mathit{min}\, F', \quad \text{for} \quad F_2 \geqslant F_1$$

The $F_3^*$ in (17a) is defined as the critical value of the F-ratio $MS_{\mathrm{T\times S\times W}}/MS_{\mathrm{T\times S}}$, $MS_{\mathrm{S\times WwT}}/MS_{\mathrm{T\times S}}$, or $MS_{\mathrm{S\times WwT}}/MS_{\mathrm{SwT}}$, whichever is appropriate for the design; the $F_3^*$ in (17b) is defined analogously, but with $MS_{\mathrm{T\times W}}$ or $MS_{\mathrm{WwT}}$ in the denominator, whichever is appropriate. The range of $F'$ given by *max* $F'$ minus *min* $F'$ will decrease as the number of subjects and words increase and as the larger

of $F_1$ and $F_2$ increases.[9] Thus, *min* $F'$ is to be recommended in cases where $F_1$ and $F_2$ are easy to calculate but $F'$ is not. The use of *min* $F'$ in these cases seems far preferable to the next best procedure, which is to report $F_1$ and $F_2$ separately with the requirement that both be significant.[10] The latter procedure will sometimes lead to judgments of "significant" where such judgments are not justifiable, because $F_1$ and $F_2$ can both be significant even though $F'$ is not.

Investigators can sometimes avoid the use of $F'$ and its restrictive assumptions altogether by choosing designs in which simpler statistics are appropriate, even when Words *is* treated as a random effect. I will illustrate with two designs discussed by Winer (1971, p. 364–365). Imagine that Cushman, a psychologist, wishes to compare the recall of concrete and abstract words. She therefore gives each of 20 subjects eight concrete and eight abstract words to recall. But with scores of 1s and 0s, Cushman does not feel that the computation of $F'$ would be legitimate. To sidestep the problem, she therefore presents each subject with a different set of eight concrete and eight abstract words. With Words (a random effect) nested within Subjects (a random effect), Winer (p. 365) shows that the Treatments effect is legitimately tested against the Treatments by Subjects interaction, that is, with the F-ratio $MS_T/MS_{T\times S}$. That is, Cushman can collapse over the Words factor altogether, compute an F-ratio by subjects (or an equivalent nonparametric test such as the Wilcoxon test or sign test), and, if the test shows significance, justifiably claim that her finding is general for both words and subjects. (See Carroll 1966 for an application of this design.) The second design is similar. Imagine that Cushman has given half her subjects only concrete words and the other half only abstract words. With Words nested within Subjects, and Subjects in turn nested within Treatments, Winer (p. 364) demonstrates that the Treatments effect is appropriately tested against the Subjects within Treatments error term, that is, by the F-ratio $MS_T/MS_{S\text{w}T}$. So again, Cushman can collapse over the Words factor entirely, compute the appropriate parametric or nonparametric tests by subjects, and legitimately generalize her findings to both words and subjects. Although both of these designs require a large number of words, they have the advantage that they simplify the statistics required, especially if parametric statistics are deemed inappropriate for the design. For other possible simplifying designs, one should consult Winer (1971) or other similar reference texts.

When should the investigator treat language as a random effect? The answer is, whenever the language stimuli used do not deplete the population from which they were drawn. Note that the answer is *not*, whenever the language stimuli used were chosen *at random* from this population. The latter requirement is, in a sense, secondary to whether or not language should be treated as a random effect. Consider, for example, the Meyer (1970) study examined in detail above. In it, Meyer explicitly considered treating Word-triples as a random effect, but rejected the idea "because of the procedure used to select the test stimuli" (p. 263). Meyer based this decision, presumably, on the fact that he had not sampled the word-triples at random, but had composed them with the aid of

[9] It should be noted that the degrees of freedom of *max* $F'$ will not necessarily be the same as those of *min* $F'$: Whereas the degrees of freedom $j$ for the denominator will not change, the degrees of freedom $i$ for the numerator can be larger for *max* $F'$ than for *min* $F'$. Nevertheless, when $F_1$ and $F_2$ are both significant, the nearest integer value of $i$ will typically be the same for both *max* $F'$ and *min* $F'$ too (for example, see footnote *b* of Table 2).

[10] In my own previous research, because of difficulties in calculating $F'$, I have relied, with lapses, on this weaker method, reporting $F_1$ and $F_2$ separately (see Clark & Begun, 1968; Clark & Card, 1969; Clark & Clark, 1968). If *min* $F'$ had been available then, it would obviously have been a more appropriate statistic to use. Fortunately, the statistical conclusions in these studies are affected very little by the change from $F_1$ and $F_2$ to *min* $F'$.

dictionaries and thesauruses while trying to minimize word ambiguities, frequency biases, and the like. Obviously, Meyer could have composed other word-triples for each comparison than the ones he did, for he composed the same type of word-triples for other comparisons in the study. But by treating Word-triples as a fixed effect, Meyer cannot legitimately generalize his findings even to the population defined by his nonrandom sampling procedure. This, in turn, allows him no possibility of generalizing his findings to the more inclusive population—the unbiased collection of all appropriate word-triples—as Meyer clearly wanted to do. So even though the investigator knows that his words were not chosen at random, he should treat Words as a random effect so long as he can think of other words he could have chosen instead. The nonrandom sampling procedure causes difficulty only later when the investigator wants to determine exactly what population he can legitimately generalize his results to.

*Choose the Appropriate Design*

On the face of it, choosing the appropriate design seems like an innocent enough problem, for everyone is presumably taught how in courses on experimental design. The introduction of Words as a second random factor, however, brings with it special considerations. In traditional designs one worries about whether to use a between-subject or a within-subject design. And as the experienced investigator knows, within-subject designs generally require fewer subjects than the comparable between-subjects designs, since each subject "serves as his own control." Furthermore, more subjects are required in those experiments in which the investigator expects smaller differences among the treatments. Increasing the number of subjects raises the power of the design. All of this traditional wisdom applies *ceteris paribus* to designs with Words as a random factor. Within-word designs are more sensitive than between-word designs, since in the former each word serves as its own control. Consequently, within-word designs generally require fewer words than between-word designs. Also, to detect small differences, the investigator must increase the number of words in his sample.

The most important rule to keep in mind, however, is this: An experimental design is only as sensitive as the less sensitive of the two subdesigns it contains—the Treatments by Subjects subdesign and the Treatments by Words subdesign. The former is produced by collapsing across Words, and it leads to the calculation of $F_1$; the latter is produced by collapsing across Subjects, and it leads to the calculation of $F_2$. As noted above, the quasi F-ratio, $F'$, must be smaller than both $F_1$ and $F_2$ (assuming both are larger than $F_3^*$, a minimal requirement). This implies that to increase sensitivity in an experiment, the investigator cannot simply add in more subjects alone. While this will increase $F_1$, it will affect $F_2$ very little. Likewise, he cannot simply add in more words alone, for this will increase $F_2$, but not increase $F_1$ appreciably. To increase his possibility of generalizing to both the language and the subject populations simultaneously, the investigator must add in more subjects and more words in comparable amounts. As illustration of what can happen if one does not follow this advice, consider the comparison of Meyer's (1970) examined in detail above. The design for this comparison contained 56 subjects, constituting a very sensitive subdesign by subjects; but it contained only eight word-triples, constituting a rather insensitive subdesign by word-triples. The former half of the design led to a significant $F_1$, namely, 5.5. But if Meyer had computed the statistics for the latter half, he would probably have found a much smaller $F_2$. It is the latter, therefore, that places the upper limit on the $F'$ he could have obtained. Generally speaking, this type of imbalance in design works against the investigator and should be avoided whenever possible.

Two concrete examples will be helpful in showing how $F'$ depends on $F_1$ and $F_2$. In

these two examples, $F_1$ and $F_2$ will both be assumed to have 1 and 20 degrees of freedom, as is appropriate for the typical experiment in which the investigator wishes to compare one treatment against another with a sample of around 20 subjects and 20 words. First, imagine that $F_1$ and $F_2$ are both significant at the .01 level, that is, $F_1 = F_2 = 8.10$. In this case, $F'$, with 1 and 40 degrees of freedom, will be somewhere between 4.05 (*min* $F'$) and 5.09 (*max* $F'$) where the entire range of $F'$ is significant at only the .05 level. From this rather typical example, we see that $F'$ will often be only marginally significant ($p < .05$) even though both $F_1$ and $F_2$ are at quite a respectable level of significance ($p < .01$). Indeed, $F'$ is most sensitive when $F_1$ and $F_2$ are about the same size. When one of the two is much larger than the other, there is a strikingly large trade-off between them. Imagine, for example, that $F_1$ is equal to 5.87, which is just significant at the .025 level. In this case $F_2$ must be at least 14.40, which is significant at far beyond the .001 level, in order for $F'$ to be significant at the .05 level, that is, for $F'$ (1, 34) to range between 4.17 and 4.78. So if either $F_1$ or $F_2$ is not terribly reliable, the other F-ratio must be highly reliable just to compensate for it.

*Sample Language by Systematic, Repeatable Procedures*

If the investigator is to treat language as a random effect, then he must draw a sample at random from the language population he wishes to generalize to. In the past this requirement has caused investigators much grief. The three main problems have been: (1) defining the language population; (2) sampling without bias from this population; and (3) sampling by a procedure that other investigators can repeat.

Many investigators of semantic memory have drawn their samples from the available word norms such as those for overall word frequency (Kučera & Francis, 1967), frequency of subordinates (Battig & Montague, 1969), frequency of superordinates (Loftus & Scheff, 1971), frequency of free associations (for example, Postman & Keppel, 1970), and concreteness and imagery value (Paivio, Yuille, & Madigan, 1968), as well as from special purpose norms such as those of Conrad's (1972). In doing so, the investigator in each case has thereby defined his population: It is that set of words found in the norms or the population they were sampled from. This population, however, may or may not be appropriate to the intents of the investigator. If he wishes to generalize to all superordinates in English, for example, the Loftus-Scheff norm are, properly speaking, inappropriate, for while they contain many nouns and their superordinates, this is only a selected subset of the possible nouns, and it does not include verbs, adjectives, and other words that have superordinates. As Landauer and Meyer (1972) have pointed out, the investigator must draw his sample from a population that encompasses every domain he wishes to generalize to. Nevertheless, if the investigator can live within these limitations on norms, they are to be recommended on several grounds. First, the investigator can reduce the possibility of unwitting sampling biases. Second, he can use such norms to stratify his word sample according to frequency or some other control factor (as Rubenstein *et al.* did) and thereby make use of more powerful analyses of variance. Third, other investigators can repeat the procedure, drawing comparable samples from the same norms. And finally, other investigators can do detailed studies of other properties of the population defined by the norms either to find related phenomena or to check on the representativeness of the norms themselves. Thus, whenever possible, investigators should use some explicit, repeatable procedure, and norms are often a handy source of words with the wanted properties.

Without such explicit procedures, sampling biases, especially very subtle ones, can easily arise, and they can lead to serious error. To give an example, Collins and Quillian (1969)

were interested in the verification of sentences with various subject–predicate (S–P) relations, sentences such as *A canary is a bird, A canary can sing*, and so on. But instead of sampling their sentences by a systematic procedure, Collins and Quillian composed the sentences themselves, informally assessing their judgments of certain S-P relations against the responses of two subjects. For the so-called "property" sentences (for example, *A canary can sing*), they reported systematically excluding instances in which the property in P was felt to be too closely "associated" with S. By doing so, of course, Collins and Quillian left themselves open to the charge that what they found was not due to the S–P relations themselves, but to the sampling biases correlated with the S–P relations. Conrad (1972) challenged the Collins and Quillian study on just these grounds and attempted to correct the situation by using a systematic sampling procedure. What she did was collect appropriate S–P pairs as composed by subjects and then select instances from among these pairs according to a stratified sampling procedure. On the basis of her results, Conrad was able to argue that Collins and Quillian had used a biased sample of sentences and that it had affected their conclusions. It should be noted, however, that Conrad too relied on inappropriate statistics, treating her word pairs as a fixed effect despite the fact that she had explicitly sampled them. So the question still remains as to how many of Conrad's results will stand up under the appropriate statistics. There are many other studies in semantic memory, such as the Rubenstein *et al.* and Meyer studies examined earlier, in which the investigators selected their stimuli by rather informal procedures. How this has affected their results is impossible to tell.

Sampling biases, generally speaking, have two effects on experimental findings: (1) They spuriously increase the differences between the treatments of interest; and (2) they spuriously reduce the error term for the treatments effect. These two effects, either alone or together, will lead to spuriously high estimates of the reliability of the treatments effect. A concrete example ought to make this clear. In a recent experiment, Smith and Schumacher (unpublished) presented word-pairs in sentences such as *A daisy is a flower* to subjects who were timed as they judged whether the sentence was true or false. Smith and Schumacher compared pairs like *daisy-flower* against pairs like *daisy-plant*, where the two pairs were generated from the word-triple *daisy-flower-plant*, just as in the Meyer (1970) and Collins and Quillian (1969) studies. Smith and Schumacher constructed the word-triples from the Loftus-Scheff norms of superordinates by selecting a word's two most frequently elicited superordinates for which one was the superordinate of the other as well. Half the word-triples they chose could be called "transitive" in that the second word (for example, *flower*) was more frequent than the third (*plant*) as a superordinate of the first (*daisy*); the other half they chose could be called "intransitive" since for them the reverse was true. From Smith and Schumacher's data we can compare what would have happened under a biased sampling procedure that selected only transitive word-triples against what would have happened under an unbiased procedure that selected half transitive and half intransitive word-triples (approximating their actual occurrence in the Loftus-Scheff norms). The results are as follows. With the biased procedure, pairs like *daisy-flower* were verified 142 msec faster than pairs like *daisy-plant*; but with the unbiased procedure, this difference was only 42 msec. As for the error terms, the standard deviation of the biased difference was only 56 msec; that for the unbiased difference was 63 msec. Had the investigator used the biased procedure, then, he would have concluded that there was a highly reliable difference where in fact it was much smaller and far less reliable than he thought. It should be noted that the difference between the two sampling procedures is a very subtle one: It could probably not have been noticed without

consulting the Loftus-Scheff norms or other similar measures. Despite its subtlety, it had profound consequences on the results. Examples like this, therefore, raise serious questions about experiments in which such sampling biases could have arisen.[11]

In practice, however, it is sometimes difficult, if not impossible, to make use of systematic sampling procedures. This is illustrated by the Bransford, Barclay, and Franks (1972) study in which they compared recognition memory errors from presented sentences like *Three turtles rested (on/beside) a floating log and a fish swam under it* to test sentences in which the final *it* had been replaced by *them*. When the study sentence contains *on*, one can readily infer that the test sentence with *them* is also true; but when the study sentence contains *beside*, one would not normally make such an inference. For this reason, Bransford *et al.* expected more false alarms on the test sentence for the former case than for the latter. But what is the population from which these sentences were drawn? Apparently, it consists of all those sentences describing situations for which people would readily draw the inference described in the test sentence. This population cannot easily be listed or specified with a rule. Yet it appears to be legitimate. The method Bransford *et al.* used to generate their sample —they composed the sentences on intuitive grounds—will therefore have to suffice until some more exact procedure is possible. Nevertheless, investigators using such intuitive procedures should be as explicit as possible about the constraints they were trying to stick to so that other investigators can construct similar samples and perhaps even refine the procedures for specifying the appropriate language population.

*The Method of Single Cases*

An entirely different approach to the study of language is to work from single cases, that is, to examine only a few words at a time, treating them as if they comprised the total population of interest. This method has had a long and respectable tradition in linguistics. It has come about because there are many critical issues in linguistics that investigators have been able to study by examining no more than one or two words at a time. For example, by examining *good* in detail, Katz (1964) was able to raise certain issues about the form of dictionary entries in a semantic theory; on the basis of *come* and *go*, Fillmore (1966) and Clark (in press) were able to argue for the importance of deixis in linguistic theory; from the properties of *even*, Fraser (1971), Horn (1969), and Anderson (1972) were able to question the adequacy of various grammars; and from his detailed examination of *remind*, Postal (1970) was able to argue for lexical decomposition in grammar. By referring to small "fields" of words, other linguists have been able to test the adequacy of various semantic theories, as in Bierwisch's (1967) study of spatial adjectives, Lehrer's (1969) study of cooking terms, Bendix' (1966) study of verbs of possession, and Fillmore's (1971) recent study of verbs of judgment. Thus, even though they examined only a few single cases, these investigators were able to make general claims about language; typically, they could do so because their cases constituted critical counter-examples to previous theories. In addition, these linguists would have found it impractical, perhaps even impossible, to examine more cases in the detail required for their argument. Indeed, investigations such as these have served as the backbone of linguistics. The knowledge accumulated from them has been fundamental to the construction of viable linguistic theories.

The method of single cases applies to psychology for the same reasons. There are many significant issues that psychologists can attack by examining only a few words in detail. For example, using only *some* and *all*, Johnson-Laird (1969a, b) was able to show that the scope of quantifiers is affected by

[11] I am indebted to Edward E. Smith and Thomas D. Schumacher for making their data available to me.

sentence voice, and Meyer (1970) was able to make significant claims about retrieval from semantic memory; from sentences containing only *above* and *below*, Clark and Chase (1972) were able to argue for the sequential nature of sentence verification; with only the pairs *good-bad* and *deep-shallow*, Clark (1969) was able to test several theories of deductive reasoning; on the basis of *ask* and *tell*, Chomsky (1969) was able to argue for the importance of grammatical complexity in language acquisition; and using *more* and *less*, and *before* and *after*, respectively, Donaldson and Balfour (1968) and Clark (1971) were able to argue for the systematic acquisition of meaning. Most of these studies would have been impractical if they had been required to examine more words, yet each was successful in raising or resolving some critical issue. To give an example, the Clark (1969) study was carried out in part to test a theory DeSoto, London, and Handel (1965) had proposed to explain why linear syllogisms containing *better* (for example, *Abel is better than Baker; and Baker is better than Charlie; therefore, Abel is better than Charlie*) are easier to solve than the equivalent syllogisms containing *worse*. DeSoto *et al.*'s theory was able to be disconfirmed by two single cases, one with syllogisms containing the relations *isn't as good as* and *isn't as bad as*, and the other with syllogisms containing *deeper* and *shallower*. The method of single cases was applicable here because it was possible to disconfirm the DeSoto *et al.* theory with just one clear counter-example.

Useful as it is for disconfirming theories and raising critical issues, the method of single cases requires care when it is used to support hypotheses. With this method, the investigator must, as always, accumulate enough single cases, presumably drawn at random from the language population, so that he can argue for generality with the usual statistical tests. Consider, for example, the so-called markedness hypothesis (Clark, 1969) which claims that unmarked, or positive adjectives (like *better, longer, faster*), are comprehended more quickly than their marked, or negative, opposites. The hypothesis was originally demonstrated with only one pair of adjectives, *good* and *bad*, which is hardly convincing evidence that the hypothesis is true for language in general. This demonstration, however, was accompanied by a list of 10 other adjective pairs found in the literature, each studied singly by various investigators, which were all consistent with the hypothesis. Taken together these 11 single cases did suggest that the hypothesis was true in general; the appropriate tests would have shown this to be a statistically significant result over both subjects and word-pairs. Since that time other confirmatory pairs have appeared in the literature, and there have been no counter-examples. So the method of single cases, used with discretion, can lead to generalizations about language as a whole.

When used in testing or supporting hypotheses, the method of single cases has one quite severe requirement: The hypotheses of interest must be applicable to single cases, and these are often rather strong hypotheses. The markedness hypothesis, for example, claims that unmarked adjectives should be easier than marked ones for each and every unmarked–marked pair that can be found. This makes it possible to disconfirm or at least force revision of the hypothesis by finding single counter-instances. In contrast, the Rubenstein *et al.* hypothesis about homographs claims that homographs take longer to recognize than nonhomographs all other things being equal. Since it is impossible to find single homograph/nonhomograph pairs identical in all other possible factors—frequency, meaning, word length, spelling difficulty, and other undetermined factors—it is only possible to test the hypothesis by looking at the central tendencies (for example, the means) of homographs versus nonhomographs. There is no single case imaginable that suffices to disconfirm the homograph hypothesis. So the method of single cases is simply not applicable to such "central-tendency" hypotheses.

Some hypotheses have not been stated clearly enough to know whether they are "every instance" hypotheses (like the markedness hypothesis) or "central tendency" hypotheses (like the homograph hypothesis). Collins and Quillian (1969), for example, proposed that for such word-triples as *ruby-stone-solid* it is easier to verify *A ruby is a stone* than *A ruby is a solid.* But was this meant to hold for every such word-triple or only for such word-triples on the average? The theory itself appeared to require it to hold for every word-triple, since there was no criterion for excluding the word-triples it should not hold for. Nevertheless, Collins and Quillian themselves conjectured that such word-triples as *dog-mammal-animal* (that is, those containing *mammal* and *animal*) were probably counterexamples, and this was later confirmed by Collins and Quillian (1971) as well as by Rips *et al.* (1973). Depending on whether the hypothesis was meant to hold for every instance or for central tendencies, this single case does or does not disconfirm the theory. Desirable as the stronger "every instance" hypotheses may be, however, it is often not possible to construct such hypotheses, and so the method of single cases will not be applicable for a large number of hypotheses.

The main purpose of the method of single cases is to shed light on individual words. Thus, it is crucial for investigators using this method to report both (1) the instances used and (2) the data for each instance separately. Without this requirement, the method reduces to the traditional design where the investigator reports only overall results. And once the investigator takes the latter option, he is obliged to treat words as a random effect and to demonstrate statistically that his results hold for the language population in general. The argument is simply this. When an investigator reports only overall means, he is implicitly assuming that these means are representative of the single instances contained in that mean. This assumption is only justifiable, of course, if he can show that his findings are consistent over both subjects and words, a demonstration that requires him to treat words as a random effect. To turn the argument around, reporting the data for single cases separately is always justifiable: The data can thereafter be used as support or disconfirmation for any number of theories, however the investigator wishes to use them. It is the lumping together of data, obliterating the single cases, that requires the strong assumption. For this to be done, the overall means must be shown to be representative of each instance. This in turn requires the statistics discussed in this paper.

Most investigators, whether they use the method of single cases or not, would be doing others a service by reporting the data for each case singly whenever feasible. One reason for this recommendation is that the sampling methods in most experiments on language are rather crude, if they exist at all. There is no guarantee that an investigator's own characterization of his sample—for example, "it is a random sample of concrete nouns"—is wholly accurate. Such characterizations often obscure sampling biases that only become obvious on close examination of the words actually used. With the full sample at hand, other investigators can look for systematic confoundings, attempt alternative characterizations of the sample, and test new hypotheses. A more important reason, however, stems from the increasing interest psychologists have shown in the properties of individual words. Not too long ago, many psychologists tended to think of words only as items varying in meaningfulness, concreteness, form class, and the like. More recently, many of these same psychologists have come to appreciate the fact that each word can be interesting in its own right. A word such as *or* is not just another function word low in meaningfulness and imagery, but is a conjunction with a specific semantic and logical function in the structure of English. To lump *or* with other function words, therefore, can only obscure its unique properties, rendering the data useless for

those investigators who want to know more about *or* by itself. To put it plainly, this is a plea to consider individual words within language as objects worthy of study for their own sake.[12]

## Who is Committing the Language-as-Fixed-Effect Fallacy?

The answer, sad to say, is almost everyone. Although I have singled out semantic memory for criticism in this paper, I could have chosen any area that makes use of words, sentences, phrases, printed letter strings, paragraphs, stories, word lists, letters, digits, nonsense syllables, or trigrams. To bring this point home I will briefly list some of the areas to which the present arguments apply.

Coleman's (1964) original criticisms were directed in part at studies in verbal learning. In paired-associates learning experiments, for example, the stimulus and response terms are typically words, nonsense syllables, or digits, and the treatments of interest often concern the association value, frequency, or imagery value of the words, where the words have been sampled from some well-defined population. Although these studies fulfill all the criteria for treating words as a random effect, few of them, to my knowledge, have ever done so, or have given any statistical evidence that the findings would generalize to other stimulus and response terms. The same applies *ceteris paribus* to serial learning studies using words or nonsense syllables. In most memory studies the items to be remembered are also words, and yet investigators of memory have never, so far as I know, treated words as a random effect or provided statistical evidence for generalizing to the word population.

Psycholinguists have been particularly guilty of the language-as-fixed-effect fallacy. Whereas investigators of paired-associates learning, serial learning, and memory have at least worked out procedures for sampling words and nonsense syllables, psycholinguists with few exceptions have not. Typically, the investigator selects his stimulus sentences with specific linguistic constraints in mind, but makes no effort to sample such sentences systematically even though he could (see Coleman, 1965). Furthermore, since the appropriate sentences are often so difficult to construct, the usual design includes only a small number of such sentences. On top of all this, psycholinguists have almost universally treated sentences as a fixed, instead of a random, factor. Taken together, these practices raise considerable doubt about many of the smaller findings in psycholinguistics.

The present criticisms apply just as forcefully to experiments on the perception of wordlike letter strings. Many of these experiments, relying on only small samples of words, have produced effects that have been rather small in terms of latencies, percent correct, or perceptual thresholds. It is under just these circumstances, as the studies in semantic memory have demonstrated, that the language-as-fixed-effect fallacy can have its most serious repercussions. Again, it has been an almost universal practice to report no statistics enabling one to generalize the results to the language population.

The wide-spread capitulation to the language-as-fixed-effect fallacy, though alarming, has probably not been disastrous. In the older established areas, most experienced investigators have acquired a good feel for what will replicate on a new language sample and what will not. They then design their experiments accordingly. It is in new, relatively unexplored areas like semantic memory that the language-as-fixed-effect fallacy has its greatest potential for damage. In these areas investigators will be less acquainted with their stimulus materials, and so their experiments will be particularly vulnerable to lack of replicability.

[12] Note that words differ from subjects in this respect. The word *or*, for example, can be studied by any investigator with English speaking subjects. The subject Joe Doakes, however, cannot. As a transient sophomore at Stanford University, he is for all practical purposes inaccessible to all other investigators.

Nonreplicability wastes valuable time and energy, of course, because other investigators are led off in wrong directions or are forced to replicate the original studies with better controls. The remedies suggested in this paper, if used consistently, can perhaps spare us some of the early tribulations in new areas like semantic memory, and they can only strengthen the experiments in the remaining well-established areas.

## Appendix

In this appendix I will explicate some simple but useful algebraic consequences of $F'$. I will derive these consequences only for the analysis of variance model in Table 1, although all the formulae below are the same for the model in Table 3 with an appropriate change of $MS_{\mathrm{WwT}}$ to $MS_{\mathrm{T\times W}}$, and $MS_{\mathrm{S\times WwT}}$ to $MS_{\mathrm{T\times S\times W}}$.

As shown in (5), the quasi F-ratio for the Treatments effect in Table 1 is given in (i):

$$\text{(i)}\quad F'(i,j) = (MS_{\mathrm{T}} + MS_{\mathrm{S\times WwT}})/(MS_{\mathrm{T\times S}} + MS_{\mathrm{WwT}}).$$

If, as before, we define $F_3 = MS_{\mathrm{S\times WwT}}/MS_{\mathrm{T\times S}}$, and assume $F_1 \geqslant F_2$, or equivalently, $MS_{\mathrm{T\times S}} \leqslant MS_{\mathrm{WwT}}$, then *max* $F'$ is given by (ii), as shown in the text:

$$\text{(ii)}\quad max\,F'(i,j) = (MS_{\mathrm{T}} + F_3^* MS_{\mathrm{T\times S}})/(MS_{\mathrm{T\times S}} + MS_{\mathrm{WwT}})$$

where $F_3^*$ is the critical value of $F_3$. To be able to examine the range of *max* $F'$, we will multiply both the numerator and denominator of (ii) by $MS_{\mathrm{T}}/MS_{\mathrm{T\times S}}\,MS_{\mathrm{WwT}}$, and this yields:

$$\text{(iii)}\quad max\,F'(i,j) = \left(\frac{MS_{\mathrm{T}}}{MS_{\mathrm{T\times S}}}\cdot\frac{MS_{\mathrm{T}}}{MS_{\mathrm{WwT}}} + F_3^*\frac{MS_{\mathrm{T}}}{MS_{\mathrm{WwT}}}\right)\Big/\left(\frac{MS_{\mathrm{T}}}{MS_{\mathrm{WwT}}} + \frac{MS_{\mathrm{T}}}{MS_{\mathrm{T\times S}}}\right)$$

and since $F_1 = MS_{\mathrm{T}}/MS_{\mathrm{T\times S}}$ and $F_2 = MS_{\mathrm{T}}/MS_{\mathrm{WwT}}$, we can simplify this formula as follows:

$$\text{(iv)}\quad max\,F'(i,j) = F_2(F_1 + F_3^*)/(F_1 + F_2).$$

There are several interesting conclusions that follow directly from (iv). First, there is only one really interesting case and that is when both $F_1$ and $F_2$ are larger than $F_3^*$; if they are not, there is no chance of *max* $F'$ being significant. But when this is the case, the following inequality holds:

$$\text{(v)}\quad max\,F' \leqslant F_2 \leqslant F_1$$

This inequality follows from (iv) because when $F_1 \geqslant F_2 \geqslant F_3^*$, the fraction $(F_1 + F_3^*)/(F_1 + F_2)$ will always be less than or equal to 1, hence $F_2$ times this fraction will always be less than or equal to $F_2$. Indeed, the first equality sign will obtain only if *max* $F' = F_2 = F_3^*$, and *max* $F'$ could not be significant, or if $F_1$ were infinitely large. Since this whole argument is symmetrical for $F_1$ and $F_2$, it follows that in all practical cases, *max* $F'$ will be less than $F_1$ or $F_2$, whichever is smaller. Second, if we again assume $F_1 \geqslant F_2$, then:

$$\text{(vi)}\quad max\,F' \leqslant \tfrac{1}{2}(F_1 + F_3^*)$$

This follows because when $F_1 = F_2$, equation (iv) reduces to *max* $F' = \frac{1}{2}(F_1 + F_3^*)$, and when $F_1$ becomes larger than $F_2$, *max* $F'$ will always be less than this value.

As shown in the text, *min* $F'$ is given by (vii):

$$\text{(vii)}\quad min\,F'(i,j) = MS_{\mathrm{T}}/(MS_{\mathrm{T\times S}} + MS_{\mathrm{WwT}})$$

Multiplying the numerator and denominator by $MS_{\mathrm{T}}/MS_{\mathrm{T\times S}}MS_{\mathrm{WwT}}$ and simplifying, we obtain:

$$\text{(viii)}\quad min\,F'(i,j) = F_1F_2/(F_1 + F_2)$$

It follows directly that *min* $F'$ will always be less than $F_1$ and $F_2$, and indeed:

$$\text{(ix)}\quad min\,F' \leqslant \tfrac{1}{2}F_1, \quad \text{for} \quad F_1 \geqslant F_2,$$
$$min\,F' \leqslant \tfrac{1}{2}F_2, \quad \text{for} \quad F_2 \geqslant F_1.$$

The degrees of freedom can also be calculated in terms of $F_1$ and $F_2$. If $F_1$ has $n$ and $n_1$ degrees of freedom, and $F_2$ has $n$ and $n_2$ degrees of freedom, then $i = n$, and $j$ is the nearest integer calculated by the following formula:

$$\text{(x)}\quad j = (MS_{\mathrm{T\times S}} + MS_{\mathrm{WwT}})^2\Big/\left(\frac{MS_{\mathrm{T\times S}}{}^2}{n_1} + \frac{MS_{\mathrm{WwT}}{}^2}{n_2}\right).$$

By multiplying both the numerator and denominator in (x) by $(MS_T/MS_{T\times S}MS_{W\times T})^2$ and simplifying ,we get:

$$\text{(xi)}\quad j = (F_1 + F_2)^2 \Big/ \left(\frac{F_1^2}{n_2} + \frac{F_2^2}{n_1}\right).$$

It should be noted that when $F_1 = F_2$, $j$ will simply be twice the harmonic mean of $n_1$ and $n_2$; in any case, $j$ will always be less than the sum of $n_1$ and $n_2$ and will equal it only when $F_1/n_2 = F_2/n_1$.

Finally, while assuming $F_1 \geqslant F_2$, we can derive the relationship between *max* $F'$ and *min* $F'$ simply by taking the ratio of (iv) to (viii). This gives:

$$\text{(xii)}\quad max\,F'/min\,F' = (F_1 + F_3^*)/F_1$$

or

$$\text{(xiii)}\quad max\,F' = (1 + F_3^*/F_1)\,min\,F'.$$

## References


Anderson, S. R. How to get *even*. *Language*, 1972, **48**, 893–906.

Battig, W. F., & Montague, W. E. Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monograph*, 1969, **80**, 1–46.

Bendix, E. H. *Componential analysis of general vocabulary: The semantic structure of a set of verbs in English, Hindi, and Japanese*. The Hague: Mouton, 1966.

Bierwisch, M. Some semantic universals of German adjectivals. *Foundations of Language*, 1967, **3**, 1–36.

Bransford, J. D., Barclay, J. R., & Franks, J. J. Sentence memory: A constructive versus interpretive approach. *Cognitive Psychology*, 1972, **3**, 193–209.

Carroll, J. B. An experiment in evaluating the quality of translations. *Mechanical Translation and Computational Linguistics*, 1966, **9**, 55–66.

Chomsky, C. *The acquisition of syntax in children from 5 to 10*. Cambridge, Mass: MIT Press, 1969.

Clark, E. V. On the acquisition of the meaning of *before* and *after*. *Journal of Verbal Learning and Verbal Behavior*, 1971, **10**, 266–275.

Clark, E. V. Normal states and evaluative viewpoints: More on *come* and *go*. *Language*, in press.

Clark, H. H. Linguistic processes in deductive reasoning. *Psychological Review*, 1969, **76**, 387–404.

Clark, H. H., & Begun, J. S. The use of syntax in understanding sentences. *British Journal of Psychology*, 1968, **59**, 219–229.

Clark, H. H., & Card, S. K. The role of semantics in remembering comparative sentences. *Journal of Experimental Psychology*, 1969, **82**, 545–553.

Clark, H. H., & Chase, W. G. On the process of comparing sentences against pictures. *Cognitive Psychology*, 1972, **3**, 472–517.

Clark, H. H., & Clark, E. V. Semantic distinctions and memory for complex sentences. *Quarterly Journal of Experimental Psychology*, 1968, **20**, 129–138.

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

Coleman, E. B. Learning of prose written in four grammatical transformations. *Journal of Applied Psychology*, 1965, **49**, 332–341.

Collins, A. M., & Quillian, M. R. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 1969, **8**, 240–247.

Collins, A. M., & Quillian, M. R. Does category size affect categorization time? *Journal of Verbal Learning and Verbal Behavior*, 1970, **9**, 432–438. (a)

Collins, A. M., & Quillian, M. R. Facilitating retrieval from semantic memory: The effect of repeating part of an inference. *Acta Psychologica*, 1970, **33**, 304–314. (b)

Collins, A. M., & Quillian, M. R. Categories and subcategories in semantic memory. Paper presented at the annual meeting of Psychonomic Society, St. Louis, MO, 1971.

Conrad, C. Cognitive economy in semantic memory. *Journal of Experimental Psychology*, 1972, **92**, 149–154.

DeSoto, C., London, M., & Handel, S. Social reasoning and spatial paralogic. *Journal of Personality and Social Psychology*, 1965, **2**, 513–521.

Donaldson, M., & Balfour, G. Less is more: A study of language comprehension in children. *British Journal of Psychology*, 1968, **59**, 461–472.

Fillmore, C. J. Deictic categories in the semantics of *come*. *Foundations of Language*, 1966, **2**, 219–227.

Fillmore, C. J. Verbs of judging: An exercise in semantic description. In C. J. Fillmore & D. T. Langendoen (Eds.), *Studies in linguistic semantics*. New York: Holt, Rinehart, & Winston, 1971.

Fraser, J. B. An analysis of "even" in English. In C. J. Fillmore & D. T. Langendoen (Eds.), *Studies in linguistic semantics*. New York: Holt, Rinehart, & Winston, 1971.

Freedman, J. L., & Loftus, E. F. Retrieval of words from long-term memory. *Journal of Verbal Learning and Verbal Behavior*, 1971, **10**, 107–115.

HORN, L. J. A presuppositional approach to *only* and *even*. *Papers from the 5th Regional Meeting, Chicago Linguistic Society*, 1969, 98–107.

JOHNSON-LAIRD, P. N. On understanding logically complex sentences. *Quarterly Journal of Experimental Psychology*, 1969, **21**, 1–13. (a)

JOHNSON-LAIRD, P. N. Reasoning with ambiguous sentences. *British Journal of Psychology*, 1969, **60**, 17–23. (b)

KATZ, J. J. Semantic theory and the meaning of "good". *Journal of Philosophy*, 1964, **61**, 739–766.

KUČERA, H., & FRANCIS, W. N. *Computational analysis of present-day American English*. Providence, RI: Brown University, 1967.

LANDAUER, T. K., & FREEDMAN, J. L. Information-retrieval from long-term memory: Category size and recognition time. *Journal of Verbal Learning and Verbal Behavior*, 1968, **7**, 291–295.

LANDAUER, T. K., & MEYER, D. E. Category size and semantic-memory retrieval. *Journal of Verbal Learning and Verbal Behavior*, 1972, **11**, 539–549.

LEHRER, A. Semantic cuisine. *Journal of Linguistics*, 1969, **5**, 38–56.

LOFTUS, E. F. Nouns, adjectives, and semantic memory. *Journal of Experimental Psychology*, 1972, **96**, 213–215.

LOFTUS, E. F. Category dominance, instance dominance, and categorization time. *Journal of Experimental Psychology*, 1973, **97**, 70–74.

LOFTUS, E. F. Activation of semantic memory. *American Journal of Psychology*, in press.

LOFTUS, E. F., & FREEDMAN, J. L. Effect of category-name frequency on the speed of naming an instance of the category. *Journal of Verbal Learning and Verbal Behavior*, 1972, **11**, 343–347.

LOFTUS, E. F., FREEDMAN, J. L., & LOFTUS, G. R. Retrieval of words from subordinate and supra-ordinate categories in semantic hierarchies. *Psychonomic Science*, 1970, **21**, 235–236.

LOFTUS, E. F., & SCHEFF, R. W. Categorization norms for 50 representative instances. *Journal of Experimental Psychology Monograph*, 1971, **91**, 355–364.

MEYER, D. E. On the representation and retrieval of stored semantic information. *Cognitive Psychology*, 1970, **1**, 242–300.

MEYER, D. E. Dual memory-search of related and unrelated semantic categories. Paper presented at the meeting of the Eastern Psychological Association, New York, April, 1971.

MEYER, D. E. Verifying affirmative and negative propositions: Effects of negation on memory retrieval. In S. Kornblum (Ed.), *Attention and Performance IV*. New York: Academic Press, 1973.

MEYER, D. E. Correlated operations in searching stored semantic categories. *Journal of Experimental Psychology*, in press.

MEYER, D. E., & ELLIS, G. B. Parallel processes in word recognition. Paper presented at the meeting of the Psychonomic Society, San Antonio, November, 1970.

MEYER, D. E., & SCHVANEVELDT, R. W. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 1971, **90**, 227–234.

PAIVIO, A., YUILLE, J., & MADIGAN, S. Concreteness, imagery, and meaningfulness value for 925 nouns. *Journal of Experimental Psychology*, 1968, **76**, Monograph Supplement, No. 1, Part 2.

POSTAL, P. M. On the surface verb "remind". *Linguistic Inquiry*, 1970, **1**, 37–120.

POSTMAN, L., & KEPPEL, G. *Norms of word associations*. New York: Academic Press, 1970.

RIPS, L. J., SHOBEN, E. J., & SMITH, E. E. Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, 1973, **12**, 1–20.

ROSCH, E. On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language*. New York: Academic Press, 1973.

RUBENSTEIN, H., GARFIELD, L., & MILLIKEN, J. A. Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, 1970, **9**, 487–494.

RUBENSTEIN, H., LEWIS, S. S., & RUBENSTEIN, M. Homographic entries in the internal lexicon: Effects of systematicity and relative frequency of meanings. *Journal of Verbal Learning and Verbal Behavior*, 1971, **10**, 57–62. (a)

RUBENSTEIN, H., LEWIS, S. S., & RUBENSTEIN, M. Evidence for phonemic recoding in visual word recognition. *Journal of Verbal Learning and Verbal Behavior*, 1971, **10**, 645–657. (b)

SCHAEFFER, B., & WALLACE, R. Semantic similarity and the comparison of word meanings. *Journal of Experimental Psychology*, 1969, **82**, 343–346.

SCHAEFFER, B., & WALLACE, R. The comparison of word meanings. *Journal of Experimental Psychology*, 1970, **86**, 144–152. (a)

SCHAEFFER, B., & WALLACE, R. Semantic interference: Obligatory or optional. *Journal of Experimental Psychology*, 1970, **86**, 335–337. (b)

SCHVANEVELDT, R. W., & MEYER, D. E. Retrieval and comparison processes in semantic memory. In S. Kornblum (Ed.), *Attention and Performance IV*. New York: Academic Press, 1973.

SMITH, E. E. Effects of familiarity of stimulus recognition and categorization. *Journal of Experimental Psychology*, 1967, **74**, 324–332.

SMITH, E. E., HAVILAND, S. E., BUCKLEY, P. B., & SACK, M. Retrieval of artificial facts from long-term memory. *Journal of Verbal Learning and Verbal Behavior*, 1972, **11**, 583–593.

SNEDECOR, G. W., & COCHRAN, W. G. *Statistical methods*. Ames: Iowa State University Press, 1967.

WILKINS, A. Conjoint frequency, category size, and categorization time. *Journal of Verbal Learning and Verbal Behavior*, 1971, **10**, 382–385.

WINER, B. J. *Statistical principles in experimental design*. New York: McGraw-Hill, 1971.

(Received December 18, 1972)